High precision Intracity Temperature Estimation based on Generated Point Clouds

GuanJie Huang, KaiQun He, Xiaoyue Lyu, Yu Zang, *Senior Member, IEEE*, Hang Shu, Lei Zhao, Jonathan Li, *Senior Member, IEEE* and Cheng Wang, *Senior Member, IEEE*

***Abstract*—Estimation of urban surface temperature is crucial for urban planning and emergency management. Due to the complexity of intracity structures, it is very challenged to acquire satisfied prediction errors of the Land Surface Temperature (LST) at very high resolution, like 60-by-60 meters. By considering this, we propose a low-cost method for generating urban point clouds via readily accessible city data. Then we design an efficient descriptor, GeoFeature Distribution Matrix (GFDM) to describe the complex intracity structure. By utilizing GFDM, we introduce a 3D Urban structure guided temperature Prediction network (3D-UP Net) to capture the complex relationship between urban structure, upper atmospheric conditions, and surface temperature. The proposed 3D-UP Net is generalizable, capable of predicting future surface temperature for existing cities and even for those that are planned. Experiments conducted in multiple regions of China demonstrate that our method's error is less than 1.5 Kelvin (in most cases) at a high resolution (60-by-60 meters).**

***Index Terms*—Land surface temperature, point cloud, deep neural network.**

I. Introduction

WITH the acceleration of global warming and urbanization, the variation in urban temperature has become a prominent research topic in recent years. Effective urban planning and infrastructure-based growth strategies are increasingly dependent on accurate high-resolution urban climate predictions[1], [2]. Cities, which concentrate large populations and infrastructures, are epicenters for significant climate-driven impacts[3], [4]. Predicting Land Surface Temperature (LST) with high resolution and precision presents considerable challenges due to the extensive heterogeneity[5] and the complex effects of human behaviours, such as seasonal variations, diurnal temperature differences, urbanization, population density, and energy structure, etc[6], [7], [8], [9]. While high-resolution urban LST data for historical periods can be readily acquired via satellite, forecasting future high-resolution urban LST remains a difficult endeavor. Seasonal forecasts or long-term projections from climate models are either at coarse resolutions(>1 km) [1], [2] or integrate the urban landscape well[10].

Previous research has approached surface temperature using dynamic downscaling and statistical downscaling methods. Dynamic downscaling, exemplified by models such as the Weather Research and Forecast (WRF) model[11], calculates the temperature field based on the urban atmospheric boundary layer. Although it is computationally demanding, the highest resolution typically achievable is between 1-2km[2], [12], [13]. Meanwhile, the performance of the statistical downscaling methods are strongly rely on the selection of the temperature-related features such as land cover, vegetation indices and other observational data[14], [15], [16], [17]. Such a way will make the model hard to get generalizable prediction results, because the choice of feature maybe arbitrary and lack of physics representation in the statistical models[18], [19], [20], [21]. Both of these methods have limitations when it comes to high-resolution LST prediction, making it challenging to predict temperatures for diverse and future urban landscapes.

Actually, high precision urban temperatures is highly related to local urban structures, the unique urban structure can cause local temperature fluctuations compare to the surroundings[22]. To address this, recent work proposed DeepUrbanDownscale(DUD)[23], a deep learning framework that leverages high-precision 3D point clouds. DUD captures the structural features of urban surface by converting 3D point clouds into the novel local spatial coefficient index (LSCI). This combination of high-resolution 3D point clouds with atmospheric data enables the physically meaningful prediction of urban LST at both high-resolution and high-precision. However, obtaining high-resolution urban 3D point clouds data and their semantic label, particularly at the extensive scale required for cities, presents challenges due to its limited availability, affecting the model's generalizability. At the same time, with the development of remote sensing, remote sensing images have been applied to scientific research[24], [25], [26], [27] and have achieved considerable results. By considering this, [28] proposed Physics Informed Hierarchical Perception (PIHP) network, which try to utilize high resolution remote sensing images to generate urban surface in a cheaper way. However, due to the lost of 3D urban texture, such a method may lead to the unsatisfied generalization of the model.

In this study, we firstly propose the generated point clouds for urban structure description in a low cost way. Specifically, the urban generated point cloud is built by Digital Surface

This work is supported by Fujian Provincial Natural Science Foundation under Grant 2023J01004, The corresponding author: Yu Zang, email: zangyu7@126.com

G. Jie, Y. Zang, K. Qun, H. Shu,and C. Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail:guanjiehuang@stu.xmu.edu.cn;zangyu7@126.com;cwang@xmu.edu.cn).

Xiaoyue Lyu, Geospatial Intelligence and Mapping Lab, University of Waterloo, Canada (e-mail: x6lyu@uwaterloo.ca).

L. Zhao is with the Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA (e-mail: leizhao.yale@gmail.com).

J. Li is with the GeoSTARS Laboratory, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, Canada (e-mail: junli@uwaterloo.ca).

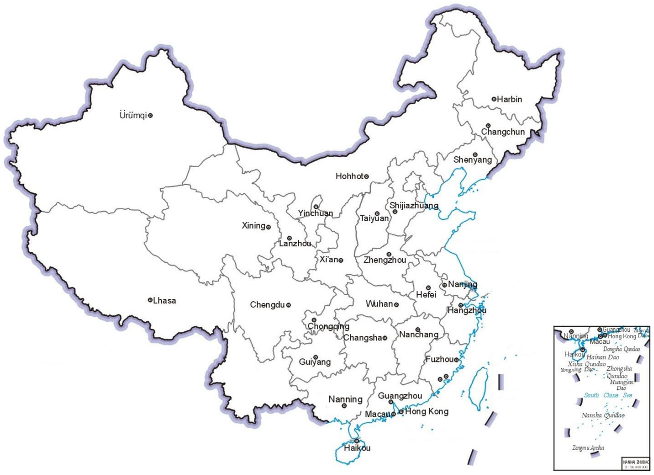


Fig. 1. Distribution of the selected urban in China.

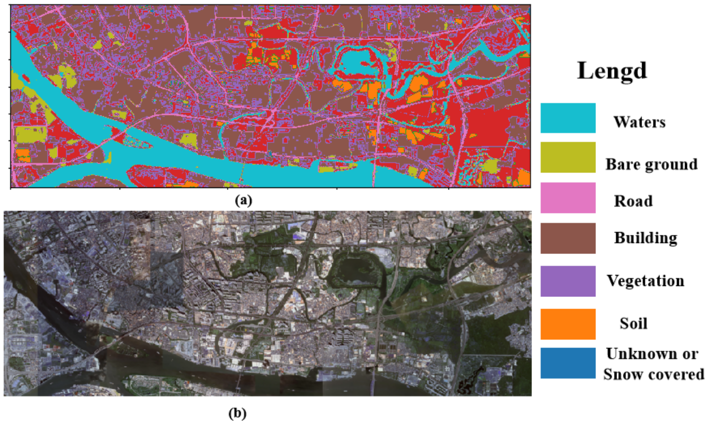


Fig. 2. Visualization of remote sensing images and their segmentation. (a) Visualized map of the labeled. Inferred semantic categories (water, buildings, vegetation, soil, roads/pavements, and unknown) with text colored by the label color. (b)Satellite remote sensing images of Guangzhou,China.

Mode(DSM) and 3D Building Model, meanwhile the 3D semantic label of the point cloud is propagated from the corresponding label of high resolution remote sensing images. Then based on this, we further design a learning based framework to predict high-resolution urban LST. Such a framework make it possible to generate high precision point clouds for various cities, thus supporting to get generalizable models for large range, like the whole China. In this framework we firstly propose GeoFeature Distribution Matrix (GFDM) to further abstract the generatred point cloud. The GFDM is able to capture the features of the point cloud at adjustable resolution, we also demonstrate mathematically that the GFDM is equivalent to the the original point clouds at the finest resolution. Based on this, we further propose 3D Urban structure guided temperature Prediction network (3D-UP Net), different from previous works[23][28] that employed the MLP framework, our 3D-UP Net utilizes a convolutional framework, which allows for more effective extraction of local urban structures. such a network comprehensively considers the affection of the atmosphere condition, terrain, and urban structure to LST, thus able to provide high precision results and better model generalization.

In the experiments, we rigorously tested our model across over 30 major or provincial capital cities in China to evaluate the performance of the model we proposed, Compared with previous work[23], our model is capable not only of predicting future LST for cities involved in the training but also for cities that were not involved in the training dataset. And we compared our 3D-UP Net with several point cloud-based methods, including PointNet, PointNet++, and Point Cloud Transformer(PCT), as well as with the current state of the art method, PIHP-net. Our results show that the prediction error for LST was reduced to below 1.7 Kelvin, representing a 14.2% decrease in error compared to PIHP-net.

II. Study area and data

In this study, 30 major or provincial capital cities across China were selected as the study areas, as illustrated in Fig.1. These cities encompass a variety of terrain types, including metropolitan areas, mountains, and bodies of water. To implement our 3D-UP Net, we collected and processed a suite of datasets from diverse sources within these regions. Each city chosen is characterized by high levels of urbanization and population density, making them particularly relevant for research that aims to enhance urban planning, contribute to carbon neutrality efforts, and improve climate prediction accuracy. Specifically, six distinct datasets are collected and processed for this paper, as described below:

The first dataset utilized in this research is the Landsat-based Ready-to-use (RTU) land surface temperature product, sourced from the CASEarth DataBank system[1]. This dataset is derived from the USGS's Landsat8 OLI/TIRS sensor, featuring a spatial resolution of 30 meters. LST retrival for this product employs a single channel algorithm [29]. Data spanning from 2014 to 2019 were collected for regions specifically encompassing urban landscapes. In line with our objective to determine block-level urban temperatures from satellite data, the urban areas within each region were divided into 60m-by-60m segments for analysis in our study. Consequently, the LST data was resampled to match a 60-meter resolution to facilitate our experimental needs.

The second dataset comprises nationwide remote sensing images of major and provincial capital cities in China which is the combination of surface reflectance image in red, green, and blue bands. Fig.2 (a) and (b) provide examples from GuangZhou, China: (a) displays the semantic labels visually, with different urban surface attributes distinguished by varying colors. while (b) presents the original remote sensing image. These multi-spectral remote sensing images, obtained from the Google Earth website[2], boast a resolution of 1 meter and cover all major provinces and key cities across China. The dataset not only includes spectral data, but also integrates pixel-level labeling of the remote sensing imagery, revealing a robust correaltion with LST and surface attributes. Each pixel has been automatically categorized into predefined classes such as

[1] http://databank.casearth.cn/

[2] https://earth.google.com/

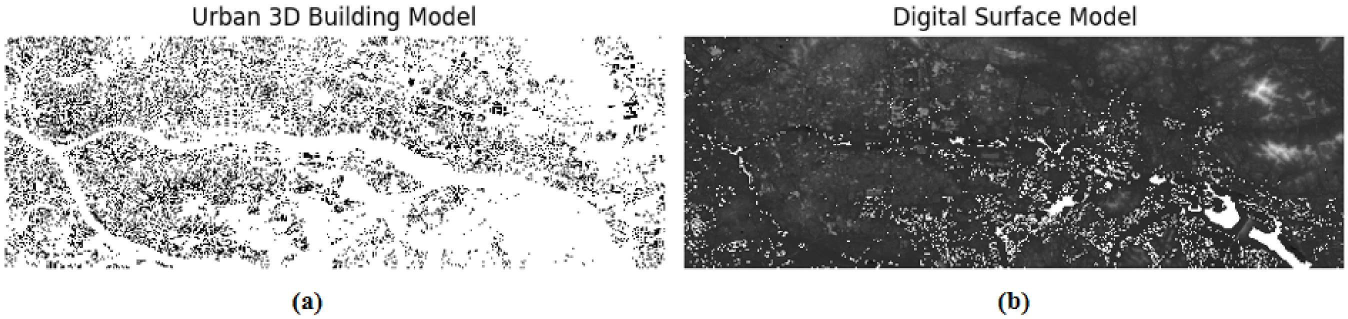


Fig. 3. (a) Visualization of Urban 3D Building Model of Guangzhou. (b) Visualized map of Digital Surface Model of Guangzhou. It should be pointed out that the blank area in (a) represents null, while the blank area in (b) represents low altitude areas

water, buildings, vegetation, soil, and roads/pavements, etc., employing a classification technique [30] with an accuracy of 96.5%.

The third dataset is the Digital Surface Model (DSM), an open-access dataset provided by JAXA, which is available for downloaded for free in[3]. The DSM is a three-dimensional representation that captures the Earth's surface, including all natural and man-made objects like buildings and vegetation. It is constructed using stereophotogrammetry techniques. This DSM dataset is particularly valuable in geographic information science, remote sensing, and urban planning for its ability to offer precise elevation of both terrain and surface features.

The fourth dataset consists of an Urban 3D Building Model, which is publicly accessible and can be downloaded from Baidu Map[4]. This dataset captures the height information of urban structures, such as streets and buildings, with a resolution of 1 meter, However, it's worth noting that the Urban 3D Building Model is relatively sparse. It records only building information and lacks details on other features like river terrain, which poses challenges for direct usage. Fig. 3 shows the disparity between the original density of urban 3D building models and the DSM.

The fifth dataset employed in our study is the atmospheric forcing data, sourced from the NASA MERRA-2 reanalysis data system [31]. This dataset is freely available through the NASA MERRA-2 website[5]. It features a spatial resolution of 0.5^o latitude $\times$ 0.625^o longitude and offers daily temporal resolution, providing a snapshot of the general weather conditions over the city for each day. The specific variables included in this dataset are detailed in Table I.

The Last dataset is the temperature measured by the weather station from the website of China's National Greenhouse Data System[6]. The dataset contains temperatures measured at weather stations in urban centers in 30 cities for each day from 2014-2019. And the kind of temperatures used is daily average air temperature at 2m.

Due to national policy restrictions, the original high-resolution multispectral satellite remote sensing imagery cannot be openly shared. Nonetheless, we have made the computed GeoFeature Distribution Matrix (GFDM) available, with further details presented in the Methods section. The complete dataset will be disseminated via an FTP server at a subsequent stage.

TABLE I
ATMOSPHERIC FORCING DATA. THE AIR TEMPERATURE IN THIS STUDY MEANS THE ATMOSPHERIC TEMPERATURE AT THE REFERENCE HEIGHT (60M ABOVE THE SURFACE CANOPY TOP) IN REANALYSIS DATA OR CLIMATE MODELS.

Type	Name
Land Surface Forcings	Surface absorbed longwave radiation
	Surface income shortwave flux
Land Surface Diagnostics	Total precipitation land
Single-Level Diagnostics	Atmospheric temperature max
	Atmospheric temperature mean
	Atmospheric temperature min
Analyzed Meteorological Fields	Surface pressure
	Atmospheric temperature at the reference height
	Eastward wind
	Northward wind
	Specific humidity

III. METHODS

In this study, we utilize the Land Surface Temperature (LST) data measured by Landsat as our training labels. Although the LST measured by Landsat does not reflect the urban surface temperature well compare to that measured by weather stations, the reason we still chose Landsat-measured LST as the training label is because not every city has the weather station, and the locations of weather station are very sparse, also lacking high-resolution LST data of whole city. But meanwhile, Landsat can measure the LST of most areas with high resolution.

Land Surface Temperature of cities is influenced by numerous factors, including urban meteorology, human activities, and the structure of the city itself. It is known from [32] that Landsat calculates the LST by regressing the Top of Atmosphere Reflectance against Surface Emissivity. And Surface Emissivity is closely related to the urban surface materials and structures[33]. Also, in [34] has proposed that based on the current observed forcing data, MERRA-2 model can utilize the 3DVAR algorithm on the GSI to predict short-term future forcing data. Based on the availability of these two sets of data and their correlation with LST, we can predict urban surface temperature based on urban structure and meteorological information.

[3] https://www.eorc.jaxa.jp/ALOS/en/aw3d30/data/
[4] https://lbs.baidu.com
[5] https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
[6] http://data.sheshiyuanyi.com/WeatherData/

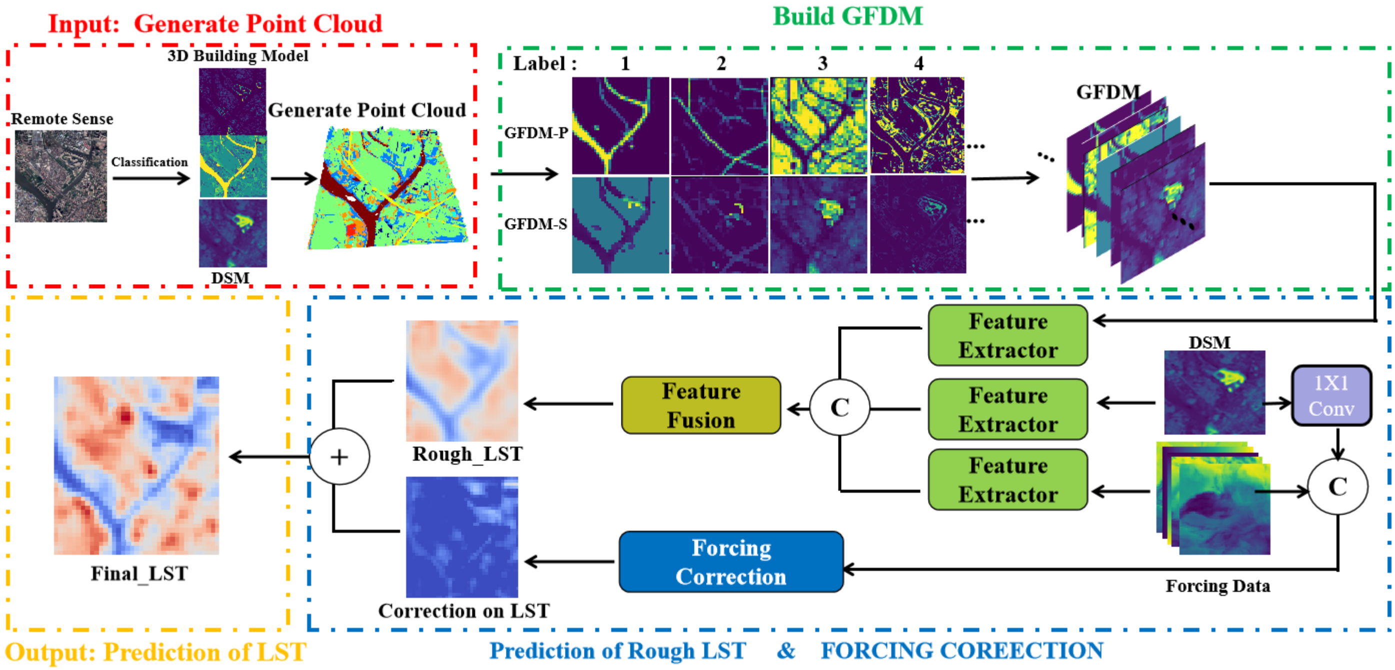


Fig. 4. Flowchart of the proposed framework for predicting the LST.

By consider of this, we have developed the 3D Urban structure guided temperature Prediction network (3D-UP Net). Our work is divided into four stages: 1) Generate Point Clouds; 2) Build GeoFeature Distribution Matrix (GFDM) 3) Prediction of rough LST; 4) Forcing Correction. The workflow for predicting LST is shown in Fig. 4. In stage 1, we generate three-dimensional point clouds of the city by combining the 3D Building Model, Digital Surface Model (DSM), and remote sensing labels. In stage 2, we convert the three-dimensional point clouds into GFDM descriptors. In stage 3, we input the GFDM, Forcing Data, and DSM into the network to predict rough LST. In stage 4, we input the Forcing Data and DSM into the Forcing Correction module to obtain the correction value for LST, which is then combined with the rough LST to obtain the final LST prediction value.

A. Generation of Point Clouds

Initially, the satellite imagery undergoes pixel-wise semantic labelling using an image ground target classification network [30], segmenting the image into geophysical categories such as water, soil, roads, buildings, vegetation, and others. The resolution of these semantic labels is 1 meter. Subsequently, in order to fuse the Digital Surface Model (DSM) with remote sensing image labels, we use the bilinear interpolation method to increase the resolution of the DSM to 1 meter. Thirdly, after utilizing these dataset to construct the 2D urban elevation image, we delineate the contour lines at different heights across the cityscape, then, we subtract the average altitude of the city to avoid unnecessary point clouds generated by the city base. Along these contours, points are added at 1-meter intervals in the vertical direction. This step results in the creation of the basic generated point cloud of the city. We then assign the corresponding semantic labels to the basic point cloud, culminating the formation of a labeled city-generated point cloud. The methodology for generating these point clouds is outlined as follows, and an example of the generated point cloud is displayed in fig. 5:

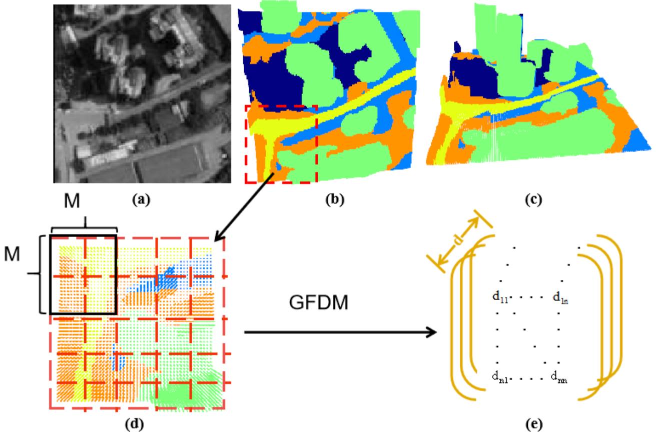


Fig. 5. The illustration of the generated point cloud and GFDM. (a) Original remote sensing images. (b) Land surface semantic label image, different colors represent different categories. (c) Generated point cloud combining labels. (d) The background is a part of (b) and the red dashed grid represents the 60-by-60 meter pixels. (e) A example of GFDM. For each pixel, we can calculate the percentage and average height of each category in one pixel. There are d categories in total, so each pixel corresponds to a vector of $1 \times 2d$.

$$H(x, y) = DSM(x, y) + BM(x, y) \tag{1}$$

$$h(i) = \min(H) + i - DSM_{avg} \quad (0 \leq i \leq \max(H) - \min(H)) \tag{2}$$

$$F(i) = f(h(i)) = (x_i, y_i) \tag{3}$$

$$P = (x, y, z, l) = (F(i), k, L(F(i))) \\ \begin{cases} \text{for } k = 0 \text{ to } h(i) \text{ k++ } h(i) > 0, \\ \text{for } k = 0 \text{ to } h(i) \text{ k- - } h(i) < 0. \end{cases} \tag{4}$$

Where $DSM(x, y)$ represents the height of DSM at coordinates (x, y), $BM(x, y)$ denotes the height of the 3D Building Model at the same coordinates and DSM_{avg} represents the average height of DSM. The term $h(i)$ signifies the height of the i^th contour line, while $F(i)$ designates the horizontal coordinate (x, y) of the i^th contour line. $L(x, y)$ corresponds to the the urban remote sensing image labels at (x, y). Finally, P symbolizes the generated point cloud.

B. GeoFeature Distribution Matrix (GFDM)

To integrate the structural information of the generated point cloud into the neural network, one straightforward approach is feed the entire 3D point cloud of a local area directly into the network through point-based networks, as suggested in studies such as [35], [36] and [37]. However, our research found that the enormous scale of the city and complexity of urban surface features, coupled with the high memory requirements of processing point cloud, compromises the overall system's feasibility. This complexity hinders the method's ability to capture the general relationship between regional average temperature and local geometric data. Moreover, the intracity LST is closely related to the local neighborhood structure. By considering these, we have designed a descriptor GFDM, which aggregates potential factors influencing local urban surface temperature. By utilizing GFDM, memory requirements can be significantly reduced and local neighborhood structure can be efficiently transmitted.

GFDM comprises two key components: the semantic proportion index P and the semantic structure index S. The semantic proportion index quantifies the proportion of each semantic labels within a pixel, whereas the semantic structure index calculates the average height of structures corresponding to each semantic labels within the same pixel. This approach facilitates an abstraction of spatial configurations, surface roughness, and building verticality, all of which are influential factors in local atmospheric turbulence on the urban surface.

The Land Surface Temperature (LST) images of urban surface are sourced from land satellites and feature a spatial resolution of 60 × 60 meters. In our method, the preliminary semantic labeling process identifies five primary temperature-related structural categories: water, buildings, vegetation, soil, and road/pavement, as shown in fig. 2. Following this, as illustrated in fig. 5 (c), we generate a point cloud of the city, where the various colors in the point cloud represent different urban surface categories. Subsequently, this generated point cloud is divided into pixel that corresponding to the pixel of the LST images. Within each pixel, we calculate the proportion of each category present in the point cloud to obtain the semantic proportion index. Similarly, we compute the average height within each pixel of the point cloud to obtain the semantic structure index.

Specifically, the GFDM is quantified by the following equation:

$$a(i, j) = \begin{cases} 1 & \text{if } L(i) = j, \\ 0 & \text{if } L(i) \neq j. \end{cases} \tag{5}$$

$$p_j = \frac{\sum_{i=1}^{n} a(i, j)}{\sum_{i=1}^{n}} \tag{6}$$

$$s_j = \frac{\sum_{i=1}^{n} H(i) * a(i, j)}{\sum_{i=1}^{n} a(i, j)} \tag{7}$$

$$d_{xy} = (p_1, p_2, \ldots, s_1, s_2, \ldots) \tag{8}$$

$$GFDM = \begin{bmatrix} d_{11} & \ldots & d_{1w} \\ \vdots & \ddots & \vdots \\ d_{h1} & \ldots & d_{hw} \end{bmatrix} \tag{9}$$

where i denotes the point number i, and n represents the total number of points. The term p_j refers to the semantic proportion index of label j within the pixel, while s_j corresponds to the semantic structure index of label j within the pixel. The variable d_{xy} defines the descriptor of GFDM for row x and column y. Lastly, w and h indicate the width and height dimensions of the GFDM. Fig. 5 demonstrates examples of the constructed GFDM.

One may question how the GFDM can effectively substitute point clouds as descriptors to capture the essence of local geometric information. This can be understood through the following equation:

$$\lim_{M \to M_0} p_j = \begin{cases} 1 & \text{if } L(x, y) = j, \\ 0 & \text{if } L(x, y) \neq j. \end{cases} \tag{10}$$

$$\lim_{M \to M_0} s_j = \begin{cases} 1/2H(x, y) & \text{if } L(x, y) = j, \\ 0 & \text{if } L(x, y) \neq j. \end{cases} \tag{11}$$

$$d_{xy} = \begin{pmatrix} \overbrace{0 \cdots 0}^{L(x,y)-1} & 1 & \ldots & \frac{1}{2}H(x, y) & \overbrace{0 \cdots 0}^{n-L(x,y)} \end{pmatrix} \tag{12}$$

$$GFDM[x, y] = d_{xy} \equiv P = (x, y, H(x, y), L(x, y)) \tag{13}$$

Let M denotes the resolution of GFDM, and M_0 represents the resolution of point cloud which is 1 meter. As the resolution of GFDM approaches that of the point cloud, we observe, according to equation 4, that each pixel has points only in vertical direction. Consequently, in such a configuration, there can only be a singular category of points within any given coordinate(x, y), which renders the value of p_j in this pixel to be either 1 or 0, as depicted in equation 10. Concurrently, the average height of the point cloud within the pixel approximates to either half of the height of the coordinate(x, y) or 0, as demonstrated in equation 11. Utilizing the values of p_j and s_j at (x, y), we can construct d_{xy} via Equation 12. Further, as shown in Equation 13, by applying $GFDM(x, y)$, we can

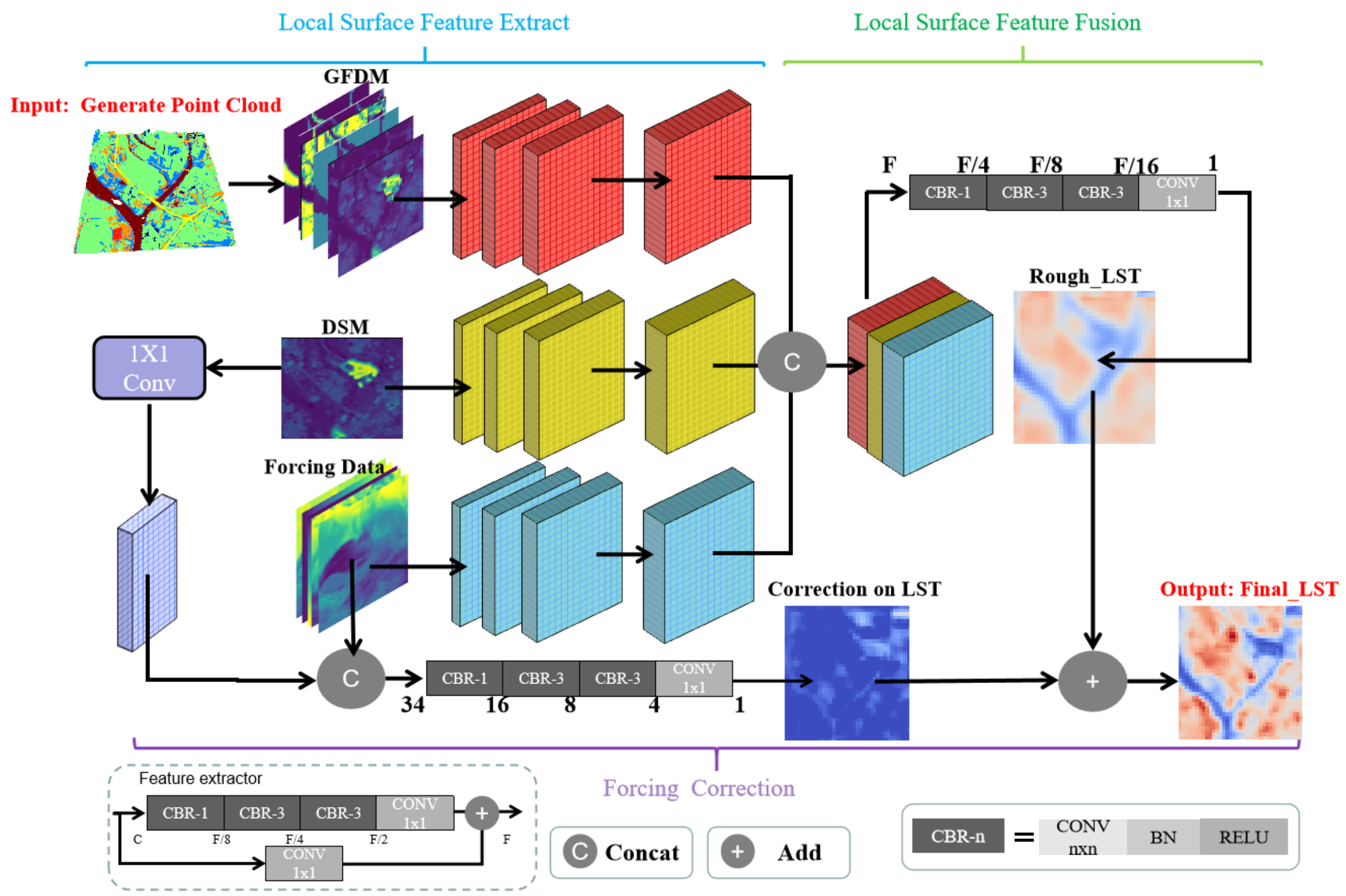


Fig. 6. An overview of 3D Urban structure guided temperature Prediction network(3D-UP Net)

deduce both $H(x, y)$ and $L(x, y)$ to generate point P by $GFDM(x, y)$. Thus, GFDM effectively replicates the impact that the generated point cloud would have.

C. 3D Urban structure guided temperature Prediction network (3D-UP Net)

The architecture of our proposed 3D-UP Net is depicted in Fig. 6. Diverging from previous approaches[23][28] that employed Multi-Layer Perceptron (MLP) to extract feature in a single pixel(60-by-60 meters), our study harnesses convolutional framework to extract feature from pixel and its surrounding pixels. This methodology shift is inspired by the strong relationship between LST and the local neighborhood structure[38]. Compare to MLP, convolutional framework can better extract spatial relationships between a pixel and its surrounding.

The input of 3D-UP Net including GFDM, DSM, Forcing Data, which is mentioned in Section II and Section III-B. And the $final_LST$ is the output of the net. And the 3D-UP Net is structured into three components: Local Surface Feature Extraction, Local Surface Feature Fusion and Forcing Correction.

1) Local Surface Feature Extraction: The initial phase in the Local Surface Feature Extraction involves fusing the urban remote sensing image labels, DSM, and 3D Building Model to form a point cloud. Using previous discussed equations, we then construct the GFDM. Features from the GFDM, DSM, and atmospheric forcing data are extracted concurrently via their respective Feature Extractors operating in parallel.

To delve into the structural details of the urban surface, we employ the structre of ResNet [39], designed to extract features that reveal local surface characteristics and spatial structure from the three types of data. The Feature Extractor is crucial for the neural network to detect high-resolution variations in urban surface temperature. It comprises a 5-stage network, where each stage contains a convolution block which is shown in Fig. 6.

2) Local Surface Feature Fusion: Upon completing the Local Surface Feature Extraction phase, the extracted features from the GFDM, DSM, and Foricng Data are concatenated, resulting a composite local surface feature. This composite feature is then input into the Local Surface Feature Fusion module to obtain the rough LST. The fusion module comprises four convolutional blocks which is shown in Fig. 6 .

3) Forcing correction: As is well known, urban LST is closely related to forcing data. However, the Forcing Data we obtain can only reflect the atmospheric environment over the city, and the surface temperature predicted based on such Forcing Data can only reflect the temperature over the city, which has a significant difference from the actual surface temperature. To quantify this difference, we collected data from

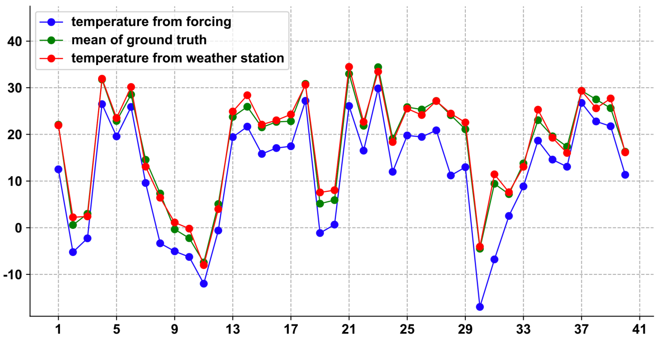


Fig. 7. Comparing the temperature from weather stations, forcing and average LST. The ordinate represents the temperature in Celsius (°C). The abscissa represents the city number.

weather stations across 30 well-known cities in various climate zones of China, covering around 40 different seasons. We then calculated the average temperatures near the surface from the weather station and the higher atmospheric temperatures from the forcing. These averages were plotted on a scatter plot, as shown in Fig. 7. Despite the differences between the LST from forcing and those measured by weather stations exist, a strong correlation exists overall (with a relatively fixed difference for unique city). However, weather station data is not available for every city or region, based on this fact, we designed a forcing correction module to simulate the correction value between the actual LST and the satellite-measured LST, to more accurately reflect near-surface conditions.

Since the Digital Surface Model (DSM) reflects the terrain structure of an area, and the forcing data which is mentioned in Table.I reflects the meteorological information of the area, using the forcing correction module is akin to solving the physical equations in a process-based model through deep learning, essentially acting as a dynamic simulator. We input the forcing data and DSM, convert the DSM's channels to match those of the forcing data using a 1x1 convolution, then add them together. After a series of convolutions, we obtain the correction value, as specifically calculated in figure 6.

4) Loss function: In contrast to previous studies[23][28] that adopted root mean square error (RMSE) as their loss function, our approach utilizes Weighted Mean Square Error (WMSE) to account for the loss. The rationale for this choice is based on the unique feature of urban LST prediction, where certain areas such as rivers, industrial zones, and roads exhibit significant deviations from the average LST. Unlike RMSE, Mean Square Error (MSE) omits the square root operation, rendering it highly sensitive to these areas with marked biases. Consequently, MSE is more effective in training the model recognize the distinctive features of such regions. Furthermore, we incorporate weights to MSE to emphasize these outliers, ensuring that data with larger prediction errors have a proportionally greater influence on the total loss. Nonetheless, to maintain consistency with prior work and ensure fairness in comparison, we employ RMSE as the performance metric for evaluation. The loss function is formulated as follows:

$$W = |Y - \hat{Y}| \tag{14}$$

$$L_i(Y_i, \hat{Y}_i) = W \times (Y_i - \hat{Y}_i)^2 \tag{15}$$

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(L_i(Y_i, \hat{Y}_i))^2 \tag{16}$$

where Y represents the ground truth matrix of LST, and $\hat{Y}$ denotes the predicted LST matrix. The term W signifies the residuals between the ground truth and predictions. By taking the absolute value of W, we ensure it is always a positive quantity, which concurrently serve as the weight in our WMSE calculation. Thus, larger prediction errors are weighted more heavily, contributing more significantly to the overall loss.

IV. Results and Discussion

A. Implementation details and baselines

1) Implementation details: In alignment with the methodologies from previous studies [40], we have quantified the performance of our models on the test datasets using the RMSE, defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 / n} \tag{17}$$

Our implementation of the 3D-UP Net leverages PyTorch [41] as the underlying framework. The training of our network is conducted end-to-end, utilizing the Adam optimizer [42] with an initial learning rate of 0.001, and we apply a decay rate of $0.5^{1/500}$ after each epoch to improve convergence.

2) Baselines: Considering the computational demands of dynamic equation based methods at this resolution, we've focused our comparative analysis on a selection of statistical models. This include linear regression [43], K-Nearest Neighbors (KNN) regression [44], and random forest regression [45], all implemented using Scikit-learn [46]. Additionally, we compare our approach to deep learning methods, specifically those based on point cloud like PointNet[36], PointNet++[47], Point Cloud Transformer(PCT)[37] and the state-of-the-art in the field, PIHP-net [28]. The specific parameter settings detailed as follows:

- **Linear Reg** means linear regression. This model uses linear regression for simulation, where the input is a concatenation of GFDM, Forcing Data, and DSM, reshaped into $1 \times d$ vectors for regression analysis. In essence, linear models are structured in the following manner:

$$y = \sum_{i} \beta_i x_i + \varepsilon \tag{18}$$

  where y is LST, and x_i is GFDM, Forcing Data, and DSM, which are reshaped to a $1 \times d$ vector. β_i indicates how LST changes linearly with each x_i, while ε is the normally distributed error.
- **KNN Reg** means KNN regression. The KNN regression model is configured with a fixed number of neighbors at 4, with the tree's maximum depth at 30, using the same input features as the Linear Regression model.

- **RF Reg** is the random forest regression, which is set up with a predefined number of 150 trees and used to predict urban temperatures.
- **PointNet** We first employ the Farthest Point Sampling(FPS) method to sample the generated point clouds in each LST pixel into 1024 points. Then we extract features from the generated point cloud by PointNet. Due to computational constraints, this model predicts temperatures for individual LST pixel, which may result in a loss of local neighborhood structure.
- **PointNet++** Due to the generated point cloud being uniform with each point spaced 1m apart, we employ PointNet++ as the backbone to extract features from the generated point clouds within each LST pixel. We have set up three layers of SSG(Single-Scale Grouping), with the first layer having a neighborhood radius of 5m and a point sampling number of 32; the second layer with a neighborhood radius of 10m and a point sampling number of 64; and the final layer samples all points.
- **PCT** means Point Cloud Transformer. In this experiment, we employ PCT to extract features from the generated point cloud, wherein we employed four layers of TransformerBlock, with the number of neighbors set to 16.
- **PIHP-net** This approach utilizes a bidimensional emprical model decomposition (BEMD) to dissect raw data into multi-scale components, building upon established signal processing techniques [48], [49], [50]. The method primarily relies on MLP for prediction, rather than convolutional operations, which may limit its ability to effectively capture spatial features from the surrounding neighborhood. And in this part of the experiment, we used the original hyperparameters.

B. Parameter Discussion and Ablation Study

In this section, we design a series of experiments to evaluate the effects of the various components in the proposed 3D-UP Net.

1) Discussion of GFDM Resolution: We have previously shown in Equation 13, that as the resolution of GFDM approaches that of the generated point cloud, GFDM can replicate the effect on the generated point clouds. In this experiment of parameter discussion, we evaluate how GFDM resolution influences LST prediction.

This parameter discussion experiment involved five cities: Guangzhou, Zhengzhou, Chongqing, Shenyang, and Xiamen, which correspond to the south, central, west, north, and east regions of China, respectively. We used data from 2014 to 2018 for training and data from 2019 for testing. All the data were processed to be cloud-free, mitigating potential anomalies in satellite temperature readings. Each city was divided into 120×120 slices for training. One-third of the dates were allocated for testing, with the remaining dates split into 70% for training and 30% for validation .

When the resolution of GFDM set 60m, GFDM and LST have the same resolution. When the resolution of GFDM set 60m, 30m, 15m, 10m, and 1m corresponding to 1, 1/2, $1/4$, $1/6$, $1/60$ the resolution of LST, respectively. In these situation, interpolation methods were employed to adjust the LST resolution. And when the resolution of GFDM set 120m, we performed downsampling processing for network training. The errors for 3D-UP Net at these resolutions are detailed in Table II. The result shows that the resolution of GFDM approch the resolution of LST (like $120m$, $60m$ and $30m$) has the minimum error ($1.41K$, $1.35K$ and $1.31K$).

However, increasing GFDM resolution does not necessarily enhance 3D-UP Net's accuracy possibility. From the Table II, we can observe that as the resolution of GFDM is set to 15m, 10m, and 1m, the prediction errors reach $1.45k$, $1.61k$, and $1.81k$, respectively. This is because that employing simple interpolation on LST to improve resolution does not increase the effective information of LST, while in the same convolutional framework with an identical receptive field, the higher the resolution of GFDM, the less local neighborhood information there is about urban structures, just like the state-of-the-art method, PIHP-net, employs an MLP framework to regress urban structures within individual pixel, neglecting a substantial amount of local neighborhood structure. This also explains why our method is able to achieve improvements.

2) Ablation of Original DSM: In 3D-UP Net, as shown in Fig. 6, the original DSM data is a significant branch. Our experiments indicate that incorporating original DSM imagery influences prediction results. To validate this, we conducted an ablation study comparing results with and without the DSM branch, as depicted in Table III. The average RMSE for various cities improved by approximately 7% when incorporating DSM data. This improvement may be due to subtracting the average altitude of the city from the construction process of the point cloud generated by the city, and adding DSM information can enable the network to relearn the altitude information of the city. At the same time, adding DSM information to the forcing correction module can enable the network to understand the relationship between forcing data and DSM, making it perform better in correction.

3) Ablation of Forcing Correction: In 3D-UP-Net, as shown in Figure 6, Forcing Correction is an important component of the network. To verify this, we conducted an ablation study comparing the results with and without this module, as shown in Table. IV. The errors for various cities were significantly reduced($2.55K$) after applying forcing correction. This is because the forcing data we obtain is measured from above the city, and the LST predicted based on this can only reflect the temperature above the city. By incorporating this module, we can correct the temperature above the city to the surface temperature, as described in in Section III-C3..

Visualization of the experiments concerning GFDM resolution, DSM branches, and point based methods is provided in Fig. 8.

C. Comparisons

To thoroughly evaluate the proposed methodology against previous statistical and deep learning-based methods, we designed two sets of experiments. The first set focuses on single city temperature prediction to assess the model performance in specific urban settings. The second set focuses on cities which is "unseen" in the training samples.

TABLE II
THE AVERAGE TEMPERATURE ERROR OF DIFFERENT GFDM RESOLUTION SCHEMES

Resolution of GFDM	**Guangzhou**	**Zhengzhou**	**Chongqing**	**Shenyang**	**Xiamen**	**Avg. Error (Kelvin)**
120m	1.31	1.27	1.34	1.57	1.55	1.41
60m	1.23	**1.18**	1.42	**1.46**	1.47	1.35
30m	**1.06**	1.22	**1.27**	1.52	**1.42**	**1.31**
15m	1.37	1.27	1.37	1.73	1.51	1.45
10m	1.54	1.46	1.58	1.82	1.63	1.61
1m	1.62	1.74	1.93	1.97	1.77	1.81

TABLE III
THE ABLATION OF HOW THE DSM BRANCH AFFECT THE RESULTS

	Guangzhou	Zhengzhou	Chongqing	Shenyang	Xiamen	Avg. Error(Kelvin)
GFDM-30M without DSM branch	1.13	1.39	1.32	1.52	1.67	1.41
GFDM-30M with DSM branch	**1.06**	**1.22**	**1.27**	**1.52**	**1.42**	**1.31**

TABLE IV
THE ABLATION OF HOW THE FORCING CORRECTION AFFECT THE RESULTS

	Guangzhou	Zhengzhou	Chongqing	Shenyang	Xiamen	Avg. Error(Kelvin)
GFDM-30M without Forcing Correction	3.47	4.13	3.72	4.34	3.61	3.86
GFDM-30M with Forcing Correction	**1.06**	**1.22**	**1.27**	**1.52**	**1.42**	**1.31**

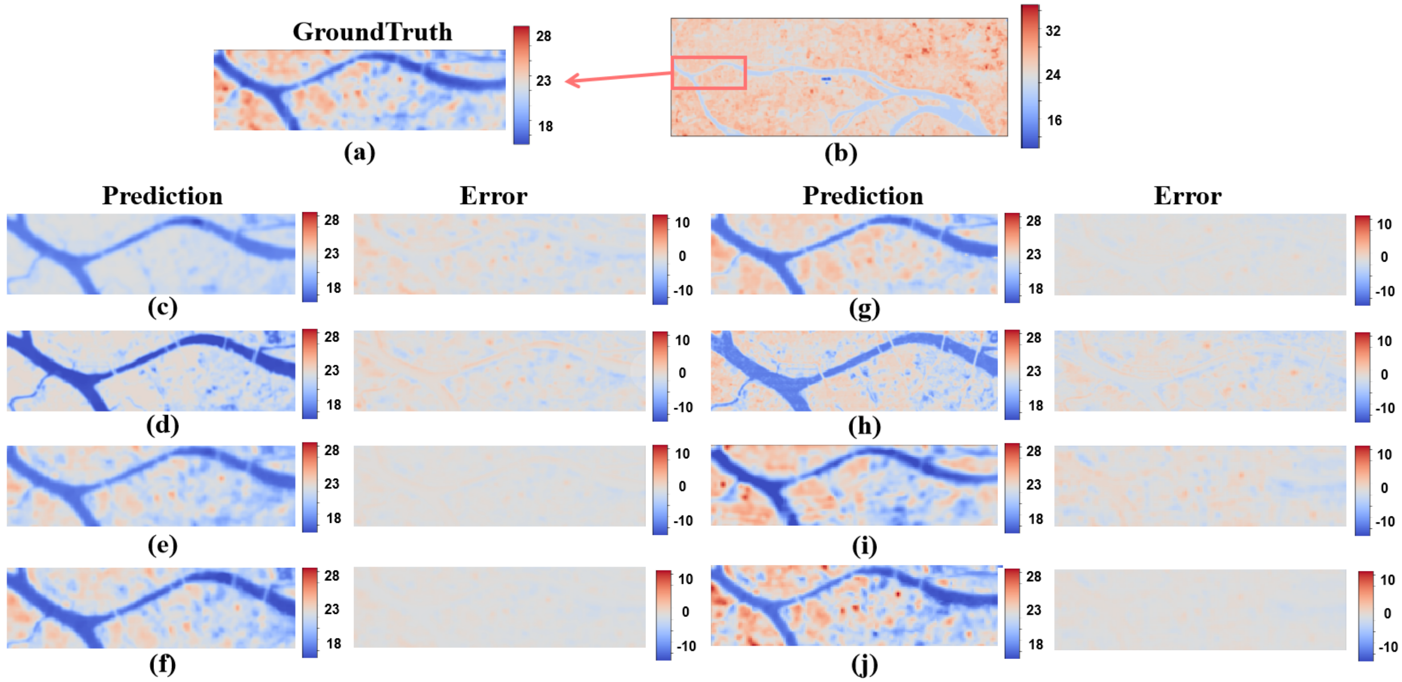


Fig. 8. Visualization of our method's ablation experiments on the Guangzhou dataset.(a) Part of Guangzhou groundtruth visualization. (b) Groundtruth visualization of Guangzhou. (c) GFDM-30M without forcing correction. (d) GFDM-30M without DSM branch. (e) GFDM-120M. (f) GFDM-60M. (g) GFDM-15M. (h) GFDM-10M. (i) PIHP-net. (j) GFDM-30M.

This experiment involved 30 major or provincial capital cities in China. However, despite using LST data measured by Landsat, there are still some cities with insufficient LST data to support prediction for single city. Therefore, we included the data from these cities as training samples in the second experiment of ”unseen” city for training .

Our study encompasses several major cities across China, each representing a typical city from different regions of the

TABLE V
OVERALL PERFORMANCE COMPARISON OF DIFFERENT APPROACHES ON THE EXPERIMENT OF SINGLE CITY TEMPERATURE PREDICTION. A SMALLER VALUE INDICATES A BETTER PERFORMANCE. THE AVERAGE ERRORS OVER ALL THE CITIES OF EACH METHOD ARE SHOWN IN THE LAST COLUMN

Model	RMSE (Kelvin)										
	Chengdu	Nanchang	Zhengzhou	Shenyang	Guangzhou	Chongqing	Hefei	Xiamen	Yinchan	Lhasa	Avg. Error
Linear Reg	4.61	5.82	5.63	3.37	5.05	4.13	4.50	5.36	6.29	6.23	5.10
KNN Reg	4.88	5.73	4.14	4.50	3.33	3.98	3.41	4.22	5.41	5.37	4.51
RF Reg	3.95	5.02	4.07	3.79	3.37	3.54	3.48	4.74	5.97	4.93	4.29
PointNet	2.28	2.05	1.79	2.13	1.63	1.76	2.31	2.05	2.64	4.24	2.28
PointNet++	1.97	1.92	1.63	1.94	1.46	1.52	1.75	2.05	2.53	4.35	2.10
PCT	1.95	2.06	1.81	1.77	1.44	1.49	2.10	1.68	2.35	4.17	2.03
PIHP-Net	1.82	1.67	2.13	1.66	1.32	1.87	**1.41**	1.58	2.19	4.12	1.97
3D-UP Net	**1.39**	**1.64**	**1.22**	**1.46**	**1.06**	**1.27**	1.57	**1.42**	**2.17**	**3.93**	**1.69**

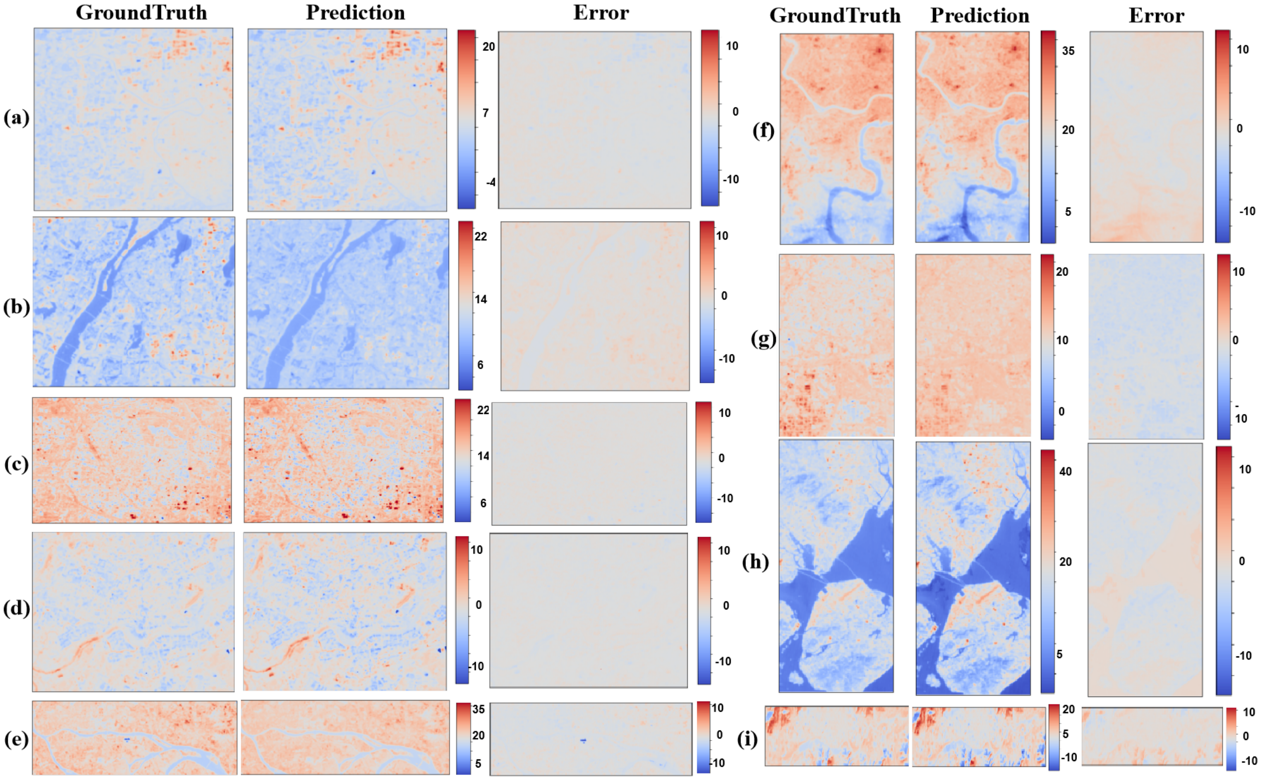


Fig. 9. Visualization of temperature predicted by 3D-UP Net for single city: (a) ChengDu, (b) NanChang, (c) Zhengzhou, (d) Shenyang, (e) Guangzhou, (f) Chongqing, (g) HeFei, (h) XiaMen, (i) Lhasa. Each colorbar's number represents the temperature in Celsius (°C). All visualization plots are averaged results from the corresponding multiple data.

country. Since the urban surface structure is minimally affected by time, the generated three-dimensional structures of cities can be used for long-term urban LST prediction. And we also selected urban forcing data from 2014 to 2017 as the input training input data, with urban LST serving as the training labels. Subsequently, we used forcing data and LST from 2018 and 2019 as the test data. We attempt to simulate and predict the future urban LST in this manner.

1) Single City Temperature Prediction: Although statistical methods may not be well-suited for predicting high-resolution urban LST, there are few existing methods for high-resolution urban LST prediction. And Statistical methods are still being used in some cases. In addition to traditional statistical methods(linear regression, K-Nearest Neighbors (KNN) regression, and random forest regression), we also compare our 3D-UP net with deep learning methods based on 3D point clouds (PointNet, PointNet++, and Point Cloud Transformer), as well as the state-of-the-art method (PIHP-Net). The parameter settings for each model are detailed in Section IV-A2.

The average error for each city is presented in Table V. Additionally, the column at the end displays the average errors for all cities combined, as derived from each analytical method. It's observed that traditional statistical model-based methods yield an average error exceeding 3.33K across all cities. Notably, the linear model exhibits the poorest performance with an error of 5.10K. This is attributed to the intricate and nonlinear nature of the interaction between Land Surface Temperature (LST), the terrestrial surface, and the upper atmosphere. KNN and Random Forest show variable performance, with errors ranging from 3.33K to 5.41K.

Conversely, 3D-UP Net achieves a much lower average error of 1.69K, closely matching the typical observational

TABLE VI
SPLIT OF TESTING AND TRAINING CITIES FOR EACH REGION

Region	City for training	City for testing
South China	Nanchang,Guangzhou,Haikou	HongKong,Nanning
west China	Guiyang,Chengdu	Changsha
NorthWest China	YinChuan,Xining	XiAn,LanZhou
Central China	Wuhan,HeFei	ZhengZhou
North China	ShenYang,JiNan,ZhengZhou,	ShiJiaZhuang
East China	Nanchang,JiNan,XiaMen	Hangzhou,Nanjing

TABLE VII
COMPARISON OF RMSE ON SIX REGIONS FOR TEMPERATURE PREDICTION OF TESTING CITIES USING 3D-UP NET AND OTHER BASELINE METHODS. THE AVERAGE ERRORS OVER ALL TESTING CITIES OF EACH METHOD ARE SHOWN IN THE LAST COLUMN

Model	South China		West China	NorthWest China		Central China	North China	East China		Avg. Error
	Nanning	HongKong	Changsha	XiAn	Lanzhou	Zhengzhou	ShiJiaZhuang	HangZhou	Nanjing	
Linear Reg	4.67	4.76	5.68	6.30	4.94	5.77	3.49	5.84	5.74	5.24
KNN Reg	4.25	3.57	4.94	5.93	4.69	4.67	2.82	5.19	5.21	4.59
RF Reg	3.98	3.44	4.82	5.77	4.55	4.26	3.72	4.72	4.77	4.44
PointNet	2.91	2.47	3.52	5.24	3.26	2.86	3.54	3.71	3.43	3.42
PointNet++	3.52	2.52	3.31	4.97	2.84	2.78	2.83	3.58	3.19	3.18
PCT	2.77	2.29	3.34	4.96	2.74	2.57	3.13	3.55	2.68	3.04
PIHP-Net	2.58	2.24	3.08	3.72	2.40	2.25	2.75	2.92	3.14	2.79
3D-UP Net	**2.29**	**1.93**	**2.47**	**3.77**	**2.45**	**2.04**	**2.38**	**2.67**	**2.39**	**2.48**

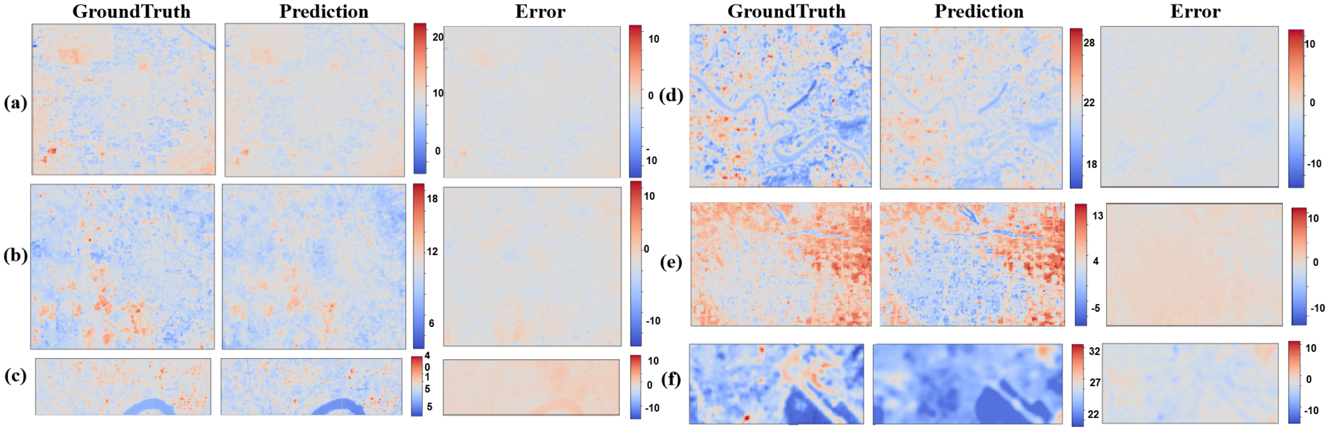


Fig. 10. Visualization of temperature predicted by 3D-UP Net for ”Unseen” city:(a) Xian, (b) Changsha, (c) Hangzhou, (d) NanNing, (e) Shijiazhuang, (f) Hongkong. Each colorbar’s number represents the temperature in Celsius (°C). All visualization plots are averaged results from the corresponding multiple data.

error of satellites in some cities like Zhengzhou, Guangzhou and Chongqing. While the point based method like PointNet records a average error of 2.28K, PointNet++ records a average error of 2.10K, PCT records a average error of 2.03K, and the state-of-the-art method PIHP-Net’s error stands at 1.97K. We found that deep learning methods based on point clouds have made significant improvements compared to statistical methods. This means that urban three-dimensional point clouds can effectively reflect the urban surface structure, thereby predicting urban land surface temperatures. Visualizations of prediction are shown in Fig. 9.

2) *”Unseen” City Temperature Prediction:* There is a situation we must concerned:when there are changes in the internal structure of the city such as urbanization. And this is precisely the reason we designed the experiments in this section. We conducted a series of experiments to evaluate the capability of the 3D-UP network to predict the temperature of cities not included in the training dataset. This study covers a range of important cities across China, divided into southern, western, northwestern, central, northern, and eastern regions. For each region, $2-4$ cities are selected for training, with 4 different seasons per city was covered (Table VI). The remaining city with $3-6$ season’s data serve as the test set.

Various methodologies, including Linear Regression, K-Nearest Neighbors Regression, Random Forest Regression, PointNet, PointNet++, PCT, PIHP-Net, and 3D-UP Net were compared. The average errors for each method across all test cities are compiled in Table VII. Notably, the increase in the scale of testing leads to higher errors, particularly for statistical models where errors in some cities exceed 7.03K. The average error for all cities is above 4K for these models.

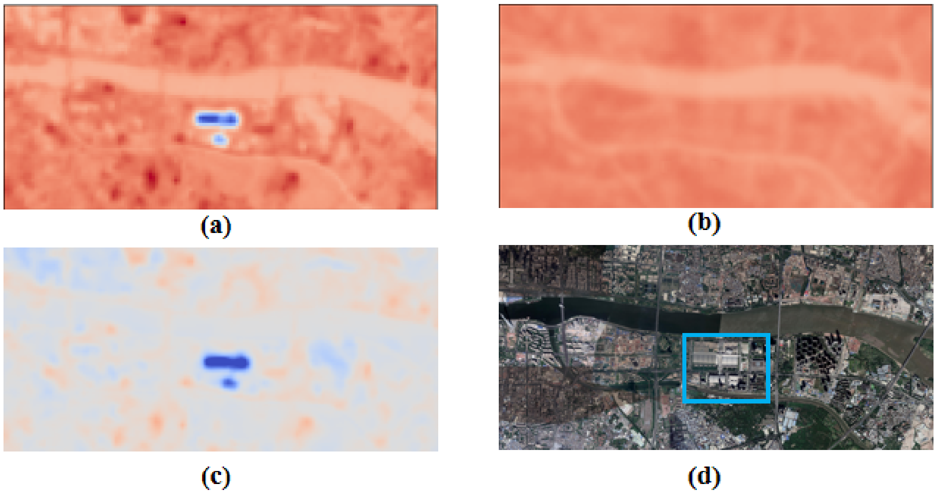


Fig. 11. Analysis of prediction accuracy of experimental results in GuangZhou. (a) The LST measure by satellite. (b) The prediction of LST by 3D-UP Net. (c) The error between (a) and (b). (d) Part of Guangzhou remote sensing, and the blue box indicates an exhibition center.

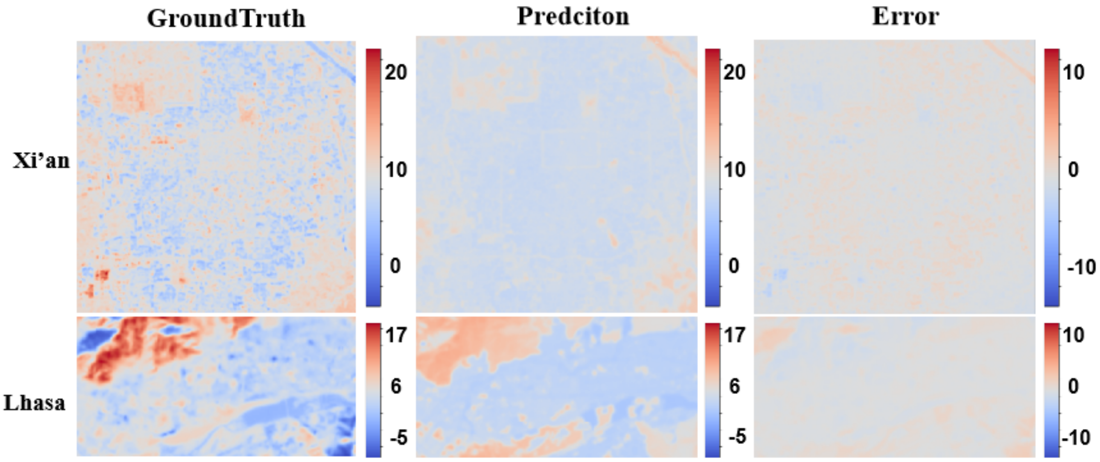


Fig. 12. blueVisualization of the LST prediction for Xi'an and Lhasa.

TABLE VIII
THE AVERAGE ALTITUDE AND ERROR OF LHASA XIAN AND XINING

	Lhasa	XiAn	XiNing
Altitude(M)	3650	1027	2275
Error(k)	3.93	3.38	3.71

D. Discussion

The section delves into the error analysis of our experimental results, highlighting not only the discrepancies between surface and atmospheric conditions but also identifying key factors influencing urban LST. Our main is to shed light on these aspects to inform and enhance future urban LST estimation systems.

1) Errors from building materials: While deep learning models show average errors of 3.42K (PointNet), 3.18K (PointNet++), 3.04K (PCT), 2.79K (PIHP-Net) and 2.48K (3D-UP Net). Visualizations for some of these cities are available in Fig. 10. In this section of comparative experiments, the point cloud based method yielded poorer results. This is attributed to the high complexity of large-scale urban point clouds, which makes it challenging for the network to learn urban structural features. Despite the challenges posed by urban complexity and various weather conditions, 3D-UP Net demonstrates commendable effectiveness. However, its superiority is less pronounced in broader-scale predictions compared to single-city forecasts.

One notable experiment was conducted in Guangzhou, where we found that the material of building roofs can significantly affect satellites measurements of surface temperatures.

A typical example of this phenomenon is shown in the fig. 11. During our single-city temperature prediction experiment in Guangzhou, we encountered a substantial error in (c).

By examining the corresponding remote sensing image (d) and focusing on the area within the blue box, we identified this location as a convention and exhiition center. The center's roof, constructed from high reflectivity materials, tends to reflect most of the solar radiation. This reflection leads to anomalously low surface temperature readings in that area, as detected by satellites, thereby complicating the model's ability to accurately capture the impact of such structures. But through our model, by learning a large amount of local neighborhood structure, we can provide a reliable LST data for reference which can be seen in (b).

2) Errors from Altitude: The average elevation of a city is another critical factor affecting the performance of the 3D-UP Net model. This is because the climate conditions in high-altitude regions is unpredictable. At the same time, monitoring data in high-altitude regions are relatively scarce, which hinders model training in these areas. Furthermore, high-altitude regions exhibit diverse terrain, with numerous mountains, canyons, and plateaus, making it challenging for the simple structure index in GFDM to capture these terrains. Additionally, high-altitude regions have varied land surface cover types, including snow, glaciers, alpine meadows, and rocks, among others. Different land surface cover types possess distinct thermal properties, rendering the simple portion index in GFDM inadequate for capturing these surface characteristics. The experiment gave examples of three cities, Lhasa, Xi'an, and Xining, with average elevations of 3650M, 1027M, and 2275M, respectively. The LST predictions for Xi'an and Lhasa are shown in the in the fig. 12.

V. Conclusion

In this research, we introduced a cost-effective approach to generate urban 3D point clouds and applied this method to create 3D point cloud dataset for major cities across China. Utilizing these, we developed the point cloud descriptor GFDM and the neural network 3D-UP Net. The 3D-UP Net leverages these descriptors for precise and high-resolution urban surface temperature prediction, integrating generated point clouds with upper atmospheric forcing data. This network is particularly valuable for forecasting future urban LST, easily incorporating data from regional climate models.

Comprehensive experiments across 30 major or provincial capital cities in China validate the superior performance of the proposed 3D-UP Net. It consistently surpasses previous methodologies, with an error generally below 1.7 Kelvin, Compared to the state-of-the-art PIHP, the error in LST prediction for a single city has been reduced by 15%, while the error in LST prediction for "unseen" cities has been reduced by 11%.

The primary contributions of this study include:

1) Propose a low-cost technique to generate urban 3D point clouds, applied to major Chinese cities.
2) Propose GeoFeature Distribution Matrix (GFDM) descriptor, derived from urban 3D point cloud, which effectively extracts urban surface structural features to support LST prediction.
3) Design the 3D-Urban structure guided temperature Prediction network(3D-UP Net), a novel learning based system which is able to provide high resolution LST prediction results under the guiding of 3D urban structures.
4) Design the 3D-Urban structure guided temperature Prediction network(3D-UP Net), a novel learning based system which is able to provide high resolution LST prediction results under the guiding of 3D urban structures.

Our research results also indicate that building materials can affect the measurement of LST in remote sensing images, and high altitude can have a certain impact on the model's extraction of urban surface structure features, but we can provide a referable LST data in this two situation by our model.

However, there are still some issues that we have not resolved: For example, since Landsat collects LST data from high altitudes, although it considers Surface Emissivity in calculating LST, which allows it to reflect LST to a certain extent, the LST obtained in this manner still cannot accurately represent the surface LST. In future work, we will collect more ground station-measured LST data for model training.

References

[1] M. Georgescu, P. E. Morefield, B. G. Bierwagen, and C. P. Weaver, "Urban adaptation can roll back warming of emerging megapolitan regions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 8, pp. 2909–2914, 2014.

[2] E. S. Krayenhoff, M. Moustaoui, A. M. Broadbent, V. Gupta, and M. Georgescu, "Diurnal interaction between urban expansion, climate change and adaptation in us cities," *Nature Climate Change*, vol. 8, no. 12, pp. 1097–1103, 2018.

[3] N. B. Grimm, S. H. Faeth, N. E. Golubiewski, C. L. Redman, J. Wu, X. Bai, and J. M. Briggs, "Global change and the ecology of cities," *science*, vol. 319, no. 5864, pp. 756–760, 2008.

[4] C. Mora, B. Dousset, I. R. Caldwell, F. E. Powell, R. C. Geronimo, C. R. Bielecki, C. W. Counsell, B. S. Dietrich, E. T. Johnston, L. V. Louis *et al.*, "Global risk of deadly heat," *Nature climate change*, vol. 7, no. 7, pp. 501–506, 2017.

[5] S. Gaffin, C. Rosenzweig, R. Khanbilvardi, L. Parshall, S. Mahani, H. Glickman, R. Goldberg, R. Blake, R. Slosberg, and D. Hillel, "Variations in new york city's urban heat island strength over time and space," *Theoretical and applied climatology*, vol. 94, no. 1, pp. 1–11, 2008.

[6] H. Li, R. Li, Y. Yang, B. Cao, Z. Bian, T. Hu, Y. Du, L. Sun, and Q. Liu, "Temperature-based and radiance-based validation of the collection 6 MYD11 and MYD21 land surface temperature products over barren surfaces in northwestern china," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1794–1807, 2021.

[7] B. Chun and J.-M. Guldmann, "Impact of greening on the urban heat island: Seasonal variations and mitigation strategies," *Computers, Environment and Urban Systems*, vol. 71, pp. 165–176, 2018.

[8] D. Zhou, J. Xiao, S. Bonafoni, C. Berger, K. Deilami, Y. Zhou, S. Frolking, R. Yao, Z. Qiao, and J. A. Sobrino, "Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives," *Remote Sensing*, vol. 11, no. 1, p. 48, 2019.

[9] X.-M. Zhu, X.-N. Song, P. Leng, D. Guo, and S.-H. Cai, "Impact of atmospheric correction on spatial heterogeneity relations between land surface temperature and biophysical compositions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2680–2697, 2021.

[10] L. Zhao, K. Oleson, E. Bou-Zeid, E. S. Krayenhoff, A. Bray, Q. Zhu, Z. Zheng, C. Chen, and M. Oppenheimer, "Global multi-model projections of local urban climates," *Nature Climate Change*, vol. 11, no. 2, pp. 152–157, 2021.

[11] H. Kusaka, H. Kondo, Y. Kikegawa, and F. Kimura, "A simple single-layer urban canopy model for atmospheric models: Comparison with multi-layer and slab models," *Boundary-layer meteorology*, vol. 101, no. 3, pp. 329–358, 2001.

[12] M. Georgescu, M. Moustaoui, A. Mahalov, and J. Dudhia, "Summertime climate impacts of projected megapolitan expansion in arizona," *Nature Climate Change*, vol. 3, no. 1, pp. 37–41, 2013.

[13] C. Ru, S.-B. Duan, X.-G. Jiang, Z.-L. Li, Y. Jiang, H. Ren, P. Leng, and M. Gao, "Land surface temperature retrieval from Landsat 8 thermal infrared data over urban areas considering geometry effect: Method and application," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[14] K. Zakšek and K. Oštir, "Downscaling land surface temperature for urban heat island diurnal cycle analysis," *Remote Sensing of Environment*, vol. 117, pp. 114–124, 2012.

[15] W. P. Kustas, J. M. Norman, M. C. Anderson, and A. N. French, "Estimating subpixel surface temperatures and energy fluxes from the vegetation index–radiometric temperature relationship," *Remote sensing of environment*, vol. 85, no. 4, pp. 429–440, 2003.

[16] S. Bonafoni, "Downscaling of Landsat and MODIS land surface temperature over the heterogeneous urban area of milan," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 5, pp. 2019–2027, 2016.

[17] I. Keramitsoglou, C. T. Kiranoudis, and Q. Weng, "Downscaling geostationary land surface temperature imagery for urban analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1253–1257, 2013.

[18] H. J. Fowler, S. Blenkinsop, and C. Tebaldi, "Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 27, no. 12, pp. 1547–1578, 2007.

[19] J. Tang, X. Niu, S. Wang, H. Gao, X. Wang, and J. Wu, "Statistical downscaling and dynamical downscaling of regional climate in china: Present climate evaluations and future climate projections," *Journal of Geophysical Research: Atmospheres*, vol. 121, no. 5, pp. 2110–2129, 2016.

[20] S. Spak, T. Holloway, B. Lynn, and R. Goldberg, "A comparison of statistical and dynamical downscaling for surface temperature in north america," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D8, 2007.

[21] W. Li, L. Ni, Z.-L. Li, S.-B. Duan, and H. Wu, "Evaluation of machine learning algorithms in spatial downscaling of modis land surface temperature," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2299–2307, 2019.

[22] A. Azhdari, A. Soltani, and M. Alidadi, "Urban morphology and landscape structure effect on land surface temperature: Evidence from shiraz, a semi-arid city," *Sustainable cities and society*, vol. 41, pp. 853–864, 2018.

[23] L. Chen, B. Fang, L. Zhao, Y. Zang, W. Liu, Y. Chen, C. Wang, and J. Li, "Deepurbandownscale: A physics informed deep learning framework for high-resolution urban surface temperature estimation via 3d point clouds," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102650, 2022.

[24] X. Sun, D. Yin, F. Qin, H. Yu, W. Lu, F. Yao, Q. He, X. Huang, Z. Yan, P. Wang *et al.*, "Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery," *Nature Communications*, vol. 14, no. 1, p. 1444, 2023.

[25] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[26] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

[27] Z. Chen, L. Deng, J. Gou, C. Wang, J. Li, and D. Li, "Building and road detection from remote sensing images based on weights adaptive multi-teacher collaborative distillation using a fused knowledge," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103522, 2023.

[28] D. Wu, W. Liu, B. Fang, L. Chen, Y. Zang, L. Zhao, S. Wang, C. Wang, J. Marcato, and J. Li, "Intracity temperature estimation by physics informed neural network using modeled forcing meteorology and multispectral satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[29] G. He, Z. Zhang, W. Jiao, T. Long, Y. Peng, G. Wang, R. Yin, W. Wang, X. Zhang, H. Liu *et al.*, "Generation of ready to use (RTU) products over china based on Landsat series data," *Big Earth Data*, vol. 2, no. 1, pp. 56–64, 2018.

[30] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[31] B. Felbo, A. Mislove, A. Sgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[32] S. L. Ermida, P. Soares, V. Mantas, F.-M. Göttsche, and I. F. Trigo, "Google earth engine open-source code for land surface temperature estimation from the landsat series," *Remote Sensing*, vol. 12, no. 9, p. 1471, 2020.

[33] Z.-L. Li, B.-H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo, and J. A. Sobrino, "Satellite-derived land surface temperature: Current status and perspectives," *Remote sensing of environment*, vol. 131, pp. 14–37, 2013.

[34] R. Gelaro, W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle *et al.*, "The modern-era retrospective analysis for research and applications, version 2 (merra-2)," *Journal of climate*, vol. 30, no. 14, pp. 5419–5454, 2017.

[35] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[37] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, p. 187–199, Apr 2021. [Online]. Available: http://dx.doi.org/10.1007/s41095-021-0229-5

[38] Y. Shi, S. Liu, W. Yan, S. Zhao, Y. Ning, X. Peng, W. Chen, L. Chen, X. Hu, B. Fu *et al.*, "Influence of landscape features on urban land surface temperature: Scale and neighborhood effects," *Science of the Total Environment*, vol. 771, p. 145381, 2021.

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

[40] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," *arXiv preprint arXiv:1710.11431*, 2017.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[44] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[45] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[47] X. Yan, "Pointnet/pointnet++ pytorch," *https://github.com/yanx27/Pointnet*$_Pointnet2_pytorch$, 2019.

[48] K. Sharma and M. Sharma, "Image fusion based on image decomposition using self-fractional fourier functions," *Signal, image and video processing*, vol. 8, no. 7, pp. 1335–1344, 2014.

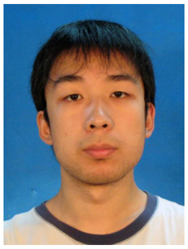
[49] A. Averbuch, D. Lazar, and M. Israeli, "Image compression using wavelet transform and multiresolution decomposition," *IEEE Transactions on Image Processing*, vol. 5, no. 1, pp. 4–15, 1996.

[50] N. E. Huang, *Hilbert-Huang transform and its applications*. World Scientific, 2014, vol. 16.

GuanJie Huang received the B.S. degree from Fuzhou University, China in 2022. He is currently pursuing the M.S. degree with the School of Informatics, Xiamen University, Xiamen, China. His research interests include remote sensing image processing and deep learning.

KaiQun He received the B.S. degree from QingDao University, China in 2021. He is currently pursuing the degree with the School of Informatics, XiaMen University, Xiamen, China. His research interests include computer vision and point cloud analysis.

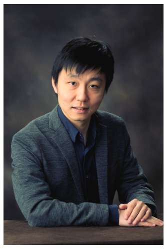
Xiaoyue Lyu received M.Eng degree from the University of Toronto, Toronto, Ontario, Canada, in 2020. From 2021 to 2022, she worked as a Machine Learning Engineer with the AI group at Luokung Technology Corporation, Beijing, China, focusing on designing intelligent transportation systems. She is currently a Research Associate in the Geospatial Intelligence and Mapping Lab at the University of Waterloo. Her research interests include point clouds, 3D scene reconstruction, smart transportation systems, deep learning for image processing, and reinforcement learning.



Yu Zang is currently a Research Associate professor at the School of Informatics, Xiamen University, China. He received his B.S. and Ph.D. degree in Xi'an Jiaotong University in 2008 and 2014. His main researches include remote sensing image processing, computer vision&graphics and mobile LiDAR data analysis.

Hang Shu received the B.A. degree from Wuhan University of Technology ,China in 2022. He is currently pursuing the M.S. degree with the School of Ocean and Earth at Xiamen University in Xiamen, China. His research interests include remote sensing image processing and deep learning.

He is currently a Postdoc with the Information and Communication Engineering Postdoctoral Research Station, and the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China. His current research interests include remote sensing, computer vision, machine learning, mobile laser scanning point cloud data processing, and augmented reality.



Lei Zhao is currently an Assistant Professor in the Department of Civil and Environmental Engineering and Assistant Professor affiliated with the National Center of Supercomputing Applications at the University of Illinois at Urbana-Champaign. He received his B.S. degree in Atmospheric Physics from Nanjing University in 2009, and Ph.D. degree from Yale University in 2015. Before joining the University of Illinois at Urbana-Champaign, Lei Zhao finished his postdoctoral training at Princeton University. His research interests include multi-scale climate modeling, remote sensing, urban climate, environmental fluid mechanics ann turbulence, machine learning and statistical modeling. He has published more than 18 peer-reviewed papers in worldwide top-ranked journals including Nature, Nature Climate Change, Nature Geoscience, Nature Communications as first author and/or corresponding author.

Jonathan Li received the Ph.D. degree in geomatics ngineering from the University of CapeTown, South Africa, in 2000. He is currently a Professor of geomatics and systems design engineer in gat the University of Waterloo, Canada. He is also Founding Member of the Waterloo Artificial Intelligence Institute. His research interests in clude AI-based informatione xtraction from mobile LiDAR point clouds and Earth observation images. He has coauthored more than 450 publications, more than 260 of which were published in refereed journals, including IEEE Transactionson Geoscience and Remote Sensing, IEEE Transactionson Intelligent Transportation Systems, ISPRS Journal of Phorogrammetry and Remote Sening, and Remote Sensing of Environment. He is currently the Editor in Chief of the International Journal of Applied Earth Observation and Geoinformation, Associate Editor of the IEEE Transactionson Intelligent Transportation Systems, IEEE Transactions on Geoscience and Remote Sensing, and Canadian Journal of Remote Sensing. He was a recipient of the ISPRS Samuel Gamble Award in 2020

Cheng Wang (M'07-SM'16) received the Ph.D. degree in signal and information processing from the National University of Defense Technology, Changsha, China, in 2002.

He is currently a Professor with the School of Informatics, and the Executive Director with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China. He has coauthored more than 150 papers in referred journals and top conferences including IEEE Transactions on Geoscience and Remote Sensing, PR, IEEE Transactions on Intelligent Transportation Systems, IEEE Conference on Computer Vision and Pattern Recognition, Association for the Advancement of Artificial Intelligence (AAAI), and International Society for Photogrammetry and Remote Sensing (ISPRS) Journal of Photogrammetry and Remote Sensing. His current research interests include point cloud analysis, multisensor fusion, mobile mapping, and geospatial big data.

Prof. Wang is a Fellow of the Institution of Engineering and Technology. He is also the Chair of the Working Group I/6 on Multi-Sensor Integration and Fusion of the International Society of Remote Sensing.