From Space to Economy: Deep Learning Fusion of Remote Sensing Data for Sub-National GDP Estimation

Thomas Stark, Michael Wurm, Simon Krenn, Wolfgang Sulzer, Eleanor Stokes, Karen C. Seto, and Hannes Taubenböck

***Abstract*—Reliable and up-to-date sub-national data on the gross domestic product (GDP) data remain scarce in many regions across the globe. This limits the ability to analyze spatial economic disparities and to design evidence-based development policies. In this study, we investigate whether high-resolution GDP estimates can be derived from multisource remote sensing and auxiliary inputs using a deep learning fusion framework. We combine optical day- and night-time satellite imagery, and auxiliary geospatial layers with several backbone architectures and fusion strategies to assess the robustness of multimodal learning for economic prediction. Across extensive experiments in Brazil, we evaluate multiple encoders (ResNet-18, EfficientNet-B3, and SwinV2-T), alternative fusion layers (concatenation, attention pooling, graph-based multilayer perceptron, and mixture-of-experts), and diverse input modality combinations. Our best-performing model, an EfficientNet-B3 encoder with concatenation fusion using all available input modalities, achieves an R^2 value of 0.87 for GDP prediction at a spatial resolution of 5 km $\times$ 5 km, demonstrating that our multimodal approach effectively captures the complex relationships between spatial patterns and economic activity. These findings highlight the potential of multimodal remote sensing to complement traditional statistical sources by providing spatially consistent, high-resolution representations of economic activity.**

***Index Terms*—Deep learning fusion, gross domestic product (GDP), multimodal data integration, remote sensing.**

I. Introduction

GROSS domestic product (GDP) is defined as the monetary value of all final goods and services produced within a country during a specified period and is widely used to measure economic activity, allocate resources, and compare economies over time and space [1], [2]. As a standardized monetary indicator, GDP enables cross-country and interregional comparisons of economic development, productivity, and living standards across national and subnational contexts [3], [4], [5]. Such comparability is central to international policy coordination, development planning, and the allocation of financial support by multilateral institutions [6]. It also underpins empirical research in growth economics, regional science, and economic geography, where harmonized indicators are required to analyze spatial disparities and convergence processes [7], [8]. At the same time, GDP is typically reported for administrative units, which vary in size and are often not directly comparable across regions [9]. This motivates the need for spatially explicit, high-resolution GDP representations that enable consistent analysis of economic activity beyond administrative boundaries [10].

However, GDP has well-documented limitations, as it captures only market transactions and omits nonmarket activities, such as household labor and environmental services [11], [12]. Moreover, GDP estimates are subject to measurement errors due to incomplete and imperfect data, often leading to inconsistencies and revisions [12], [13], [14], [15], [16], [17]. These challenges are particularly pronounced in developing and emerging economies, where large informal sectors, undocumented populations, and incomplete census and administrative records complicate the accurate estimation of economic activity [17], [18]. As a result, official GDP statistics may systematically misrepresent local economic conditions, especially in regions characterized by informal settlements and limited statistical coverage.

Recent research has explored the potential of remotely sensed data as a proxy for economic activity. While individual data sources, such as nighttime lights or optical imagery, have been shown to correlate with GDP, it remains unclear how multiple heterogeneous geospatial inputs can be effectively combined within a unified deep learning framework [19], [20].

First, we address the question of which types of remote sensing and geospatial data are most informative for estimating spatially explicit economic output. Second, we investigate how fundamentally different input modalities ranging from optical day- and night-time imagery to auxiliary layers describing land use, infrastructure, and settlement characteristics can be fused in a data-driven manner to capture complementary signals of economic activity [21].

We evaluate this methodological framework in an emerging market context. We select Brazil as a data-rich yet representative setting that provides reliable reference GDP data

Received 19 December 2025; revised 15 April 2026; accepted 22 May 2026. Date of publication 29 May 2026; date of current version 15 June 2026. *(Corresponding author: Thomas Stark.)*

Thomas Stark, Michael Wurm, and Simon Krenn are with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany (e-mail: thomas.stark@dlr.de).

Wolfgang Sulzer is with the Institute of Geography and Geology, University of Würzburg, 97074 Würzburg, Germany.

Eleanor Stokes is with the University of Graz, Geography and Regional Sciences, 8010 Graz, Austria.

Karen C. Seto is with NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA.

Hannes Taubenböck is with German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany, and also with the Yale School of the Environment, Yale University, New Haven, CT 06511 USA.

Digital Object Identifier 10.1109/JSTARS.2026.3697872

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

while still exhibiting challenges such as informal economic activity. This allows for a controlled assessment of model performance while reflecting real-world complexities. Although the primary objective is not to replace official statistics in well-documented regions, the proposed framework is inherently multimodal and particularly suited for settings, where GDP data are sparse or unreliable, but remote sensing and auxiliary geospatial data are readily available.

II. Related Work

The spatial distribution of economic activity has long been a key focus in urban and regional research, with numerous studies examining how urban form and spatial organization shape development outcomes [19], [20]. Empirical evidence shows that specific spatial structures, such as monocentric or polycentric configurations and functional land-use differentiation, are closely associated with productivity and regional GDP patterns [22], [23]. More generally, remotely observable urban form has been shown to provide a meaningful proxy for underlying economic processes, particularly in contexts where conventional economic statistics are limited or delayed [8].

Drawing on advances in remote sensing and machine learning, a growing body of work has explored approaches to estimate GDP and economic activity from spatial data. A human–machine collaborative ranking framework has been used to derive economic development scores directly from 10 m Copernicus imagery without relying on ground-truth GDP data [24]. Nighttime light products, particularly VIIRS Black Marble, are widely employed as proxies for human activity and urban infrastructure, and are frequently combined with additional geospatial datasets to model economic patterns [25], [26], [27], [28].

Several studies integrate multisource data to improve spatial GDP estimation. For example, corrected nighttime-light datasets have been combined with Sentinel-2 multispectral imagery, where machine-learning models such as random forests are used to extract built-up areas and derive indicators of economic activity [29]. Similarly, land-use information and nighttime lights have been fused using statistical and machine-learning approaches, including factor-based aggregation, random forests, and gradient boosting, to generate gridded GDP products [30]. Other approaches incorporate additional data modalities, such as street-view imagery and points of interest (POI), to capture fine-scale socioeconomic characteristics [31]. Deep-learning-based fusion architectures have also been proposed, where multiple data streams such as remote sensing imagery and POI data are processed and integrated to produce high-resolution GDP maps [32].

Beyond national-scale studies, global and regional analyses have applied combinations of nighttime lights, land-cover data, and environmental variables within both machine-learning and spatial-statistical frameworks. These include the use of unsupervised anomaly detection to derive global economic indicators [33], geographically weighted regression to model spatially varying relationships between lights and GDP [34], and multisource machine-learning models combining nighttime lights with agricultural and climate variables [35]. Additional studies have used nighttime lights in combination with environmental and socioeconomic covariates to analyze temporal dynamics of economic activity, for example, during the COVID-19 pandemic or in data-scarce regions [36], [37].

Within the context of Brazil, recent work has explored the integration of optical satellite imagery and nighttime light data on regular grids, where individual data modalities are processed separately and subsequently fused to capture complementary physical and socioeconomic signals of economic activity [21].

Multimodal input data fusion has become a key strategy in remote sensing classification over the last few years. Recent work shows that integrating complementary data modalities, such as hyperspectral (HSI), multispectral (MSI), radar (SAR), LiDAR, and mobile sensing improves classification accuracy and robustness. An X-FCN architecture is used in [38] to construct a multimodal input datacube integrating optical imagery, spectral indices, and OpenStreetMap (OSM) layers. The study illustrates how combining heterogeneous geospatial data sources can substantially improve the characterization of complex urban landscapes. In [39], a fusion approach combining a MobileNet encoder with an artificial neural network (ANN) and a concatenation-based fusion layer was used to predict urban public green spaces from remote sensing imagery and OSM data. Another approach is, TGF-Net, which decomposes modality-specific features and uses a transformer-based spectral feature extraction module and a gist-based spatial module, thereby enhancing joint HSI–SAR classification [20]. Urban functional zone mapping converts mobile signalling and POI data into images and fuses them via deep learning; the fusion achieves about 95.13% accuracy and outperforms single-modality classifiers [40]. MGSViT, a self-supervised gated multimodal transformer, pretrains a vision transformer on BigEarthNet and uses intra- and intermodal gated units to fuse MS and SAR features, addressing heterogeneity and limited labels [41]. Other studies have proposed fusion at multiple resolutions or scales. A collaborative fusion method combines HSI, MSI, and SAR data using spectral–spatial weighted modulation and spectral compensation, improving classification metrics [42]. Cross-attention-induced transformers and dual-branch transformers emphasise interactions between HSI and LiDAR or MSI, capturing global–local and long-range dependencies for enhanced classification [43], [44]. LiDAR-HSI fusion using residual networks or random forests leverages spectral richness and canopy height information, yielding high accuracy for ground object and tree species classification [45], [46]. In [47] and [48], multisensor remote-sensing data were combined with machine-learning techniques to derive detailed information on urban structures in data-scarce environments. Optical and radar imagery were jointly processed to extract, select, and classify spectral–spatial features using supervised algorithms such as random forests and support vector machines. The integration of complementary sensor information was shown to enhance the characterization of complex urban morphologies, demonstrating the effectiveness of data-driven fusion frameworks for large-scale mapping tasks. Finally, cross-modal cross-attention transformers use pyramid architectures to exploit multiscale features and cross-modal interactions for HSI-LiDAR

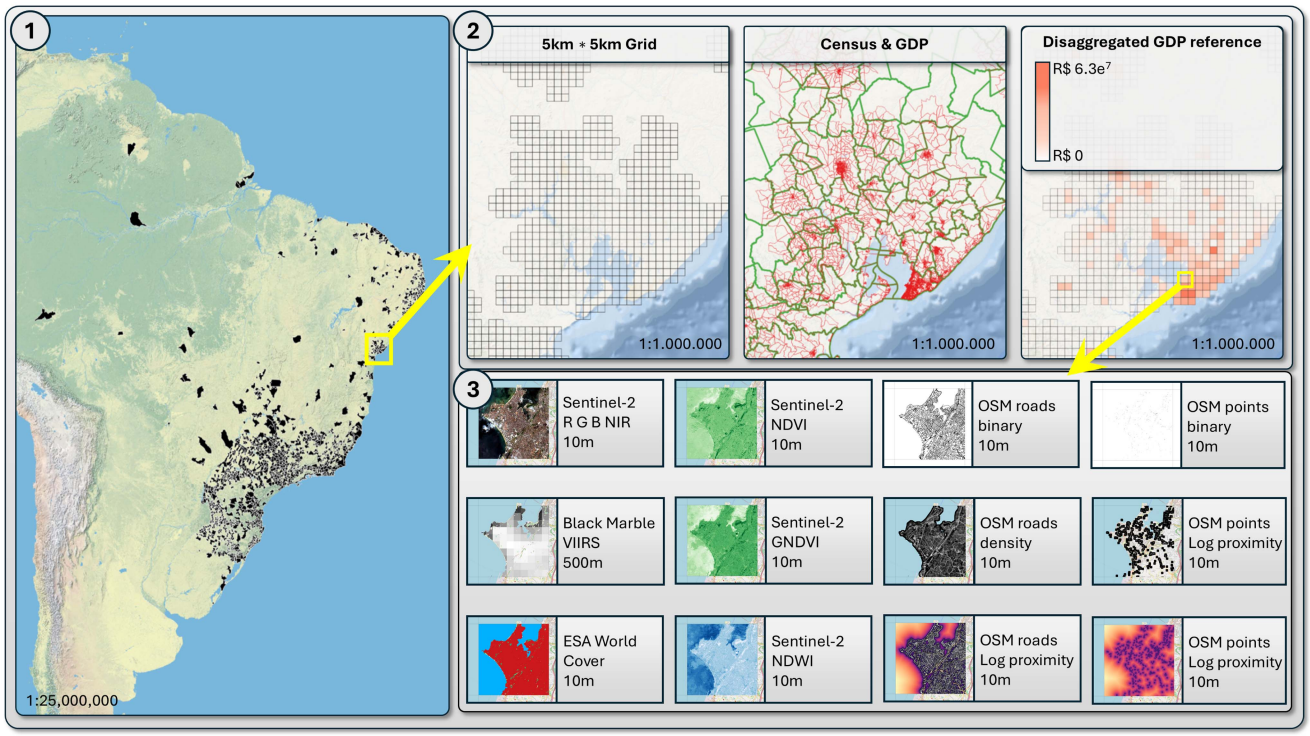

Fig. 1. Dataset overview. On the left is the complete area of interest within Brazil with the 44 652 selected 5 km × 5 km grid cells. On the top right the metropolitan area of Salvador with the 5 km × 5 km grid, the census and GDP municipal data and the disaggregated GDP reference data used to train the deep learning models. Bottom right are all available input modalities for one sample 5 km × 5 km grid cell.

classification [49]. These studies collectively demonstrate that appropriately modeling cross-modal complementarities, combining global and local features, and addressing scale differences are crucial for effective remote-sensing classification tasks.

III. Methods and Materials

A. Dataset

Brazil, one of the world's largest emerging market economies, presents a highly diverse socioeconomic landscape and significant regional disparities in economic activity [50]. As seen in Fig. 1.1 on the left side approximately 9.06% of Brazil's total area (8 515 760 km^2) was selected, representing regions that together contribute about 74.75% of the national GDP (6.267 $TR\$$).

To identify these regions, we ranked all municipalities by their GDP values in descending order and progressively aggregated them following a constraint-driven approach rather than a predefined threshold. Municipalities were included until 1) all major regions of Brazil, including economically weaker and sparsely populated areas such as the Amazonas region, were represented, and 2) the total spatial extent remained manageable for data acquisition and processing. The resulting threshold of 74.75% of national GDP therefore emerges from these constraints.

This results in an area of interest of 771 605 km^2 with a GDP value of 4.684 $TR\$$. This sampling strategy ensured a balance between computational efficiency and data representativeness, while covering all major economic sectors and including both urban centers and rural areas of productive activity.

All datasets were transformed into a common coordinate reference system SIRGAS 2000/Brazil Polyconic (EPSG: 5880) and a regular 5 km × 5 km grid was created. This spatial resolution was selected to ensure compatibility across all input and target data modalities.

The reference data consist of various economic indicators provided by the Brazilian Institute of Geography and Statistics (IBGE). Two data levels were used: 1) GDP at the municipal level from 2016 [51], and 2) income data at the neighborhood level from the 2010 census data [52].

The data sources include Sentinel-2 multispectral imagery [53], nighttime light (NTL) data from NASA's Black Marble product [54], land cover information from ESA's WorldCover dataset [55], and OSM features representing roads and POIs [56].

The optical remote sensing component is based on Sentinel-2 Level-2 A imagery acquired between 2019 and 2021. To ensure consistent radiometric conditions and minimize seasonal variation, mosaics were created using only observations from the summer months (December to March). The imagery was processed to 8-bit RGB and near-infrared (NIR) composites at a spatial resolution of 10{ m. From these bands, three vegetation and water indices were calculated per pixel: the normalized difference vegetation index (NDVI), the Green NDVI (GNDVI), and the normalized difference water index (NDWI). These indices provide complementary information on vegetation vigor and surface moisture conditions relevant for economic and land-use characterization.

Land cover information was derived from the ESA WorldCover 2021 dataset, which offers a global 10 m classification

Algorithm 1: Weighted GDP Disaggregation Based on Intersected Income Polygons.

Require: GDP polygons G, income polygons I, target grid T
Ensure: GDP estimates for all grid cells in T
Compute the intersection of I and T
Recalculate the area of each intersected income sub-polygon
Redistribute original income values proportionally to the area share of each sub-polygon
Convert intersected income sub-polygons to centroid points
Assign each centroid to its corresponding GDP polygon by spatial join
for each GDP polygon **do**
  Sum all rescaled income values of the assigned centroids
  Compute the relative income share of each centroid
  Allocate the polygon GDP proportionally according to these income shares
end for
Aggregate all allocated centroid GDP values to the 5×5 km grid cells
returngridded GDP surface

into 11 categories: Tree Cover, Shrubland, Grassland, Cropland, Built-Up, Bare/Sparse Vegetation, Snow and Ice, Herbaceous Wetland, Mangroves, and Moss/Lichen. The Built-Up class was particularly used to spatially constrain urban analyses and to mask noninhabited areas for subsequent data fusion.

NTL was obtained from the NASA Black Marble VNP46A4 annual composite for 2016 at its original 452 m resolution, resampled to 500 m. The NTL data were clipped to the Built-Up extent from WorldCover to focus on economically active regions and to reduce illumination bias from nonurban sources.

Finally, OSM data were used to represent human infrastructure and accessibility. Road networks and POIs were converted into binary raster layers and further processed to compute proximity rasters, capturing spatial relationships between economic activity and infrastructure density.

B. GDP Data and Its Disaggregation Into Grids

To harmonize all spatial datasets, both GDP and income polygons were intersected with a uniform 5×5 km grid covering the entire study area (44 000 cells) as seen in Fig. 1.1. This ensures a consistent spatial reference for all subsequent analyses. Following the intersection, the resulting sub-polygons were recalculated in terms of their area. Income values were then proportionally redistributed based on the share of each sub-polygon relative to the original polygon area, resulting in a new variable ("rescaled income"). For spatial linkage, the intersected income polygons were converted into centroids and assigned to the corresponding GDP polygons using a spatial join. The disaggregation of GDP was then performed using a weighted allocation approach, where income served as a proxy distribution variable. Specifically, each centroid received a share of GDP proportional to its relative contribution to the total income within the corresponding administrative unit. Finally, the disaggregated GDP values were aggregated to the 5×5 km grid cells, resulting in a spatially consistent gridded GDP dataset. The exact diasaggregation precess can be followed in Algorithm 1.

For validation, the values within the grid cells were reaggregated following the same principle. The results showed that the information loss of the economic values resulting from this aggregation remained below 1%. Moreover, a strong correlation between income and GDP is reported in [57], [58], [59], confirming the suitability of income as a reliable disaggregation variable. The availability of income data at the neighborhood level further enabled a high-resolution spatial representation of economic patterns, ensuring that the downscaling approach preserved both the internal consistency and the spatial heterogeneity of the underlying socioeconomic information.

C. Multimodal Data Fusion

The proposed multimodal regression framework, as seen in Fig. 2, integrates several input streams, each standardized to a spatial dimension of $224 \times 224 \times 3$ pixels, within a single deep learning architecture. Every modality is first processed independently using a dedicated encoder $f_n(x)$. An example of each three layerd input stack can be seen in Fig. 1.3 on the bottom right. Multiple backbone variants were evaluated, including ResNet-18, EfficientNet-B3, and SwinV2-T, to extract high-level feature representations from the corresponding input tensor.

ResNet-18 [60], selected due to its widespread adoption and well-understood performance characteristics in image-based learning; EfficientNet-B3 [61], chosen for its high parameter efficiency and strong accuracy computational trade-off; and SwinV2-T [62], included as a hierarchical Vision Transformer capable of modeling long-range spatial dependencies. Together, these encoders enable a balanced comparison between conventional CNNs and modern transformer-based representations.

To address robustness and reduce overfitting, a data-augmentation pipeline is applied during training. It includes basic, color-based, geometric, and mixed augmentations, supporting generalization across varying imaging conditions and acquisition settings [63], [64].

Modality-specific feature vectors $\mathbf{z}$ are subsequently combined using a configurable fusion module that supports multiple strategies, including concatenation, attention pooling, graph-based multilayer perceptron (MLP), and mixture-of-experts gating configurations. Each fusion module is illustrated in the center of Fig. 2, where the inputs $x^{(n)}$ are taken directly from the final layer of the feature encoder $f_n(x)$.

Concatenation in (1) provides the baseline, merging features along the channel dimension to form a single high-dimensional vector. This preserves all information without interaction between modalities, allowing the downstream regressor to learn

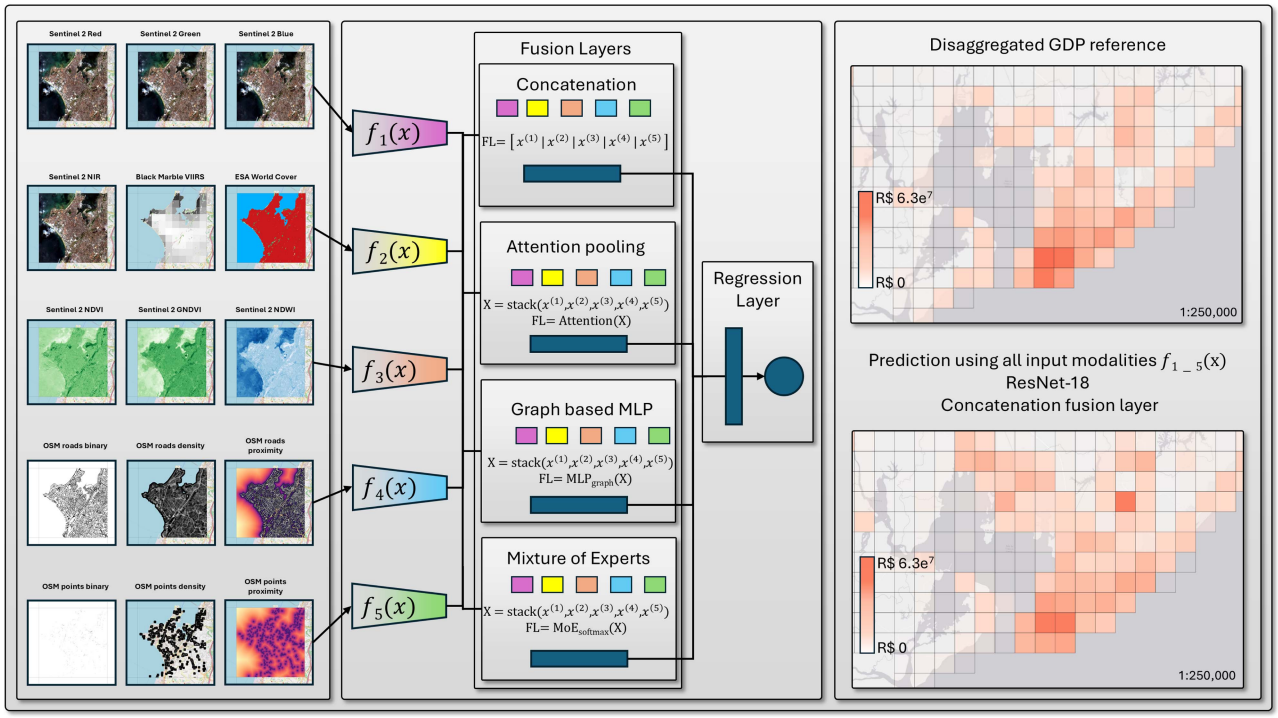

Fig. 2. Proposed method integrates 15 input modalities, grouped into three input channels. Each modality is encoded by a dedicated feature extractor (ResNet-18, EfficientNet-B3, or SwinV2-T). The resulting feature representations are combined using one of four fusion strategies, concatenation, attention pooling, graph-based MLP, or mixture-of-experts. The fused feature vector is subsequently passed to a regression head to estimate normalized GDP values, which are finally transformed back to their real currency units ($R\$$).

cross-modal relationships implicitly [65]

$$\mathbf{z} = \left[\mathbf{x}^{(1)} \,\|\, \mathbf{x}^{(2)} \,\|\, \mathbf{x}^{(3)} \,\|\, \mathbf{x}^{(4)} \,\|\, \mathbf{x}^{(5)} \right] \tag{1}$$

$$\hat{y} = f_{\text{reg}}(\mathbf{z})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^D$ is the feature vector of modality i

$[\cdot\|\cdot]$ denotes vector concatenation

$\mathbf{z} \in \mathbb{R}^{5D}$ is the fused representation

$f_{\text{reg}}(\cdot)$ is the regression head

$\hat{y} \in \mathbb{R}$ is the predicted continuous target. (2)

Attention pooling in (5) introduces a learnable weighting mechanism inspired by self-attention. Here, modality features are stacked and queried using their mean representation, enabling the model to assign higher importance to more informative inputs and produce a context-aware fused vector [66].

$$\mathbf{X} = \text{stack}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(5)}\right) \in \mathbb{R}^{5\times D} \tag{3}$$

$$\mathbf{q} = \frac{1}{5}\sum_{i=1}^{5} \mathbf{x}^{(i)} \tag{4}$$

$$\mathbf{z} = \text{Attn}(\mathbf{q}, \mathbf{X}, \mathbf{X}) \tag{5}$$

$$\hat{y} = f_{\text{reg}}(\mathbf{z})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^D$ is the feature of modality i

$\mathbf{X}$ stacks all modality features into a matrix

$\mathbf{q} \in \mathbb{R}^D$ is the averaged query vector

$\text{Attn}(\cdot)$ is a multi-head attention operation

$\mathbf{z} \in \mathbb{R}^D$ is the attention-pooled feature

$\hat{y}$ is the regression output. (6)

Graph-based fusion in (9) treats modalities as nodes in a latent graph, applying a lightweight MLP to model pairwise dependencies before averaging. This encourages information exchange while retaining scalability for a small number of inputs [67].

$$\mathbf{X} = \text{stack}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(5)}\right) \in \mathbb{R}^{5\times D} \tag{7}$$

$$\mathbf{H} = \text{MLP}_{\text{graph}}(\mathbf{X}) \tag{8}$$

$$\mathbf{z} = \frac{1}{5}\sum_{i=1}^{5} \mathbf{H}_i \tag{9}$$

$$\hat{y} = f_{\text{reg}}(\mathbf{z})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^D$ is the input feature of modality i

$\mathbf{X}$ stacks all features as graph nodes

$\text{MLP}_{\text{graph}}(\cdot)$ models inter-modality interactions

$\mathbf{H}_i \in \mathbb{R}^D$ is the transformed node representation

$\mathbf{z}$ is the mean-pooled fused embedding

$\hat{y}$ is the final regression prediction. (10)

Finally, *mixture of experts* in (13) applies a gating network that predicts modality-specific weights based on the concatenated features. The weighted sum adaptively emphasizes relevant modalities, supporting dynamic fusion in heterogeneous data scenarios [68].

$$\mathbf{g} = \left[\mathbf{x}^{(1)} \,\|\, \mathbf{x}^{(2)} \,\|\, \mathbf{x}^{(3)} \,\|\, \mathbf{x}^{(4)} \,\|\, \mathbf{x}^{(5)} \right] \tag{11}$$

$$\mathbf{w} = \text{softmax}(f_{\text{gate}}(\mathbf{g})) \in \mathbb{R}^5 \tag{12}$$

$$\mathbf{z} = \sum_{i=1}^{5} w_i \mathbf{x}^{(i)} \tag{13}$$

$$\hat{y} = f_{\text{reg}}(\mathbf{z})$$

where $\mathbf{g} \in \mathbb{R}^{5D}$ is the gating input

$f_{\text{gate}}(\cdot)$ is the gating network

$\mathbf{w} = (w_1, \ldots, w_5)$ are normalized modality weights

$$\sum_{i=1}^{5} w_i = 1$$

$\mathbf{z} \in \mathbb{R}^D$ is the weighted fused feature

$\hat{y}$ is the regression output. (14)

Optional dropout and layer normalization are incorporated to stabilize optimization, reduce co-adaptation between inputs, and improve convergence. The fused feature representation is finally passed through a two-layer regression head composed of a linear transformation, ReLU activation, and a final output layer predicting a single continuous target value. Thus, this approach provides multimodal learning in high-resolution regression tasks.

D. Experimental SetUp

The experimental setup follows a stratified fourfold evaluation scheme to ensure a balanced distribution of economic conditions across all splits.

First, the full dataset is sorted in descending order based on the target GDP values. Samples are then assigned sequentially into four equally sized subsets, resulting in each fold containing a comparable proportion of high- and low-GDP samples. A k-fold procedure is applied in which, for every run, two folds are used for training, one fold for validation, and the remaining fold for testing. This rotation is repeated until all folds have served as the test set once, ensuring that results are independent of a specific data split.

All models are trained using identical hyperparameters for comparability. Training is performed with an initial learning rate of (1×10^{-4}) using the AdamW optimizer and Huber loss. A learning-rate scheduler reduces the rate by a factor of 0.5 if the validation loss does not improve for five epochs, with a minimum threshold of (1×10^{-7}). A dropout rate of 0.3 is applied to reduce overfitting. Each model is trained for a maximum of 100 epochs, with early stopping triggered if the validation mean absolute error does not improve for more than 20 consecutive epochs. The best model for each fold is selected based on the lowest validation mean absolute error and subsequently evaluated on the held-out test set.

All input modalities were preprocessed using z-score normalization based on the standard ImageNet channel statistics to ensure consistent scaling across heterogeneous data sources. The disaggregated GDP reference labels were transformed using a square-root stabilisation to reduce the influence of extreme high-value tails, followed by min–max scaling to the [0,1] interval. During inference, predictions were inversely transformed, first by applying the inverse min–max scaling and subsequently squaring the values to recover the original GDP estimates in $R\$$.

IV. Results

In this section, we present the results of all experiments. We first evaluate the influence of different fusion strategies using a single backbone feature extractor (ResNet-18) to isolate the effect of the fusion design. Based on the best-performing fusion methods, we then compare the three backbone architectures across the selected fusion strategy. Finally, leveraging the insights from the optimal backbone–fusion combination, we assess the contribution of the different input modality configurations.

To assess model performance across all experiments, we report accuracy metrics for both the normalized target space and the reconstructed GDP values in real currency $(R\$)$. We use the coefficient of determination (R^2) to quantify how well the predicted variance aligns with the observed variance, providing a scale-invariant measure suitable for comparing models across different configurations. Complementary to this, the root mean squared error (RMSE) captures the average magnitude of prediction errors and is particularly important because it reflects deviations directly in real $(R\$)$ units. Given the strongly long-tailed GDP distribution, with few very high values and many low ones, RMSE is naturally influenced by errors in the upper tail.

A. Fusion Methods

In Table I, we present the performance of all evaluated fusion strategies, detailing their parameter counts as well as their predictive accuracy for both normalized and absolute $(R\$)$ GDP targets. Across all methods, the best performance for the normalized output was achieved by the *Graph-based* fusion strategy, yielding the lowest RMSE (.0135) and the highest R^2 (.8812). The Concatenation approach followed closely with an RMSE of. 0136 and an R^2 of. 8805, while Attention Pooling performed slightly weaker. Mixture of Experts showed substantially reduced predictive accuracy for the normalized target.

For the absolute $R\$$ predictions, *Concatenation* produced the strongest results, reaching an RMSE of **329,971** $R\$$ and an R^2 of **. 8767**. Graph-based fusion obtained an RMSE of 400 008 $R\$$ (approximately 21.2% higher than Concatenation) and an R^2

TABLE I
COMPARISON OF DIFFERENT DATA FUSION STRATEGIES FOR GDP ESTIMATION

Fusion Method	Backbone	Backbone- and Fusion Layer		Normalized		R$	
		Parameters	Parameters	RMSE	R^2	RMSE	R^2
Concatenation	ResNet-18	56M	5K	.0136	.8805	**329,971**	**.8767**
Attention Pooling	ResNet-18	56M	1M	.0141	.8714	443,267	.7774
Graph-Based	ResNet-18	56M	526K	**.0135**	**.8812**	400,008	.8187
Mixture of Experts	ResNet-18	56M	13K	.0162	.8295	521,548	.6919

The table lists the backbone network, model and fusion layer parameters, and performance metrics (RMSE and R^2) for normalized and absolute (R$) outputs.

TABLE II
ENCODER COMPARISON ACROSS MODEL SIZE, BATCH SIZE, RUNTIME, AND PERFORMANCE

Encoder	Parameters	Training			Inference			Normalized		R$	
		Batch	Total	it/s	Batch	Total	it/s	RMSE	R^2	RMSE	R^2
ResNet-18	56M	64	1:02:28	1.02	64	00:40	4.31	.0136	.8805	**329,971**	**.8767**
EfficientNet-b3	57M	16	3:39:40	1.06	64	01:10	2.48	**.0132**	**.8863**	365,493	.8487
SwinV2-T	139M	8	5:59:56	1.30	64	01:41	1.73	.0158	.8385	446,533	.7741

Training/inference columns report batch size, total wall-clock time, and throughput (it/s). Metrics reported for normalized and absolute (R$) targets.

of. 8187. Attention Pooling showed a larger deviation, with an RMSE of 443 267 R (approximately 34.3% higher) and an R^2 of. 7774. Mixture of Experts resulted in the weakest R performance, with an RMSE of 521,548 R (approximately 58.1% higher than Concatenation) and an R^2 of. 6919.

Finally, comparing the fusion-layer parameter counts shows that *Concatenation* provides the most efficient configuration, requiring only 5 k parameters while still achieving the best absolute-scale performance. In contrast, Attention Pooling and Graph-based fusion introduce substantially larger fusion layers (1 M and 526 k parameters, respectively), while Mixture of Experts remains lightweight but does not translate its efficiency into competitive accuracy.

B. Encoders

In Table II, we present the performance of all evaluated encoder architectures, reporting their parameter counts, training and inference runtimes, and predictive accuracy for both normalized and absolute (R) GDP targets. Across all models, the best performance for the normalized output was achieved by *EfficientNet-b3*, yielding the lowest RMSE (.0132) and the highest R^2 (.8863). ResNet-18 followed closely with an RMSE of. 0136 and an R^2 of. 8805, while SwinV2-T showed noticeably lower accuracy for the normalized predictions.

For the absolute R outputs, *ResNet-18* produced the strongest results, reaching an RMSE of **329,971** R and an R^2 of **. 8767**. EfficientNet-b3 yielded an RMSE of 365 493 R (approximately 10.8% higher than ResNet-18) and an R^2 of. 8487, whereas SwinV2-T showed a larger deviation, with an RMSE of 446 533 R (approximately 35.3% higher) and an R^2 of. 7741.

Regarding computational efficiency, ResNet-18 provided the shortest training time (1:02:28) at a batch size of 64 and achieved the highest inference throughput of 4.31 it/s. EfficientNet-b3 required longer training time (3:39:40) with a smaller batch size of 16 and achieved 2.48 it/s during inference. SwinV2-T was the most computationally demanding model, requiring 5:59:56 for training with a batch size of 8 and achieving 1.73 it/s at inference. Overall, these results highlight the tradeoff between model capacity, runtime efficiency, and predictive performance.

C. Feature Importance

In Table III, we present the performance of all evaluated feature-set configurations, reporting their parameter counts, training and inference runtimes, and predictive accuracy for both normalized and absolute (R) GDP targets. Across all input variations, the best performance for the normalized output was achieved when using the full feature set (*S2 + All*), which produced the lowest RMSE (.0136) and the highest R^2 (.8805). The configuration combining S2 with NIR, Landcover, and NTL achieved the second-best normalized performance (RMSE. 0144, R^2. 8652), while the remaining feature sets showed weaker predictive accuracy.

For the absolute R outputs, the *S2 + All* configuration again yielded the strongest results, reaching an RMSE of **329,971** R and an R^2 of **. 8767**. The S2 + NIR, Landcover, NTL configuration followed with an RMSE of 350,284 R (approximately 6.2% higher than the full feature set) and an R^2 of. 8610. The remaining variations showed larger deviations: S2 alone resulted in an RMSE of 403,559 R (approximately 22.3% higher), S2 + NDVI/GNDVI/NDWI reached 428,051 R (approximately 29.8% higher), and S2 + OSM Points and Roads yielded 436,254 R (approximately 32.2% higher), each accompanied by lower R^2 values.

In terms of runtime, the smallest model (S2 only, 11 M parameters) achieved the fastest training time (22:41) and the highest inference throughput (5.13 it/s). Increasing the number of input modalities consistently raised model size and training time. The full feature-set model (56 M parameters) required the longest training time (1:02:28) and showed reduced throughput (1.02 it/s), reflecting the computational cost associated with integrating all available data streams.

TABLE III
FEATURE-SET COMPARISON FOR GDP ESTIMATION: IMPACT OF INPUT VARIATIONS ON PERFORMANCE AND RUNTIME

Input Variation	Parameters	Training		Inference		Normalized		R$	
		Total	it/s	Total	it/s	RMSE	R^2	RMSE	R^2
S2	11M	22:41	2.57	00:34	5.13	.0145	.8638	403,559	.8155
S2 + NDVI, GNDVI, NDWI	22M	34:29	1.80	00:40	4.31	.0164	.8253	428,051	.7924
S2 + NIR, Landcover, NTL	22M	30:75	1.86	00:39	4.43	.0144	.8652	350,284	.8610
S2 + OSM Points + OSM Roads	33M	40:80	1.41	00:38	4:59	.0159	.8364	436,254	.7844
S2 + All	56M	1:02:28	1.02	00:40	4.31	**.0136**	**.8805**	**329,971**	**.8767**

Metrics are reported for normalized and absolute (R$) targets; runtimes report total wall-clock and throughput (it/s).
Abbreviations: S2 = Sentinel-2 RGB; NDVI/GNDVI/NDWI = spectral indices based on S2; NIR = S2 near-infrared band; Landcover = ESA WorldCover; NTL = Black Marble VNP46A4; OSM = OpenStreetMap.

V. DISCUSSION

A. Multimodal Input Data Fusion

Overall, the experiments show that deep learning–based fusion of multiple remote sensing and geospatial data sources enables accurate estimation of GDP at the subnational level. The consistent alignment between predicted and official values indicates that spatial patterns of economic activity can be reliably inferred from remote sensing data. The approach therefore offers a complementary, data-driven means to support economic monitoring, particularly in data-scarce and rapidly changing regions.

Across the three experiments, several findings were noteworthy. First, as seen in Table I, concatenation performed quite strongly, even outperforming more advanced mechanisms such as attention pooling, graph-based fusion, or mixture-of-experts. One possible explanation is that, given sufficiently expressive backbone encoders, the fused feature space may already contain enough complementary information, making additional fusion complexity unnecessary or even detrimental in smaller datasets. This also highlights the strong efficiency–performance tradeoff, as concatenation introduces almost no additional parameters.

A similar pattern appears in Table II in the encoder comparison, where ResNet-18 achieved the best absolute (R\$) performance and nearly matched EfficientNet-b3 for normalized predictions. This suggests that, in multimodal fusion settings, representational gains from larger or transformer-based encoders may be limited. Transformers like SwinV2-T also required substantially longer training times, and the available dataset may be too small for such models to fully leverage their capacity.

Finally, the feature-set analysis seen in Table III shows that Sentinel-2 alone already yields competitive results, confirming its strong predictive value. As expected, adding NIR information, land cover, and especially night-time lights improves performance, and the full multimodal feature set achieves the best results overall.

B. Predicting High Resolution Sub National GDP Values

To illustrate the spatial behavior of the model across contrasting economic contexts, Fig. 3 presents prediction results of the ResNet-18 using concatenation of all input modalities for three representative areas of interest in Brazil. The examples span a governmental centre, a tourism- and service-oriented coastal city, and the country's largest metropolitan economy, thereby covering a wide range of GDP structures. Each region is shown at a consistent spatial scale, enabling a direct visual comparison between the GDP reference labels and the model predictions.

The first area of interest in Fig. 3 at the top illustrates the capital region of Brasília, where governmental, administrative, and institutional activities are the dominant drivers of economic output. The model accurately reproduces the spatial concentration of GDP around the federal district, capturing the structured urban layout and the distinct separation between high-value administrative zones and the surrounding lower-density residential areas. These patterns demonstrate the model's ability to reflect the highly centralized economic geography characteristic of Brasília.

The second example in Fig. 3 in the middle focuses on Salvador, a coastal metropolis shaped by a diverse service economy and a strong tourism sector. Here, economic activity is more spatially heterogeneous, influenced by historical neighborhoods, coastal attractions, and commercial districts. The predictions closely follow these variations, with elevated GDP values along key service corridors and central urban districts. The model also captures the smoother gradients toward lower-income peripheries, reflecting the complex socio-economic structure of the city.

The final case in Fig. 3 at the bottom highlights São Paulo, Brazil's largest metropolitan region and the nation's most significant economic hub. The model successfully identifies the extensive high-GDP core spanning the city's commercial and financial centres, as well as the pronounced economic intensity along major transport axes. Despite São Paulo's vast scale and heterogeneity, the predicted patterns align well with the observed GDP reference, demonstrating robustness in dense, economically diverse environments.

Despite the overall strong spatial alignment, some systematic limitations are visible. In some cases, the model tends to over-emphasise extremes, producing sharper transitions between high- and low-GDP regions than those present in the disaggregated reference labels. This suggests a reduced ability to capture gradual economic gradients, especially in transitional urban–periurban zones, where true GDP tends to vary smoothly rather than abruptly.

C. Evaluating Model Prediction With Official Municipal GDP Data

To evaluate how well the model reproduces officially reported economic figures, we compare our predictions with municipal-level GDP data from the IBGE for the year 2016. These official values are provided for each municipal polygon. For comparison, the predictions of the ResNet-18 concatenation model using all input modalities were spatially summed within each

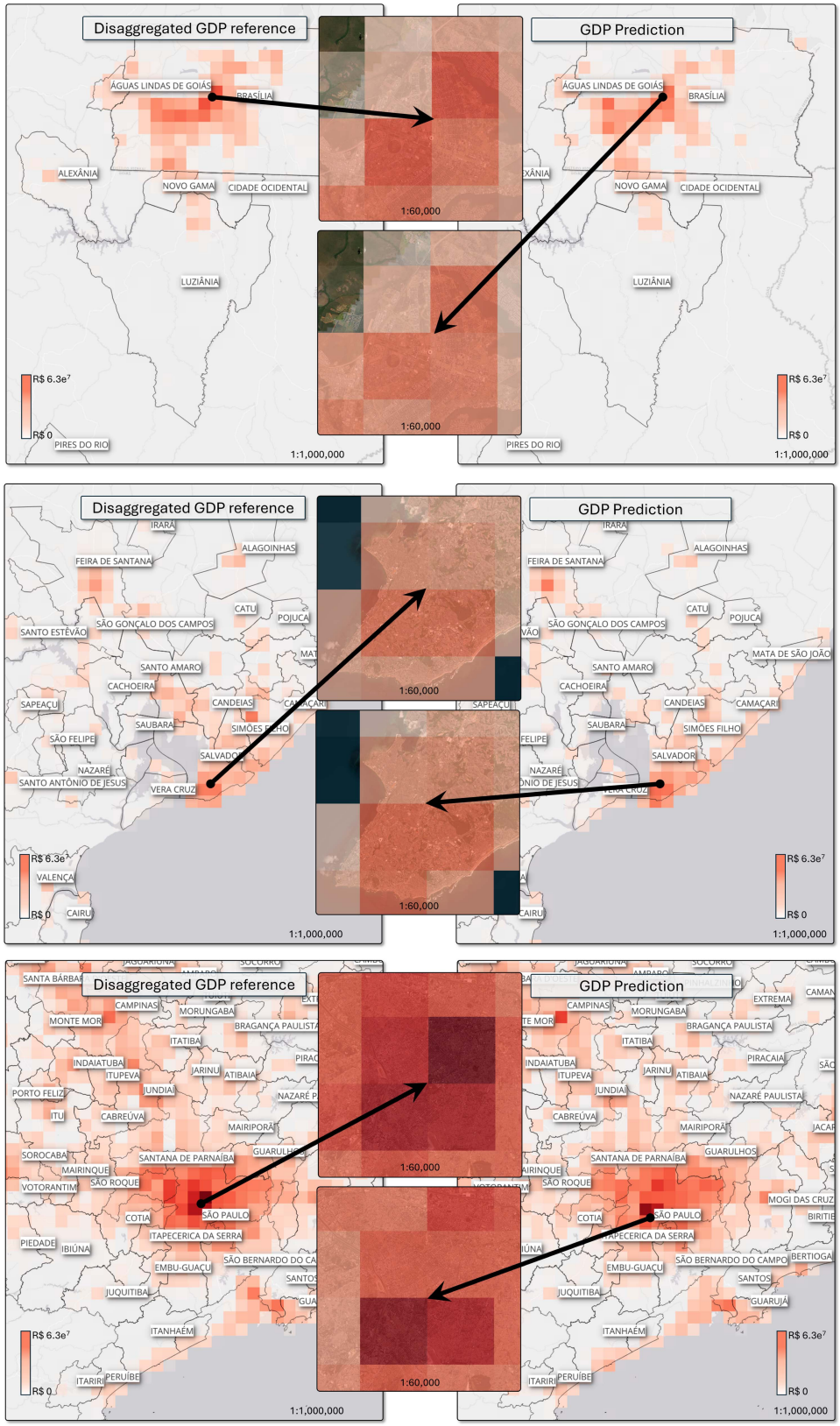

Fig. 3. Example results for three representative areas of interest, displayed at a common spatial scale and using a unified colorbar to indicate GDP values per 5 km × 5 km grid cell for both the GDP reference label and the model prediction. The top row shows the capital region of Brasília, where governmental activities contribute substantially to GDP. The middle row depicts Salvador, characterized by a strong service economy and tourism sector. The bottom row presents São Paulo, Brazil's largest metropolitan area and the region with the highest overall GDP.

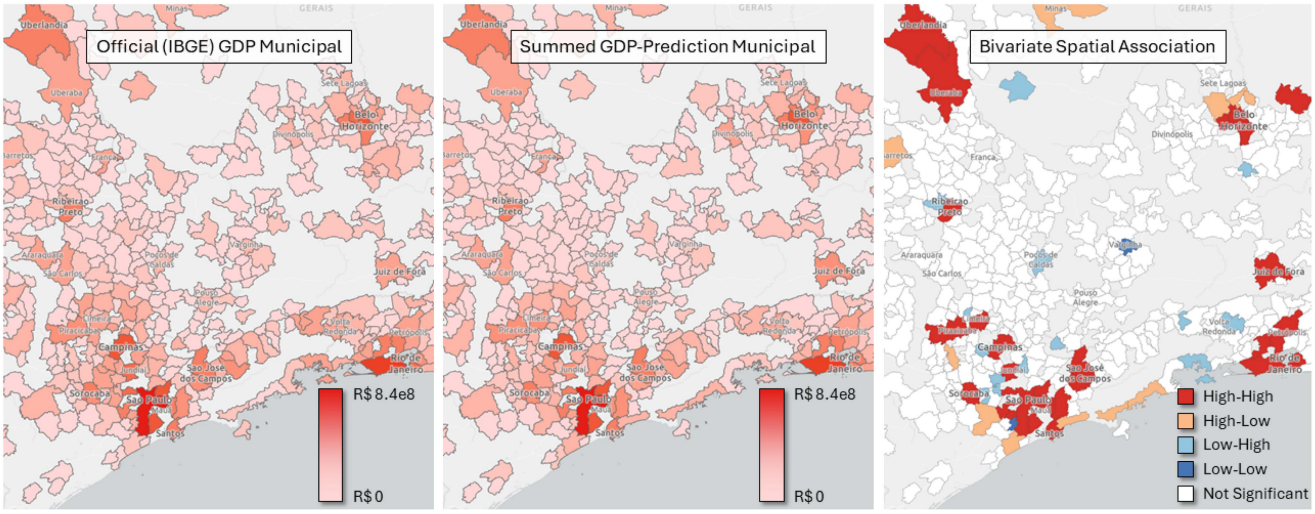

Fig. 4. Comparison of official municipal GDP from the IBGE with reaggregated GDP predictions from the multimodal fusion model for São Paulo and Rio de Janeiro; the right panel shows the corresponding bivariate spatial association.

polygon. In cases, where multiple municipalities fall within a single 5 km × 5 km prediction grid cell, the model output is proportionally allocated according to the relative area of each municipality. Fig. 4 presents an example for Brazil's two largest metropolitan regions, São Paulo and Rio de Janeiro. The visual comparison reveals that the multimodal fusion model captures the dominant spatial patterns of economic activity well, though the predictions occasionally show sharper transitions, with values leaning slightly too high or too low relative to the more gradual gradients seen in the official statistics.

To quantitatively assess spatial pattern consistency and cross-structure between predicted and reference GDP, we applied a bivariate spatial association analysis based on Lee's L as seen in Fig. 4. The results indicate a significant positive spatial association ($L = 0.1901$, $p < 0.01$), confirming that areas of high (low) predicted GDP spatially correspond to areas of high (low) reference GDP. This is further supported by very high Pearson correlations both at the raw value level ($r = 0.9896$) and for spatially smoothed neighborhood averages ($r = 0.9885$), demonstrating that the model preserves both the numerical distribution and the spatial structure of economic activity.

A quantitative comparison across all 1363 selected municipalities within our dataset is shown in Fig. 5, using both the real-valued $R\$$ domain and a log–log representation to better visualize the strong skew in Brazil's GDP distribution. In the linear plot at the top of Fig. 5, the predicted and official values exhibit a tight clustering around the regression line, with only a small number of high-GDP municipalities deviating more noticeably. This behavior is expected in datasets where economic output is concentrated in a few metropolitan areas, and it highlights that absolute errors grow with scale even when relative performance remains stable. Nevertheless, the linear fit yields an R^2 of 0.979, indicating that the fused multimodal model is able to reproduce the majority of the variance in official GDP when predictions are aggregated back to municipal level.

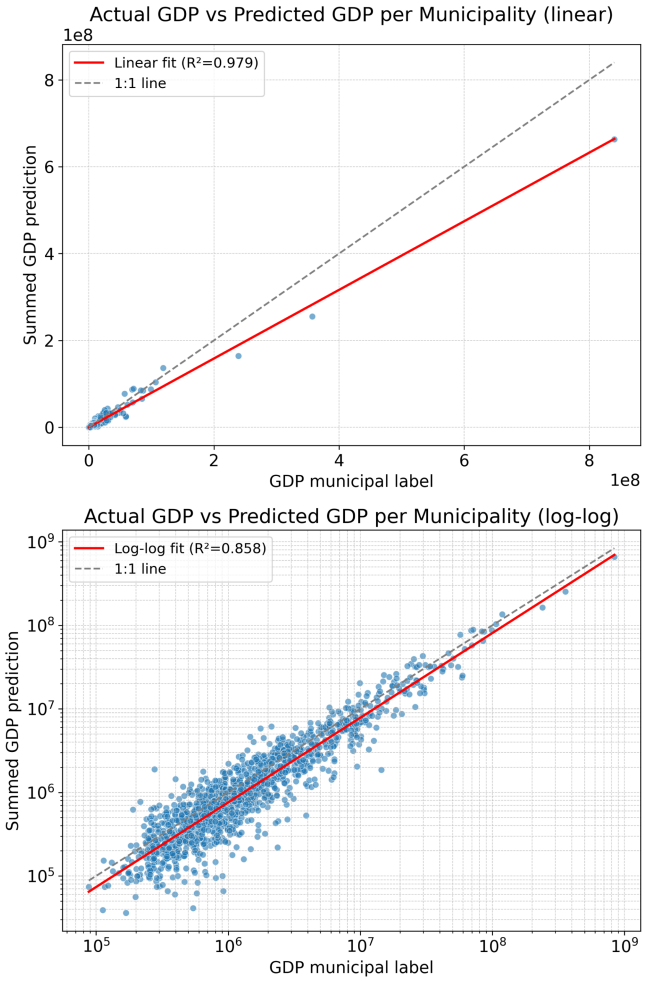

Fig. 5. Comparison of official and predicted municipal GDP values. Top: Scatter plot in real $R\$$ units to assess absolute agreement between predicted and official GDP. Bottom: Log–log scatter plot to evaluate performance across the full long-tailed distribution of municipal GDP, highlighting model behaviour from small to large economies.

The log–log plot at the bottom of Fig. 5 provides a complementary perspective by visually equalising the density of municipalities across the full economic spectrum. Here, the model captures the overall scaling relationship between predicted and observed GDP well, suggesting that it generalises across both small rural municipalities and large urban economies. The slightly lower log–log R^2 value of 0.858 reflects the increased sensitivity of logarithmic space to deviations among the numerous low-GDP municipalities, but the alignment with the 1:1 line across several orders of magnitude demonstrates robust relative ranking performance. Importantly, the log–log view makes clear that model accuracy is not limited to the upper tail, where most remote-sensing-based GDP studies traditionally perform best, but extends to the long distribution of smaller municipalities that dominate the national landscape [19], [21], [69].

Together, the two representations highlight complementary aspects of model behavior: the linear scale emphasises absolute accuracy in real monetary units, while the log–log scale reveals consistency across orders of magnitude under a highly long-tailed target distribution. Both perspectives confirm that integrating multisource remote sensing, NTLs, and auxiliary geodata layers with an efficient fusion architecture enables reliable reconstruction of municipal-level economic activity.

D. Comparison to Related Work

Our results (best model: $R^2 \approx 0.87$) are competitive with, and in some cases exceed, existing GDP spatialization studies. [29] report strong performance (typically $R^2 \sim 0.7{-}0.85$) using NTL and Sentinel-2 features, but focus primarily on built-up areas, whereas our approach generalizes across heterogeneous urban–rural contexts. Similarly, [30] achieve R^2 values around 0.8 using classical machine learning methods, but rely on fewer modalities and less expressive fusion strategies.

[34] report lower performance ($R^2 \sim 0.6{-}0.75$), highlighting the limitations of dual-source inputs (NTL + land cover), which our results confirm through ablation experiments. In contrast, [32] show improvements ($R^2 \sim 0.8{-}0.86$) using multiscale fusion and POI data, closely aligning with our findings that multimodal integration is key.

Overall, our results slightly outperform or match the state-of-the-art, particularly due to the combination of diverse data sources and efficient deep learning based fusion, while also demonstrating robustness across a larger and more diverse study area.

E. Limitations of the Current Fusion Framework

While our approach evaluates several feature-level fusion mechanisms, all models ultimately optimize a single regression loss. More advanced multitask or hierarchical fusion schemes, where each input modality receives its own dedicated loss function before joint aggregation, could help disentangle complementary signals and reduce modality-specific overfitting [70], [71]. Likewise, the current framework does not quantify predictive uncertainty. Incorporating Bayesian layers, Monte-Carlo dropout, or deep ensembles would offer calibrated uncertainty estimates, which are essential when models are applied in policy-relevant economic contexts [63], [64], [72]. Another challenge stems from the extreme long-tailed distribution of GDP, where a small number of metropolitan regions dominate national output. Even with normalization strategies, the model tends to underfit the upper tail and over-smooth rare but economically important high-value regions [73]. Finally, GDP as a single aggregate hides the heterogeneous structure of local economies. Remote sensing and other auxiliary geodata data offer the opportunity to move beyond aggregate GDP and estimate sector-specific contributions, such as agriculture, manufacturing, services, or government activity, which are likely to manifest differently in spatial patterns [74]. Future work should therefore aim at multimodal, uncertainty-aware, and sector-resolved economic mapping to fully leverage remote sensing data for economic analysis.

A limitation of the validation strategy is that the fourfold cross-validation was performed at the grid-cell level rather than using a strict spatial or administrative holdout. Consequently, some municipalities may contribute cells to both training and test folds, which reduces fold independence due to shared local context and spatial autocorrelation.

A conceptual limitation of the GDP disaggregation approach in Algorithm 1 arises from the use of income as a proxy for GDP distribution. While both variables are closely related, they are recorded at different spatial reference points: income is typically associated with place of residence, whereas GDP reflects economic activity at the place of production [75]. This introduces a potential spatial mismatch, particularly in regions with strong commuting patterns or spatial separation between residential and employment centers [76], [77]. As a result, the disaggregated GDP surfaces may partially reflect residential structures rather than purely production-based economic geography. While this assumption is common in similar downscaling approaches, it may introduce systematic biases at finer spatial resolutions and should be considered when interpreting the results.

Another limitation of this study concerns the use of the auxiliary input datasets. OSM, which is known to exhibit spatial incompleteness and thematic inconsistencies depending on region and contributor activity [78], [79]. In addition, we did not differentiate between specific categories of OSM features: both road networks and point-based data were used in their aggregated form without semantic grouping. Future work could address this by incorporating more refined categorizations of OSM features. Furthermore, integrating alternative or complementary geospatial datasets with more consistent coverage could help mitigate these limitations and improve model robustness. In particular, region-specific or authoritative data sources may provide more reliable representations of infrastructure and human activity patterns (e.g., MAPBIOMASS).

A further limitation arises from temporal and thematic inconsistencies across the input datasets. The census data represent conditions in 2010, the official municipal GDP statistics refer to 2016, while Sentinel-2 imagery was acquired between 2019 and 2021, and OSM remains a continuously updated but inherently incomplete volunteer dataset. These temporal gaps and accuracy variations introduce systematic noise that the model cannot fully resolve, potentially weakening the learned relationships between observed land patterns and underlying economic activity.

VI. Conclusion

This study demonstrates that high-resolution GDP estimation can be improved by jointly leveraging diverse remote-sensing and auxiliary data through tailored fusion strategies. Across all experiments, the interaction between backbone architecture, fusion layer design, and input modality composition proved to be important to model performance. Efficient and small backbones such as ResNet-18 and EfficientNet-B3 extracted rich spatial–contextual information, while fusion by simple concatenation remained highly effective and computationally fast, particularly when integrating many heterogeneous modalities. In parallel, the combination of optical imagery, nighttime lights, and auxiliary geodata layers provided complementary signals that strengthened predictions across both high- and low-GDP municipalities. The proposed approach requires only openly accessible datasets and lightweight model components, enabling reproducible large-scale GDP estimation without dependence on proprietary inputs. When municipal GDP statistics or census benchmarks are available, as in the case of Brazil, the framework can be trained and validated robustly, and then transferred to other emerging-market contexts with comparable data infrastructures. This capability is especially relevant for countries where official economic reporting remains infrequent or spatially coarse. By integrating multimodal remote sensing data with flexible deep learning fusion, the presented workflow provides a scalable framework for generating spatially explicit, high-resolution estimates of economic activity to support empirical analysis.

References

[1] R. W. England, "Measurement of social well-being: Alternatives to gross domestic product," *Ecological Econ.*, vol. 25, no. 1, pp. 89–103, Apr. 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921800997000980

[2] I. M. Fund, *Back to Basics: Economic Concepts Explained*. International Monetary Fund, Apr. 2019, Google-Books-ID: zBeWDwAAQBAJ.

[3] D. Landau, "Government expenditure and economic growth: A cross-country study," *Southern Econ. J.*, vol. 49, pp. 783–792, 1983.

[4] T. Geiger, "Continuous national gross domestic product (GDP) time series for 195 countries: Past observations (1850–2005) harmonized with future projections according to the shared socio-economic pathways (2006–2100)," *Earth Syst. Sci. Data*, vol. 10, no. 2, pp. 847–856, 2018.

[5] R. Dědeček and V. Dudzich, "Exploring the limitations of GDP per capita as an indicator of economic development: A cross-country perspective," *Rev. Econ. Perspectives*, vol. 22, no. 3, pp. 193–217, 2022.

[6] *World Bank Development Data Group, World Development Indicators*. World Bank, 1978, Accessed: Dec. 04, 2025.

[7] R. J. Barro and X. Sala-i Martin, "Convergence," *J. Political Economy*, vol. 100, no. 2, pp. 223–251, 1992.

[8] R. Mahtta et al., "Urban land expansion: The role of population and economic growth for 300 cities," *Npj Urban Sustainability*, vol. 2, no. 1, Feb. 2022, Art. no. 5. [Online]. Available: https://www.nature.com/articles/s42949-022-00048-y

[9] P. C. Cheshire and S. Magrini, "Urban growth drivers and spatial inequalities: Europe–a case with geographically sticky people," *LEQS Paper*, no. 11, pp. 1–41, 2009.

[10] M. Kummu, M. Taka, and J. H. Guillaume, "Gridded global datasets for gross domestic product and human development index over 1990–2015," *Sci. Data*, vol. 5, no. 1, 2018, Art. no. 180004.

[11] J. E. Stiglitzet al., A. Sen, and J. P. Fitoussi, "Report by the commission on the measurement of economic performance and social progress," 2009.

[12] J. E. Stiglitz, "GDP is the wrong tool for measuring what matters," Aug. 2020. [Online]. Available: https://www.scientificamerican.com/article/gdp-is-the-wrong-tool-for-measuring-what-matters/

[13] M. Jerven, *Poor Numbers: How We are Misled by African Development Statistics and What to Do About It*. Cornell Univ. Press, Jan. 2013, Google-Books-ID: cwljCCv50SIC.

[14] A. Deaton and A. Heston, "Understanding PPPs and PPP-based national accounts," *Amer. Econ. Journal: Macroeconomics*, vol. 2, no. 4, pp. 1–35, 2010.

[15] A. Subramanian, "India's GDP mis-estimation: Likelihood, magnitudes, mechanisms, and implications," Center for International Development Working Paper, Jun. 2019. [Online]. Available: https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37366391

[16] M. Jerven, "The relativity of poverty and income: How reliable are african economic statistics?," *Afr. Affairs*, vol. 109, no. 434, pp. 77–96., doi: 10.1093/afraf/adp064.

[17] A. A. Oloyede, N. Faruk, N. Noma, E. Tebepah, and A. K. Nwaulune, "Measuring the impact of the digital economy in developing countries: A systematic review and meta- analysis," *Heliyon*, vol. 9, no. 7, Jul. 2023, Art. no. e17654. [Online]. Available: https://www.cell.com/heliyon/abstract/S2405-8440(23)04862-4

[18] Y. Hu and J. Yao, "Illuminating economic growth," *J. Econometrics*, vol. 228, no. 2, pp. 359–378, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304407621001767

[19] N. Jean, M. Burke, M. Xie, W. M. Alampay Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, Aug. 2016. [Online]. Available: https://www.science.org/doi/full/10.1126/science.aaf7894

[20] H. Wang, H. Wang, and L. Wu, "TGF-Net: Transformer and gist CNN fusion network for multi-modal remote sensing image classification," *PLoS One*, vol. 20, no. 2, Feb. 2025, Art. no. e0316900. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0316900

[21] T. Stark, M. Wurm, E. Stokes, K. C. Seto, and H. Taubenböck, "GDP estimation using a deep learning fusion model for multi-source remote sensing data," in *Proc. Joint Urban Remote Sens. Event*, May 2025, pp. 1–4., iSSN: 2642-9535. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11076044

[22] F. Lan, H. Da, H. Wen, and Y. Wang, "Spatial structure evolution of urban agglomerations and its driving factors in mainland China: From the monocentric to the polycentric dimension," *Sustainability*, vol. 11, no. 3, 2019, Art. no. 610. [Online]. Available: https://www.mdpi.com/2071-1050/11/3/610

[23] A. S. M. Abozeid and T. A. AboElatta, "Polycentric vs monocentric urban structure contribution to national development," *J. Eng. Appl. Sci.*, vol. 68, no. 1, 2021, Art. no. 11, doi: 10.1186/s44147-021-00011-1.

[24] D. Ahn et al., "A human-machine collaborative approach measures economic development using satellite imagery," *Nature Commun.*, vol. 14, no. 1, Oct. 2023, Art. no. 6811. [Online]. Available: https://www.nature.com/articles/s41467-023-42122-8

[25] E. C. Stokes, M. O. Román, Z. Wang, R. M. Shrethsa, T. Yao, and G. Kalb, "Urban applications of NASA's black marble product suite," in *Proc. Joint Urban Remote Sens, Event*, May 2019, pp. 1–4., iSSN: 2642-9535. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8809074

[26] J. F. Martinez, K. MacManus, E. C. Stokes, Z. Wang, and A. de Sherbinin, "Suitability of NASA's black marble daily nighttime lights for population studies at varying spatial and temporal scales," *Remote Sens.*, vol. 15, no. 10, Jan. 2023, Art. no. 2611. [Online]. Available: https://www.mdpi.com/2072-4292/15/10/2611

[27] E. C. Stokes and M. O. Román, "Tracking COVID-19 urban activity changes in the middle east from nighttime lights," *Sci. Rep.*, vol. 12, no. 1, May 2022, Art. no. 8096. [Online]. Available: https://www.nature.com/articles/s41598-022-12211-7

[28] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003442571100160X

[29] Z. Chen, W. Wang, H. Zong, and X. Yu, "Precise GDP spatialization and analysis in built-up area by combining the NPP-VIIRS-like dataset and Sentinel-2 images," *Sensors*, vol. 24, no. 11, Jan. 2024, Art. no. 3405. [Online]. Available: https://www.mdpi.com/1424-8220/24/11/3405

[30] S. Liu, W. Liu, Y. Zhou, S. Wang, F. Wang, and Z. Wang, "Mapping gridded GDP distribution of China based on remote sensing data and machine learning methods," *Remote Sens.*, vol. 17, no. 10, Jan. 2025, Art. no. 1709. [Online]. Available: https://www.mdpi.com/2072-4292/17/10/1709

[31] Z. Chen, C. Zhang, S. Qiu, and J. Lin, "GDP estimation by integrating Qimingxing-1 nighttime light, street-view imagery, and points of interest: An empirical study in Dongguan city," *Remote Sens.*, vol. 17, no. 7, Jan. 2025, Art. no. 1127. [Online]. Available: https://www.mdpi.com/2072-4292/17/7/1127

[32] N. Wu, J. Yan, D. Liang, Z. Sun, R. Ranjan, and J. Li, "High-resolution mapping of GDP using multi-scale feature fusion by integrating remote sensing and POI data," *Int. J. Appl. Earth Observation Geoinformation*, vol. 129, May 2024, Art. no. 103812. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843224001663

[33] X. Wang, P. C. Sutton, and B. Qi, "Global mapping of GDP at 1 km2 using VIIRS nighttime satellite imagery," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 12, Dec. 2019, Art. no. 580. [Online]. Available: https://www.mdpi.com/2220-9964/8/12/580

[34] N. Xu, S. Zhang, and S. Jiang, "Estimating GDP by fusing nighttime light and land cover data," *Can. J. Remote Sens.*, vol. 50, no. 1, Dec. 2024, Art. no. 2377641. [Online]. Available: https://doi.org/10.1080/07038992.2024.2377641

[35] H. Suleiman, T. Nguyen, and C. Mendez, "Predicting subnational GDP in Vietnam with remote sensing data: A machine learning approach," *Lett. Spatial Resour. Sci.*, vol. 18, no. 1, p. 5, Mar. 2025.

[36] M. F. P. Del Castillo, T. Fujimi, and H. Tatano, "Estimating medium-term regional monthly economic activity reductions during the COVID-19 pandemic using nighttime light data," *Int. J. Appl. Earth Observation Geoinformation*, vol. 135, Dec. 2024, Art. no. 104223. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156984322400579X

[37] P. McSharry and J. Mawejje, "Estimating urban GDP growth using nighttime lights and machine learning techniques in data poor environments: The case of south Sudan," *Technological Forecasting Social Change*, vol. 203, Jun. 2024, Art. no. 123399. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162524001951

[38] T. Stark, M. Wurm, X. X. Zhu, and H. Taubenböck, "Satellite-based mapping of urban poverty with transfer-learned slum morphologies," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5251–5263, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9174807

[39] M. Weigand et al., "Are public green spaces distributed fairly? A nationwide analysis based on remote sensing, OpenStreetMap and census data," *Geocarto Int.*, vol. 38, no. 1, Dec. 2023, Art. no. 2286305, doi: 10.1080/10106049.2023.2286305.

[40] T. Liu et al., "Urban functional zone classification via advanced multi-modal data fusion," *Sustainability*, vol. 16, no. 24, Jan. 2024, Art. no. 11145. [Online]. Available: https://www.mdpi.com/2071-1050/16/24/11145

[41] N. Liu, Y. Yuan, G. Wu, S. Zhang, J. Leng, and L. Wan, "Multi-label remote sensing classification with self-supervised gated multi-modal transformers," *Front. Comput. Neurosci*, vol. 182024. https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1404623

[42] Y. Yuan et al., "Multi-resolution collaborative fusion of SAR, multispectral and hyperspectral images for coastal wetlands mapping," *Remote Sens.*, vol. 14, no. 14, Jan. 2022, Art. no. 3492. [Online]. Available: https://www.mdpi.com/2072-4292/14/14/3492

[43] Z. Li, R. Liu, L. Sun, and Y. Zheng, "Multi-feature cross attention-induced transformer network for hyperspectral and LiDAR data classification," *Remote Sens.*, vol. 16, no. 15, Jan. 2024, Art. no. 2775. [Online]. Available: https://www.mdpi.com/2072-4292/16/15/2775

[44] Q. Wang, B. Zhou, J. Zhang, J. Xie, and Y. Wang, "Joint classification of hyperspectral images and LiDAR data based on dual-branch transformer," *Sensors*, vol. 24, no. 3, Jan. 2024, Art. no. 867. [Online]. Available: https://www.mdpi.com/1424-8220/24/3/867

[45] Z. Chang, H. Yu, Y. Zhang, and K. Wang, "Fusion of hyperspectral CASI and airborne LiDAR data for ground object classification through residual network," *Sensors*, vol. 20, no. 14, Jan. 2020, Art. no. 3961. [Online]. Available: https://www.mdpi.com/1424-8220/20/14/3961

[46] C. L. Norton, K. Hartfield, C. D. H. Collins, W. J. D. van Leeuwen, and L. J. Metz, "Multi-temporal LiDAR and hyperspectral data fusion for classification of semi-arid woody cover species," *Remote Sens.*, vol. 14, no. 12, Jan. 2022, Art. no. 2896. [Online]. Available: https://www.mdpi.com/2072-4292/14/12/2896

[47] C. Geiß et al., "Estimation of seismic building structural types using multi-sensor remote sensing and machine learning techniques," *ISPRS J. Photogrammetry Remote Sens.*, vol. 104, pp. 175–188, Jun. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271614002007

[48] P. Aravena Pelizari, K. Spröhnle, C. Geiß, E. Schoepfer, S. Plank, and H. Taubenböck, "Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements," *Remote Sens. Environ.*, vol. 209, pp. 793–807, May 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425718300312

[49] H. Guo, B. Tian, and W. Liu, "CCFormer: Cross-modal cross-attention transformer for classification of hyperspectral and LiDAR data," *Sensors (Basel, Switzerland)*, vol. 25, no. 18, Sep. 2025, Art. no. 5698. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12473655/

[50] World Bank, "World bank country classifications by income level for 2024-2025," 2024, Accessed: Oct. 16, 2025. [Online]. Available: https://blogs.worldbank.org/en/opendata/world-bank-country-classifications-by-income-level-for-2024-2025,

[51] Instituto Brasileiro de Geografia e Estatística (IBGE), "Produto interno bruto dos municípos: Pib a preços correntes e componentes (referência 2016) — sidra, tabela 5938," 2016, download via SIDRA (IBGE Automatic Retrieval System), table includes municipal GDP and sector value added; select year 2016 in the interface. [Online]. Available: https://sidra.ibge.gov.br/tabela/5938

[52] Instituto Brasileiro de Geografia e Estatística (IBGE), "Censo demográfico 2010: Resultados do universo — agregados por setores censitários (trabalho e rendimento / renda)," 2010, fTP directory with sector-level aggregate tables (by UF/municipality/sector), including income-related variables from the 2010 Census universe results. Landing page: https://www.ibge.gov.br/estatisticas/sociais/habitacao/9662-censo-demografico-2010.html. [Online]. Available: https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/

[53] European Space Agency (ESA), "Sentinel-2 multispectral instrument (msi) level-1c and level-2a data," 2015, global multispectral optical imagery at 10–60 m spatial resolution; accessed via Copernicus Data Space Ecosystem. [Online]. Available: https://dataspace.copernicus.eu/explore-data/collections/sentinel-data/sentinel-2

[54] NASA Goddard Space Flight Center, "Viirs nighttime lights — black marble (vnp46) daily and monthly products," 2017, radiance-calibrated nighttime light data corrected for atmospheric, lunar, and stray light effects; derived from VIIRS day/night band. [Online]. Available: https://blackmarble.gsfc.nasa.gov

[55] European Space Agency (ESA), "Esa worldcover 10 m land cover map," 2021, global land cover classification at 10 m resolution based on Sentinel-1 and Sentinel-2 imagery. [Online]. Available: https://worldcover2021.esa.int

[56] OpenStreetMap contributors, "Openstreetmap planet dump" 2024, global, crowdsourced geospatial database of roads, buildings, points of interest, and other vector features, released under the ODbL. [Online]. Available: https://planet.openstreetmap.org

[57] J. A. Ohlson and S. H. Penman, "Disaggregated accounting data as explanatory variables for returns," *J. Accounting, Auditing Finance*, vol. 7, no. 4, pp. 553–573, Oct. 1992, doi: 10.1177/0148558X9200700407.

[58] M. Linden and D. Ray, "Aggregation bias-correcting approach to the health–income relationship: Life expectancy and GDP per capita in 148 countries, 1970–2010," *Econ. Modelling*, vol. 61, pp. 126–136, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264999316307933

[59] P.-E. Diacon and L.-G. Maha, "The relationship between income, consumption and GDP: A time series, cross-country analysis," *Procedia Econ. Finance*, vol. 23, pp. 1535–1543, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212567115003743

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Dec. 2015, *arXiv:1512.03385*.

[61] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Sep. 2020, *arXiv:1905.11946*.

[62] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," Apr. 2022, *arXiv:2111.09883*.

[63] T. Stark, M. Wurm, H. Debray, X. X. Zhu, and H. Taubenböck, "Uncertainty aware slum mapping in 55 heterogeneous cities," *Int. J. Appl. Earth Observation Geoinformation*, vol. 145, Art. no. 104979, Dec. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843225006260

[64] T. Stark, M. Wurm, V. Ştefan, F. Wolf, H. Taubenböck, and T. M. Knight, "Utilizing CNNs for classification and uncertainty quantification for 15 families of European fly pollinators," *PLoS One*, vol. 20, no. 9, Sep. 2025, Art. no. e0323984. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0323984

[65] X. Liang, P. Hu, L. Zhang, J. Sun, and G. Yin, "MCFNet: Multi-layer concatenation fusion network for medical images fusion," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7107–7119, Aug. 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8698840

[66] Y. Wang, H. Dong, H. Wu, W. Wang, and J. Zhang, "A neural network model based on attention pooling and adaptive multi-level feature fusion for arrhythmia automatic detection," *Comput. Methods Biomech. Biomed. Eng.*, pp. 1–15, Mar. 2025, doi: 10.1080/10255842.2025.2480264.

[67] Y. Kang et al., "MGMFN: Multi-graph and MLP-mixer fusion network for chinese social network sentiment classification," *Multimedia Tools Appl.*, vol. 83, no. 24, pp. 64989–65010, Jul. 2024, doi: 10.1007/s11042-023-17857-7.

[68] Y. Liu, Z. Zhang, J. Zhu, J. Li, and Y. Zhang, "DynMixMFusion: Dynamic mixture of experts multimodal fusion model," *J. Physics: Conf. Ser.*, vol. 3036, no. 1, Jun. 2025, Art. no. 012004, doi: 10.1088/1742-6596/3036/1/012004.

[69] C. Yeh et al., "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature Commun.*, vol. 11, no. 1, May 2020, Art. no. 2583. [Online]. Available: https://www.nature.com/articles/s41467-020-16185-w

[70] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9336293

[71] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8219–8264, Aug. 2023, doi: 10.1007/s10462-022-10366-3.

[72] V. Hertel, O. Wani, C. Geiß, M. Wieland, and H. Taubenböck, "BayeSiamMTL: Uncertainty-aware multitask learning for post-disaster building damage assessment," *Int. J. Appl. Earth Observation Geoinformation*, vol. 143, Sep. 2025, Art. no. 104759. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843225004066

[73] C. Zhang et al., "A systematic review on long-tailed learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 8, pp. 13670–13690, Aug. 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10905046

[74] J. L. Martıń-Ortega and M. González-Sánchez, "Sectoral composition of GDP and greenhouse gas emissions: An empirical analysis in EU27," *Environ., Develop. Sustain.*, vol. 27, no. 2, pp. 3899–3926, Feb. 2025, doi: 10.1007/s10668-023-04048-5.

[75] "O. for economic co-operation and development," in *OECD Regional Outlook 2014: Regions and Cities: Where Policies and People Meet*. OECD Publishing, 2014.

[76] M. Birkin and M. Clarke, "The generation of individual and household incomes at the small area level using synthesis," *Regional Stud.*, vol. 23, no. 6, pp. 535–548, 1989.

[77] K. Harland, A. Heppenstall, D. Smith, and M. H. Birkin, "Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques," *J. Artif. Societies Social Simul.*, vol. 15, no. 1, 2012.

[78] M. Minghini and F. Frassinelli, "Openstreetmap history for intrinsic quality assessment: Is OSM up-to-date?," *Open Geospatial Data, Soft. Standards*, vol. 4, no. 1, pp. 1–17, 2019.

[79] G. Yeboah et al., "Analysis of openstreetmap data quality at different stages of a participatory mapping process: Evidence from slums in Africa and Asia," *ISPRS Int. J. Geo- Inf.*, vol. 10, no. 4, 2021, Art. no. 265.