Predict Social Economic Outcomes by Transferred Knowledge with Satellite Imagery

***Abstract*—Traditional machine learning methods have play-ed a crucial role in the field of data mining, particularly in the prediction of socioeconomic indicators. Unfortunately, it needs large amounts of high-quality socioeconomic data for model training, which is prohibitively expensive and unaffordable. With the advancement of remote sensing technology, access to high-resolution satellite imagery has become increasingly affordable and abundant, providing a rich data source for model training. However, socioeconomic information is unable to be directly extracted from them. In this paper, we propose using road network types as a proxy for socioeconomic factors, which is more effectively and stably than using nightlight, and leveraging transfer learning to predict socioeconomic indicators through satellite imagery. We have aggregated 11 distinct road topological structures to generate reasonable road network types. Given the unique characteristics of road networks, we have adopted and fine-tuned a hybrid pre-trained model that combines ResNet50 and Vision Transformer architectures for the transfer learning task. Through extensive experiments conducted across multiple cities, we have demonstrated that our approach outperforms state-of-the-art method in this field. This work highlights the potential of leveraging road network types as a proxy for socioeconomic information and the effectiveness of our transfer learning-based framework in extracting valuable insights from satellite imagery to support socioeconomic policy decisions. The code and data are available at GitHub repository: https://github.com/xiachan254/PredSocecOut.**

***Index Terms*—Transfer learning, Data mining, Pre-trained model, Remote sensing**

I. Introduction

Data mining using machine learning has found widespread application in various domains, particularly in the task of forecasting socioeconomic indicators. For example, as in [1], it performed forecasts of macroeconomic factors like gross domestic product (GDP), people consumption level and poverty rate, and was able to capture complex nonlinear relationships more effectively compared to traditional econometric models. The correlation information obtained by machine learning can guide decision making, for instance, the conclusions of the study [2] point that analyzing consumer purchasing behavior to uncover consumption patterns and predict socioeconomic costs, can provide enterprises and government with valuable support for targeted marketing decisions.

However, most traditional machine learning approaches depend on the suppose that validation and training dataset share the same input feature space and data distribution, and require a wealth of labeled training samples to establish the model [3]. In the context of socioeconomic indicator forecasting, it is extremely difficult and costly to acquire such great amounts of high-quality data. As shown in [4], socioeconomic data is typically collected and released periodically by government statistical agencies, but often suffers from issues such as short time series and limited coverage. Moreover, some important socioeconomic indicators may be subject to privacy or security restrictions, making them difficult to access. Even when large amounts of socioeconomic data are available, their quality is often compromised by problems like missing values, high noise levels, and inaccurate labeling, which can negatively impact model training and performance [5]. Furthermore, when forecasting socioeconomic indicators for different regions or industries, a single training set may struggle to comprehensively capture the characteristics of the entire target population due to regional differences and industry-specific factors, leading to weak generalization ability. Therefore, to address these difficulties, it is crucial to find data that is easily accessible, feature-rich, globally representative, to predict socioeconomic indicators through machine learning. High-resolution satellite imagery data possesses these advantages.

The high-speed development of remote sensing technology has enabled high-resolution satellite imagery increasingly accessible and cost-effective. The latest high-precision satellites, such as WorldView-3 and GeoEye-2, can provide spatial resolutions of 0.3-0.5 meters, resulting in satellite images with detailed land feature information [6]. Moreover, these satellite data can be obtained on a large scale across the global domain, providing a powerful data foundation for making machine learning work in various fields. However, despite the satellite imagery owns those advantages, it is inherently highly unstructured, making it challenging to directly extract useful socioeconomic information for sustainable development policy-making using traditional machine learning methods and data mining techniques. Fortunately, transfer learning can help mitigate the reliance on learnable samples in the socioeconomic prediction task by transferring knowledge gained from a related task [7,8].

Transfer learning involves two main challenges: learning relevant domain knowledge for transfer and fine-turning a suitable pre-trained model to mine transferable knowledge for downstream tasks [9,10]. For the first challenge, the transferred features should not interfere negatively with the learning process of the target task. Previous research has attempted to use nighttime light intensity as a data-rich proxy [11], transferring the knowledge learned from classifying nighttime light intensity on satellite images to predict regional economic conditions, with reasonable results. However, using nighttime light intensity as a proxy has some drawbacks. A study [12] has shown that in urban core areas, which have high bright-

ness, the light intensity is saturated and indistinguishable from relatively low-brightness areas. Additionally, the halo effect of nighttime lights can spill over to surrounding suburban and rural areas, leading to an overestimation of the light intensity in these regions. Furthermore, the "bloom" operation during nighttime light image processing can introduce bias in the brightness distribution, making it an imperfect representation of the original uniform distribution. All these shortcomings make it inaccurate to extract socioeconomically relevant features from satellite images, thus degrading the prediction performance. Fortunately, other unsupervised learning artificial facilities features [13,14], such as road network, also can be used as a data-rich proxy. Compare with using nighttime light as a data-rich proxy, road network proxy is more representative since that the development of road networks and the process of urbanization often occur simultaneously [15], which play a crucial role in predicting socioeconomic outcomes, such as GDP prediction.

For the second challenge to select a suitable model, one potential approach is to fine-tune everything obtainable pre-trained models and adopt one with the greatest results [16]. However, this can be quite time-consuming and expensive. Alternatively, we can refer to the applicable fields and advantages of the pre-trained model itself. Traditional convolutional neural network (CNN) [17], which owns inherent characteristic of inductive biases, like local receptive field and hierarchical feature extraction, can excellent in extracting local features and generalize well in data-scarce scenarios. But, it performs poorly in mining global features. Correspondingly, the vision transformer (ViT) [18] can model long-range dependencies and overall properties in input sequence through its attention mechanism, capturing the global semantic information of images. But it lacks the ability to extract local fine-grained features. It shows that using either CNN or ViT alone cannot fully meet the requirements to sufficiently mine both local and global features.

Considering the constraints of the above challenges, it is important to determine the appropriate data-rich proxy and pre-trained model. Towards this end, we propose a method to Predict Social economic Outcomes by transferred knowledge with satellite imagery, PredSocecOut. Specifically, we use the extracted topology of regional road network for clustering, identifying four road network types, and use these types as a data-rich proxy for socioeconomic development patterns. Furthermore, considering these types own both local features like cornering angles and holistic properties such as connectivity, we adopt a hybrid pre-trained model with ResNet50 and ViT, which use a ResNet50 as backbone to process the satellite images and feed the resulting feature vectors as an input sequence to the ViT after fine-tuning setting. Base on this new data-rich proxy and the hybrid pre-trained model, PredSocecOut significantly improves the performance of predicting socioeconomic outcomes from satellite imagery.

To sum up, main contributions in our paper are as follows:

- We propose a novel method for predicting socioeconomic outcomes. To the best of our knowledge, this is the first endeavor to use clustered road networks as a data-rich proxy, in the absence of sufficient socioeconomic data, effectively predicts socioeconomic outcomes by using rich satellite imagery and transfer learning knowledge.
- We adopt and fine-tune a hybrid pre-trained model, fully utilizing the respective advantages of ResNet50 and ViT, to achieve high-precision classification of satellite imagery based on road network types.
- We extract and compute 11 types of road topological structures to generate road network, which serve as proxies for socioeconomic factors, enabling more effective knowledge transfer.
- We have conducted extensive experiments across multiple regions in China and the results demonstrate that our approach outperforms compelling state-of-the-art baselines in this field. Also, we set up ablation experiments to validate our hybrid pre-trained model achieving significantly better performance. The code is highly reusable and universal, allowing for convenient assessment of socioeconomic conditions in other developing countries where data is lacking, providing policymakers with feasible insights for formulating sustainable development policies.

The rest of our paper is organized as follows. Section II provides an overview of related work in pre-trained models and transfer learning methods. Section III introduces our PredSocecOut in detail, including various data preprocess, model structures, the road network types classification and society economy indicators prediction. In Section IV, we conduct extensive experiments, including the datasets in different area and provide detailed discussions on the results. Finally, Section V consists of a brief conclusion and work will be conducted in future.

II. Related Work

A. Traditional Methods for Predicting Socioeconomic

In the field of socioeconomic forecasting, traditional methods primarily rely on statistical models and historical data, such as time series analysis, regression models, and econometric models. These methods predict socioeconomic outcomes and social phenomena by analyzing past data trends. Time series analysis like the ARIMA model [19], captures data trends and cyclic patterns through autoregression, difference and moving averages. Regression analysis [20] uses linear or multiple regression models to find relationships between variables. Econometric models, such as Structural Equation Models (SEM) [21] and Vector Autoregressions (VAR) [22], combine statistical methods and economic theory to explain and predict economic phenomena. However, these traditional methods have limitations in handling complex nonlinear relationships and adapting to rapidly changing economic environments. In contrast, traditional machine learning methods, like Support Vector Machines (SVM), CNN and Random Forests [23]-[26], excel at managing non-linear relationships and high-dimensional data. Nevertheless, traditional machine learning

methods usually need numerous labeled data as the model input, and their performance can degrade significantly when data is insufficient. Therefore, we employ a transfer learning approach and design a pre-trained model for specific task according to fine-tuning setting.

B. Pre-trained Models in Predicting Socioeconomic

Pre-trained models have obtained great achievement in the field of computer vision, and other domains like predicting socioeconomic outcomes. They can effectively capture the latent features in data and transfer the learned knowledge to other tasks, significantly improving model performance and generalization capabilities. Early pre-trained models like AlexNet [27] and VGG [28], based on CNN, were able to learn rich visual feature representations and transferred them to other visual tasks. Later, CNN such as ResNet [14] and DenseNet [29] further optimized network structures, improving accuracy and robustness. However, pre-trained CNN models excel at efficiently capturing local features through convolution and pooling operations, but they lack the ability to model global relationships and overall semantic information. Fortunately, with the successful application of Transformer [30] in different fields, Transformer-based pre-trained models such as BERT [31] and GPT [32] have emerged, which can learn richer contextual representations. To further optimize the efficiency and performance of pre-trained models, improved versions like RoBERTa [33] and ALBERT [34] were proposed. Recently, Vit [15] and CLIP [35] have demonstrated powerful cross-modal understanding capabilities in the computer vision domain, using attention mechanisms to model long-range dependencies in input sequences and capture the global semantic information of images, unlocking more holistic feature representations. However, they are relatively weaker in extracting local features. Therefore, using a single model, CNN or Transformer, cannot fully extract features. Such as in [11,16,17], which use CNN as the pre-trained model, extract features to evaluate poverty based on identifying nightlight intensity from satellite images, ignoring global information. Our hybrid pre-trained model combines the traditional CNN model ResNet50 and the attention-based ViT, which can balance the capture of both local and global features from satellite images based on road network types, significantly better predicting socioeconomic outcomes.

C. Transfer Learning for Predicting Socioeconomic

Transfer learning is a machine learning approach that leverages the knowledge learned by a pre-trained model on one task to improve performance on a related task. In the medical and healthcare domain, researchers have attempted to transfer the universal visual features learned from natural image pre-training to medical image diagnosis tasks [36], effectively mitigating the problem of data scarcity in the medicine and reducing cost of disease images annotation. Furthermore, in the domain of financial risk prediction, transfer learning methods based on natural language processing have been applied. These methods utilize language models pre-trained on large-scale text

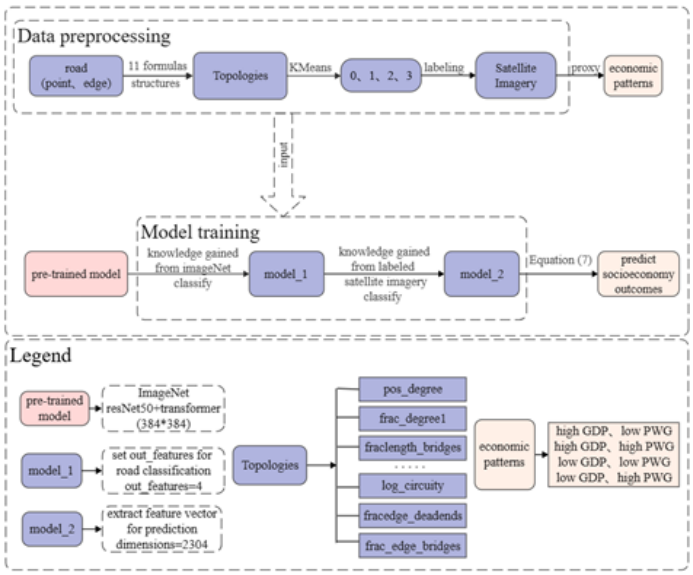

Fig. 1. The overall framework of the PredSocecOut and legends of some modules.

data to extract rich semantic representations, which are then used as inputs for supervised learning on financial risk prediction tasks, allowing the model to learn from the vast amount of available financial text data [37]. Similarly, transfer learning has been widely used in predicting socioeconomic indicators. Such as, nightlight had been applied as a data-rich proxy for socioeconomic development patterns, then transferred the know-ledge of nighttime light detection from satellite images to predict socioeconomic conditions [11,16,17]. However, this approach is limited by the fact that building occlusion can significantly reduce light intensity measurements, and indoor lighting cannot be observed. Additionally, previous study has pointed out the drawbacks of nighttime light [12], such as the light saturation, halos, and light spillover, which can affect the accuracy of the measurements. Therefore, in our study, we use road network topology as a new data-rich proxy to implement transfer learning and improve the pre-trained model to enhance the performance of socioeconomic prediction.

III. METHODOLOGY

In this part, we systematically describe the PredSocecOut we designed, which uses the topology of road networks as a data-rich proxy. We first tackle the task of classifying road network types from satellite imagery, and then transfer the knowledge gained from this task to mine features that can be used to predict socioeconomic conditions.

A. Framework Overview

The total framework of our proposed PredSocecOut is illustrated in Fig.1, which consists of two parts, data preprocessing and model training. In the first part, to annotate the highly unstructured satellite imagery data, we compute the eleven different topological structures of road networks based on the basic point and edge information of the roads,

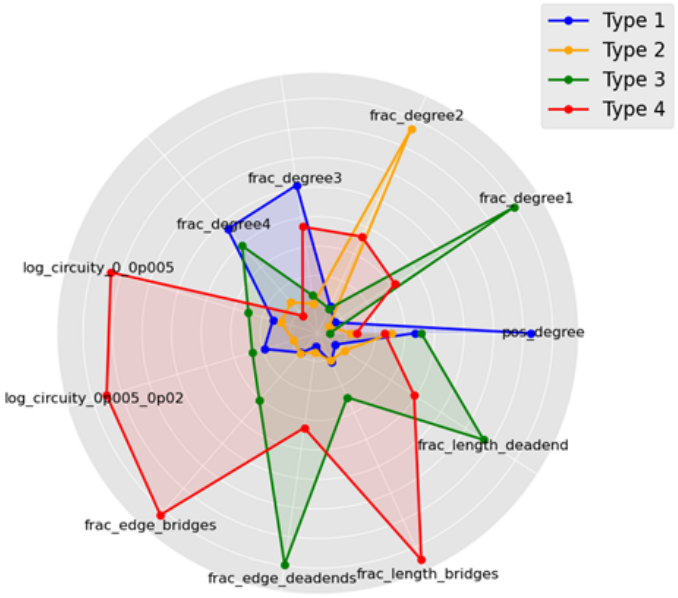

Fig. 2. The overall framework of the PredSocecOut and legends of some modules.

and then cluster them into four road network types, which are then used to label raw satellite images according to the corresponding geographical locations. In the second part, the preprocessed satellite imagery dataset is used as the input to train the model. This training process involves two stages of transfer learning. In the first stage, our goal is to transfer the knowledge learned from the ImageNet-21k object classification task, such as low-level visual features like edges, to the task of recognizing road network types from satellite images by designing model_1. Similarly, in the second stage, we aim to transfer the knowledge gained from the road network types classification task to the final task of socioeconomic outcomes prediction via constructing model_2. By leveraging this two-step transfer learning strategy, we can effectively utilize the rich information contained in the road network topology to improve the performance of socioeconomic outcomes (GDP) prediction, especially when the labeled socioeconomic data is limited.

IV. Methodology

A. Data Preprocessing

In this section, we will specifically clarify how to preprocess road topology data, socioeconomic data, and satellite image data to cope with the standards for subsequent model training.

a) Road topology: We download the basic structural information of the road on [1]OSMOpenStreetMap (OSM), and follow the formulas provided in [15] to calculate the measurement results of eleven road network topological structures, which are average degree, fraction of degree $= \{1, 2, 3, 4\}$ nodes, logarithmic circuity ($r \leq 0.5\,\mathrm{km}$), logarithmic circuity ($r > 0.5\,\mathrm{km}$), fraction of bridge edge length, bridge edge number, dead-end edge length and dead-end edge number.

TABLE I
OUR HYBRID MODEL STRUCTURES AND FINE-TUNING BLOCK

Model	modules	Output size
	input	384×384×3
ResNet50	7x7 Convolution	192×192×64
	Batch Normalization	192×192×64
	ReLU	192×192×64
	3x3 Max Pooling	96×96×64
	16 Bottleneck Blocks	12×12×2048
	Patch size	16×16
ViT	Patch Embedding	768
	Transformer Encoder	768
	Classification Head (Fine-tune)	final output size(adopted)

Moreover, the road network types aggregated by topological structures can map four socioeconomic development patterns [15], which are characterized by GDP and population mobility growth (PMG): high GDP, low PMG; high GDP, high PMG; low GDP, low PMG and low GDP, high PMG. Based on the prior knowledge, we fulfill K-means clustering on the measured values of these eleven topological structures with the K value set to four, and obtain four road network types. As shown in the radar chart in Fig. 2, we perform principal component analysis on these four road network types clustered, and we can clearly observe the proportions of various road topology in each type. For example, type 1 mainly includes frac_degree3, frac_degree4, pos_degree (average degree).

b) Socioeconomic data: The obtained socioeconomic data is in the original image file format with the suffix ADF, which is not conducive to the extraction of pixel metadata information, so we use the ArcMap tool to convert it to the raster data format with the suffix TIF in order that the values can be easily extracted. Of course, due to the resolution of the image, there will be multiple values in a sample area, so we directly perform the averaging process. At the same time, the experiment involves geospatial location information and requires a unified geographical projection coordinate system. We use the GCS_WGS_1984 geographic coordinate system, which is the standard defined by World Geodetic System 1984 (WGS_1984). In this way, the three-dimensional information on the earth surface can be converted into a coordinate system on a two-dimensional plane, ensuring that the areas located by the same longitude and latitude coordinates are consistent.

c) Satellite imagery: We download through the key provided by [2] Bing Maps, and allocate the downloaded satellite images to the category folder according to the road network topology type. The organizational structure is $root/class/image\_name.png$, where $root$ is the root directory of the satellite image data set, $class$ is the folder of the road network topology type to which the image belongs, and $image\_name.png$ is the file name of the image. So that we can read images and their corresponding labels in the folder structure through ImageFolder to meet the needs of model training. We separate the satellite images data into the training set and the validation set in a percentage of 7:3.

Each sample area collects 50 satellite images and if the all images are divided together, it will cause an imbalance in the data distribution. The reason is that satellite images in the same sample area may whole be split into training set or validation set. Therefore, our division strategy is changed to divide training dataset and validation dataset in a percentage of 7:3 in each 1km^3 sample area, and then merge them. This ensures that each sample area is distributed in the training dataset and validation dataset in proportion to ensure data distribution balance.

B. Model Structure

The proposed hybrid pre-trained model consists of two modules, ResNet50 and ViT, which combines a CNN and a transformer network model that introduces an attention mechanism. The model adopts a ResNet50 stem block, which consists of a 7×7 convolutional layer followed by batch normalization, a ReLU activation function and a 3×3 max pooling operation. After these blocks are bottleneck blocks in a variable number. It owns the representation $Rl + Ti, S, T/m$, where l is quantity of convolutional layers, and m means input patch resolution. In ViT [18], it reshapes the input feature map from ResNet50 into series of the unfolded two-dimensional map $X^M \in \mathbb{R}^{N\times\left(M^2\cdot C\right)}$, where C counts the quantity of channels, (M, M) represents every patch size divided by feature map, and N represents the quantity of patches, that also is used as the suitable input array head for the ViT. Moreover, a trainable linear projection layer is used to adjust the input array to the fixed latent vector size δ before feeding it into the transformer. This layer can learn to compress or expand the input feature map to the target dimension δ, allow model share parameters and ensure input and output dimensions are consistent across different layers, as follows:

$$z_0\left[x_{class}; x_m^1 E; x_m^2 E; \cdots; x_m^N E\right] + E_{pos} \tag{1}$$

$$E \in \mathbb{R}^{\left(\mathrm{W}^2 \bullet C\right)\times\delta}, \qquad E_{pos} \in \mathbb{R}^{(N+1)\times\delta} \tag{2}$$

where $z_0^0 = x_{class}$ represents a learnable embedding into the array of these embedded patches, whose mode at the output of the ViT encoder (z_L^0) serves as the image representation y, defined by (3). In fine-tuning setting, it adds a classification head, which is performed by a Multilayer Perceptron (MLP), to z_0 for suitable classification, then pre-train the whole model.

$$y = LN\left(z_L^0\right) \tag{3}$$

During the above process, to make the transformer more suitable to model, it commonly converts the two-dimensional map into a one-dimensional array vector, and input the results to the encoder. It includes alternating layers of multiheaded self-attention (MSA) and MLP blocks, shown in (4), (5). At the beginning of each block, model owns Layernorm (LN) and after it adds residual connections.

$$z'_\ell = MSA\left(LN\left(z_{\ell-1}\right)\right) + z_{\ell-1}, \qquad \ell = 1\ldots L \tag{4}$$

$$z_\ell = MLP\left(LN\left(z'_\ell\right)\right) + z'_\ell, \qquad \ell = 1\ldots L \tag{5}$$

where ℓ is alternating layers of MSA, range from 1 to L.

C. Road Network Types Classification

ImageNet-21k is a large-scale object classification image dataset jointly developed by Stanford University and Princeton University. It contains about 14 million diverse images from 21000 WordNet synonym sets, covering rich semantic concepts such as animals, plants, and vehicles. Compared with the well-known ImageNet-1k (containing 1000 categories), its categories are even greater, reaching 21843 types. Using ImageNet-21k for transfer learning can identify low-level and medium-level features such as edges and corners, greatly improving trained model generalization ability for other tasks. Our method first uses ImageNet-21k as original task dataset to train the hybrid model. We define this process, the ImageNet-21k object classification task, as $\mathcal{T}_1$. Then, we transfer the knowledge gained from $\mathcal{T}_1$ to the classification task of identifying road network types on satellite images, which is defined as $\mathcal{T}_2$.

In $\mathcal{T}_1$, the training data $\mathcal{D}_1 = \{(x_{1i}, y_{1i})\}$ consists of the basic images $x_{1i} \in \mathcal{X}_1$ and object category labels from ImageNet-21k. Similarly, in $\mathcal{T}_2$, the training data $\mathcal{D}_2 = \{(x_{2i}, y_{2i})\}$ consists of satellite images $x_{2i} \in \mathcal{X}_2$ and road network type labels $y_{2i} \in \mathcal{Y}_2$. Since satellite images are taken from an overhead perspective compared to the object-centered ImageNet-21k, and their data scales are different, so $P(\mathcal{X}_1) \neq P(\mathcal{X}_2)$, $P(\mathcal{X}_1)$ represent a marginal probability distribution of $\mathcal{X}_1$. Specifically, we first need to download the pre-trained hybrid model trained by the dataset $\mathcal{D}_1$ and parameters to the local as initialization. Then, referring to the road network category labels $\mathcal{Y}_2$, we fine-tune the final fully connected layer of the above hybrid model or attach a classification head to the above hybrid model to construct model_1, i.e., a fine-tuned hybrid model including ResNet50 and ViT. Finally, we input $\mathcal{D}_2$ to model_1 and use the SGD optimizer with momentum to train.

D. Society Economy Indicators Prediction

The goal of final task is to predict socioeconomic development status through satellite images, which is a typical regression task. We define this task as $\mathcal{T}_3$. The training data available at this stage is very limited, so we transfer the knowledge learned in $\mathcal{T}_2$, a data-rich task. In $\mathcal{T}_3$, the training data $\mathcal{D}_3 = \{(x_{3i}, y_{3i})\}$ consists of satellite images of the specified area $x_{3i} \in \mathcal{X}_3$ and a limited number of socioeconomic indicators values labels $y_{3i} \in \mathcal{Y}_3$. $\mathcal{T}_2$ and $\mathcal{T}_3$ own similar feature spaces of the input satellite images, which are collected in same area, that is $\mathcal{X}_2 = \mathcal{X}_3$, $P(\mathcal{X}_2) \approx P(\mathcal{X}_3)$. The tasks $\mathcal{T}_2$ (road network) and $\mathcal{T}_3$ (socioeconomic indicators) both own economic elements, but are different task. The process of $\mathcal{T}_3$ can be divided into two steps. In the first step, we load the model_1 and parameters in $\mathcal{T}_2$, then we fine-turn the model_1 classification head by $\mathcal{Y}_3$ to form model_2, and input $\mathcal{D}_3$ to train. Specifically, for $\mathcal{D}_3$, we first preprocess it to obtain an expanded $h \times w \times d -$ dimensional input $x \in \mathbb{R}^{hwd\times 1}$, then feed them to model_2, we can get a matrix vector as follows:

$$\hat{x} = f\left(Wx + b\right) \tag{6}$$

Since the first passing module of model_2 is ResNet50, with k convolutional filters of size $h \times w$, then in Equation (6), $W \in \mathbb{R}^{k \times hwd}$ represents a weight matrix, $b \in \mathbb{R}^k$ is a bias term, f is a non-linearity function, and $\hat{x} \in \mathbb{R}^k$ represents the output, the same size as the input. So, in a scalar output for every filter, the results are $\hat{x} \in \mathbb{R}^{1 \times 1 \times k}$. Next the second passing module of model_2 is ViT, whose input is the output of the previous network. We train the model based on the updated classification head. Finally, we obtain a trained model_2, whose output is the feature vector in a specified dimension, details in Section IV. Note that this process only extracts features and does not need to be updated on the parameters. In the second steps, we use the features with ridge regression model to fit predicted values of socioeconomic indicators as (7).

$$min_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \quad (7)$$

where y_i is target variable, in task is actual indicators values; x_{ij} is the j feature of i sample; β_j is the regression coefficient of j feature; λ is the regularized strength parameter.

V. Experiments

In this section, we will detail how to set up experiments to evaluate the performance to predict the regional economy from satellite imagery. Experiments use the mentioned hybrid model and the transfer learning method, which uses road network types as proxies for socioeconomic development patterns. Then, we will give conclusions from experiments and analyze them.

A. Experimental Settings

a) Datasets: The experiment mainly involves three types of data sets, namely road topology text datasets saved in JSON format, socioeconomic datasets saved in CSV format, and satellite image datasets in picture format. Below we will introduce these three datasets in detail.

- Road topology dataset: In the experiment, we select four cities in China, namely Beijing, Guiyang, Henan and Shanxi, as sample areas. Within every 1km^2, we download the basic structural information, like road nodes and edges on OSM, located by the central longitude and latitude coordinates of the respective areas. And then we calculate the corresponding feature values through the road topology formula mentioned in the data preprocessing section, and combine the longitude and latitude coordinates to generate road topology attribute data for each 1km^2 area. Lastly, we add another column, which is the road network type generated based on the road topology attribute clustering during the data preprocessing stage. Combine this data frame to get the final dataset. The detailed central longitude and latitude and the entire region range are shown in Table I.
- Socioeconomic dataset: In the experiment, we use GDP value to characterize the status of socioeconomic development. We download the GDP kilometer grid data in 2019 from [3]Environmental Science and Data Center of the Chinese Academy of Sciences. The data type is grid with a resolution of 1km^2 lattice data. After using the above socioeconomic data preprocessing method, the geographical space projection coordinate system is unified and converted into raster data in TIF format. Therefore, it is easy to extract GDP attribute value in the pixel and integrate it with the city name, regional longitude and latitude to form the socioeconomic dataset.
- Satellite image dataset: We register an API key with Bing Maps to download satellite images. Locate the area by city name, latitude and longitude to download it. The image size is 1000×1000 pixels. At the same time, in order to enrich the image data used for model training, we set a position offset of 20m in each sample area and collect 50 satellite images in every 1km^2 area. We name these images with the actual offset longitude and latitude, and save as a satellite image dataset.

For all datasets, we extract the road network type in the road topology dataset and the GDP in the socioeconomic dataset based on the shared longitude and latitude data frame, and combine the image name, original longitude and latitude, offset longitude and latitude, and city name to form attribute information file of the satellite image dataset. The file can solve the shortcomings of satellite images being highly unstructured.

b) Evaluation Metrics: To evaluate the performance of the classification task of identifying road network types on satellite images, we use two common evaluation metrics, precision and cross-entropy loss function $H(p, q)$, defined by (8), where p,q represent true class label and actual predicted class label respectively.

$$H(p, q) = -\sum_x p(x) \, logq(x) \quad (8)$$

To evaluate the regression task of predicting GDP attribute values from satellite imagery, we use the common evaluation metrics coefficient of determination (R^2). R^2, defined by Equation (9), is an indicator applied to evaluate the degree of fitting the predicted GDP by the regression model. Its value ranges from $-\infty$ to 1, where 1 represents a best fit, 0 means that degree of fit is equivalent to a simple average model, and a negative number indicates that degree of fit is worse than a simple average model. R^2 is calculated by comparing difference between model predictions and actual observations to difference between overall mean and actual observations.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (9)$$

As in (9), SSE represents the residual sum of squares, which is the sum of squares of the value calculated by y_i minus $\hat{y}_i$; SST represents the total sum of squares, which is the sum of squares of the value calculated by y_i minus $\bar{y}$. y_i is the value

to be fitted, $\bar{y}$ is the average of actual observed values, and $\hat{y}_i$ represents model prediction value.

B. Baselines and Experimental Details

a) Baselines: In this paper, to demonstrate the rationale and the effectiveness behind our choice of a hybrid pre-trained model combining traditional CNN (ResNet50) and transformer network (ViT), we select several baselines as alternative pre-trained models for comparison. They can be created via pip timm library and are described as follows:

- **VGG-11** [28] is a classical CNN framework, it uses ImageNet-1k as pre-trained dataset and the input images resolution is 224×224.
- **ResNet50** [14] solves the gradient vanishing problem in classical CNN by introducing residual connections, making the network deeper and easier to train. The dataset and input are like as VGG-11.
- **ViT** [38] is also a computer vision model that can be used to classify images, based on the transformer architecture. It is different from traditional CNN, ViT directly processes sequences of image patches and uses a self-attention mechanism to capture long-range dependencies and overall features in images. It uses ImageNet-21k as pre-trained dataset and the resolution of input images is 384×384.

ResNet50 and ViT can also be used separately as ablation experiments of hybrid models. Furthermore, we used nighttime light intensity as a proxy for economic patterns [11,16,17], as comparative experiments to validate the effectiveness and advantages of using the four road network types clustered as proxy.

b) Experimental Details: For our experiments, we download the models hosted on the Hugging Face Hub. This only requires installing the timm library via pip and then using the create_model method to download and instantiate the models. Since we are using a Hybrid ViT in PyTorch, with the hybrid object being a combination of ResNet50 and ViT, we set the create_model parameter to 'vit_base_resnet50_384', where '384' indicates that the model requires input images of 384×384 resolution. For the road network type classification task, in both the fine-tuning and pre-training stages, we first download the pre-trained hybrid model and its parameters to leverage the transferred knowledge. We then modify the final fully connected classification layer of the model, setting the output size to 4, which corresponds to the number of road network types we aim to classify. We adopt a gradual fine-tuning strategy for this task. During the first 10 epochs, we only train the newly added fully connected layer, while keeping the parameters of pre-trained layers frozen. In these subsequent epochs, we unfreeze all the model parameters and train the entire model using a learning rate of 1e-4 and a weight decay of 0.001. We choose the SGD optimizer with a momentum of 0.9 to facilitate faster convergence. For the satellite image-based GDP prediction regression task, we load the model trained on the previous road network classification task. Similarly, we adjust the output feature dimension of the fully connected layer, setting it to 2304, as this configuration exhibits the best performance. We then perform forward propagation to extract the feature vectors from the input satellite images and build the Ridge Regression model to predict GDP values. The model training process is carried out in the Google Colab environment, which provides a T4 GPU for accelerated computation.

As to set up comparative experiments, we first download the global average annual nighttime light intensity gridded dataset from 2019. We then crop the dataset to the approximate geographical coordinates covering the China regions so that it will reduce the computational and storage demands in subsequent processing steps. Next, we match the latitude and longitude coordinates from the socioeconomic dataset to the nighttime light intensity dataset, and add the corresponding light intensity values as a new column. We then categorize the light intensity values into distinct levels, following the binning strategies outlined in [13,14].

C. Performance Analysis

In our experiments, we first conduct a comparative study between the fine-tuned hybrid model and several other pre-trained network models. The primary focus is to analyze their validation accuracy and cross-entropy loss on the satellite image-based road network type classification task. To ensure fairness, we use the prepared dataset in all the experiments. The comparative study also includes the results of the ablation experiment, as displayed in Tables II and III. Through these results, we can draw several observations. First of all, for the three pre-trained models, the classification precision varies across different city datasets. This is also true in practice. Road planning in various regions is inconsistent. Areas with well-developed road networks or simple road structures tend to have higher classification precision. Secondly, these results clearly demonstrate that our hybrid model achieves the highest classification precision across all four cities. Compared to VGG-11, the hybrid model has an 10% - 16% higher precision. It also outperforms ResNet50 by 9% - 10% and ViT by 4% - 6%. Additionally, when we combine the data from all four cities and conduct this experiment, the hybrid model still exhibits the best performance for road network type classification.

Finally, we conduct a comparative analysis between our experimental method, which uses road network as a proxy, and the commonly adopted approach that employs nighttime light intensity as a proxy for socioeconomic development. We use R^2 as the evaluation metric for this comparison. As shown in Fig. 3, we can observe that for the Shanxi and Henan regions, our experimental method using the road network proxy outperforms the nighttime light intensity approach. Specifically, our method improves the R^2 by 3% in Shanxi and nearly 2% in Henan. Across all four cities, the results obtained from our experimental approach consistently outperforms the comparative method. The best improvement is observed in the Guiyang region, where our method achieves an 8% higher R^2 value. These findings lead us to the conclusion

that, while the performance may vary across different cities, using the road network topology as a proxy for socioeconomic factors can better help the model learn meaningful representations of socioeconomic development from satellite imagery. The superior performance of our method, compared with the nighttime light intensity proxy, suggests that the road network structure contains more informative cues about the underlying socioeconomic conditions in these regions.

D. Topology Structures Correlation Analysis

As explained above, the road network we use is obtained by clustering 11 road topology based on prior knowledge and principal component analysis. Considering the close relationship between road topology and economic activities, we hypothesized that it might be possible to bypass the step of generating road network types as a proxy and instead directly use the 11 topological structures to perform transfer learning for GDP prediction. To test this hypothesis, we set up a cor-relation experiment. We use a regression model to fit the value and leverage the R^2 as the evaluation metric. We also generate a correlation matrix and visualize the results, as shown in Fig. 4. It reveals that the correlation between individual topological structures and GDP is quite low. The results of fitting the 11 topological structures directly to the GDP data are extremely poor. However, we do observe that some structures exhibit high correlations with each other, such as the 0.86 correlation between frac_edge_bridges and frac_length_bridges. This suggests that it is the combination of different proportions of topological structures that can effectively represent various socioeconomic development patterns, thus indirectly justifying the validity of our clustering approach in the previous experiments.

These findings indicate that the road network types derived from the clustering process capture more nuanced and holistic information about the underlying socioeconomic conditions compared to the individual topological structures alone. The indirect relationship between the road network topology and GDP highlights the importance of our step-wise approach in using the road network as a proxy for socioeconomic analysis, rather than a direct mapping between individual topological features and economic metrics.

VI. Conclusion

In our work, we propose a approach for predicting socioeconomic outcomes, named PredSocecOut. Different from previous method, we first time use the road network types as a data-rich proxy clustered by eleven topological structures and adopt a novel hybrid pre-trained model, combining ResNet50 and ViT. The proposed method achieves significantly effective and accurate prediction. And through extensive experiments on various city datasets, the results demonstrate that the proposed approach outperforms several state-of-the-art baselines. The result can also be used as a reference for government departments to formulate policies for sustainable development or resource coordination and allocation.

In the future work, we plan to pursue two main directions. Firstly, we will attempt to predict other representative socioeconomic outcomes, such as population mobility and people consumption level, to verify the universality of using road network types as a data-rich proxy for socioeconomic development patterns. Secondly, we will conduct experiments using datasets sampled globally, especially focusing on underdeveloped countries where socioeconomic data is scarce and difficult to obtain, considering the current experimental data limited in geographical scope of China. This will allow us to prove generalization of the method and gain a more comprehensive understanding of the relationship between road network topology and socioeconomic activity across a diverse set of regions and economic contexts.

References

[1] Chakraborty, Chiranjit, and Andreas Joseph. "Machine learning at central banks." (2017).

[2] Ghose, Anindya, and Panagiotis G. Ipeirotis. "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics." IEEE transactions on knowledge and data engineering 23.10 (2010): 1498-1512.

[3] Zhou, Zhi-Hua. "A brief introduction to weakly supervised learning." National science review 5.1 (2018): 44-53.

[4] Athey, Susan. "The impact of machine learning on economics. The Economics of Artificial Intelligence." (2017).

[5] Mullainathan, Sendhil, and Jann Spiess. "Machine learning: an applied econometric approach." Journal of Economic Perspectives 31.2 (2017): 87-106.

[6] Pu, R., Gong, P., & Xu, B. "Advances and applications of very high spatial resolution satellite remote sensing". Remote Sensing of Environment, 2023, 281, 113293.

[7] Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109.1 (2020): 43-76.

[8] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.

[9] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning." Journal of Big data 3 (2016): 1-40.

[10] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q."A comprehensive survey on transfer learning." Proceedings of the IEEE 109.1 (2020): 43-76.

[11] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping", AAAI, vol. 30, no. 1, Mar. 2016.

[12] Zhao, Min, Yuyu Zhou, Xuecao Li, Wenting Cao, Chunyang He, Bailang Yu, Xi Li, Christopher D. Elvidge, Weiming Cheng, and Chenghu Zhou. 2019. "Applications of Satellite Remote Sensing of Nighttime Light Observations: Advances, Challenges, and Perspectives", Remote Sensing 11, no. 17: 1971.

[13] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. "Combining satellite imagery and machine learning to predict poverty." Science 353.6301 (2016): 790-794.

[14] Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. "Using satellite imagery to understand and promote sustainable development. " Science 371.6535 (2021): eabe8628.

[15] Xue, J., Jiang, N., Liang, S., Pang, Q., Yabe, T., Ukkusuri, S. V., & Ma, J. "Quantifying the spatial homogeneity of urban road networks via graph neural networks." Nature Machine Intelligence 4.3 (2022): 246-257.

[16] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L."How to train your vit? data, augmentation, and regularization in vision transformers." arXiv preprint arXiv:2106.10270 (2021).

[17] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770-778.

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N."An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[19] Xu, Z., Lv, Z., Li, J., Sun, H., & Sheng, Z. "A novel perspective on travel demand prediction considering natural environmental and socioeconomic factors." IEEE Intelligent Transportation Systems Magazine 15.1 (2022): 136-159.

[20] Arokiasamy, P., Selvamani, Y., Jotheeswaran, A. T., & Sadana, R. "Socioeconomic differences in handgrip strength and its association with measures of intrinsic capacity among older adults in six middle-income countries." Scientific Reports 11.1 (2021): 19494.

[21] Loh, X. M., Lee, V. H., Leong, L. Y., Aw, E. C. X., Cham, T. H., Tang, Y. C., & Hew, J. J. "Understanding consumers' resistance to pay with cryptocurrency in the sharing economy: A hybrid SEM-fsQCA approach." Journal of Business Research 159 (2023): 113726.

[22] Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino. "Capturing Macro-Economic Tail Risks with Bayesian Vector Autoregressions." Journal of Money, Credit and Banking (2020).

[23] Mihoub, A., Snoun, H., Krichen, M., Salah, R. B. H., & Kahia, M. "Predicting covid-19 spread level using socio-economic indicators and machine learning techniques." 2020 first international conference of smart systems and emerging technologies (SMARTTECH). IEEE, 2020.

[24] Luo, Y., Yan, J., McClure, S. C., & Li, F. "Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model." Environmental Science and Pollution Research (2022): 1-13.

[25] Li, T., Xin, S., Xi, Y., Tarkoma, S., Hui, P., & Li, Y. "Predicting multi-level socioeconomic indicators from structural urban imagery." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

[26] Fan, X., Wang, X., Zhang, X., & Yu, X. B. "Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors." Reliability Engineering & System Safety 219 (2022): 108185.

[27] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

[28] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv: 1409.1556 (2014).

[29] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 4700-4708.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[31] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[32] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. "Improving language understanding by generative pre-training." (2018).

[33] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[34] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

[35] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021, pp. 8748-8763.

[36] Raghu, Maithra, and Eric Schmidt. "A survey of deep learning for scientific discovery." arXiv preprint arXiv:2003.11755 (2020).

[37] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. "Revisiting unreasonable effectiveness of data in deep learning era." Proceedings of the IEEE international conference on computer vision. 2017. pp. 843-852.

[38] Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., ... & Pavetic, F. "Flexivit: One model for all patch sizes." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. pp. 14496-14506.