Variational Autoencoders for Unsupervised Object Counting from VHR Imagery: Applications in Dwelling Extraction from Forcibly Displaced People Settlement Areas

Getachew Workineh Gella, Hugo Gangloff, Lorenz Wendt, Dirk Tiede, Stefan Lang

VARIATIONAL AUTOENCODERS FOR UNSUPERVISED OBJECT COUNTING FROM VHR IMAGERY: APPLICATIONS IN DWELLING EXTRACTION FROM FORCIBLY DISPLACED PEOPLE SETTLEMENT AREAS

Getachew Workineh Gella, Hugo Gangloff**, Lorenz Wendt*, Dirk Tiede*, Stefan Lang**

*Paris Lodron University of Salzburg (PLUS)
Christian Doppler Laboratory for Geospatial and EO-Based Humanitarian Technologies (GEOHUM)
Department of Geoinformatics—Z_GIS, Salzburg, Austria
**Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France

ABSTRACT

Even though computer vision models are excellent for automatic scene segmentation and object identification from remotely sensed imagery, they demand a huge corpus of annotated data for the training and validation which is a huge challenge in humanitarian emergency response. To tackle this problem, we propose unsupervised dwelling object counting combining Variational Autoencoder (VAE) with an anomaly detection approach. The approach is tested in six Forcibly Displaced People (FDP) settlement areas situated in different parts of the world. Using an anomaly map computed with the VAE model, we demonstrated the possibility of properly locating dwelling objects using anomaly maps. Dwelling counts are obtained by further segmenting anomaly maps. Results show that, though it has strong spatio-temporal variation, the VAE model exhibits promising potential for locating and counting dwellings. It is also observed that in FDP settlements with dense buildings and extremely low contrast between buildings and ground or environment, the performance is relatively lower than the performance achieved in settlement areas with regularly spaced and less complex building structures.

Index Terms— Anomaly detection, Dwelling extraction, Emergency response, Variational Autoencoder

1. INTRODUCTION

A growing proportion of the global population is facing displacement from their home, either staying in the Internally Displaced Persons (IDP) sites or crossing international borders mostly staying in refugee camps. Hereafter we use the inclusive term Forcibly Displaced People (FDP) settlement sites both for IDPs and refugee settlements. Dwelling information is crucial to monitor camp and temporary settlement expansion, estimate residing populations, and provide respective humanitarian emergency responses and long-term socio-economic planning. For the last decade, the proliferation of high-resolution Earth Observation (EO) imaging has enabled closer monitoring of FDP sites [1]. Key elements for efficient information retrieval workflows are speed and automation.

In this aspect, advances in computer vision, especially deep learning, have paved new opportunities for automatic information retrieval about various aspects of FPD settlements. For example, [2] quantified the spatial expansion and internal densification of FDP settlements from time series high-resolution satellite images. More specifically, [3, 4] used instance and semantic segmentation models for dwelling extraction. Similarly, [5] used an instance segmentation model for rapid mapping in complex urban settings in response to a pandemic alert, which has helped humanitarian emergency response operations.

Although deep learning models have proven performance for the detection and segmentation of objects from remote sensing images, the traditional supervised deep learning approaches require an extensive amount of training and validation data, which is often impractical during emergency response because of two reasons, (1) short response time required for humanitarian service delivery, (2) frequent spatial monitoring and extensive geographic coverage is required. This constrains the operational utility of deep learning models in EO-based operational humanitarian response. As a solution, label-efficient strategies like transfer learning [4] and domain adaptation approaches could be employed which both assume the presence of some amount of annotated data for fine-tuning of pretrained models or a sufficient amount of source labeled data for joint training. Beyond the assumption of data availability, sometimes it also fails to transfer relevant representations to undertake intended detection and object counting tasks under different data distributionsand indeed, the scene and object characteristics vary a lot.

Recently, generative models, especially Variational Autoencoders [6] (VAE), have shown significant performance in

PLUS authors are funded by Christian Doppler Research Association and Doctors without Borders-Section Austria.

image reconstruction and anomaly detection. They have been successfully applied in wild animal detection from aerial images [7], ship detection [8], anomaly localization and segmentation of medical images [9], and defect identification [10] tasks. Based on the work of [7], we propose an unsupervised dwelling object localization and counting by combining a VAE [6] and anomaly detection approaches [7, 9, 11]. To the best of our knowledge, this is the first study to address building detection in general and dwelling counting in particular as an unsupervised anomaly detection approach using EO data.

2. METHODS

2.1. Datasets and processing

This study uses very high-resolution satellite imagery taken from six different FDP settlements [12]((See Table 1). The preparation of the training and testing datasets follows the conceptual definition of anomalous and normal images in unsupervised anomaly detection [13]. Accordingly, we first define normal images as patches expected to have high probability of soil and various land cover types other than dwellings. This set of image patches are taken from image areas outside of the FDP settlement premises. Conversely, we define anomalous images as containing dwelling objects: these are thus image chips within FDP settlement premises. For anomalous images, the annotations for evaluation of the model performance are obtained from an in-house [12] database, generated as long-term engagement in EO-based humanitarian emergency response tasks. Both anomalous and normal images were converted to small image chips of size 256×256 pixels.

2.2. VAE for dwelling object localization and counting

As indicated in Fig. 1 a VAE model is trained on normal images for image reconstruction. At testing time we expect the dwellings to be missing from the reconstructed anomalous images, which thus enables to properly locate and undertake further segmentation. To this end, a normal image is fed into the encoder network which acts as the feature extraction module. Variational sampling is done in the compressed latent space and fed into a decoder module where an image is reconstructed back. As indicated in Eq. 1, given the anomalous input x the encoder produces compressed latent code $q_\phi(z|x)$. The latent code is fed into the decoder and reconstructed $\hat{x}$ which is $p_\theta(z|x)$. The model is optimized by maximizing Evidence Lower Bound (ELBO) [6],

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

where the first term can be interpreted as a reconstruction loss between the input x and reconstructed $\hat{x}$ and the second term is a Kullback–Leibler divergence. Classically, $p_\theta(z)$ is chosen as a standardized Gaussian. Further details can be found in [6, 7]. Once the VAE training has converged, prediction is done on anomalous image patches containing dwellings taken within the premises of FDP settlements. Anomaly scores are then computed using image structural similarity index [14] which is computed as:

$$SSIM(x, \hat{x}) = SSIM(r_i, p_i), \\ = \frac{(2\mu_r\mu_p + C_1)(2\sigma_{pr} + C_2)}{(\mu_p^2 + \mu_r^2 + C_1)(\sigma_r^2 + \sigma_r^2 + C_2)} \quad (2)$$

where μ and σ^2 indicate the mean and the variance of reconstructed r and predicted p images, respectively, at pixel location i in a certain window, whose size is a model hyperparameter (see Table 1). C_1 and C_2 are constants set to 0.01 and 0.03 respectively [14]. These anomaly score maps are normalized to values between 0 and 1. Then, based on the anomaly scores, dwelling objects are segmented using a combination of binary and non-parametric Otsu's thresholding [15] and morphological opening operator [16]. Based on our intuition use of coarser image could constrain proper reconstruction of very bright dwellings, we also obtained the best results by downscaling input images to coarser resolution with a scale factor of 8. The anomaly score is then generated with the original image (see Table 1). Finally, the performance of the proposed anomaly detection approach is evaluated on the unsupervised tasks of locating and counting the dwelling objects. The unsupervised dwelling segmentation task is evaluated by pixel-wise area under the Receiver Operating Characteristic curve (AUC) [17] while unsupervised dwelling counting is evaluated using Mean Absolute Error (MAE) between the model output count and the reference dwelling counts. Our approach is also compared with a state-of-the-art anomaly detection approach based on a VAE with an anomaly attention mechanism [10]. Implementation code is provided at `https://github.com/HGangloff/getch-geohum`

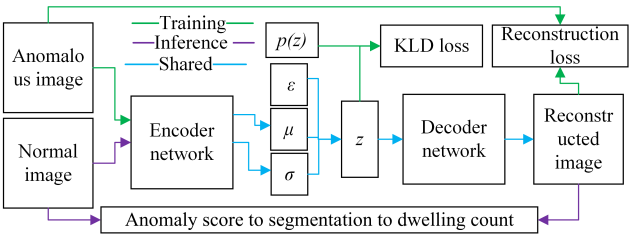

Fig. 1: Implementation workflow.

3. RESULTS AND DISCUSSION

The analysis of the results indicates that even though there is variation between different FDP settlements and seasons, the combined use of the VAE with an unsupervised anomaly detection approach shows promising results for locating and counting of dwellings(Table 1 and Fig. 2). The best dwelling location score is observed in Nguenygiel followed by Nduta and Minawao datasets with AUC values of 0.97 and 0.91

respectively. These areas exhibit relatively less complex dwelling structures: they are well-spaced and contrast with the background. On the other hand, the lowest scores are achieved in Kutupalong and Dagahaley datasets with AUC values of 0.64 and 0.77 respectively. Dwellings that are not well detected in those FDP areas are very similar to the vegetation and bare land. Therefore, the decoder treats dwellings as normal elements and thus yields weak anomaly scores. Based on the MAE values predicted dwelling counts deviate from reference counts with MAE values ranging from 8 to 72 dwellings, depending on dwelling complexity per dataset, the SSIM window size and the downscaling at input. Anomaly scores created with smaller neighborhood window (Eq. 2) has enabled relatively better delineation of individual dwellings (Fig. 3) and better dwelling location scores (Table 1). The less favorable cases for the counting task are observed in settlements dominated by very complex dwelling structures characterized by either contiguous and dense dwellings (*e.g.* Kutupalong), small, or extremely low-contrast-to-background features. Note that such a dataset represents a tough task even for manual delineation. Moreover, some dwellings are easily reconstructed by the VAE, thus they do not appear as anomalies and are missed by our approach. All in all, the absence of detection or the detection of contiguous dwellings as only one resulted in underestimation of dwelling counts. Our VAE model has achieved better results than re-implementation of [10] (see Table 1), especially in locating dwellings. The anomaly attention maps we obtained were not explicitly strong on dwelling but on the entire neighbourhood of dwellings. For complex Dagahaley and Kutupalong datasets, it failed to yield meaningful results.

4. CONCLUSIONS

In this research, we have demonstrated the potential of VAE for unsupervised dwelling location and counting from different FDP sites. VAE can properly localize and count dwellings properly in settlements with less spatial and spectral complexity. Despite promising results, there is a spatio-temporal variation of the results which leaves room for improvement of the proposed approach.Specifically: (1) the inherent complexity of dwelling features in terms of density and contrast with the background, (2) the VAE model which can sometimes easily reconstruct bright dwellings which resulted in poor anomaly score to localize and further detect dwellings (3) the dependence on the window size of the SSIM anomaly maps. For comprehensive object location and counting, further work will focus on latent space conditioning with self-supervision to get a strong anomaly score, or combination of different anomaly generation approaches. In a nutshell, for operational humanitarian emergency response, our new approach could help generate critical information in real-time and in highly dynamic situations.

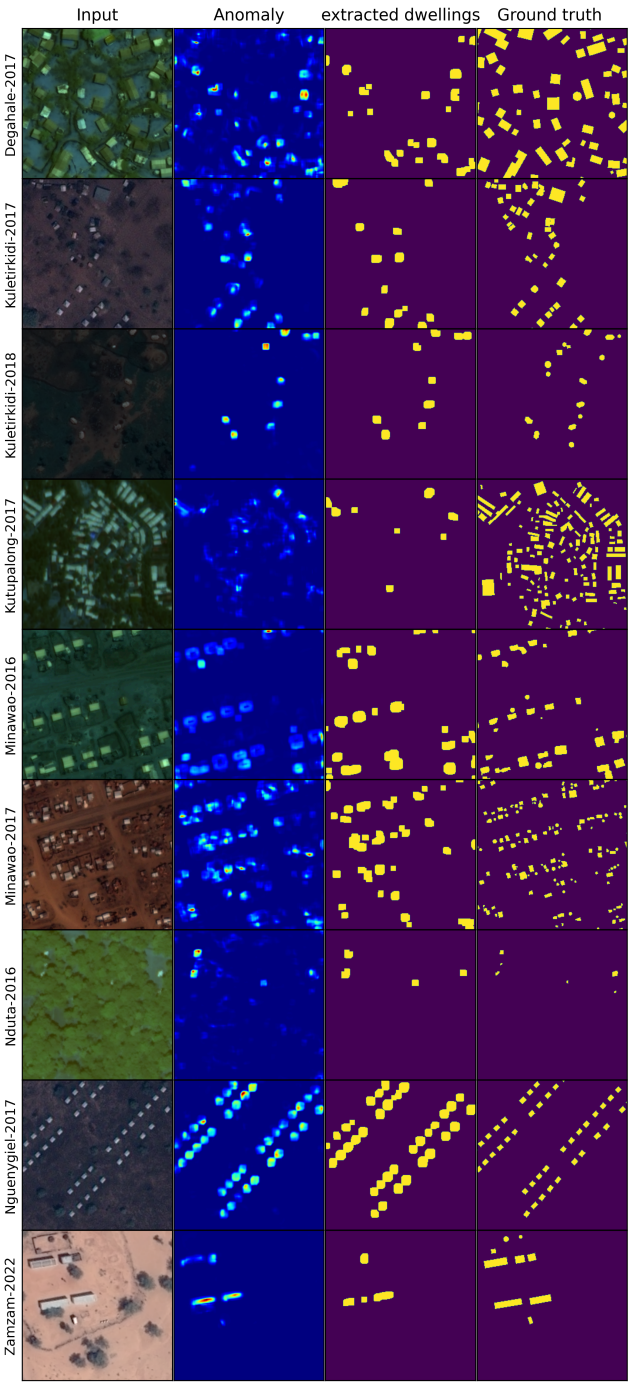

Fig. 2: Results from randomly selected image patches

5. REFERENCES

[1] S. Lang et al., "Earth observation tools and services to increase the effectiveness of humanitarian assistance," *Eur. J. Remote Sens.*, vol. 53, no. sup2, pp. 67–85, 2020.

[2] S. Benz, H. Park, J. Li, D. Crawl, J. Block, M. Nguyen, and I. Altintas, "Understanding a rapidly expanding refugee camp using convolutional neural networks and

Table 1: Results for dwelling localization and counting

Dataset	VAE [10]	$SSIM_5^*$	$SSIM_{11}^*$	$SSIM_d^*$
Dagahaley-2017	-	0.73	0.77	0.76
	-	27	26	20
Kuletirkidi-2017	0.62	0.90	0.92	0.93
	30	35	36	26
Kuletirkidi-2018	0.62	0.88	0.90	0.91
	28	37	34	23
Kutupalon-2017	-	0.69	0.64	0.69
	-	72	72	67
Minawao-2016	0.44	0.92	0.91	0.95
	13	16	15	8
Minawao-2017	0.54	0.96	0.95	0.96
	42	37	37	34
Nguenygiel-2017	0.56	0.98	0.97	0.98
	24	23	25	22
Nduta-2016	0.43	0.91	0.91	0.93
	23	17	16	12
Zamzam-2022	0.52	0.78	0.81	0.82
	46	40	40	52

$*$ 5 & 11 are SSIM window sizes, d donwscaling; 1^{st} and 2^{nd} rows for locating (AUC) and count (MAE) values respectively.

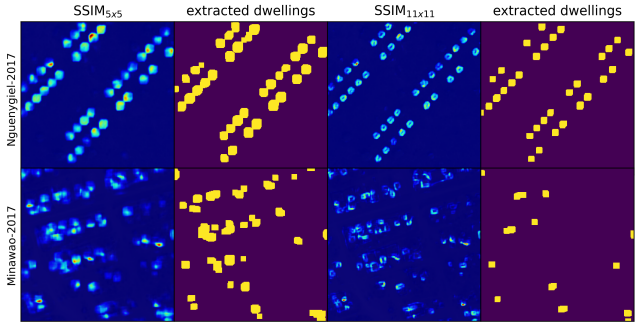

Fig. 3: Sensitivity to SSIM window size.

satellite imagery," in *2019 15th International Conference on eScience (eScience)*. IEEE, 2019, pp. 243–251.

[3] 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, 2021.

[4] 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," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 376, no. 2128, pp. 20170363, 2018.

[5] 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, 2021.

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

[7] H. Gangloff, M.T. Pham, L. Courtrai, and S. Lefèvre, "Variational autoencoder with gaussian random field prior: application to unsupervised animal detection in aerial images," *hal-03774853*, 2022.

[8] N. Ferreira and M. Silveira, "Ship detection in sar images using convolutional variational autoencoders," in *IGARSS 2020-2020*. IEEE, 2020, pp. 2503–2506.

[9] 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*, vol. 69, pp. 101952, 2021.

[10] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R.J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proceedings of the CVPR*, 2020, pp. 8642–8651.

[11] L. Ruff, J.R. Kauffmann, R.A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T.G. Dietterich, and K.R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[12] 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.

[13] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[14] 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, 2004.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[16] B. Preim and C.P. Botha, *Visual computing for medicine: theory, algorithms, and applications*, Newnes, 2013.

[17] 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, 1997.