Self-Supervised Variational Autoencoder for Unsupervised Object Counting From Very-High-Resolution Satellite Imagery: Applications in Dwelling Extraction in FDP Settlement Areas

Getachew Workineh Gella, *Graduate Student Member, IEEE*, Hugo Gangloff, Lorenz Wendt, Dirk Tiede, *Member, IEEE*, and Stefan Lang, *Member, IEEE*

***Abstract*— In supervised learning, deep learning models demand a large corpus of annotated data for object detection and classification tasks. This constrains their utility in humanitarian emergency response. To overcome this problem, we have proposed an unsupervised dwelling counting from very-high-resolution (VHR) satellite imagery by combining a variational autoencoder (VAE) with an anomaly detection approach. When VAE applied in earth observation images for dwelling localization and counting, we observed two critical limitations: 1) the balance between reconstruction and good latent code, where the favor of good reconstruction of dwellings leads to weak anomaly score maps that fail to properly localize dwellings and 2) limited spatiotemporal invariance of the learned latent code. When the model is trained with datasets obtained from different geography and time, it fails to properly localize dwellings. For the first problem, we introduced self-supervision by creating synthetic anomalies. For the second problem, we introduced latent space conditioning. The approach is tested on nine VHR images obtained from six forcibly displaced people settlement areas. Results indicate that combining VAE with an anomaly detection approach has reached an area under the receiver operating characteristic curve value ranging from 0.70 at complex settlements to 0.98 at relatively less complex settlement areas. Similarly, a mean absolute error (MAE) value of 56.67 toward 5.03 is achieved for dwelling counting. Joint training of combined datasets with latent space conditioning and self-supervision enabled the achievement of results better than classical VAE, with improved spatiotemporal transferability of the model with more crisp and strong anomaly maps. Overall implementation code will be available at https://github.com/getch-geohum/SSL-VAE.**

***Index Terms*— Anomaly, cutPaste, dwelling counting, latent space conditioning, localization, self-supervision, unsupervised learning, variational autoencoder (VAE).**

Manuscript received 21 September 2023; revised 21 November 2023; accepted 15 December 2023. Date of publication 19 December 2023; date of current version 24 January 2024. The work of Getachew Workineh Gella, Lorenz Wendt, Dirk Tiede, and Stefan Lang was supported in part by the Christian Doppler Research Association; in part by the Doctors Without Borders-Section Austria; in part by the Austrian Federal Ministry of Labour and Economy; and in part by the National Foundation for Research, Technology and Development. *(Corresponding author: Getachew Workineh Gella.)*

Getachew Workineh Gella, Lorenz Wendt, Dirk Tiede, and Stefan Lang are with the Christian Doppler Laboratory for Geospatial and EO-Based Humanitarian Technologies (GEOHUM), Paris Lodron University of Salzburg (PLUS), 5020 Salzburg, Austria (e-mail: getachewworkineh.gella@plus.ac.at).

Hugo Gangloff is with Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, 91120 Palaiseau, France.

Digital Object Identifier 10.1109/TGRS.2023.3345179

I. Introduction

A HUGE number of the global population has been displaced from their home and staying either in temporary settlement areas for internally displaced persons (IDPs) or in refugee camps, which we hereafter inclusively term forcibly displaced population (FDP) settlement sites. According to the United Nations Higher Commission for Refugees (UNHCR), by the end of the year 2022, there were around 108.4 million forcibly displaced people [1]. It is reported that those FDPs are hosted in more than 13 000 FDP settlement areas [2], distributed across the globe with different geographical settings. Hence, dwelling information is crucial to monitor camp and temporary settlement expansion by FDP influx, estimate residing populations, and provide adequate humanitarian emergency assistance. For the past decade, Earth observation (EO) technology has played a significant role in providing first-hand information to support humanitarian emergency assistance [3], [4], [5]. In this aspect, the proliferation of various sensors, especially in the optical domain, has enabled precise monitoring of FDP settlement areas with fine spatial granularity without much temporal latency. Notable methods include camp mapping and settlement expansion using an object-based image analysis (OBIA) and rule set approaches [6], [7]. This has enabled information retrieval to the level of detecting individual dwelling instances [4], [8], [9], classification of dwelling types and corresponding counts [4], [6], [9], and further estimation of the resident population using dwelling information as proxy indicator variable [4]. Despite the performance of these semiautomatic approaches, skilled expert knowledge and context-specific curated rule sets are still needed. The latter is challenged by the short response time required for information generation and delivery to assist operational humanitarian emergency response. This challenge gains complexity since, nowadays, monitoring needs to be done more frequently in time and at larger geographic scales

(e.g., due to increased VHR data availability and the dynamic nature of FDP settlement areas).

Currently, deep learning models have shown great performance in classification [10], segmentation [11], [12], and object detection tasks [13], [14], [15], which paves the way for automatic information retrieval pipelines. Benefiting from advances in deep learning for computer vision, there are promising works dedicated to dwelling extraction from temporary settlements for humanitarian emergency response [16], [17], [18], [19], [20], [21]. Despite the strong performance of supervised computer vision models in various fields, they have known limitations that constrain their full-fledged usage in operational humanitarian emergency response. The first one is intensive data demand for model training and testing in a supervised setting. One of the critical elements in operational humanitarian emergency response is the speed of information retrieval for decision-making. Supervised models demand a bulk of annotated data, which is time-consuming and, sometimes, quite challenging and impractical to prepare under time pressure. More importantly, even after supervised training of the model using images and corresponding annotations obtained from a specific geography with a given time stamp, the model would fail to have similar skills to undertake intended tasks on datasets obtained from different times and geography. This lack of generalization under distribution shift is caused by changing object, and scene characteristics are another well-identified limitation of deep learning.

In situations where ground-truth annotations exist in a specific geography and or time and there are no annotations in the target site, model skills can be transferred using unsupervised domain adaptation [22], [23], [24] and multisource training approaches, such as transfer learning by fine-tuning (retraining) [16] and metalearning [25]. These strategies assume the availability of a sufficient amount of data either for joint training or fine-tuning during the transfer of the model. Situations, where no annotations are available, constitute the worst case scenario; they are nonetheless very common during the occurrence of unexpected natural and man-made disasters that foster the establishment of new FDP settlements. Under operational humanitarian response, data-related challenges demand vision models that can quickly train and scale with unavailable or very scarce annotations. For this, (self)-unsupervised learning strategies were found as ideal workflow options, especially for image reconstruction, an objective, which is useful for other downstream tasks.

Self-supervised learning approaches in EO focus on learning representation from unlabeled data as pretraining following contrastive [26], [27], [28], [29], [30] and generative [31], [32], [33] approaches. Learned representations can be used for downstream tasks supported by fine-tuning with curated labels in scene classification [26], [27], [29], [31], [32] and semantic segmentation [28], [29], [30] tasks. This two-step strategy is also leaning toward semisupervision and knowledge distillation [34]. These workflows could not be directly applied to unsupervised localization and counting dwellings because transferring the learned representations to downstream tasks still requires supervision (annotations). This motivates the need for strategies that do not demand annotations to learn the image representation and, at the same time, have a strong reconstruction ability suitable for automatic visual anomaly detection (the downstream task in our study). Therefore, variational autoencoders (VAEs) seem the ideal candidates. Other innovative approaches implement self-supervision for detecting visual anomalies (e.g., [35], [36] on non-EO images) with specific patterns and textures that could easily be leveraged as contextual information for self-supervision. Even though those approaches are valid to adopt, they have limitations that restrict their application to EO datasets. First, contextual information generation is mainly driven by patterns in the image; EO datasets (especially dwellings) do not have any distinct spatial pattern. Second, plain autoencoders are not robust enough to provide proper reconstruction for out-of-distribution samples.

VAE [37] is one of the powerful models for unsupervised learning with various application domains. Autoencoders and VAEs [38] are used for various remotely sensed SAR images for classification, scene understanding, and detection. In scene classification, as summarized in [39], the main utility of autoencoders with various architectures, especially sparse autoencoders, was for learning proper feature representation as a pretext task and further scene classification. By leveraging the nature of VAEs for proper representation of images in a compressed latent space, Xu et al. [40] and de Oliveira et al. [41] performed SAR data compression, while Ferreira and Silveira [42] used VAEs for automatic ship detection from SAR images. The VAE model is also used for learning low-dimensional satellite image representation [43] for multispectral images, spectral feature extraction [44], [45], and useful feature generation [46] from hyperspectral images and wider application for further classification of those images [43], [43], [44], [45], [46], [47]. For a similar task, Chen et al. [48] have further combined adversarial training using a self-attention mechanism to augment the performance of VAEs for hyperspectral image classification. VAEs have also been combined with a reinforcement learning strategy, for satellite image captioning [49].

By leveraging the power of VAEs for proper learning of normal image latent space representation and its capability to reconstruct the input image from compressed latent space, there are recent studies that used VAE for anomaly detection. To note, Sinha et al. [50] implemented VAE to detect avalanche deposits as anomalies from SAR imagery, while Zhang et al. [51] has used VAE for anomaly detection and background suppression from hyperspectral imagery. The approach implemented in [51] for anomaly detection from hyperspectral imagery is further improved by Wei et al. [52] who combined adjacency matrix from graph regularized learning with VAE latent code. These studies, however, have not addressed the localization and counting of individual objects from high-resolution satellite imagery.

In this respect, a recent study by Gangloff et al. [53] has addressed the proper localization of wild animals from aerial imagery. As an extension of this work, the main objective of this study is the unsupervised localization and counting of dwelling objects from very-high-resolution (VHR) optical satellite images obtained from FDP settlement areas. The main contribution of this study is outlined as follows.

1) We have combined anomaly detection and VAEs for unsupervised localization and counting of dwelling

TABLE I
DATASET DESCRIPTION

FDP site	Country	Sensor	Date	Resolution*
Dagahaley	Kenya	WorldView-3	08/04/2017	0.5
Kuletirkidi	Ethiopia	Pléiades	22/06/2018	0.5
Kuletirkidi	Ethiopia	Pléiades	24/03/2017	0.5
Kutupalong	Bangladesh	WorldView-2	25/09/2017	0.5
Minawao	Cameroon	WorldView-2	12/02/2017	0.5
Minawao	Cameroon	WorldView-3	03/0/2016	0.5
Nguenygiel	Ethiopia	Pléiades	24/03/2017	0.5
Nduta	Tanzania	Pléiades	21/10/2016	0.5
Zamzam	Sudan	Pléiades-1B	26/02/2022	0.5

objects from VHR optical satellite images. To the extent of our knowledge, no study conceptualized buildings or dwellings as an anomaly for unsupervised object detection and counting.

2) Inspired by the concepts of the cutPaste algorithm [35], we have customized synthetic anomaly generation for remotely sensed data and implemented self-supervised conditional variational training.
3) We have tried a joint training of multiple datasets with latent space conditioning and ensure dataset invariance of learned representation, which ensures the spatial and temporal transferability of the model.

Given these contributions, the remainder of this article is organized as follows. Detailed implementation of methods related to data and study site together with synthetic anomaly generation is provided in Section II. The implementation of self-supervised conditional variational learning is provided in Section III-C. Details about the model and experimental setup obtained results are presented in Section IV. Results were further discussed in Section V followed by conclusions and remarks for further work, which are provided in Section VI.

II. METHODS

In this section, details about the data, study sites, and implemented approach to synthetic anomaly generation are provided.

A. Data and Test Sites

The study used multisource and multitemporal VHR satellite imagery sensed from FDP settlements located in different geographic areas (continent, climate, and background characteristics). The images and corresponding annotations are stored in an in-house image database [54], which is built as part of a long-term engagement in EO-based humanitarian emergency response. Annotations were generated both with manual digitization and OBIA approaches. The overall details of utilized imagery concerning the sensor, date, and resolution are indicated in Table I. The FDP settlements are purposely selected to test the performance of our proposed approach in areas that have diverse backgrounds environmental characteristics and varying levels of dwelling object complexity in terms of the object properties (size and shape), spatial patterns (density and distance to neighboring buildings), and spectral characteristics, which mostly are governed by the material they built from and location-specific biophysical factors.

As shown in Fig. 1, the dwelling objects in Minawao and Nduta FDP settlements are dominated by standard UNHCR tents that have round- or dome-shaped dwellings. In Kutupalong, the FDP site is characterized by complex terrain dominated by densely populated dwellings with diverse spectral characteristics. With a relatively similar level of complexity, a Degahaley FDP site is characterized by complex dwellings occupying the landscape with an irregular spatial pattern where vegetation is also an integral part of the background environment. Trees are planted as fences, and some parts of dwelling rooftops are also covered by tree crowns. Attenuation of dwelling objects with single-standing trees is common in Nguenygiel and Nduta FDP settlements. Zamzam has unique dwelling structures where a cluster of dwellings is situated within a fence or wall. The dwellings were also a mix of a few bright dwellings with corrugated iron sheet rooftops and dominant low-contrast dwellings. For each image patch used for testing, corresponding annotations were obtained from the same database as images [54]. These annotations are made either by manual digitization by people with domain expertise or the OBIA approach followed by proper postprocessing and quality control.

Pléiades images obtained from Airbus are ortho-ready images with proper geometric and radiometric calibration [25]. Those images were passed through correction for local terrain effect and pan-sharpening to make use of high-resolution panchromatic images. The images are processed by this corrective pipeline before being stored in the database.

Using a VAE for anomaly detection assumes that a trained VAE can properly learn the latent representation of normal images and reconstruct them back from latent code. During the inference phase, the model is expected to yield less reconstruction error on normal images and higher reconstruction images in the anomalous part of the unseen images during the training phase. Therefore, the proper definition of the normality or abnormality of image chips is essential for data preparation. Accordingly, by adopting the conceptualization presented in [42], in this study, dwelling objects in FDP settlements are anomalies, and image patches containing dwellings are considered anomalous images. Based on this, training images are images without any dwelling objects, while testing images are images of chips with dwelling objects. Training image chips were generated from empty areas near or on the outskirts of FDP settlement areas. It should be noted that there is a nonnull probability of encountering houses on the outskirts of refugee settlements; therefore, at most care is given to exclude any houses and building structures in those areas. Testing data are prepared from images within FDP settlements. Both testing and training image chips have a dimension of 256×256 pixels. For test datasets, annotations obtained in the Environmental Systems Research Institute (ESRI) shape file format were converted to binary raster with a similar raster profile and coregistered with corresponding VHR imagery. Then, both VHR imagery and rasterized annotations were converted to image chips with similar dimensions to training chips.

B. Synthetic Anomaly Generation for Self-Supervision

Unfortunately, VAEs might be able to reconstruct dwellings although the latter has been hidden during the training phase.

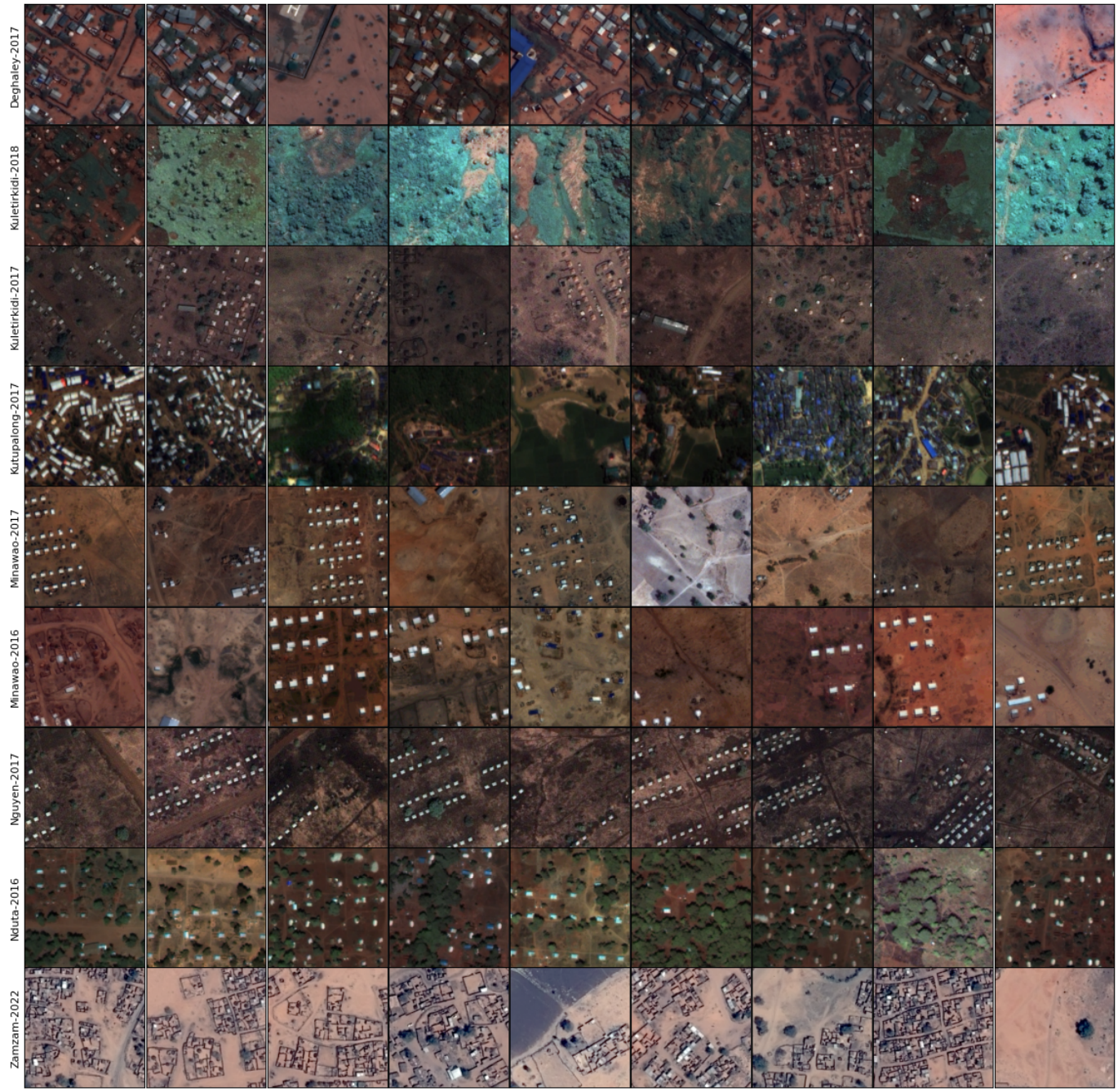

Fig. 1. Variations of dwelling properties within and between various FDP settlements. For visual quality, each image chip is scaled using channelwise (min–max normalization).

We try to overcome this issue with self-supervised learning using synthetic masks. The use of synthetic anomalies in self-supervised anomaly detection has shown promising results. In this regard, Li et al. [35] have implemented the cutPaste algorithm for the creation of synthetic anomalies. Inspired by the cutPaste algorithm, Bauer [36] has created synthetic anomalies that mimic defects. Both approaches have proven their performance on the MVtec dataset [55] for anomaly detection. However, direct implementation of this strategy in EO datasets, especially for unsupervised dwelling object counting, is not straightforward. This is mainly because dwellings in the EO datasets neither have any distinctive spatial pattern nor uniform shape and size, which makes it hard to employ synthetic anomaly generation approaches based on a distortion of the pattern by deformation [36] and masking out some image section [55]. We were required to create synthetic anomalies from the test dataset without using any ground-truth annotations.

Therefore, based on expert knowledge, weakly annotated synthetic masks were generated using the thresholding approach on spectral bands and derived indices. For images that have near-infrared (NIR) bands, first, the normalized difference vegetation index (NDVI) was generated. As can be understood from [56], bare land and built-up surfaces exhibit a lower NDVI, mostly lower than 0.3 depending on the season and vegetation characteristics. By enforcing this threshold, a

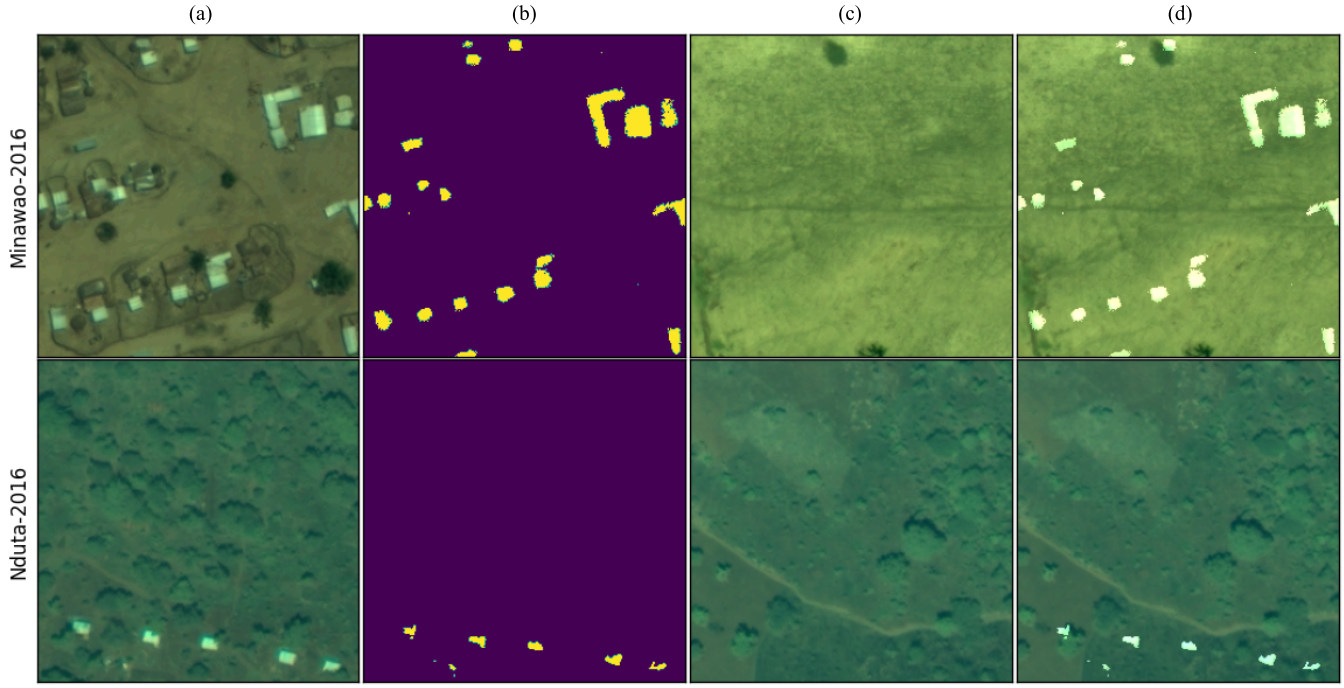

Fig. 2. Synthetic anomaly creation. (a) First image patch is an anomalous input image randomly selected from the test set. (b) Second image patch is a binary mask created from the anomalous image using NDVI thresholding. (c) Normal image without any dwellings, which is also randomly selected from a train set obtained from image regions out of the premises of the FDP settlement. (d) Modified image where masked and cut-out dwelling pixels from anomalous images are pasted on a normal image. During the training phase, synthetic masks and modified images can be used as a conditioning variable. Please note that these are very easy and ideal use cases created with highly tuned NDVI threshold values and selected for visual clarity of the workflow.

2-D binary mask with coarse quality indicating the location of dwellings in test datasets could easily be identified. Then, dwelling objects will be cut out from the test image using this mask and pasted on the normal training image chips as follows.

Assume that we have a random anomalous image chip X_a and a normal image chip X_n. The binary mask $\mathcal{M}$ indicates the background pixels with 1; therefore, its complementary $\bar{\mathcal{M}}$ indicates anomalous locations with 1. Then, the image with synthetic anomalies $\tilde{X}$ is obtained as $\tilde{X} = (X_a \times \bar{\mathcal{M}}) + (X_n \times \mathcal{M})$. As the position where cut-out dwelling objects were pasted is known from $\mathcal{M}$, $\mathcal{M}$ can be considered as a ground truth obtained with self-supervision (see Fig. 2). For datasets that did not have the NIR band, the binary thresholding could be done on a grayscale image either on a selected single channel of the image or any derived index that did not require an NIR channel, which makes the approach versatile to apply on any remotely sensed image. It should be noted that the masks at this level are not expected to be crisp or show exact dwelling footprints. However, even with inexact synthetic anomalies, we can still resort to self-supervised training, which forces the model to not properly reconstruct dwelling objects during the training phase (see Section III-A for formulations and usage of synthetic anomalies).

III. SSCVAE for Unsupervised Object Counting From VHR Imagery

In this section, we will describe a new model, called self-supervised conditional VAE (SSCVAE), and how it is used in the context of unsupervised object localization and counting. We follow the traditional unsupervised anomaly detection approach. In deep learning-based anomaly detection, the procedure requires learning the structure of "normality" from image examples without anomalies [57]. Hence, as we introduced in Section II-A, we have defined normal image patches as image regions that contained either background soil or any other land object except dwelling objects. On the other hand, anomalous image patches are image regions that contain both background and dwelling objects. It should be noted that in a very strict sense, this could be assumed as weak annotations but as far as these annotations were created on the fly as contextual information using strategies detailed in Section II-B, we chose to label the workflow as a self-supervised approach. As indicated in [37], a classical VAE can be trained on normal images to have proper reconstruction skills by maximizing the evidence lower bond (ELBO) loss, which is given as

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta KL\left(q_\phi(z|x)||p_\theta(z)\right) \tag{1}$$

where the first and second terms are the reconstruction and Kullback–Leibler terms, respectively. During the training phase, a normal image patch x is fed into an encoder network parameterized with ϕ, which then produces a compressed latent code z, sampled by reparametrization trick. z is then fed to the decoder network, parameterized with θ, which reconstructs $\hat{x}$. Note that in the original VAE, the Kullback–Leibler divergence (KLD) term, which enforces a prior distribution $p_\theta(z)$ on the latent space, is introduced with $\beta = 1$. However, in most real-world applications, the balance between the reconstruction term and the Kullback–Leibler term needs to

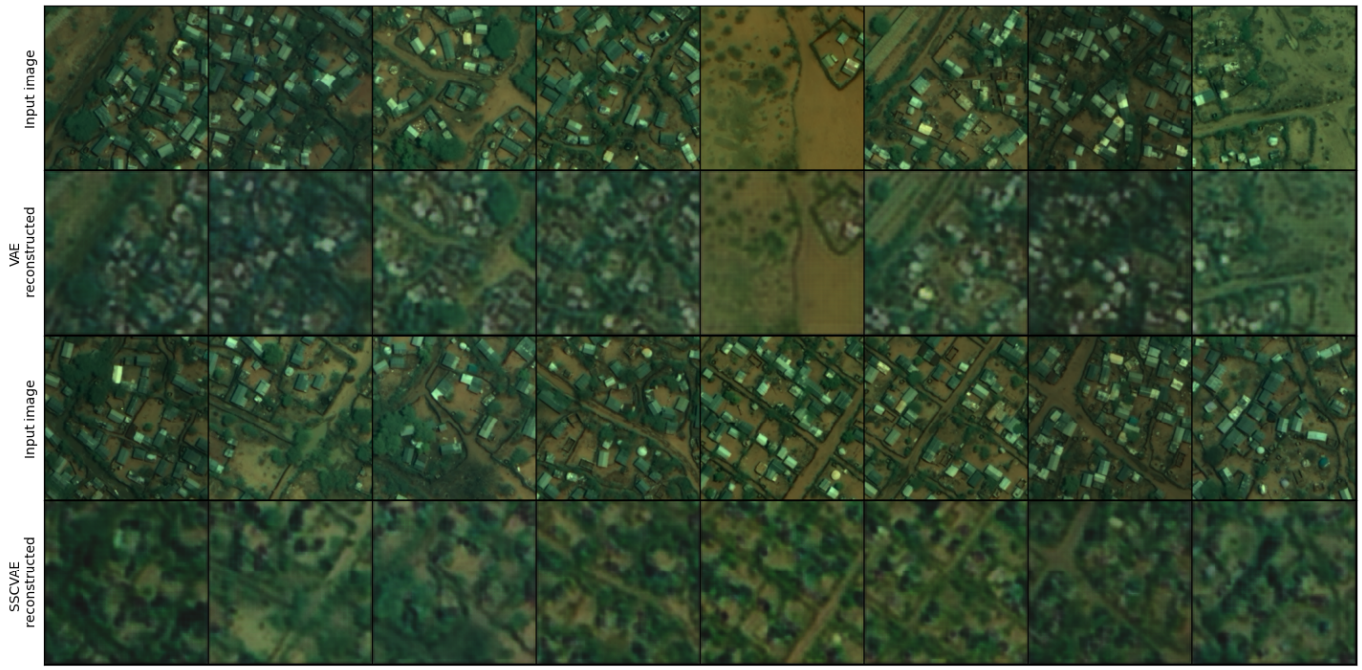

Fig. 3. Image reconstruction with and without self-supervision for randomly selected image chips from the FDP site with complex dwelling structures—Dagahaley. The first image resulted from reconstructions made with classical VAE, while the second image resulted from self-supervised VAE using synthetic anomalies and resultant masks. The first rows are for raw images, while the second rows are for reconstructed images.

be carefully studied [58]. In our experiments, we empirically chose the optimal β value by a grid search approach on a logarithmic scale.

As indicated in Section I, when applied to remotely sensed images, we have observed that classical VAE has two limitations.

1) Despite the latent space compression, the VAE can easily reconstruct relatively bright dwellings that yield anomaly score maps with poor quality to properly localize dwelling objects (see Fig. 3 for comparison).
2) When the VAE is trained with the data collected from different geographic areas and/or different times (see Fig. 1 for some randomly selected image chips taken at different times and places), the model is not properly learning good latent code. In this work, we propose a VAE model that tends toward spatial and temporal invariance. This invariant nature of the model is intended to leverage datasets across different geographic areas and time stamps when there is no sufficient number of normal images for a specific dataset.

For the first problem, we resort to self-supervision in a conditional VAE model. For the second problem, we have implemented latent space conditioning using dataset-level labels (IDs) as a covariable. We now present how these two approaches are combined. The concise graphical summary of the overall pipeline is indicated in Fig. 4.

A. Self-Supervised Conditioned VAE Model

We now refer to our new model as SSCVAE, and we present it mathematically. The model is composed of four main variables.

1) t, a fixed covariable that represents the dataset the image belongs to, with $t \in \{1, \ldots, T\}$ when the model is trained on a total dataset that mixes T (sub)-datasets.
2) x, a vector of an independent random variable with real values whose realizations correspond to the pixel intensities; we have $p(x) = \sum_{i=1}^{N} p(x_i)$, where i describes the N image pixel.
3) z, a vector random variable with real values whose realizations correspond to the hidden units of the latent space; we have $p(z) = \sum_{j=1}^{M} p(z_j)$, where j describes the M latent hidden units.
4) y, a vector of independent Bernoulli random variable (discrete values in $\{0, 1\}$); its realizations describe the mask of the anomalies that provide auxiliary information during training. As earlier, $p(y) = \sum_{i=1}^{N} p(y_i)$, where $y_i = 0$ at a normal pixel position and $y_i = 1$ at an abnormal pixel position.

The SSCVAE is trained according to the following ELBO that we have to maximize with respect to θ and ϕ:

$$\mathcal{L}_{x,y|t}(\theta, \phi) = \mathbb{E}_{z\sim q_\phi(z|y,x,t)}\left[\log p_\theta(x|y,z)p_\theta(y)\right] - \beta KL\left(q_\phi(z|x,y,t)||p(z)\right) \quad (2)$$

where β is a hyperparameter, which has been discussed in the previous paragraph. In the previous formula, it appears that we will not directly use the conditioning on t in $q_\phi(z|x, y, t)$. The model is then composed of the generative network

$$p_\theta(x, y, z) = p(y)p(z)p_\theta(x|y, z) \quad (3)$$

with

$$\begin{cases} p(y) = \prod_{i=1}^{N} \mathcal{B}(y_i; \pi_i) \\ p(z) = \mathcal{N}(z; 0_M, I_M) = \prod_{j=1}^{M} \mathcal{N}(z_j; 0, 1) \\ p_\theta(x|y,z) = \prod_{i=1}^{N} \mathcal{CB}(x_i;\, f_\theta(y,z)) \end{cases} \tag{4}$$

where $\mathcal{B}, \mathcal{N}$, and $\mathcal{CB}$ refer, respectively, to the Bernoulli, Gaussian, and Continuous Bernoulli distributions, while π_i indicates the conditional probability of a pixel to be foreground or background. Notably, the output of f_θ, a neural network parameterized by θ, will play the role of the parameter of the conditional likelihood $\mathcal{CB}$ distribution.

The inference network is written as

$$q_\phi(z|x,y,t) = q_\phi(y|x,t) q_\phi(z|x,y,t) \tag{5}$$

with

$$\begin{cases} q_\phi(y|x,t) = \prod_{i=1}^{N} \mathcal{B}\big(y_i; h_\phi(x)\big) \\ q_\phi(z|y,x,t) = \prod_{j=1}^{M} \mathcal{N}\Big(z_j; \big(g_\phi(y,x)\big)_j\Big) \end{cases} \tag{6}$$

where h_ϕ is a neural network that parameterizes the Bernoulli distribution and g_ϕ is a neural network with two outputs in $\mathbb{R}^M$ and $(\mathbb{R}_*^+)^M$ representing, respectively, the vectors of the means and the variances of the independent Gaussian distributions. Now that we have introduced the full model; we focus successively on its two particular properties.

B. Self-Supervision

The main idea with self-supervision in VAE is to generate contextual information by leveraging prior knowledge from vast unlabeled data during the training or pretraining phase [59]. In related studies, this has been implemented in different forms such as transformation [60], corruption [61], partial masking [62], [63] of input images, and the introduction of random noise and crafting of pseudolabels [64] that all are designed to fit a specific application. In this study, self-supervision is used to expose the model to anomalies as early as during the training step. To do so, we augment the training samples with real anomalies taken from the test set by adapting the workflow from [35]; the details about this step are described in Section II-B. Such an approach exhibits two main advantages. The first is that the model has already been exposed to the anomalies during the training phase; therefore, the model outputs during the test phase will be more stable and are expected to produce better anomaly score maps. The second advantage is that we can theoretically force the model to poorly reconstruct the anomalies that it has been exposed to by adapting the loss function. This approach has been studied in [36] in the context of simple autoencoders.

Interestingly, we can draw a link between the autoencoder loss function [36] and the SSCVAE loss function [see (2)]. In particular, the first term reads $\mathbb{E}_{z\sim q_\phi(z|y,x,t)}[\log p_\theta(x|y,z)p_\theta(y)]$, and it will be classically approximated by Monte Carlo sampling

$$\begin{aligned} &= \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y) p\big(x|y,z^l\big), \quad \text{where } z^l \sim q_\phi(z|x,y,t) \\ &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{M} \log \mathcal{B}(y_i;\pi_i) + \log p\big(x_i|y_i,z_i^l\big) \end{aligned} \tag{7}$$

with

$$\mathcal{B}(y_i;\pi_i) = \begin{cases} 1-\pi_i, & y_i = 0 \\ \pi_i, & y_i = 1. \end{cases} \tag{8}$$

We can further write the reconstruction term as

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{M} \begin{cases} \log(1-\pi_i) + \log p_\theta\big(x_i|y_i,z_i^l\big), & y_i = 0 \\ \log(\pi_i) + \log p_\theta\big(x_i|y_i,z_i^l\big), & y_i = 1 \end{cases} \tag{9}$$

with $z^l \sim q_\phi(z|x,y,t)$. By considering $L = 1$ as it is classically the case in the VAE literature [37], the reconstruction term becomes

$$\sum_{i=1}^{M} \big[1_{y_i=0} \log(1-\pi_i) p_\theta\big(x_i|y,z^l\big) + 1_{y_i=1} \log(\pi_i) p_\theta\big(x_i|y,z^l\big)\big]. \tag{10}$$

As we are dealing with images, we can rewrite the previous expression by introducing the matrices $\text{Rec}_{i,j} = \log p_\theta(x_{i,j}|y,z^l)$ and

$$P_{i,j} = \begin{cases} \log\big(1-\pi_{i,j}\big), & y_{i,j} = 0 \\ \log\big(\pi_{i,j}\big), & y_{i,j} = 1. \end{cases} \tag{11}$$

Finally, as indicated in Section II-B, let $\mathcal{M}$, respectively, $\bar{\mathcal{M}}$, be the binary mask indicating the normal image pixels, respectively, modified image pixels, i.e.,

$$\mathcal{M}_{i,j} = \begin{cases} 0, & y_{i,j} = 0, \\ 1, & y_{i,j} = 1 \end{cases} \quad \text{and } \bar{\mathcal{M}}_{i,j} = 1 - \mathcal{M}_{i,j}. \tag{12}$$

We can write

$$\begin{aligned} &\mathbb{E}_{z\sim q_\phi(z|y,x)}\big[\log p_\theta(x|y,z) p_\theta(y)\big] \\ &\qquad = \|\bar{\mathcal{M}} \odot (P + \text{Rec})\|_1 + \|\mathcal{M} \odot (P + \text{Rec})\|_1. \end{aligned} \tag{13}$$

We fall back on a self-supervised loss similar to that of [36], up to the KLD term. The latter is a regularizing term proper to the VAE model, which is missing in autoencoders.

C. Latent Space Conditioning Using Covariable

In this section, we describe, to the best of our knowledge, an original use of a Conditional VAE model. The latter family of models has been introduced in [65] and [66] and can take very diverse forms. In our case, we want to address the problem of anomaly detection when training the model on a dataset that is an aggregate of subdatasets, i.e., satellite images with different spatial locations and time stamps.

Recall the conditioning covariable t in the model [see (2)] available at training time; it represents the subdataset that the

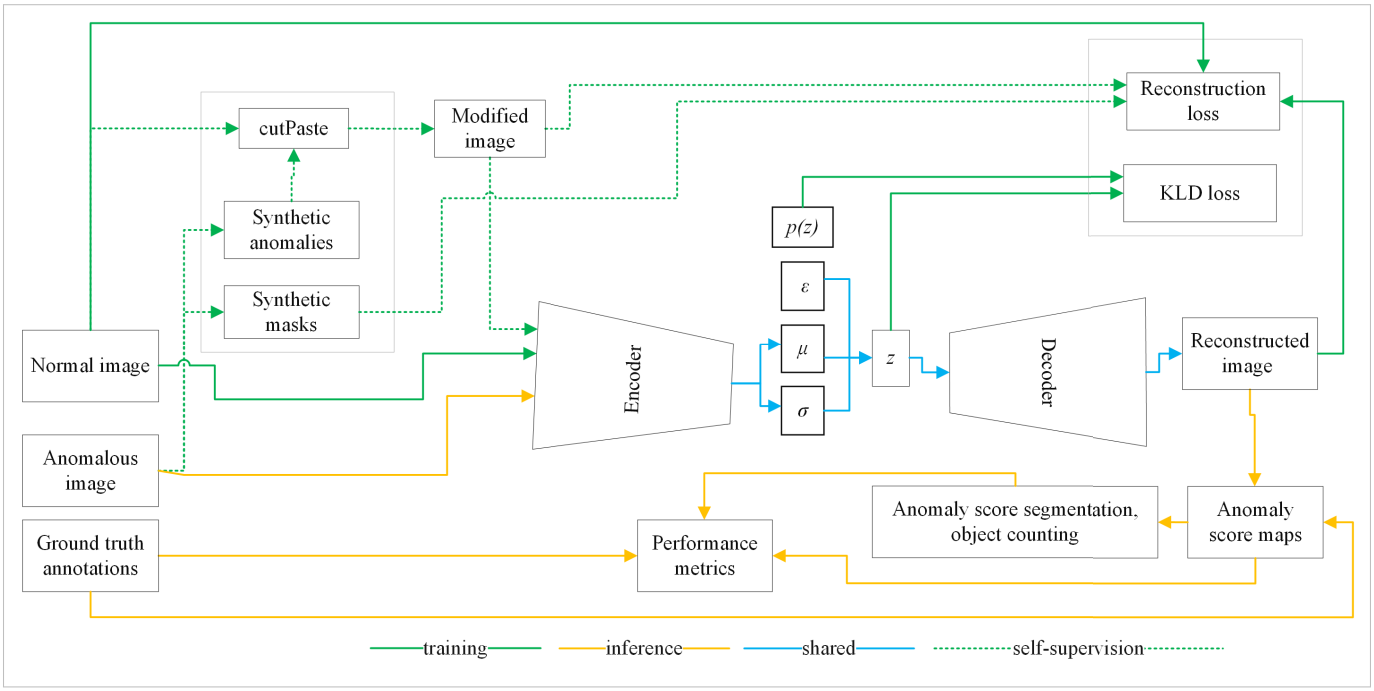

Fig. 4. Overall workflow for self-supervised VAE for dwelling detection. Please note that the dotted green lines are also for the training phase during self-supervision.

sample belongs to. Before going further, we need to be explicit about the structure of the latent space. The SSCVAE has a convolutional latent space, which means that the latent space is composed of M_C latent space images of size $M_W \times M_H$, with $M = M_C \times M_W \times M_H$ (similar architectures have been proposed in [53] and [67]). Now, if we have T subdatasets, we then propose to divide the M_C dimensions of the latent space such that a training step of the SSCVAE model only updates a particular subset M_t of latent dimensions with size M_C/T. This is what we call *latent space conditioning*.

Later on, at test time, the second step of our original approach consists of *test time averaging*. The test sample is fed into the encoder, and the corresponding latent space is computed for the M_C latent images, but then, before reconstructions, all the M_C latent images have their value set to their average value (average computed on the latent pixel level). We empirically found out that this approach enabled reconstruction in a unified latent space, which seemed to exhibit space and time invariance despite the heterogeneity of the complete training dataset. However, the reconstructions after this operation of the test time average still enable us to recover the anomalies that we want to detect. At the same time, it yields lower reconstruction error, and the latent space exhibits a more Gaussian structure. As we will see, compared with a classical VAE trained on such a heterogeneous dataset, the SSCVAE shows performance gain both in localization and counting.

D. Anomaly Score Generation and Dwelling Count

In this section, we describe how dwellings are localized as anomalies using the input and the reconstructed image. We here resort to the classical approach of unsupervised anomaly detection: in the reconstructed image, dwelling objects are supposed to disappear or be poorly reconstructed as their representation is not learned. Anomaly scores were generated using the structural similarity index (SSIM) [68], which is provided in (14). The structural similarity accounts for the contrast, brightness, and texture of the input and reconstructed images at the specified sliding windows and is provided as

$$\begin{aligned}\mathrm{SSIM}(x, \hat{x}) &= \mathrm{SSIM}(r_i, p_i) \\ &= \frac{(2\mu_r\mu_p + C_1)(2\sigma_{pr} + C_2)}{(\mu_p^2 + \mu_r^2 + C_1)(\sigma_r^2 + \sigma_r^2 + C_2)}\end{aligned} \tag{14}$$

where μ, σ, and σ^2 indicate the mean, standard deviation, and covariance of reconstructed r and predicted p images, respectively, at pixel location i with a certain window size. The proper window size is selected to balance the strength of the anomaly score and crisp dwelling boundaries (see Section III-D and Fig. 10 for visual understanding of anomaly score sensitivity for window size w). It should be noted that as SSIM values and image similarities are linear, the anomalies are further computed as dissimilarities ($1 - \mathrm{SSIM}$).

Provided that structural similarity works in a sliding window, it fails to yield distinct dwelling boundaries in patches that contain densely packed dwellings and very small dwelling structures. Plausible anomaly scores were also generated by using channelwise mean absolute deviation (MAD) between input and reconstructed images

$$\mathrm{MAD} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \hat{x}_i\right| \tag{15}$$

where N is the number of channels and x and $\hat{x}$ are input and reconstructed images for specific channel i. For visual quality

TABLE II
DWELLING LOCALIZATION OF DIFFERENT VAE VARIANTS BASED ON AUC VALUE

FDP site	[83]	[82]	[36]	MAD			SSIM		
				VAE	SSCVAE	SSCVAE$_{lc}$	VAE	SSCVAE	SSCVAE$_{lc}$
Dagahale-2017	0.88	0.78	0.84	0.65	0.73	**0.80**	0.53	0.76	**0.79**
Kuletirkidi-2018	0.91	0.75	0.69	0.60	0.66	**0.71**	0.74	0.85	**0.88**
Kuletirkidi-2017	0.88	0.83	0.61	0.66	0.66	**0.72**	0.56	0.90	**0.92**
Kutupalong-2017	0.73	0.71	0.48	0.58	0.63	**0.65**	0.63	**0.70**	0.69
Minawao-2017	0.90	0.90	0.48	0.81	0.76	**0.82**	0.66	**0.96**	**0.96**
Minawao-2016	0.95	0.94	0.56	0.67	**0.91**	0.90	0.25	**0.96**	**0.96**
Nguenygiel-2017	0.97	0.93	0.70	0.87	0.80	**0.89**	0.57	0.97	**0.98**
Nduta-2016	0.92	0.93	0.65	0.74	0.87	**0.89**	0.44	**0.93**	**0.93**
Zamzam-2022	0.83	0.73	0.79	**0.58**	0.36	0.33	0.51	**0.81**	0.74
Average	0.86	0.82	0.66	0.68	0.71	**0.75**	0.54	**0.87**	**0.87**

TABLE III
DWELLING COUNTING PERFORMANCE OF DIFFERENT VAE VARIANTS BASED ON MAE VALUE

FDP site	[83]	[82]	[36]	MAD			SSIM		
				VAE	SSCVAE	SSCVAE$_{lc}$	VAE	SSCVAE	SSCVAE$_{lc}$
Dagahale-2017	4.73	20.91	17.75	35.91	19.99	**16.40**	**11.66**	23.90	17.21
Kuletirkidi-2018	12.66	22.61	23.16	39.84	**26.52**	30.76	23.30	**22.73**	23.12
Kuletirkidi-2017	15.10	26.89	25.14	34.56	42.36	**35.65**	28.34	24.13	**22.60**
Kutupalong-2017	59.64	68.39	53.19	**37.27**	44.41	43.46	**54.79**	61.84	56.67
Minawao-2017	13.20	35.83	22.71	32.26	36.00	**30.22**	**25.35**	52.01	50.57
Minawao-2016	3.30	9.36	15.75	25.63	7.04	**6.76**	12.34	5.89	**5.03**
Nguenygiel-2017	2.56	21.76	7.78	**29.66**	51.80	38.36	16.53	**16.40**	16.65
Nduta-2016	5.80	11.56	22.00	32.27	15.29	**14.38**	16.79	11.99	**11.00**
Zamzam-2022	32.39	46.34	30.62	52.29	**18.89**	21.42	56.14	43.94	**40.36**
Average	16.59	31.79	24.22	35.52	29.14	**26.38**	27.25	29.20	**27.02**

and fair comparison of different VAE implementations, the MAD and SSIM anomaly scores were standardized to value range [0, 1].

Dwelling object counting is done after converting anomaly scores to dwelling object instances. As anomaly score maps are continuous intensity maps, the main challenge in counting objects is drawing a single decision boundary that segregates anomalies from the background from the foreground (dwellings) pixels. Approaches such as threshold-based could be applied but are highly challenging to optimize a single threshold for the entire dataset. From the intention to create a more objective approach, in this study, we have followed a threshold-free, unsupervised clustering using a Gaussian mixture model (GMM) [69], [70], which is also common for unsupervised land cover classification [71], [72], [73] and target detection [74] from the EO dataset. Initial cluster centers were determined using k-means clustering with two classes (background and dwelling). The implementation is forked from [75]. Then, small false positives were removed using morphological binary opening. This morphological operation was mainly selected as there are false positives from the background especially for dwelling instances obtained from MAD-based anomaly score maps. Then, the cleaned binary mask is converted to dwelling instances where dwelling counting is performed.

E. Evaluation Metrics

The quality of anomaly maps to localize dwellings is evaluated with a threshold-free area under the receiver operating characteristic (ROC) curve (AUC) [76], which is also a common evaluation metric for unsupervised anomaly detection [77], [78]. The performance of dwelling count is evaluated using mean absolute error (MAE) between counts from the model and reference data, which is provided as

$$\text{MAE} = \frac{1}{N}\sum_{i}^{N}\left|y_i - \bar{y}_i\right| \tag{16}$$

where y_i and $\bar{y}_i$ indicate the reference and predicted counts per image chip, while N is the total number of image chips in a specific dataset.

IV. EXPERIMENTS AND RESULTS

A. Model and Experimental Setup

A VAE is a model with an encoder–decoder architecture that we now describe. For feature extraction, an encoder network with ResNet [79] architecture with layer depth 34 is used as follows. The first four layers of ResNet34 are stacked between the entry and encoder exit layers. The entry layer consists of a 2-D convolution with a kernel size of 7 followed by batch normalization, Rectified Linear Unit (ReLU) activation, and maxpooling. Then, in the classical VAE implementation that we propose and in the VAE with self-supervision only, the exit layer consists of both a unit kernel and strides with padding size of 0 and yields a feature map with a spatial dimension of $M_W \times M_H = 32 \times 32$ with depth of $2 \times M_C = 2 \times 256$ sized feature; with $M = M_C \times M_W \times M_H$ being the total number of hidden random variables. The latter feature map is then a vector representing the mean and variance of a convolutional latent space (see Fig. 4). In the proposed SSCVAE with latent space conditioning, we set that one input from a particular dataset updates only $M_t \times M_W \times M_H$ latent random variables with half of them accounting for the mean and the other half for the variance. Therefore, for the SSCVAE, we have $M = T \times M_t \times$

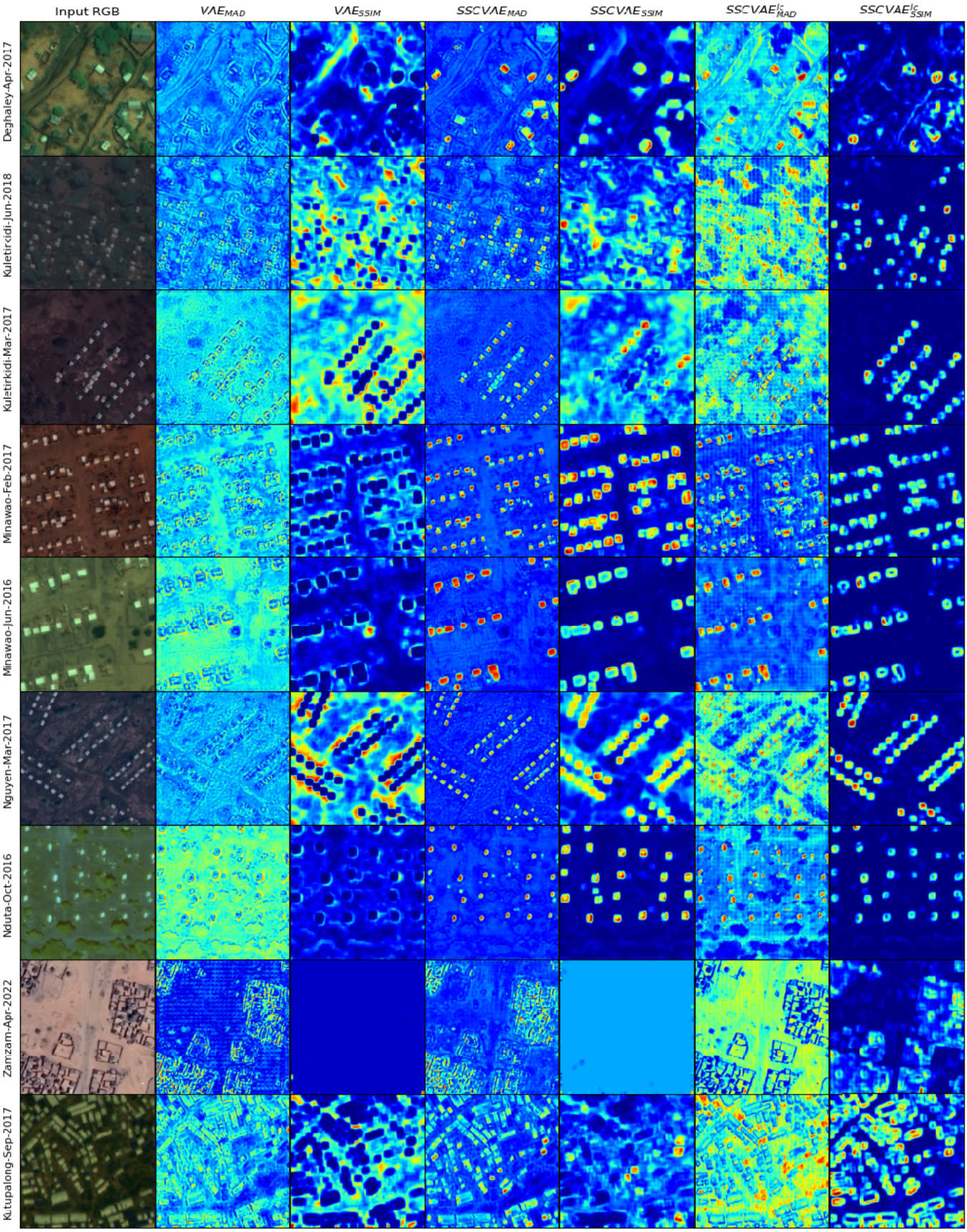

Fig. 5. Spatial plots for unsupervised localization of dwellings using VAE and its variates using anomaly scores created from MAD and SSIM; SSCVAE^{lc} stands for SSCVAE with latent space conditioning and the subscripts indicates the anomaly score generation approach followed.

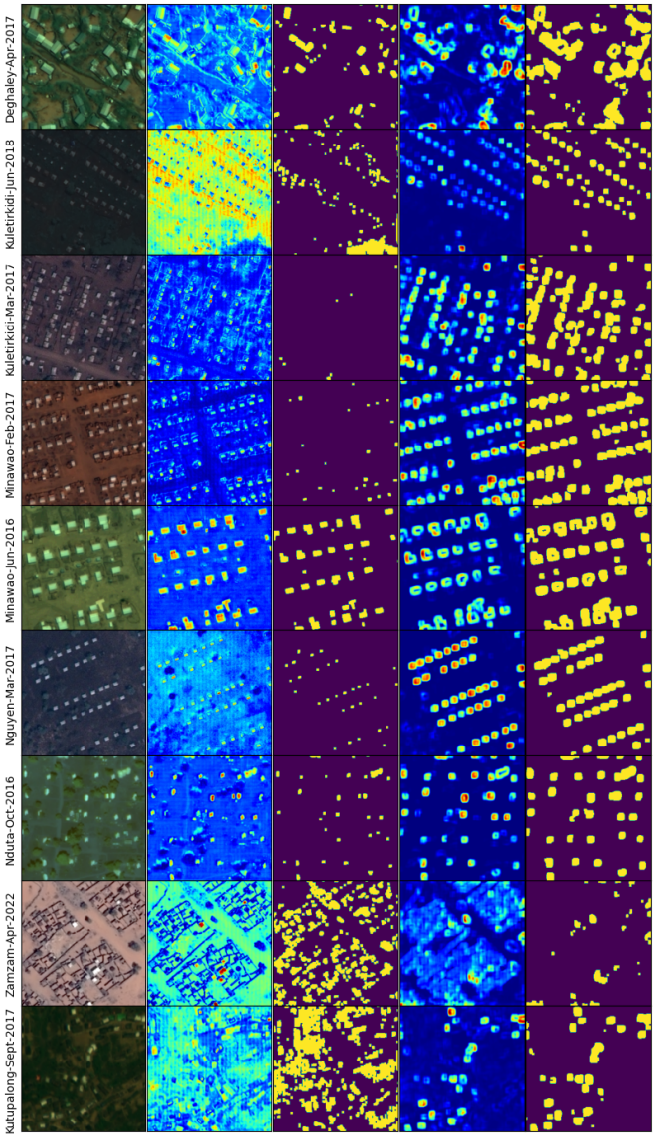

Fig. 6. Dwelling hard binary classes obtained after unsupervised clustering of anomaly score maps. The first column is the input image, and the second and third columns are anomaly scores from MAD and respective dwellings, while the fourth and fifth columns are anomaly scores from SSIM and respective obtained binary dwelling classes. The obtained mask was a result of a postprocessing operation with a morphological opening for two iterations.

$M_W \times M_H$; hence, the relation $M_C = T \times M_t$ given in Section III-C. If we decided the depth of $M_t = 4$; for the number of datasets $T = 9$, the dimension of latent space would become $M_C = 36$. Then, the latent space fed into the decoder would have a size of $36 \times 32 \times 32$.

For the decoder network, the same number of layers with transposed convolution is used. This is mainly for spatial upsampling and reconstructing an output image with the same spatial dimension as an input image. The entry layer of the decoder is composed of transposed convolution with a unit kernel and stride and padding size of 0. Except for the second layer that is only a transposed convolution, the rest of the decoder layers are composed of transposed convolution followed by batch normalization and ReLU activation.

The model is optimized with a learning rate of 10^{-4} and stochastic gradient descent (SGD) optimizer. Following the works of Fu et al. [80], we have tried to optimize β in (1) using cyclic annealing. Though cyclic annealing produced better performance than setting $\beta = 1$, annealing to a value less than 0.01 does not improve the results (see Fig. 12). We found that using a smaller and fixed $\beta = 10^{-4}$ is better, and the subsequent results presented in Section IV-B are obtained using this value. Similar to our choice, setting minimal KLD weight is also reported effective to yield minimum reconstruction on the application of conditional VAE for language translation [81]. It should be noted that all experiments are run with a mixed dataset setup where each dataset obtained from different geographies and time stamps (see Table I) are combined as one.

To compare our results with other models in the literature, we have selected one model from each of the following anomaly detection approaches: reconstruction-based [36], embedding similarity-based [82], and supervised density-based [83], [84]. Then, we reimplemented the models with some marginal modifications required to obtain the best results. To train the supervised density-based approach, we followed an iterative leave-one-dataset approach where the model is trained on eight datasets and predicted on a left-out dataset. Recall from Section II-A that we have at our disposal the ground-truth locations for dwellings; however, the supervised model is the only one in which the ground truths are used. Note that while implementing [36], we did not apply Gaussian filtering of anomaly scores as it reduced the performance. Except for [82] that is run on a machine equipped with six Intel Xeon Processor CPUs, the overall experiment is done on a computer equipped with a single NVIDIA GeForce RTX 3090 GPU.

B. Results

1) Anomaly Localization: When trained and tested on individual datasets, VAE could properly localize dwellings. Such a study (VAE trained with and tested on individual datasets) has been made in [85]. When datasets obtained from different places and times are combined and subjected to joint training, the classical VAE fails to reach localization performances obtained from training a model using individual datasets. This is the setting studied in this article; in all the following, the results are from mixed dataset training.

As can be indicated in Table II, except for the Zamzam-2022 dataset, classical VAE has yielded lower localization. This is true for both anomaly score maps obtained from MAD and SSIM. Then, the localization performance is improved by SSCVAE and further improved by the introduced latent space conditioning using the dataset label as a covariable. Beyond performance deviations among models, it is also clearly observed that there is performance variation between localization results obtained from different anomaly score generation approaches: MAD and SSIM. For classical VAE, for most datasets, MAD provided better AUC values than SSIM, while for SSCVAE, AUC values obtained from anomaly score maps generated from SSIM have yielded relatively better localization results. It should also be noted that the SSIM values could vary with SSIM (w) hyperparameter; therefore, the results presented in Table II are using w values of

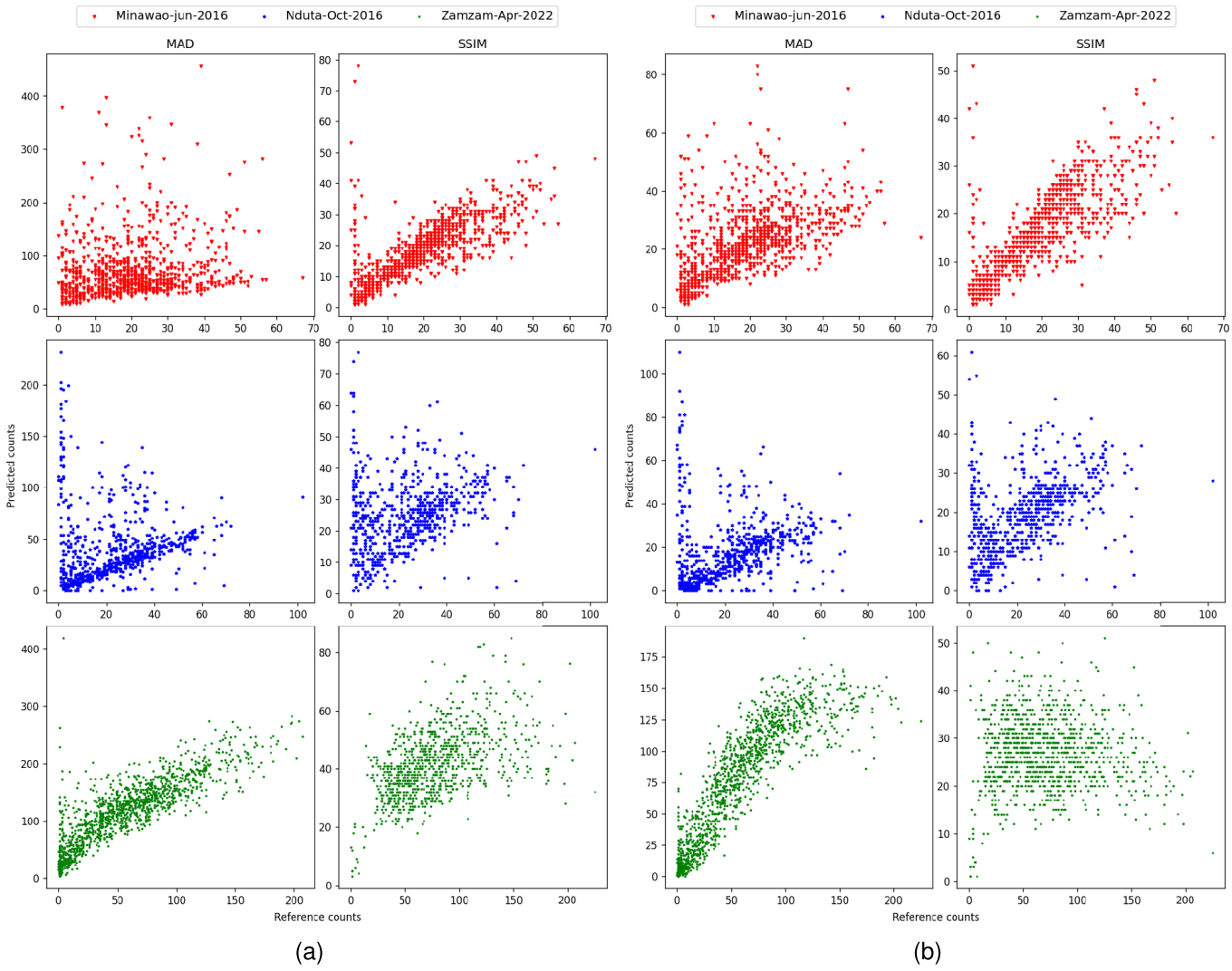

Fig. 7. Comparison of reference and predicted dwelling counts for selected FDPs using (a) and (b) two anomaly score generation approaches for before and after postprocessing, respectively. Please note that results were generated from SSCVAE with latent space conditioning. Overestimation and underestimation patterns among SSIM and MAD are almost the same for VAE and SSCVAE.

11×11 pixels. The best localization results were obtained from SSCVAE$_{lc}$, which range from the AUC values of 0.74 for the Zamzam-2022 dataset to 0.98 for the Nguenygiel-2017 dataset.

As can be seen from Fig. 1, the Zamzam-2022 dataset has fences and dwellings mostly made from mud (including most dwellings) confused with an environment dominated by dry bare soil, which has a spectral resemblance with houses. This makes it quite hard for the models to properly localize the dwellings. The minimum localization performance for MAD-based anomaly score maps is also observed in the same dataset. As can be seen from the spatial plot map (see Fig. 5), the anomalies both from dwellings and fences are quite strong. This also has its implication on the proper counting of dwellings from anomaly score with further clustering.

From anomaly score maps presented in Fig. 5, we can see three clear patterns.

1) An important variation in the anomaly score quality between classical VAE and the SSVAE. The anomaly scores from classical VAE were poor; especially, the one obtained from SSIM misses very bright dwellings, while the one from MAD is not strong enough to identify dwelling objects from the background counterpart. In addition to this, there is a strong attenuation from the background.
2) A variation between MAD and SSIM-based anomaly score maps where they have tradeoffs between strong anomaly with background suppression and merging closer dwellings in one way and anomaly scores with clear boundaries between neighbor dwellings but with relatively higher false alarm anomaly signals from the background. This could have a strong implication on the quality of generating hard binary classes useful for counting dwellings (see Fig. 6).
3) An anomaly score variation due to the heterogeneity of datasets. In datasets taken from relatively less complex sites such as Minawao, Nduta, and Nguenygiel, obtained anomaly scores from SSCVAE and latent space conditioning counterparts using SSIM are way better than

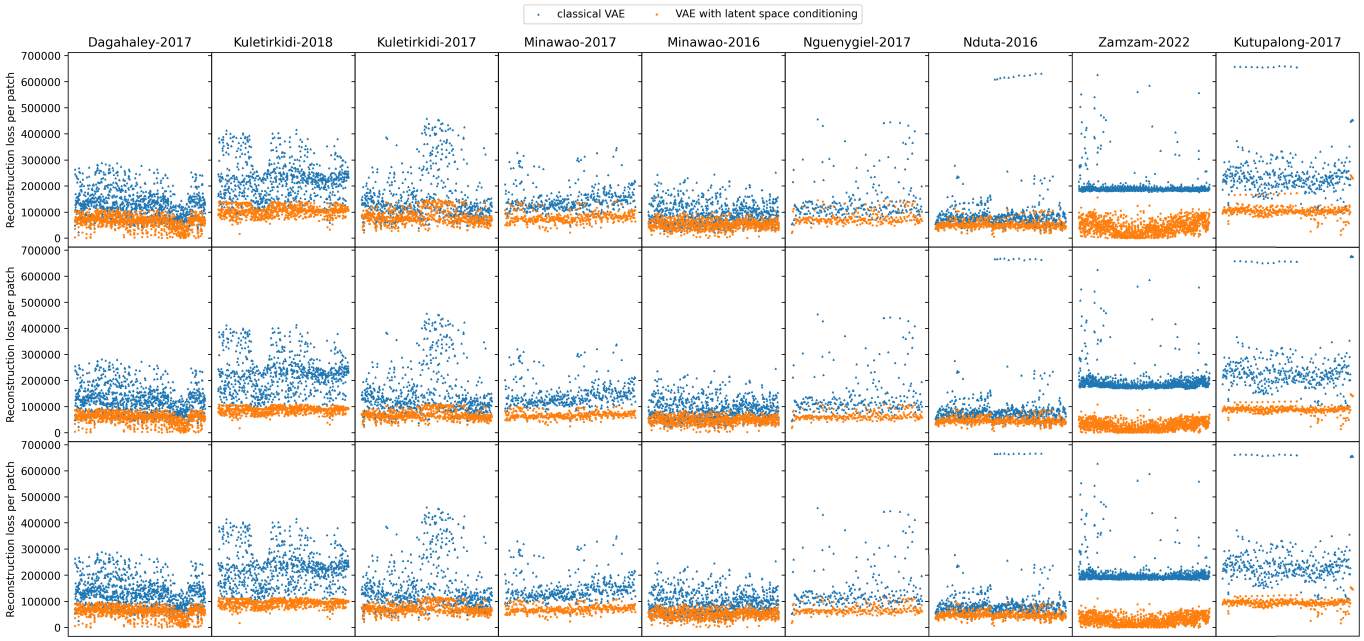

Fig. 8. Reconstruction error term for classical VAE and VAE with latent conditioning and test time averaging.

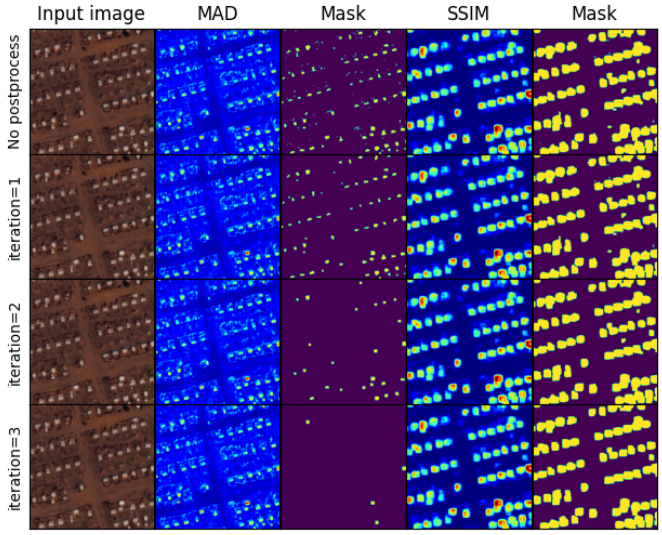

Fig. 9. Effect of postprocessing with the morphological opening on dwelling output quality.

scores obtained in complex sites such as Dagahaley, Kuletirkidi, and Zamzam.

2) Anomaly Counting: As indicated in Table III, similar to localization results presented in Table II, dwelling counts also vary among anomaly score generation strategies, datasets, and the implemented VAE model variates. As expected, the best count performance is obtained from a supervised density-based approach [83], [84]. It should be noted that this approach demands a large number of dot annotations. For example, the Nguenygiel-2017 dataset that achieved an MAE of 2.56 (see Table III) has used 6224 dot annotated image chips with varying numbers of dwelling objects per scene. Except, in a few cases, the introduced self-supervision and latent space conditioning have yielded better results than classical VAE for dwelling object counting. The best object count is obtained in less complex test sites where an MAE of 5.03 in Minawao-2016 followed by an MAE of 11.0 in Nduta-2016 is achieved using SSCVAE with latent space conditioning. Similarly, as indicated in Fig. 6, spatial plots obtained from clustering of anomaly score maps exhibit quality variations in terms of detecting all dwelling objects, proper delineation of individual dwellings, and appearances of false positive predictions.

3) Postprocessing: Note that the postprocessing step also induces some bias in the final results. Accordingly, dwelling counts obtained from score maps generated using SSIM provided relatively smaller MAE values than their counterparts obtained from scores generated with MAD.

Beyond performance deviations, as can be seen from scatter presented in Fig. 7 experiments, before postprocessing, the MAD-based anomaly score yielded an overestimation of counts, which is the result of attenuation from false positive anomalies from the background. Counts obtained from SSIM have relatively smaller underestimations mainly caused by the windowing effect of SSIM, which merges dwellings in very close proximity. Once the postprocessing is introduced, the MAD-based anomaly scores provided almost closer MAE values as morphological opening clears smaller false positive dwelling objects. Even though postprocessing improves counts from MAD-based anomaly scores by a big margin, its effect on counts obtained from SSIM-based anomaly scores is not as favorable as MAD counterparts. It degrades the performance of dwelling count performance. For example, if we compare the results presented in Tables III and V that are with and without postprocessing, respectively, for some datasets (e.g., Dagahale-2017 and Kutupalong-2017), SSIM-based dwelling count performance is slightly lower than

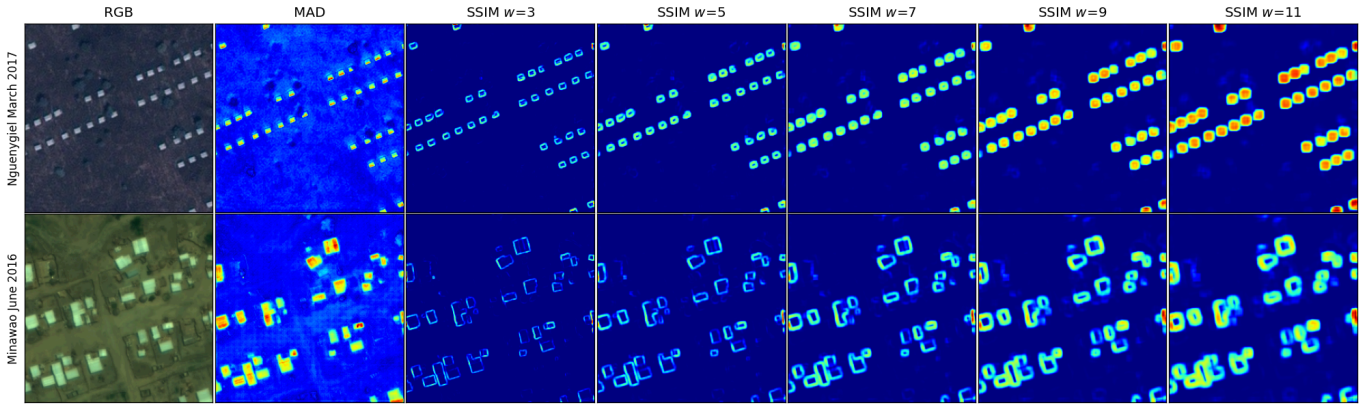

Fig. 10. Sensitivity of anomaly score maps for SSIM window (w) size. As window size increases, the anomaly score is stronger and clearer in dwelling locations, but the boundary between adjacent and closer dwellings becomes closer and closer where finally merging of closer dwellings happens.

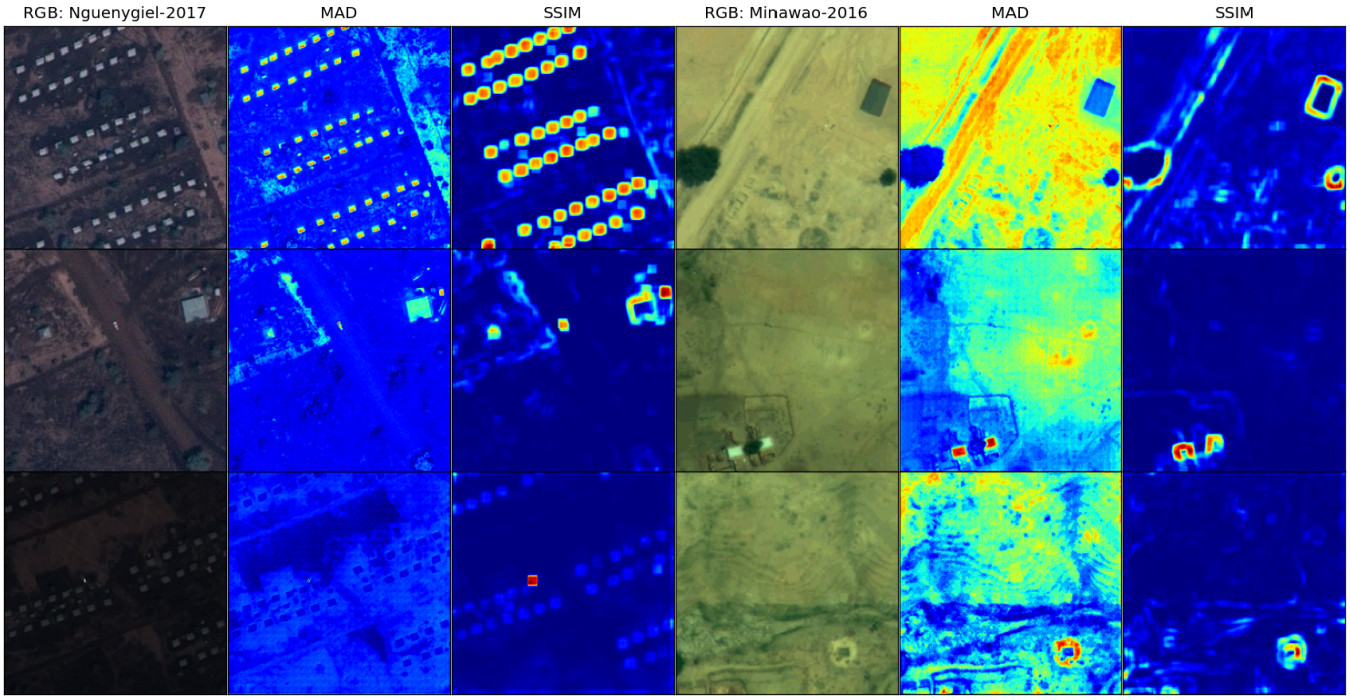

Fig. 11. Examples of poor performance of the VAEs at bare bright surfaces, footpaths, and low-contrast dwellings for anomaly scores both for MAD and SSIM taken from the Nguenygiel March 2017 and Minawao June 2016 datasets.

without postprocessing. Similar behavior could also be seen from Fig. 7(b) for the Zamzam-2022 dataset where counts from the SSIM-based approach after postprocessing are not highly associated with the reference.

V. Discussion

As can be understood from Table II and Fig. 5, the implemented approaches, especially self-supervision and latent space conditioning, have yielded better localization and counting of dwellings. Compared to recent anomaly detection works [86], [87], [88], though the datasets and anomaly localization pipelines are different from this study, our approach has reached the best localization performance where AUC values reach approximately 98% for some datasets (see Table II for Nguenygiel-2017 and Minawao-2016 datasets).

The impact of the quality of the anomaly score map has a direct implication on the counting of individual dwelling instances. As can be seen in Fig. 5, though the strength of anomaly score maps obtained from SSIM is strong enough to properly localize, during the segmentation, the closer dwelling instances get aggregated and become larger dwellings. This, in turn, leads to an underestimation of obtained dwelling instances (see Fig. 7). It is observed that this spatial aggregation of closer dwellings is mainly attributed to the sensitivity of the SSIM-based anomaly score generation approach (see Table IV and Fig. 10). The converse has happened for anomaly score maps obtained from MAD. Even though anomaly score maps look visually good, the signal from the background is not as weak as anomalies in the background of SSIM. This fosters false positive dwelling classes and overestimation

TABLE IV
SENSITIVITY OF AUC VALUES BASED ON SSIM WITH DIFFERENT WINDOW SIZES

	MAD	SSIM				
dataset	-	w=3	w=5	w=7	w=9	w=11
Nguenygiel 2017	0.76	0.91	0.92	0.95	0.96	0.97
Minawao 2016	0.90	0.92	0.93	0.94	0.95	0.95

TABLE V
DWELLING OBJECT COUNTING WITHOUT MORPHOLOGICAL POSTPROCESSING

	MAD			SSIM		
FDP site	VAE	SSCVAE	SSCVAE$_{lc}$	VAE	SSCVAE	SSCVAE$_{lc}$
Dagahale-2017	164.28	26.00	80.56	**13.16**	30.00	14.37
Kuletirkidi-2018	122.11	82.17	105.37	43.51	23.61	**24.27**
Kuletirkidi-2017	238.82	177.33	132.97	32.20	24.31	**22.54**
Kutupalong-2017	67.39	52.16	104.73	52.27	58.83	**51.40**
Minawao-2017	147.73	115.00	132.89	**18.93**	97.00	81.64
Minawao-2016	199.00	12.81	18.89	13.26	5.82	**5.17**
Nguenygiel-2017	307.72	219.67	136.83	19.50	16.48	**16.08**
Nduta-2016	155.35	26.78	28.00	22.13	14.68	**12.49**
Zamzam-2022	190.67	51.42	56.78	55.63	33.86	**24.48**

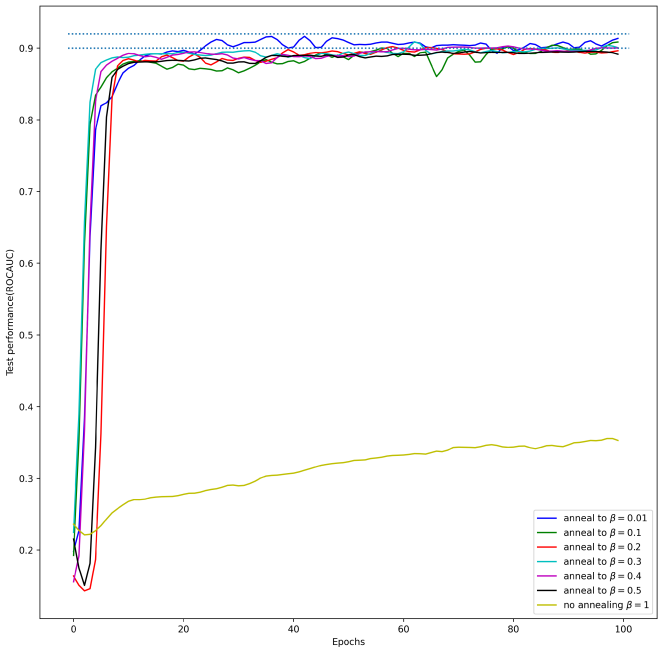

Fig. 12. KLD annealing with different β ceiling values.

(see Fig. 7) of counts. The anomaly score quality variation among l_p norm-based and SSIM anomaly score maps is also reported by Bergmann et al. [78]. They have indicated that SSIM provides salient and strong anomaly score maps with observable boundary distortion of object edges but with better localization results. Contrary to our finding, though they noted minor sensitivity of SSIM values for window size, they, in general, reported the stability of the SSIM approach for other data and model hyperparameters such as input image patch size and model latent dimension. Lack of crisp boundary in anomalies and further segmented results for anomaly detection is also reported in deep learning approaches that did not utilize VAEs [57].

Here, it should be noted that dwelling count results could also be affected by the type of postprocessing with the morphological operation, including the kernel size for smoothing, the number of iterations the operator could be applied, and its interaction with specific dwelling types within a particular dataset. A notable example is provided in Fig. 9 where the quality and number of dwelling objects that appear in the final layer vary as per the postprocessing number of iterations. When the number of cycles the opening is applied increases, the tiny dwellings cease to exist, while, without postprocessing, more noisy false positive predictions appear. This is mostly prevalent in outputs obtained from MAD-based anomaly score maps. Therefore, keeping the right balance between the number of iterations and window size with output quality is an essential aspect.

The overestimation and the underestimation of dwelling counts are also associated with the anomaly generation approach but also with inherent characteristics and spatial heterogeneity of dwelling objects across space and time (see Fig. 1). The small dwellings that have lower contrast with the background environment failed to be spotted in the anomaly score (see Fig. 11). In addition to this, even though contagious buildings appear with strong anomaly score maps, they get merged and appear as single dwelling objects. More importantly, as the training involves datasets from different places with different time stamps, there is also confusion between dwellings in one dataset and resembling nondwelling instances in another dataset. Then, these nondwelling targets such as very bright surfaces and footpaths are flagged as anomalous surfaces (see Fig. 11). False positive localization of nonanomalous targets (better to say targets that are not of interest) is also observed in recent anomaly detection works [86]. We also understood the contribution of the inherent complexity of dwelling structures across space–time by only running the experiment on relatively less complex datasets (Minawao-2017, Minawao-2017, Nguenygiel-2017, and Nduta-2016), and the performance of the model has shown moderate improvement both in localization and counting (achieved MAE of 4.81).

As argued in Section I and provided in Section II, even though the anomaly score maps did not change by a large margin, the latent space conditioning has enabled the creation of a unified feature space. In addition to this, at the same time, the reconstruction error has been reduced for the latent space conditioning (see Fig. 8).

VI. CONCLUSION

In this study, we have demonstrated the application of a VAE for unsupervised dwelling localization and counting using VHR imagery obtained from FDP sites situated in different geographical settings. The critical limitation of VAE for joint training of images obtained from different places at different time stamps, which is the easy reconstruction of bright dwellings, is reduced by the introduction of the synthetic anomaly as a conditioning variable during self-supervision. As datasets obtained from different geographical settings exhibit disparities in the latent space, the spatiotemporal invariance of the VAE latent code is achieved by

introducing a latent space conditioning, i.e., using dataset labels as a covariable. Although we observe a variation in the results, mainly due to each dataset's intrinsic complexity, the introduced self-supervision and latent space conditioning have yielded localization and counting that achieve an AUC value of 98% and MAE value of 5.03 in less complex datasets, respectively. Even though localization results are quite remarkable, dwelling counts suffer from the conversion from anomaly scores to dwelling instances. Therefore, in an operational setting of humanitarian emergency response, our approach could help generate first-hand estimates for the number of dwellings. Then, these results need to be enhanced by proper quality control and postprocessing such as filling of false negative dwellings and cleaning merged dwellings (especially for a product obtained from SSIM-based anomaly score maps).

Given these promising results, further work could focus on the following issues. As introduced self-supervision using synthetic anomalies is proven to detect easy reconstruction of bright dwellings, the introduction of semisupervision with few curated labels that account for contrast and spectral diversity of dwellings could boost the results. Second, to reinforce the spatiotemporal invariance of learned latent code, we will look into recent developments that assume the EO dataset's hierarchical nature with the motivation to model it in hyperbolic latent space (for example, readers can consult a recent work by Hamzaoui et al. [89] on the classification task and [90] for concise review on current advances on hyperbolic neural networks). Third, by leveraging semisupervision, the introduction of a segmentation network to have an end-to-end workflow that also learns the segmentation process while optimizing the VAE could help to bypass the hurdles during unsupervised clustering of anomaly score maps.

ACKNOWLEDGMENT

The authors are very grateful to anonymous reviewers for their insights on the draft paper.

REFERENCES

[1] UNHCR. (2023). *Unhcr Refugee Data Finder: 2022 Refugee Statistics*. [Online]. Available: https://www.unhcr.org/refugee-statistics/

[2] UNHCR. (2023). *Operational Data Portal: Unhcr People of Concern*. [Online]. Available: https://data.unhcr.org/en/geoservices/

[3] S. Lang and P. Füreder, "Earth observation for humanitarian operations," *GI_Forum—J. Geograph. Inf. Sci.*, vol. 3, pp. 384–390, 2015. [Online]. Available: http://austriaca.at/?arp=0x00324a8b, doi: 10.1553/giscience2015.

[4] S. Lang, D. Tiede, D. Hölbling, P. Füreder, and P. Zeil, "Earth observation (EO)-basedex postassessment of internally displaced person (IDP) camp evolution and population dynamics in Zam Zam, Darfur," *Int. J. Remote Sens.*, vol. 31, no. 21, pp. 5709–5731, Nov. 2010.

[5] S. Lang et al., "Earth observation tools and services to increase the effectiveness of humanitarian assistance," *Eur. J. Remote Sens.*, vol. 53, no. 2, pp. 67–85, Jul. 2020.

[6] K. Spröhnle, D. Tiede, E. Schoepfer, P. Füreder, A. Svanberg, and T. Rost, "Earth observation-based dwelling detection approaches in a highly complex refugee camp environment—A comparative study," *Remote Sens.*, vol. 6, no. 10, pp. 9277–9297, Sep. 2014.

[7] D. Tiede, S. Lang, D. Hölbling, and P. Füreder, "Transferability of OBIA rulesets for IDP camp analysis in Darfur," in *Proc. GEOBIA*, 2010, pp. 1–6.

[8] D. Tiede and S. Lang, "Distributed computing for accelerated dwelling extraction in refugee camps using VHSR satellite imagery," in *Proc. GI_Forum*, Salzburg, Austria, 2008, pp. 1–4.

[9] G. Laneve, G. Santilli, and I. Lingenfelder, "Development of automatic techniques for refugee camps monitoring using very high spatial resolution (VHSR) satellite imagery," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, Jul. 2006, pp. 841–845.

[10] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[11] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.

[12] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2D images," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, Art. no. 2032924.

[13] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103910.

[14] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[15] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[16] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, and M. Luengo-Oroz, "Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 376, no. 2128, Sep. 2018, Art. no. 20170363.

[17] Y. Lu, K. Koperski, C. Kwan, and J. Li, "Deep learning for effective refugee tent extraction near Syria–Jordan border," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1342–1346, Aug. 2021.

[18] O. Ghorbanzadeh, D. Tiede, L. Wendt, M. Sudmanns, and S. Lang, "Transferable instance segmentation of dwellings in a refugee camp–integrating CNN and OBIA," *Eur. J. Remote Sens.*, vol. 54, no. 1, pp. 127–140, Feb. 2021.

[19] D. Tiede, G. Schwendemann, A. Alobaidi, L. Wendt, and S. Lang, "Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to COVID-19 response in Khartoum, Sudan," *Trans. GIS*, vol. 25, no. 3, pp. 1213–1227, Jun. 2021.

[20] L. Wickert, M. Bogen, and M. Richter, "Lessons learned on conducting dwelling detection on VHR satellite imagery for the management of humanitarian operations," *Sensors Transducers*, vol. 249, no. 2, pp. 45–53, 2021.

[21] G. W. Gella et al., "Mapping of dwellings in IDP/refugee settlements from very high-resolution satellite imagery using a mask region-based convolutional neural network," *Remote Sens.*, vol. 14, no. 3, p. 689, Feb. 2022.

[22] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[23] M. Xu, M. Wu, K. Chen, C. Zhang, and J. Guo, "The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data," *Remote Sens.*, vol. 14, no. 17, p. 4380, Sep. 2022.

[24] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[25] G. W. Gella, D. Tiede, S. Lang, L. Wendit, and Y. Gao, "Spatially transferable dwelling extraction from multi-sensor imagery in IDP/refugee settlements: A meta-learning approach," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, Mar. 2023, Art. no. 103210. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843223000328

[26] P. Jain, B. Schoen-Phelan, and R. Ross, "Multi-modal self-supervised representation learning for Earth observation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3241–3244.

[27] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, "Contrastive self-supervised learning with smoothed representation for remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] W. Li, H. Chen, and Z. Shi, "Semantic segmentation of remote sensing images with self-supervised multitask representation learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6438–6450, 2021.

[29] K. Heidler et al., "Self-supervised audiovisual representation learning for remote sensing data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103130.

[30] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1421–1430.

[31] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.

[32] V. Stojnić and V. Risojević, "Evaluation of split-brain autoencoders for high-resolution remote sensing scene classification," in *Proc. Int. Symp. ELMAR*, Sep. 2018, pp. 67–70.

[33] S. Vincenzi et al., "The color out of space: Learning self-supervised representations for Earth observation imagery," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3034–3041.

[34] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[35] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9659–9669.

[36] A. Bauer, "Self-supervised training with autoencoders for visual anomaly detection," 2022, *arXiv:2206.11723v1*.

[37] D. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[38] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 415–419.

[39] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[40] Q. Xu et al., "Synthetic aperture radar image compression based on a variational autoencoder," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[41] V. A. de Oliveira et al., "Reduced-complexity end-to-end variational autoencoder for on board satellite image compression," *Remote Sens.*, vol. 13, no. 3, p. 447, Jan. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/3/447

[42] N. Ferreira and M. Silveira, "Ship detection in SAR images using convolutional variational autoencoders," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 2503–2506.

[43] S. Valero, F. Agulló, and J. Inglada, "Unsupervised learning of low dimensional satellite image representations via variational autoencoders," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2987–2990.

[44] W. Yu, M. Zhang, and Y. Shen, "Spatial revising variational autoencoder-based feature extraction method for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1410–1423, Feb. 2021.

[45] Y. Zerrouki, F. Harrou, N. Zerrouki, A. Dairi, and Y. Sun, "Desertification detection using an improved variational autoencoder-based approach through ETM-Landsat satellite data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 202–213, 2021.

[46] C. Tao, H. Wang, J. Qi, and H. Li, "Semisupervised variational generative adversarial networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 914–927, 2020.

[47] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "CVA2E: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5676–5692, Aug. 2020.

[48] Z. Chen, L. Tong, B. Qian, J. Yu, and C. Xiao, "Self-Attention-Based conditional variational auto-encoder generative adversarial networks for hyperspectral classification," *Remote Sens.*, vol. 13, no. 16, p. 3316, Aug. 2021, doi: 10.3390/rs13163316.

[49] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, "Remote sensing image captioning via variational autoencoder and reinforcement learning," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 105920. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120302586

[50] S. Sinha et al., "Variational autoencoder anomaly-detection of avalanche deposits in satellite SAR imagery," in *Proc. 10th Int. Conf. Climate Informat.*, Sep. 2020, pp. 113–119.

[51] J. Zhang, Y. Xu, T. Zhan, Z. Wu, and Z. Wei, "Anomaly detection in hyperspectral image using 3D-convolutional variational autoencoder," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2512–2515.

[52] J. Wei, J. Zhang, Y. Xu, L. Xu, Z. Wu, and Z. Wei, "Hyperspectral anomaly detection based on graph regularized variational autoencoder," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[53] H. Gangloff, M.-T. Pham, L. Courtrai, and S. Lefèvre, "Unsupervised anomaly detection using variational autoencoder with Gaussian random field prior," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 1620–1624, doi: 10.1109/ICIP49359.2023.10222900.

[54] S. Lang et al., "Multi-feature sample database for enhancing deep learning tasks in operational humanitarian applications," *GI_Forum*, vol. 9, no. 1, pp. 209–219, 2021.

[55] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1038–1059, Apr. 2021.

[56] H. Hashim, Z. A. Latif, and N. A. Adnan, "Urban vegetation classification with NDVI threshold value method with very high resolution (VHR) Pleiades imagery," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 237–240, Oct. 2019.

[57] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors*, vol. 18, no. 2, p. 209, Jan. 2018.

[58] I. Higgins et al., "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[59] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[60] I. Gatopoulos and J. M. Tomczak, "Self-supervised variational auto-encoders," *Entropy*, vol. 23, no. 6, p. 747, Jun. 2021.

[61] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[62] J. Chen, M. Hu, B. Li, and M. Elhoseiny, "Efficient self-supervised vision pretraining with local masked reconstruction," 2022, *arXiv:2206.00790*.

[63] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 645–654.

[64] Y. Min and Y. Li, "Self-supervised railway surface defect detection with defect removal variational autoencoders," *Energies*, vol. 15, no. 10, p. 3592, May 2022.

[65] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3581–3589.

[66] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3483–3491.

[67] P. Berg, D. S. Maia, M.-T. Pham, and S. Lefèvre, "Weakly supervised detection of marine animals in high resolution aerial images," *Remote Sens.*, vol. 14, no. 2, p. 339, Jan. 2022.

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[69] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. New York, NY, USA: Springer-Verlag, 2006.

[70] A. Berge and A. H. S. Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, Nov. 2006.

[71] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using Gaussian mixture models and Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 153–157, Jan. 2014.

[72] S. Prasad, M. Cui, W. Li, and J. E. Fowler, "Segmented mixture-of-Gaussian classification for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 138–142, Jan. 2014.

[73] B. Zhao, Y. Zhong, A. Ma, and L. Zhang, "A spatial Gaussian mixture model for optical remote sensing image clustering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5748–5759, Dec. 2016.

[74] Ç. Ari and S. Aksoy, "Detection of compound structures using a Gaussian mixture model with spectral and spatial constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6627–6638, Oct. 2014.

[75] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[76] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[77] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9584–9592.

[78] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," 2018, *arXiv:1807.02011*.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[80] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," 2019, *arXiv:1903.10145*.

[81] A. Pagnoni, K. Liu, and S. Li, "Conditional variational autoencoder for neural machine translation," 2018, *arXiv:1812.04405*.

[82] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 475–489.

[83] 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*, vol. 9, no. 1, pp. 1–10, Dec. 2021.

[84] T. Singh, H. Gangloff, and M.-T. Pham, "Object counting from aerial remote sensing images: Application to wildlife and marine mammals," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 6580–6583.

[85] G. W. Gella, H. Gangloff, L. Wendt, D. Tiede, and S. Lang, "Variational autoencoders for unsupervised object counting from VHR imagery: Applications in dwelling extraction from forcibly displaced people settlement areas," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 1162–1165.

[86] R. Chan et al., "SegmentMeIfYouCan: A benchmark for anomaly segmentation," 2021, *arXiv:2104.14812*.

[87] B. Ge, C. Hou, Y. Liu, Z. Wang, and R. Wu, "Anomaly detection of power line insulator from aerial imagery with attribute self-supervised learning," *Int. J. Remote Sens.*, vol. 42, no. 23, pp. 8819–8839, Dec. 2021.

[88] J. Li, X. Wang, H. Zhao, S. Wang, and Y. Zhong, "Anomaly segmentation for high-resolution remote sensing images based on pixel descriptors," 2023, *arXiv:2301.13422*.

[89] M. Hamzaoui, L. Chapel, M.-T. Pham, and S. Lefèvre, "Hyperbolic variational auto-encoder for remote sensing scene embeddings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 5391–5394.

[90] W. Peng, T. Varanka, A. Mostafa, H. Shi, and G. Zhao, "Hyperbolic deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10023–10044, Dec. 2022.

Getachew Workineh Gella (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in applied geoinformatics with the Department of Geoinformatics, Paris Lodron University of Salzburg, Salzburg, Austria.

He has a broader interest in the integration of Earth observation and deep learning for automatic information retrieval, label-efficient learning strategies, and spatiotemporal transferability of deep learning models.

Hugo Gangloff received the Ph.D. degree in statistical signal processing from the University of Strasbourg, Strasbourg, France, in 2020.

He is currently a Research Engineer with Institut national de recherche pour l'agriculture (INRAE), Unité Mixte de Recherche Mathématique et Informatique Appliquées Paris-Saclay, Université Paris-Saclay, Palaiseau, France. His research interests include optimization and inference in probabilistic models, physics-informed machine learning, and deep learning.

Lorenz Wendt is currently a remote sensing and geographic information system (GIS) specialist and has been part of the Humanitarian Remote Sensing Team, Department of Geoinformatics, Paris Lodron University of Salzburg, Salzburg, Austria, since 2013. He has been involved in a variety of Earth observation (EO)/GI projects in humanitarian and development contexts, including water exploration, water infrastructure planning, population estimation, 3-D data analysis, and data preparedness, as a researcher or a consultant.

Dirk Tiede (Member, IEEE) is currently an Associate Professor and the Deputy Head of the Department of Geoinformatics, Paris Lodron University of Salzburg, Salzburg, Austria, where he is also the Co-Head of the research laboratory EO Analytics. His research focuses on methodological developments in image analysis using optical Earth observation (EO) data, object-based methodologies, and process automation in the context of Big EO data analysis. His research fields include environmental monitoring and support of humanitarian relief operations.

Dr. Tiede received the Christian Doppler Award of the Federal State of Salzburg in 2014.

Stefan Lang (Member, IEEE) is currently an Associate Professor and a specialist in geographic information system (GIS) and remote sensing with the Paris Lodron University of Salzburg, Salzburg, Austria, where he is also the Vice-Dean of the Faculty for Digital and Analytical Sciences and leads the Earth Observation Division, Department of Geoinformatics. He is the Head of the Christian-Doppler Laboratory (GEOHUM), Paris Lodron University of Salzburg, with research interests in geospatial artificial intelligence (GeoAI), object-based image analysis (OBIA), multisource data integration and assimilation, spatial analysis, and regionalization. He initiated a range of collaborative research projects with industry and Non-Governmental Organization (NGO) and has coordinated a Horizon 2020 project in support of the Copernicus Academy.

Dr. Lang chairs the Erasmus+ Joint Master Copernicus in Digital Earth and is an Academic Coordinator of the Erasmus+ Sector Skills Alliances project EO4GEO.