Fine-Scale Population Estimation Using a Variational Autoencoder-Based Approach Integrating Geospatial Data

Issa Nasralli
DES Unit, Faculty of Sciences of Sfax
University of Sfax
Sfax, Tunisia
aissanasralli@gmail.com

Imen Masmoudi
DES Unit, Faculty of Sciences of Sfax
University of Sfax
Sfax, Tunisia
masmoudi.imene@gmail.com

Hassen Drira
ICube Laboratory
University of Strasbourg, CNRS
Strasbourg, France
hdrira@unistra.fr

***Abstract*—This paper presents a novel approach for population estimation by focusing on the comparison between two deep learning models: popVAE, which integrates a Variational Autoencoder (VAE) for latent spatial contextual feature extraction, and popCNN, which relies on conventional convolutional layers for spatial contextual feature extraction. The models were evaluated on a geospatial dataset from Tunisia. Our results show that popVAE outperforms popCNN in terms of predictive accuracy, as evidenced by higher Coefficient of Determination (R^2) of 0.8760 and lower Mean Squared Error (MSE) values. The popVAE performance was also compared to baseline models achieved a competitive accuracy. These findings suggest that VAE offer significant advantages in population estimation tasks by capturing complex spatial dependencies in the data. The code and training dataset are available at: https://github.com/IssaNasralli/popVAE.**

***Index Terms*—Convolutional Neural Network, Variational Autoencoder, Feature Extraction and Latent Representation.**

I. Introduction

Accurate population estimation is crucial for urban planning, resource management, and policy development [1], supporting evidence-based decisions for identifying at-risk populations and generating health and development indicators [2]. Advances in geospatial data (e.g., GeoTIFF, Shapefile) and machine learning [3] have greatly improved population estimation methods.

Census data are provided at coarse resolutions to protect privacy, limits detailed local planning. To address this, fine-scale population estimation has been explored at building, pixel, and region levels. However, challenges persist, especially in data-scarce regions [4], where very high-resolution imagery, though accurate, is costly and limited in availability [5].

Population mapping approaches include bottom-up methods, which estimate population through local surveys, and top-down methods, which use census data combined with high-resolution geospatial data. Techniques like dasymetric mapping refine estimates by redistributing population counts based on spatial relationships and data quality [5].

Remote sensing data has been instrumental in analyzing Land Use and Land Cover (LULC) [6], forming a foundation for population estimation techniques [7]. Points of Interest (POIs) have demonstrated a strong relationship with population density [8] and can be utilized in both continuous format, through Kernel Density Estimation (KDE) [9]–[11] , or in discrete format [8], [12]. Additionally, building footprint datasets have gained significant attention in this field [13], as they can be used with [14] or without building height data [15]. However, approaches that lack building height data often struggle with accuracy. Nighttime Light (NTL) imagery, road networks, and Digital Elevation Models (DEM) are also commonly utilized for fine-scale population estimation.

While datasets like LandScan [16] and GPWv4 [17] provide valuable insights, their inherent limitations—such as low spatial detail, reliance on proprietary data, or modeling assumptions—highlight the need for advanced methodologies. Techniques like dasymetric mapping have partially addressed these limitations by redistributing population counts based on spatial relationships. However, the rise of deep learning approaches has offered a more flexible and powerful alternative, leveraging complex architectures to capture spatial contextual features for fine-scale population estimation. For instance, [18] integrates spatial contextual features by using a VGG architecture with $74 \times 74 \times 7$ input images. [19] employs a VGG model combining Landsat-8 and Sentinel-1 imagery to capture diverse features of human settlements, while [10] integrates spatial and attribution features to analyze complex spatial patterns by the fusion of Convolutional Neural Network (CNN) and Multi-Layer Perceptron (MLP) models. [20] investigates the impact of neighboring patches on model performance using advanced CNN models. [21] introduces InceptionResNet2 with Local and Global Attention Modules to effectively capture both local and global features, and [22] utilizes a modified ResNet-based architecture with separate branches for spatial features and pixel features, enhancing model interpretability. On the other hand, some approaches deliberately avoid incorporating spatial contextual features. For example, [23] excludes neighboring urban context features to achieve generalization across different urban environments by applying three 1D convolutional layers, while [24] treat each pixel independently using CNNs with 1×1 convolutional layers, focusing solely on pixel-level

feature extraction. Similarly, [5] uses 1 × 1 kernels, avoiding spatial interactions in their model. Table I provides a detailed comparison of various deep learning approaches for population estimation, highlighting their estimation level and resolution (only for pixel level), study areas, ancillary data used, training populations, model architecture, and corresponding prediction accuracy (R^2).

TABLE I
COMPARISON OF DEEP LEARNING APPROACHES.

Method	Estimation Level and Resolution	Study Area	Ancillary Data	Training Data	Model Architecture	R^2
[18]	pixel (33m)	USA	Satellite Imagery	Science Center 33m	Modified VGG	0.9365
[19]	region	Rural Villages, India	Satellite Imagery	Census Village	Modified VGG	0.7720
[10]	pixel (100m)	Shenzhen, China	NTL, NDVI, LULC, DEM, Roads, POI	Worldpop100m	Regular CNN and MLP	0.7700
[20]	pixel (1km)	Atlanta, USA	Satellite Imagery	LandScan 1km	DenseNet	0.9100
[21]	pixel (1km)	Hebei, China	NTL	Worldpop 1km	Inception ResNet2 and Attention Mechanism	N/A
[22]	pixel (100m)	98 Cities, Europe	Satellite Imagery, DEM, Climate, LU, NTL	So2Sat 100m	ResNet	0.8630
[23]	pixel (1km)	15 Cities, Europe	OSM	GPW 1km	Regular CNN	N/A
[24]	pixel (N/A)	13 Cities, USA	Satellite Imagery	2010 Census	Regular CNN	N/A
[5]	building	Zambia	Building Footprints, Transport, LULC, DEM, NTL	Census at the administrative boundaries level.	Regular CNN	0.8790

Despite the advancement of deep learning approaches, a common limitation is their reliance on CNN architectures for extracting contextual features, which may be insufficient as CNNs primarily focus on image features like edges and textures. Another critical limitation is the reliance on proprietary or restricted datasets and country-specific data dependencies which not only limit the widespread applicability of proposed methods but also make fair comparisons with other approaches and reproducibility infeasible. The lack of accessible and standardized datasets hinders the ability to validate and benchmark these approaches to assess their performance on a global scale.

This research proposes the use of freely available, open-access datasets with low spatial resolution and accessible globally to ensures scalability, reproducibility, and broad applicability. Our main contribution is the proposed model popVAE which integrates spatial contextual features into a novel deep learning architecture, combining a VAE [25] with a CNN for pixel-level population prediction based on continuous POI, building footprint with height data, and other various geospatial data, aiming to enhance model accuracy. To support reproducibility and encourage future research, our code and the final assembled dataset are made available online.

For our study area, we chose Tunisia (Fig 1-a), a North African moderate country divided into 24 Governorates, each offering distinct characteristics and significance. Notably, the capital Tunis and Sfax Governorate stand out due to their economic importance. Spanning an area of 163,610 km², Tunisia provides a manageable yet diverse region, enabling a thorough analysis of population distribution across the entire country. Fig 1-b illustrates the workflow schema of our method.

II. METHODOLOGY

A. Data Assembly

Table II details each data category, the respective sources, the data type, the spatial resolution (only for raster data), and the year of the dataset.

TABLE II
SUMMARY OF DATA ASSEMBLY.

Data Category	Data Source	Data Type	Resolution	Year
Population Data	WorldPop	Raster	100 meters	2020
	INS	Tabular	-	2020
POI	NODP	Tabular	-	2023
LU	OSM	Vector	-	2024
Building Footprint	WSF	Raster	10 meters	2019
Satellite Imagery	MODIS	Raster	500 m	2020
Road Network	OSM	Vector	-	2024
LULC	DynamicWorldV1	Raster	30 m	2020
DEM and Slope	SRTM	Raster	90 m	2020
NTL	VIIRS	Raster	463 m	2020
Boundaries	GAUL	Vector	-	2015

For population data, we used the WorldPop dataset, for the year 2020 with a resolution of 100 meters [26]. We also utilized population data at the governorate level obtained from the National Institute of Statistics (INS), the official agency in Tunisia tasked with conducting censuses and generating demographic, social, and economic statistics (https://www.ins.tn/statistiques/111).

For ancillary data, we sourced 5,965 educational POIs from the National Open Data Platform (NODP) [27]. Land Use (LU) and road network data were extracted from OpenStreetMap (OSM). Building footprints were acquired from the World Settlement Footprint (WSF) dataset [28]. The WSF2019 dataset provides information on the presence or absence of buildings, while the WSF3D dataset offers details on building heights. For satellite imagery, we opted for Moderate Resolution Imaging Spectroradiometer (MODIS) due to its high temporal resolution, ensuring more reliable, cloud-free data [29]. The LULC data was sourced from the Dynamic World V1 dataset [30], while the DEM data was obtained from the Shuttle Radar Topography Mission (SRTM) [31], and slope data was derived from this DEM. NTL data was gathered from the VIIRS Nighttime Lights dataset. Boundaries was obtained from the the Global Administrative Unit Layers (GAUL) dataset [32], which included one Shapefile for the entire country and 24 Shapefiles corresponding to each governorate.

MODIS, DEM, Dynamic World V1, VIIRS, and GAUL data were acquired through Google Earth Engine [33], and OSM data via the HOT EXPORT TOOL (https://export.hotosm.org). The remaining datasets were obtained directly from the official websites of the organizations that produced them.

B. Data Pre-processing and Transformation

The NTL, DEM, and Slope data are considered raw ancillary data, the remaining ancillary data have undergone specific processing steps using the ArcGIS Pro 3.4:

- POI: The KDE layer of educational POIs was generated using the ArcGIS Kernel Density Tool with a search

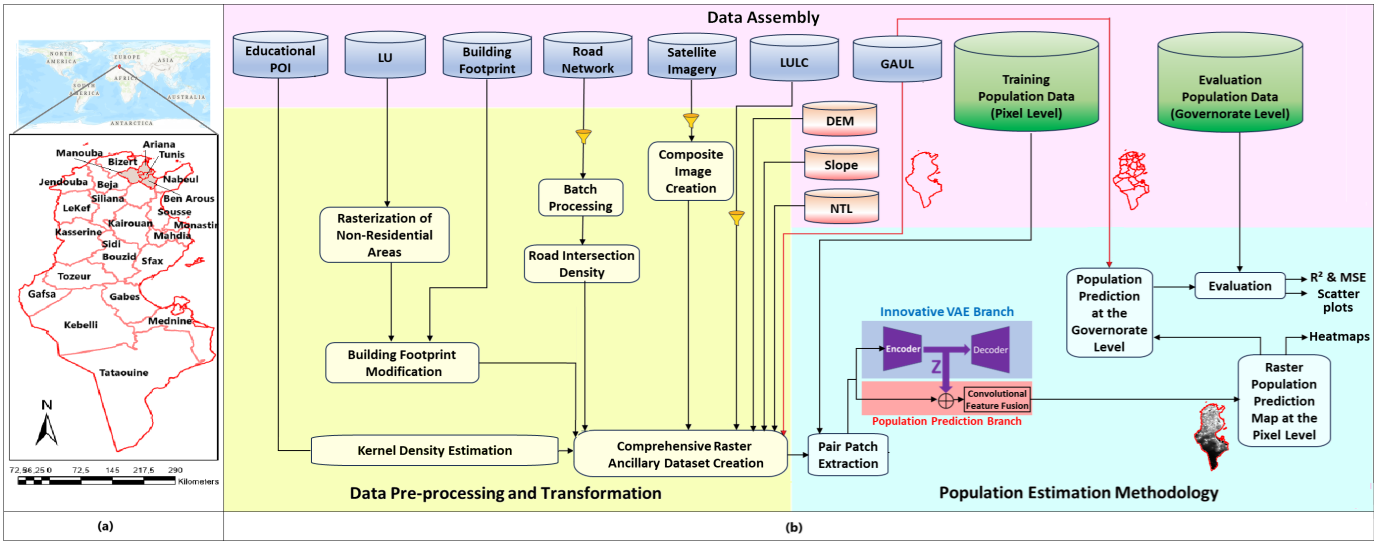

Fig. 1. Tunisia Location and Flowchart illustrating the methodology for estimating population in the study.

radius (bandwidth) of 5 km, using the planar method and area units in square kilometers.

- Rasterization of Non-Residential Areas: We created a spatial index for the LU polygon and selected non-residential areas based on provided attributes. These areas were then rasterized into a binary mask, with non-residential areas assigned a value of 0 and all other areas a value of 1.
- Building Footprint Modification: The WSF2019 dataset was checked against the non-residential binary mask. Building heights in the WSF 3D data were reclassified by dividing pixel heights by 3 meters to estimate the number of floors, which was then multiplied with the checked WSF2019 dataset.
- Road Intersection Density: We focused on major road types and excluded "residential" and "service" roads. Using the Batch Processing tool, we snapped road segments to eliminate gaps. The Intersect tool identified intersection points, and the Point Density tool created a raster layer of intersection density.
- Composite Image Creation: The preprocessing involved selecting pertinent bands such as Nadir Reflectance Bands 1, 3, and 4 to capture visible light data. Following this, a median composite image was created spanning the defined timeframe (January 1 to December 31, 2020). This step aimed to mitigate noise and variability across individual images.
- LULC: The Dynamic World V1 dataset includes 11 bands: one indicating the land cover class with the highest probability and ten showing the estimated probabilities for each class. We used only the band representing the highest probability label.
- Boundaries: The Shapefile representing each governorate boundary was converted to raster mono band data with dimensions of $w = 3,807$ by $h = 8,116$ pixels. The pixel values inside the governorate boundary were set to 1, otherwise it were set to 0.

The ancillary datasets were normalized, resampled to a resolution of 100 meters, clipped to the boundary of Tunisia using the Shapefile of the entire country, and converted to the WGS 1984 UTM Zone 32N coordinate system. Subsequently, these datasets were overlaid to create a comprehensive raster ancillary dataset with dimensions of $w = 3,807$ by $h = 8,116$ pixels and $n = 10$, and n is the number of bands.

It is important to note that GAUL data was not included in the comprehensive raster ancillary dataset. Instead, it was used to extract or cilp data specific to Tunisia (Shapefile of the entire country), while the raster governorate boundary will be used to aggregate population as mentioned in section III-B.

C. Model Development

1) Overview: We developed a two-branch model: one branch for population prediction and an innovative VAE branch to enhance spatial contextual feature extraction for the prediction branch (Fig. 2).

We traversed the raster ancillary dataset created in II-B, denoted X_c, horizontally and vertically to extract patches. Then we regrouped ancillary data bands into a set $X_c^{(i,j)}$, then leveraging the VAE to generate a latent representation of the spatial contextual features, enabling the model to capture implicit spatial characteristics that may influence population density in the pixel with coordinates (i, j).

Compared to a traditional CNN, the VAE provides regularization via its latent space, allowing the model to avoid overfitting while capturing implicit spatial hierarchies. This regularization effect, achieved through the Kullback-Leibler (KL) divergence term, ensures a structured latent representation that facilitates improved generalizability for population prediction.

2) Mathematical Formulation:

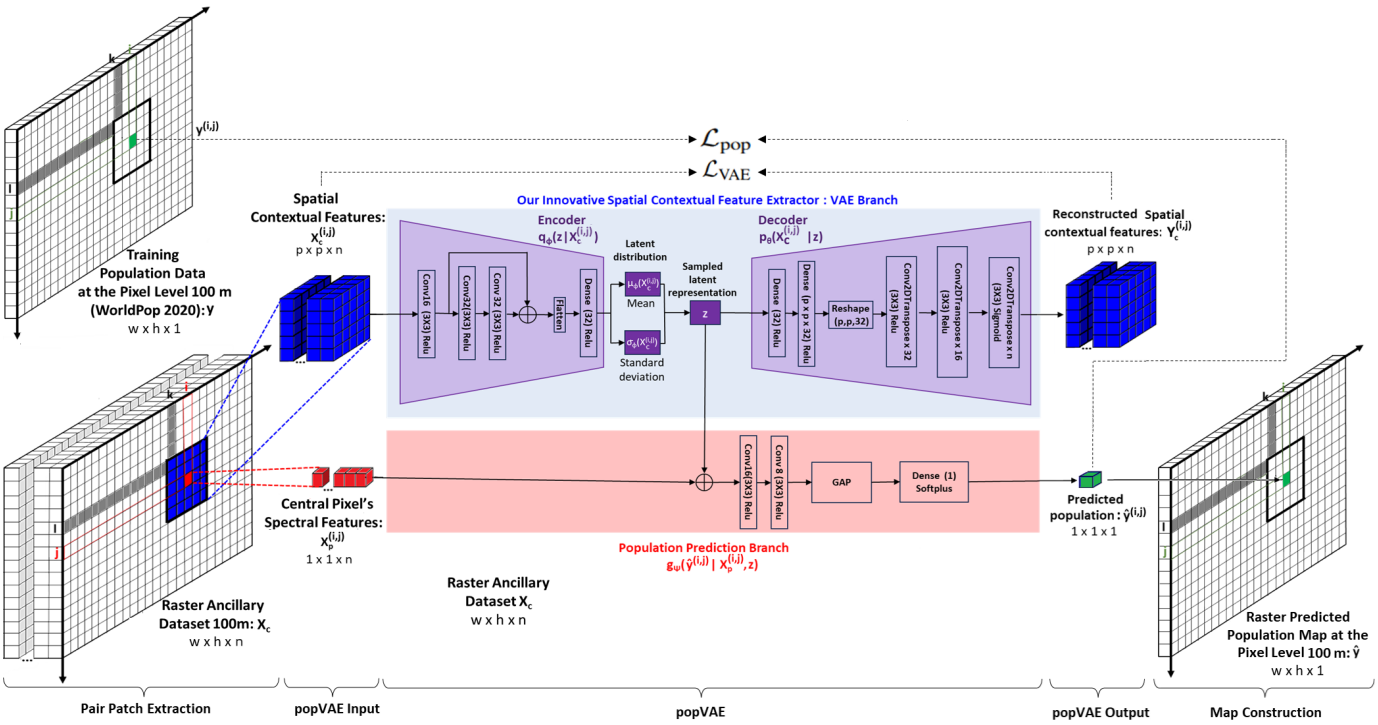

Fig. 2. The architecture of popVAE and popuplation prediction map construction (For this illustration : p=5, w=25 and h=15).

The notion **Dense (b) f** refers to a dense layer with $\mathbf{b}$ neurons and activation function $\mathbf{f}$.
The notion **ConvN (k x k) f** means that a convolution with a kernel of size $\mathbf{k} * \mathbf{k}$ and activation function $\mathbf{f}$ is applied, where $\mathbf{N}$ denotes the number of filters.
The notion **Conv2DTranspose x N (k x k) f** indicates the application of transposed convolutional layers with $\mathbf{N}$ filters, a kernel of size $\mathbf{k} * \mathbf{k}$, and activation function $\mathbf{f}$.

a) Encoder: $X_c^{(i,j)} \in R^{p \times p \times n}$ is the input data tensor to VAE, where p is the size of the extracted patch (with p being an odd number). This tensor captures both the spatial context and spectral features within a patch of size $p \times p$. The central pixel's spectral features is isolated as $\mathbf{X}_p^{(i,j)} \in \mathbb{R}^{1\times1\times n}$.

For each patch, we define the central pixel coordinates as (i, j), where $i = \lfloor k+\frac{p}{2} \rfloor+1$, $j = \lfloor l+\frac{p}{2} \rfloor+1$ and $\lfloor . \rfloor$ denotes the integer part of a number, and (k, l) represents the coordinates of the top-left corner of the patch. The target population $y^{(i,j)}$ (obtained from the WorldPop dataset denoted y) at this central pixel is associated with the coordinates (i, j).

The encoder network, parameterized by ϕ, processes the input data $\mathbf{X}_c^{(i,j)}$ to generate parameters for the latent distribution. Specifically, it outputs the mean $\mu_\phi(\mathbf{X}_c^{(i,j)})$ and the standard deviation $\sigma_\phi(\mathbf{X}_c^{(i,j)})$, which define the distribution over the z space. The encoder's objective is to approximate the prior distribution $q_\phi(\mathbf{z} \mid \mathbf{X}_c^{(i,j)})$.

The vector $\mathbf{z}$ is sampled from the latent distribution characterized by $\mu_\phi(\mathbf{X}_c^{(i,j)})$ and $\sigma_\phi(\mathbf{X}_c^{(i,j)})$. Given the intractability of the true posterior $p_\theta(\mathbf{z} \mid \mathbf{X}_c^{(i,j)})$, variational inference is employed to approximate it with $q_\phi(\mathbf{z} \mid \mathbf{X}_c^{(i,j)})$, modeled as a Gaussian distribution $\mathcal{N}(\mu_\phi(\mathbf{X}_c^{(i,j)}), \sigma_\phi^2(\mathbf{X}_c^{(i,j)}))$.

The dimensionality of $\mathbf{z}$, often referred to the latent dimension d, is a crucial hyperparameter that influences the model's capacity to capture and represent the underlying structure of the input data $\mathbf{X}_c^{(i,j)}$. A higher-dimensional $\mathbf{z}$ space allows for more complex representations, potentially capturing more intricate features from the input. However, a very high-dimensional $\mathbf{z}$ may lead to overfitting, while a too-low dimension might result in an oversimplified model that fails to capture all relevant features.

b) Decoder: The decoder network, parameterized by θ, takes $\mathbf{z}$ and reconstructs the input data $\mathbf{X}_c^{(i,j)}$ producing a reconstructed output denoted as $\mathbf{Y}_c^{(i,j)}$. This process defines the likelihood $p_\theta(\mathbf{X}_c^{(i,j)} \mid \mathbf{z})$, representing how well z explains the observed data. The output $\mathbf{Y}_c^{(i,j)}$ is crucial during the training phase as it significantly impacts the learning of $\mathbf{z}$. However, during inference, $\mathbf{Y}_c^{(i,j)}$ will be ignored because the primary objective at this stage is to utilize the learned latent representations ($\mathbf{z}$) for population prediction as mentioned in section II-C2d.

c) The VAE Training: The objective of training the VAE is to maximize the Evidence Lower Bound (ELBO), which is given by:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi; \mathbf{X}_c^{(i,j)}) = \sum_{i,j} \Big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X}_c^{(i,j)})}[\log p_\theta(\mathbf{Y}_c^{(i,j)} \mid \mathbf{z})] - D_{\text{KL}}[q_\phi(\mathbf{z} \mid \mathbf{X}_c^{(i,j)}) \parallel p_\theta(\mathbf{z})] \Big] \quad (1)$$

where: $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X}_c^{(i,j)})}[\log p_\theta(\mathbf{Y}_c^{(i,j)} \mid \mathbf{z})]$ represents the reconstruction likelihood, measuring how well the decoder can reconstruct the input data $\mathbf{X}_c^{(i,j)}$ from z, and $D_{\text{KL}}[q_\phi(\mathbf{z} \mid \mathbf{X}_c^{(i,j)}) \parallel p_\theta(\mathbf{z})]$ is the KL divergence between the approximate

posterior $q_\phi(\mathbf{z} \mid \mathbf{X}_c^{(i,j)})$ and the prior distribution $p_\theta(\mathbf{z})$, typically assumed to be a multivariate unit Gaussian $\mathcal{N}(0, I)$. This means that the prior distribution has a mean vector of 0 and a covariance matrix that is an identity matrix I. The KL divergence term acts as a regularizer, encouraging the learned latent distribution to be close to the prior.

The VAE training process involves optimizing both the encoder and decoder parameters (ϕ and θ, respectively) to jointly maximize the ELBO. This optimization simultaneously improves the reconstruction quality of the decoder and aligns the approximate posterior with the prior distribution.

d) Population Prediction Branch: This branch, parameterized by Ψ, is designed to predict population density denoted $\hat{y}^{(i,j)}$, using $X_p^{(i,j)}$ combined with the z. The branch can be represented as a function: $g_\Psi(\hat{y}^{(i,j)} \mid \mathbf{X}_p^{(i,j)}, \mathbf{z})$. The performance of this branch is evaluated using a loss function, defined as: $\mathcal{L}_{\text{pop}}(\Psi; y^{(i,j)})$.

e) The Total Loss Function: Combining both VAE and population prediction losses, for each patch centered at a pixel with coordinate (i, j) the overall loss function is defined as:

$$\mathcal{L}_{\text{pixel}}^{(i,j)} = \alpha \mathcal{L}_{\text{pop}}(\Psi; y^{(i,j)}) + \beta \mathcal{L}_{\text{VAE}}(\theta, \phi; \mathbf{X}_c^{(i,j)}) \quad (2)$$

where α and β are weight factors for the VAE and population prediction losses, respectively. The training process is designed to jointly optimize both VAE training and population prediction training, enabling the model to learn robust spatial contextual features while simultaneously improving population prediction accuracy.

Finally, the total loss function, generalized over the entire training dataset, is defined as:

$$\mathcal{L}_{\text{tot}} = \frac{1}{w \times h} \sum_{(i,j)} \left| \mathcal{L}_{\text{pixel}}^{(i,j)} \right| \quad (3)$$

D. Model Implementation

1) Encoder: The VAE branch processes the spatial contextual features $(X_c^{(i,j)})$ by encoding it into latent representation, z. The encoder network begins with a convolutional layer of 16 filters and a 3x3 kernel size, applying ReLU activation and He Normal initialization to stabilize weight initialization. This is followed by a residual block consisting of two sequential convolutional layers, each with 32 filters and a 3x3 kernel. The residual block includes a shortcut connection that bypasses these layers, enabling efficient gradient flow and facilitating the learning of complex spatial hierarchies. The encoder's final dense layer reduces dimensionality to 32 units and outputs the mean and the standard deviation to parameterize the latent distribution.

2) Decoder: The decoder reconstructs the spatial contextual features $(X_c^{(i,j)})$ by transforming this latent representation back into the input shape, utilizing Conv2DTranspose layers with 16 and 32 filters, each with a 3x3 kernel size, to restore the original spatial dimensions.

3) Population Prediction Branch: Here, we flatten and reshape z to align with the spatial dimensions of the primary input, $X_p^{(i,j)}$), which contains pixel-level features. After concatenating z and $X_p^{(i,j)}$), the model enters the Convolutional Feature Fusion stage, where it applies additional convolutional layers, each with a reduced filter count (16 filters for the first layer and 8 filters for the second layer) and a 3x3 kernel size, to capture fine-grained population-relevant features from both pixel-level and spatial contextual features. Finally, a global average pooling layer reduces the spatial dimensions, followed by a dense layer with a softplus activation function, which outputs the predicted population value as a positive scalar.

III. Experimental Evaluation and Results

A. Configuration

The experiments were conducted on a machine running Ubuntu 22.04 LTS 64-bit, equipped with an AMD Epyc 7443p 24-core processor (x8), 32 GiB of memory, and a Tesla K80 GPU. The model was implemented using TensorFlow, Keras, and Scikit-learn.

The raster ancillary dataset, created in section II-B was used for training. This dataset was randomly reduced using the thumb rule, where the number of training samples is approximately ten times the number of trainable parameters in the model. After applying this rule, the dataset was split into Train (50%), Validation (10%), and Test (40%) sets. The MSE loss function was chosen for both population prediction and as the reconstruction likelihood, while the KL loss was computed as defined in [25] (equation 10). We utilized the Adam optimizer with a default learning rate and trained the model for up to 1000 epochs. EarlyStopping was implemented with a patience of 20 epochs to prevent overfitting. Cross-validation was performed to identify the optimal hyperparameters. Specifically, we fixed α and β to 1, and p to 11. The z dimension d was tested with values ranging from 100 to 1000 with a step of 100, and batch sizes were varied across [32, 64, 128, 256]. Increasing d means that our model is more influenced by the spatial contextual features, while decreasing d indicates that the model is less influenced by the spatial contextual features. The best-performing d and batch size were found to be 100 and 256, respectively. Notably, when $d = 100$, the model had a total of 247,811 parameters. Consequently, the training dataset was reduced to 12.46%. During the popVAE inference across the entire country of Tunisia, 87.54% of the data comprised completely unseen regions. Testing or evaluating our model on this portion of the dataset, which was not part of the training process, demonstrates our approach ability to generalize to new unseen regions.

B. Evaluation Methodology

Following several approach studies, such as [18], [24], and [5], which train models to predict raster population map then rely on coarse (region-level) population counts performed by official agencies or organizations (e.g., census data or advanced studies), we evaluated the predicted final raster population map based on actual population counts at the

governorate level, provided by the INS. The performance of the model was evaluated using two key metrics MSE and R²:

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(w_t - \hat{w}_t)^2, R^2 = 1 - \frac{\sum_{t=1}^{N}(w_t - \hat{w}_t)^2}{\sum_{t=1}^{N}(w_t - \bar{w})^2} \quad (4)$$

Here, N is the number of governorates, which in our case is 24. w_t represents the actual population at the governorate level, as provided by the INS, and $\hat{w}_t$ represents the predicted population at the governorate level.

MSE measures the average of the squares of the errors—that is, the average squared difference between $\hat{w}_t$ and w_t. Lower MSE values indicate better model performance.

R^2 is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1 indicates a better fit, meaning that the model accounts for a larger proportion of the variance. Additionally, R^2 can be negative, indicating a poor fit.

For our case, we construct the Raster Predicted Population Map at the pixel level (mono-band with dimensions of $w = 3,807$ by $h = 8,116$ pixels) by applying the popVAE inference to the entire raster ancillary dataset. To obtain $\hat{w}_t$, we aggregate the pixel-level predictions using the raster governorate boundary as an element-wise multiplier ($\odot$). Fig. 3 illustrates an example showing how we calculated the predicted population for the governorate of Siliana. In this example, an overestimation error of 10,688 so the Relative Error (RE) for Siliana is approximately $\frac{10,688}{227,992} \times 100 \approx 4.69\%$. Additionally, for the capital Tunis, the aggregated predicted population is 1,014,544, while the actual population provided by the INS is 1,073,249, resulting in an underestimation error of 58,705. The RE for Tunis is approximately 5.47%.

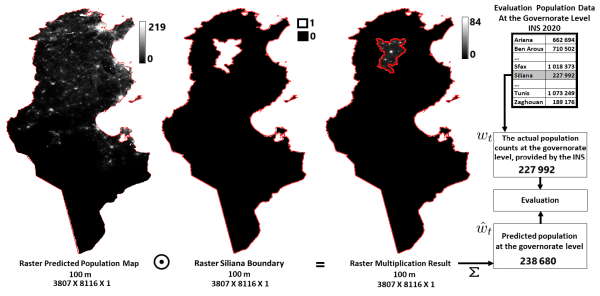

Fig. 3. Evaluation and Population Prediction at The Governorate Level.

C. Comparison with CNN-Based Contextual Feature Extraction

To evaluate the effectiveness of popVAE in incorporating spatial contextual features, we compare it with a modified architecture, denoted as popCNN, which is similar to the approaches proposed by [10]. This model relies solely on convolutional layers for spatial contextual feature extraction. In this alternative architecture, instead of a VAE for encoding latent spatial features, we employ a CNN as the feature extractor for spatial context.

The popCNN model processes X_c using the same convolutional and residual blocks to capture local and spatial dependencies within the data. The flattened features are then passed through a dense layer, producing a 32-dimensional feature representation, which serves as the spatial context features. To integrate this contextual features into the population prediction branch, we reshape the output to match the dimensions of the primary input X_p and concatenate them. Finally, the population prediction branch remain unchanged.

The primary goal of this comparison is to assess the performance and efficiency of popVAE's latent representation against a more conventional CNN-based spatial contextual feature extractor, popCNN. By comparing these two architectures, we aim to quantify the benefits of using a VAE for spatial contextual feature extraction in terms of predictive accuracy.

Our model, popVAE, achieves a significantly higher R^2 (0.8760) than popCNN (0.8154), indicating popVAE's stronger fit to population distribution patterns. Additionally, popVAE shows a lower MSE (8.20×10^9 for popVAE and $12.20 \times 10^9$for popCNN), suggesting improved accuracy in estimating population densities across diverse regions.

D. Visualization of Predicted Population Maps

We generated heatmaps (Fig. 4-a) to evaluate the population density predictions of popVAE and popCNN models, using WorldPop as the reference for comparison. The heatmaps feature a color scale ranging from pink for low-density areas to red for highly populated regions. To provide a comprehensive analysis, we selected five diverse regions: (A) Tunis, the capital; (B) a northern region; (C) Sfax, a key economic hub; (D) a southern region; and (E) an uninhabited desert area.

In Region A, both popVAE and popCNN are showing a broad red spot, indicating that both models can capture highly populated areas. However, the red spot in popVAE is closer to the WorldPop red spot, while the red spot in popCNN is larger, suggesting an overestimation by popCNN in dense urban settings. This highlights an improved accuracy of popVAE in such areas. Additionally, in Region C, popVAE's estimation closely matches Worldpop in populated areas, while popCNN indicates a very high population density at the governorate center, which is not accurate.

For Region B, both popVAE and popCNN fail to capture the small, lower-density purple zones, reflecting their reduced sensitivity to sparsely populated areas in the north. However, in Region D, both two models successfully capture the small, lower-density purple zones, demonstrating improved sensitivity to sparsely populated areas in the south.

For Region E, both models successfully capture the extensive pink areas, demonstrating accurate estimation in uninhabited desert regions.

E. Predicted vs Actual Population by Governorate

Figure 4-b presents scatter plots comparing the predicted population values from the popCNN and popVAE models

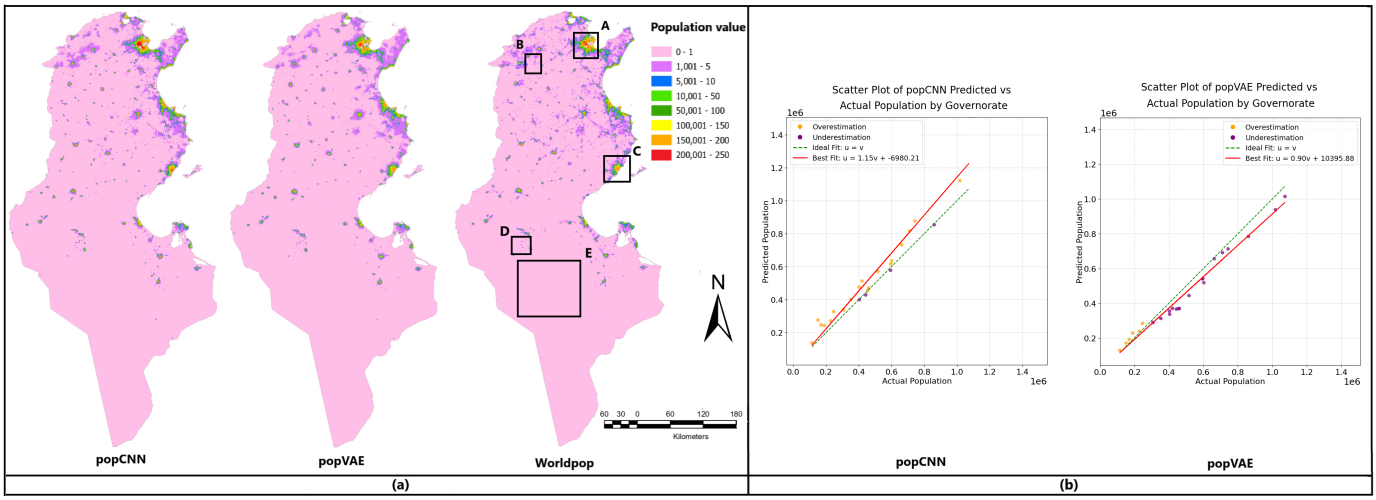

Fig. 4. Comparison of Predicted Population Metrics: Heatmaps and Scatter Plots for popCNN and popVAE Models Against References

against the actual population values provided by INS for each governorate. The comparison aims to evaluate the predictive performance of both models by examining their alignment with the ideal fit. In each scatter plot, the actual population values are shown on the horizontal axis, while the predicted population values are plotted on the vertical axis. The identity line or the line of equality $u = v$ represents the ideal fit (green), where predictions perfectly match the actual values. A line of best fit is also included for each model to indicate the trend of its predictions (red). The individual data points, representing governorates, are color-coded based on the prediction's accuracy relative to the actual values: overestimation is marked in orange, while underestimation is marked in purple.

The scatter plots reveal that the line of best fit for popVAE predictions is closer to the identity line compared to popCNN. Specifically, the slope of the best fit line for popVAE is 0.9, while the slope for popCNN is 1.15. This suggests that popVAE predictions have a better overall alignment with the actual values, while popCNN shows a tendency to overestimate the population. Overall, this comparison highlights that popVAE achieves a better balance between predicted and actual values, as indicated by its closer alignment to the ideal fit line.

F. Comparison with Baseline Models

TABLE III
COMPARATIVE PERFORMANCE OF POPULATION ESTIMATION MODELS

Model	R^2	MSE
popVAE	0.8760	8.20×10^9
LandScan	0.9647	2.34×10^9
GPWv4	0.9633	2.05×10^9
Cheng et al.	0.8369	1.05×10^{10}

In Table III, we compare the performance of popVAE with LandScan [16] and GPWv4 [17] at the pixel level for the year 2020 with a resolution of 1 km (https://landscan.ornl.gov/ and https://www.earthdata.nasa.gov/data/catalog/sedac-ciesin-sedac-gpwv4-popdens-r11-4.11), as well as the Cheng et al. approach [10]. The Cheng model was implemented following their paper and trained on the same reduced 12.46% of the raster ancillary dataset described in Section II-B, with 87.54% of the data comprising unseen regions to evaluate generalization.

We selected the Cheng et al. approach for its similarity to our architecture, as both use a dual-branch structure: one for spatial contextual feature extraction and the other for combining these features with pixel-level data to predict population. Cheng relies on CNNs for feature extraction, whereas popVAE uses a VAE to learn robust latent representations.

Our model, popVAE, achieves an R^2 of 0.8760, which, while lower than LandScan's 0.9647 and GPWv4's 0.9633, outperforms Cheng et al.'s 0.8369. This indicates that popVAE effectively captures population distribution patterns, though it does not reach the precision level of LandScan and GPWv4. In terms of MSE, popVAE (8.20×10^9) also falls between LandScan (2.34×10^9) and GPWv4 (2.05×10^9), and Cheng et al. (1.05×10^{10}). These results demonstrate that popVAE's prediction errors are moderate compared to the high accuracy of LandScan and GPWv4, and the higher errors of Cheng et al. and.

G. Analysis of Findings

The higher R^2 value achieved by popVAE compared to popCNN and Cheng et al. indicates that the VAE architecture is more effective at capturing population density patterns, particularly in regions with varying densities. The VAE's robust latent representations enhance predictive accuracy.

In contrast, CNN-based approaches like popCNN struggle with broader spatial contexts, leading to overestimation in dense urban areas such as Region A and C. While effective for feature extraction, CNNs may be less suited for tasks requiring complex spatial dependencies.

Compared to baseline models, popVAE achieves lower MSE than popCNN and Cheng et al., though slightly higher than LandScan and GPWv4, demonstrating reasonable accuracy.

In summary, VAEs offer significant advantages in extracting spatial contextual features, improving performance and adaptability for complex spatial relationships.

IV. Conclusion

The proposed popVAE model effectively captured complex spatial dependencies, improving population density representation in both urban and rural areas. Its VAE architecture ensured flexibility across heterogeneous contexts, while integrating diverse geospatial datasets for scalable, national-scale applications like population forecasting and disaster management.

PopVAE outperformed Cheng et al. approach, and showed competitive performance with LandScan and GPWv4, demonstrating its potential for real-world use.

Future work will focus on refining spatial contextual integration to better address challenges in dense urban regions.

References

[1] Nurrokhmah Rizqihandari and Satria Indratmoko, "Using openstreetmap data for population distribution model," in *1st International Conference on Geography and Education (ICGE 2016)*. Atlantis Press, 2016, pp. 244–247.

[2] T Grippa, C Linard, M Lennert, S Georganos, N Mboga, S Vanhuysse, and A Gadiaga, "Improving urban population distribution models with very-high resolution satellite information. data. 2019; 4: 13," .

[3] Martin Breunig, Patrick Erik Bradley, and Jahn, "Geospatial data management research: Progress and future directions," *ISPRS International Journal of Geo-Information*, vol. 9, no. 2, pp. 95, 2020.

[4] Casper Samsø Fibæk, Carsten Keßler, Jamal Jokar Arsanjani, and Marcia Luz Trillo, "A deep learning method for creating globally applicable population estimates from sentinel data," *Transactions in GIS*, vol. 26, no. 8, pp. 3147–3175, 2022.

[5] N Metzger, JE Vargas-Muñoz, RC Daudt, et al., "Fine-grained population mapping from coarse census counts and open geodata, sci rep, 12, 20085," 2022.

[6] Srishti Gaur and Rajendra Singh, "A comprehensive review on land use/land cover (lulc) change modeling for urban development: current status and future prospects," *Sustainability*, vol. 15, no. 2, pp. 903, 2023.

[7] Delmar E Anderson and Philip N Anderson, "Population estimates by humans and machines," *Photogrammetric Engineering*, vol. 39, no. 2, pp. 147–154, 1973.

[8] Caiyun Zhang and Fang Qiu, "A point-based intelligent approach to areal interpolation," *The Professional Geographer*, vol. 63, no. 2, pp. 262–276, 2011.

[9] Mohamed Bakillah, Steve Liang, et al., "Fine-resolution population mapping using openstreetmap points-of-interest," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1940–1963, 2014.

[10] Luxiao Cheng, Lizhe Wang, Ruyi Feng, and Jining Yan, "Remote sensing and social sensing data fusion for fine-resolution population mapping with a multimodel neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5973–5987, 2021.

[11] Yunchen Wang, Chunlin Huang, Minyan Zhao, Jinliang Hou, Ying Zhang, and Juan Gu, "Mapping the population density in mainland china using npp/viirs and points-of-interest data based on a random forests model," *Remote Sensing*, vol. 12, no. 21, pp. 3645, 2020.

[12] Kangning Li, Yunhao Chen, and Ying Li, "The random forest-based method of fine-resolution population spatialization by using the international space station nighttime photography and social sensing data," *Remote Sensing*, vol. 10, no. 10, pp. 1650, 2018.

[13] Stefan Leyk, Andrea E Gaughan, Adamo, et al., "The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use," *Earth System Science Data*, vol. 11, no. 3, pp. 1385–1409, 2019.

[14] Rachel H Swanwick, Quentin D Read, et al., "Dasymetric population mapping based on us census data and 30-m gridded estimates of impervious surface," *Scientific Data*, vol. 9, no. 1, pp. 523, 2022.

[15] Xin Huang and Ying Wang, "Investigating the effects of 3d urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of wuhan, central china," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 119–131, 2019.

[16] Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley, "Landscan: a global population database for estimating populations at risk," *Photogrammetric engineering and remote sensing*, vol. 66, no. 7, pp. 849–857, 2000.

[17] Erin Doxsey-Whitfield, Kytt MacManus, et al., "Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4," *Papers in Applied Geography*, vol. 1, no. 3, pp. 226–234, 2015.

[18] Caleb Robinson, Fred Hohman, and Bistra Dilkina, "A deep learning approach for population estimation from satellite imagery," in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, 2017, pp. 47–54.

[19] Wenjie Hu, Jay Harshadbhai Patel, Zoe-Alanah Robert, and Novosad, "Mapping missing population in rural india: A deep learning approach with satellite imagery," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 353–359.

[20] Xiao Huang, Di Zhu, Fan Zhang, Tao Liu, Xiao Li, and Lei Zou, "Sensing population distribution from satellite imagery via deep learning: Model selection, neighboring effects, and systematic biases," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5137–5151, 2021.

[21] Yanxiao Jiang, Zhou Huang, Linna Li, and Quanhua Dong, "Local–global dual attention network (lganet) for population estimation using remote sensing imagery," *Resources, Environment and Sustainability*, vol. 14, pp. 100136, 2023.

[22] Sugandha Doda, Matthias Kahl, and Ouan, "Interpretable deep learning for consistent large-scale urban population estimation using earth observation data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 128, pp. 103731, 2024.

[23] Luciano Gervasoni, Serge Fenet, Régis Perrier, and Peter Sturm, "Convolutional neural networks for disaggregated population mapping using open data," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2018, pp. 594–603.

[24] Nathan Jacobs, Adam Kraft, et al., "A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018, pp. 33–42.

[25] Diederik P Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] "Worldpop. tunisia 100m population, version 2. https://hub.worldpop.org/doi/10.5258/soton/wp00536," Accessed: May 09, 2024.

[27] National Open Data Platform, "https://catalog.data.gov.tn/fr/dataset/liste-des-coordonnees-gps-relatives-aux-etablissements-scolaires," Accessed: May 09, 2024.

[28] "Eoc geoservice," https://geoservice.dlr.de/web/maps, Accessed on May 09, 2024.

[29] Maninder Singh Dhillon, Carina Kübert-Flock, Thorsten Dahms, et al., "Evaluation of modis, landsat 8 and sentinel-2 data for accurate crop yield predictions: a case study using starfm ndvi in bavaria, germany," *Remote Sensing*, vol. 15, no. 7, pp. 1830, 2023.

[30] Christopher F Brown, Steven P Brumby, and Guzder-Williams, "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, no. 1, pp. 251, 2022.

[31] Tom G Farr, Paul A Rosen, Edward Caro, Ladislav Crippen, et al., "The shuttle radar topography mission," *Reviews of geophysics*, vol. 45, no. 2, 2007.

[32] The Food and Agriculture Organizatio, "https://data.apps.fao.org/catalog/dataset/global-administrative-unit-layers-gaul," Accessed: May 17, 2024.

[33] Haifa Tamiminia, Bahram Salehi, and Mahdianpari, "Google earth engine for geo-big data applications," *ISPRS journal of photogrammetry and remote sensing*, vol. 164, pp. 152–170, 2020.