Measuring Fine-Grained Urban Air Temperature with Satellite Imagery

Minhyuk Song[1], Sungwon Han[1], Seungeon Lee[1], Donhyun Ahn[2], Jihee Kim[1], Meeyoung Cha[2,1]

[1]Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
[2]Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany

Abstract

Recent studies on the urban heat island phenomenon reveal how rapid urbanization intensifies temperature disparities in urban cores, highlighting the need for sustainable urban planning solutions. Analyzing the problems caused by these effects requires high-resolution climate data; however, physical weather stations often lack sufficient regional coverage and resolution. Proposals for alternative methods have attempted to bridge this gap, but they fall short in capturing regional characteristics adequately or necessitate obtaining difficult-to-get input data. This research proposes to use satellite data, where the visual spectrum provides rich information about the degree of human development and is easy to obtain, to measure urban air temperature. Our model, UrbanHeat, uses multi-resolution satellite imagery and employs land surface temperature and global climate data as proxy labels to predict air temperature at a granular scale. The results show that the model provides predictions at a much finer scale while showing superior performance in measuring ordinal relationships between points by capturing both local and broad land cover details of the region. Our case studies demonstrate how predictions at high resolution can help protect vulnerable populations from extreme heat (e.g., elders or developing countries) and contribute to sustainable urban development worldwide.

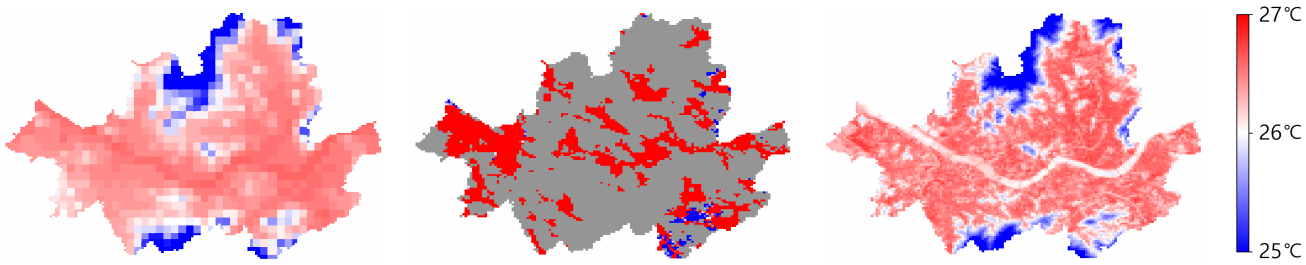
Introduction

With rapid urbanization, more cities worldwide are filled with complex human-made structures like high-rise buildings and roads that can change airflow dynamics and environmental impact within urban cores. The term *urban microclimate* is being used to investigate and address the diverse effects resulting from the growing complexity of urban climates. In addition to the heat generated by human activities and built environments, urban areas retain heat for substantially longer periods than natural areas, leading to the urban heat island (UHI) phenomenon (Rizwan, Dennis, and Chunho 2008). Abnormally high temperatures due to the UHI effect contribute to burdening public health systems and increasing energy consumption in the urban core. These issues significantly threaten sustainable urban growth, particularly in developing countries, which is highlighted in the United Nations' Sustainable Development Goal (SDG) #11.

However, it is difficult to obtain air temperature data at a high spatial resolution that is needed to accurately understand urban microclimates and the impacts of UHI effects. The traditional approach to gathering weather data involves the use of automatic weather stations (AWS). However, weather stations often lack comprehensive regional coverage (Menne et al. 2012; Muller et al. 2013), and their spatial resolution cannot be used for fine-grained measurements. This limitation is further exacerbated by factors such as funding constraints and geographical accessibility, making it even more challenging to obtain high-quality measurements in developing countries. Moreover, although land surface temperature (LST) data from infrared channel of satellite imagery provides high-resolution insights into extreme heat emissions from industrial activities (Portela et al. 2020), its scale often differs significantly from air temperature due to complex air dynamics influenced by human-made structures and traffic. Therefore, LST data requires adjustment or supplementation to accurately study urban heat conditions.

At the macro level, global climate datasets (GCD) have been constructed to approximate air temperature at an arbitrary point, either by combining observational weather data with climate models or by simulating global climate by physical processes of the atmosphere and ocean. However, the resolution of GCD is often low: commonly used global climate data (Hersbach et al. 2020; Eyring et al. 2016) are at resolutions lower than 0.25° (i.e., 27.75km at the equator), which are insufficient to capture regional characteristics essential for policy-making. Moreover, while there is extensive literature on producing high-resolution climate datasets, a field often referred to as *climate downscaling*, these approaches are less suitable for studying urban microclimates because they do not fully incorporate the human-made changes prevalent in urban regions.

This work introduces UrbanHeat, a framework that predicts air temperature at a fine-grained resolution, enabling urban heat measurement at the scale of human activities and artifacts like buildings. By combining daytime satellite imagery at multiple resolutions with LST data, our approach offers four key advantages: capturing rich visual information about land cover and human activities, analyzing interactions between core and surrounding areas of regions by utilizing multiple-resolution images, achieving high-resolution predictions with the aid of LST, and provid-



(a) GCD (CHELSA) (b) LST (c) UrbanHeat

Figure 1: Visualization of the air temperature predictions for Seoul by the UrbanHeat compared to other baseline models. Our method can provide predictions at a much more fine-grained scale by considering the landscape of the region.

ing a cost-effective solution with extensive global coverage that does not require active measurement efforts.

Using multi-resolution satellite imagery and topological data (i.e., latitude and elevation) as inputs, UrbanHeat employs two visual encoders and cross-attention modules to uncover the relationships between the inputs. We used two proxy labels, LST and GCD, to train the model to overcome the lack of real air temperature measurements. The model learns to rank the points based on the LST data while keeping its prediction scale consistent with the coarse-grained GCD value. We confirm model predictions have high correlations with the ground-truth air temperature measured at real-life weather stations, while further extrapolating the data to fine-grained levels (i.e., 0.25km×0.25km). The heatmap in Figure 1 shows that the model enables predictions at a granular scale. For instance, the model accurately captures temperature differences near the Han river in Seoul, whereas previous methods often overestimated temperatures by focusing solely on elevation data. This shows that UrbanHeat effectively incorporates the region's landscape in temperature prediction.

Lastly, we offer case studies where we apply UrbanHeat in Dhaka, Bangladesh, and Seoul, South Korea for spatial and temporal assessments of urban heat. Specifically, we show how UrbanHeat can be utilized to assess spatial disparities and track time changes in urban heat at a fine-grained level, uncovering unequal temperature rises within a city. When coupled with other socio-economic data, this fine-grained measurement can help real-world policymaking and urban planning by identifying vulnerable populations exposed to extreme heat. Accurate measurement of urban heat is particularly relevant for the future as more people migrate to cities, with an estimated 68% of the global population expected to live in urban regions by 2050.[1] Rapid urbanization in developed and developing nations suggests that many cities will struggle to accommodate the increasing population while maintaining the same standard of well-being for the vulnerable and less equipped. In this respect, UrbanHeat can aid in tailored policy-making and urban planning to prepare for changes in urban microclimates, enhancing the quality of life in the future. Our code and appendix are available at https://github.com/archive-cs-minhyuk/UrbanHeat.

[1]UN Report 2018, https://tinyurl.com/y4s8vbek

Related Work

Climate Downscaling There have been many approaches to downscaling the global climate data, including regional climate models (RCM) and statistical downscaling methods (SDM). These two methods differ in that the RCM tries to make a prediction based on physical principles, while the SDM utilizes statistical relationships between various local climate variables. SDMs are recognized for their relatively low computational complexity and flexible applicability to GCD (Sunyer, Madsen, and Ang 2012). SDMs now commonly employ deep learning; for instance, the super-resolution (SR) technique in computer vision is used to downscale climate variables such as surface temperature or seasonal precipitation (Park et al. 2022; Baño-Medina et al. 2022). Others have developed new augmentation strategies for climate downscaling to enhance the performance of deep learning-based models (Yoo, Ahn, and Sohn 2020). These methods have successfully achieved downscaling on a broader scale. The current study complements these works by focusing on local and granular analysis.

Measuring UHI Effect Station-based air temperature measurements are often insufficient to capture the complex urban temperature systems, thereby introducing the need to employ other metrics. For example, LST has been utilized as a proxy for air temperature to measure the UHI intensity (Estoque et al. 2020; Wei et al. 2023). For the cities that have a sufficient number of weather stations, supervised approaches are used to predict temperature over various environments such as population, traffic data (Oukawa, Krecl, and Targino 2022), built-up area percentage, and wind speed (Yoo et al. 2023). They provide useful insights related to UHI effect measurement, but they primarily rely on dynamic climatic inputs, which require data collection over time. This requirement significantly limits the scalability of these approaches in terms of time and cost.

Satellite Imagery Deep learning approaches have utilized satellite imagery to address various social issues. For example, there are studies that estimate socioeconomic variables (Jean et al. 2016; Han et al. 2020a), or detect natural disasters (Amit and Aoki 2017). These studies suggest that satellite images contain rich information about human life, which can be extracted by computer vision models to provide meaningful insights into social patterns.

Method

Problem Definition

UrbanHeat predicts fine-grained urban air temperatures by using LST and GCD values as proxy labels, without requiring station-based ground-truth labels. Consider a set of points $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ within a city with a distance of approximately 0.25km. For each point c_i, we collected high-resolution images $\mathbf{x}_i^{(h)}$ with resolution of 0.25km (same with area of c_i) and low-resolution images $\mathbf{x}_i^{(l)}$ with resolution of 0.5km. Latitude (a_i) and elevation (e_i) information were collected based on the center of each point c_i. Additionally, LST values and GCD predictions were collected as heat-related proxy labels, denoted respectively as $\tilde{y}_i^{(LST)}$ and $\tilde{y}_i^{(GCD)}$. Our goal is to train the model f that predicts the air temperature $\hat{y}_i$ at each point c_i by using the above data (i.e., $\hat{y}_i = f(\mathbf{x}_i^{(h)},\ \mathbf{x}_i^{(l)},\ a_i,\ e_i)$), where the prediction $\hat{y}_i$ is desired to be close to real temperature y_i.

Framework Overview

We propose UrbanHeat, a framework for predicting fine-grained urban air temperatures. UrbanHeat utilizes multi-resolution satellite images to capture both local and broad landscapes of the region and employs two different knowledge sources for effective optimization. The proposed model includes two distinct encoders and cross-attention modules to extract a comprehensive embedding that incorporates the interactions between the core and surrounding areas using satellite images at two different resolutions. This embedding is combined with basic topological information (i.e., latitude, elevation) and used to predict air temperature. Then, using multi-task learning, we optimize the model using information from GCD and LST. We use regression for GCD to adjust the absolute scale of the model prediction, and employ a ranking-based training objective for LST to learn the ordinal relationship of temperature between regions. Figure 2 visualizes the overall framework.

Model Architecture

With input variables of $\mathbf{x}_i^{(h)}, \mathbf{x}_i^{(l)}, a_i, e_i$ for each point c_i, we design a structure to capture relevant information affecting air temperature and to identify the relationships among these variables. By utilizing encoders $g^{(h)}$ and $g^{(l)}$, the module first maps satellite images $\mathbf{x}_i^{(h)}$ and $\mathbf{x}_i^{(l)}$ to latent representations $\mathbf{z}_i^{(h)}$ and $\mathbf{z}_i^{(l)}$ with the shape $\mathbb{R}^{(W\times H)\times D}$, respectively.

$$\mathbf{z}_i^{(h)} = g^{(h)}(\mathbf{x}_i^{(h)}),\ \ \mathbf{z}_i^{(l)} = g^{(l)}(\mathbf{x}_i^{(l)}), \tag{1}$$

where W, H and D refer to the width, height, and hidden dimension of the representations.

To enhance the interactions between core and surrounding information presented in multi-resolution satellite imagery, we incorporate these considerations into the architecture design, rather than simply utilizing the concatenation of two image embeddings. Consequently, we introduce a multi-view cross-attention module to efficiently highlight the critical information depicted in both satellite images. Each cross-attention module treats one image as the query image and the other as the value image. In the query image, max-pooling is used to extract a representative understanding of the image, and a linear projection $P_{\mathbf{q}}$ is applied to create the query vector $\mathbf{q} \in \mathbb{R}^D$. In the value image, linear projections $P_{\mathbf{K}}$ and $P_{\mathbf{V}}$ are applied to the latent representation $\mathbf{z}$, generating the key matrix $\mathbf{K} \in \mathbb{R}^{(W\times H)\times D}$ and the value matrix $\mathbf{V} \in \mathbb{R}^{(W\times H)\times D}$, respectively. For instance, when the query image is $\mathbf{x}_i^{(h)}$ and the value image is $\mathbf{x}_i^{(l)}$ (i.e, $h{\rightarrow}l$), these concepts are defined as follows.

$$\begin{aligned}
\mathbf{q}_{h\rightarrow l} &= P_{\mathbf{q}}(\textit{maxpool}(\mathbf{z}_i^{(h)}) \\
\mathbf{K}_{h\rightarrow l} &= P_{\mathbf{K}}(\mathbf{z}_i^{(l)}) \\
\mathbf{V}_{h\rightarrow l} &= P_{\mathbf{V}}(\mathbf{z}_i^{(l)})
\end{aligned} \tag{2}$$

For the case where $\mathbf{x}_i^{(l)}$ is the query image and $\mathbf{x}_i^{(h)}$ is the value image, the process can be reversed and executed in parallel using separate linear projection matrices. With these values, we apply scaled dot-product attention (Vaswani et al. 2017) which are defined as the following equation.

$$\textit{Att}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \textit{softmax}(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{3}$$

To better utilize the information contained in the original image, we added the original image embedding to the applied attention values, finalizing the embedding for each image.

$$\begin{aligned}
\mathbf{r}_i^{(l)} &= \textit{Att}(\mathbf{q}_{h\rightarrow l}, \mathbf{K}_{h\rightarrow l}, \mathbf{V}_{h\rightarrow l}) + avgpool(\mathbf{z}_i^{(l)}) \\
\mathbf{r}_i^{(h)} &= \textit{Att}(\mathbf{q}_{l\rightarrow h}, \mathbf{K}_{l\rightarrow h}, \mathbf{V}_{l\rightarrow h}) + avgpool(\mathbf{z}_i^{(h)})
\end{aligned} \tag{4}$$

Finally, we concatenate $\mathbf{r}_i^{(l)} \in \mathbb{R}^D$ and $\mathbf{r}_i^{(h)} \in \mathbb{R}^D$ with latitude a_i and elevation e_i, and then use MLP layers to predict the air temperature $\hat{y}_i$.

$$\hat{y}_i = \text{MLP}(Concat(\mathbf{r}_i^{(h)}, \mathbf{r}_i^{(l)}, a_i, e_i)) \tag{5}$$

Learning with Proxy Labels

Since the spatial resolution of weather stations is often too sparse, many regions are not covered by stations, making it challenging to determine temperatures in these uncovered areas. We hence consider two proxy labels from GCD and LST values for multi-task learning. Both data are accessible globally, enabling these steps applicable anywhere on Earth.

Global Climate Dataset GCD predictions are likely to have a similar scale to that of real-world air temperatures (Mistry et al. 2022). Therefore, we utilize a *GCD-based scaling loss*, which calculates the Mean Squared Error (MSE) between our model's prediction and the GCD value, $\tilde{y}_i^{(GCD)}$, to match the prediction scale. With a training batch $\mathcal{B}_{GCD}$, the equation for calculating L_{GCD} is written as the following equation:

$$L_{GCD} = \frac{1}{|\mathcal{B}_{GCD}|} \sum_{i\in\mathcal{B}_{GCD}} ||\hat{y}_i - \tilde{y}_i^{(GCD)}||^2 \tag{6}$$

However, typical GCD methods fail to consider regional anthropogenic heat sources crucial for understanding complex

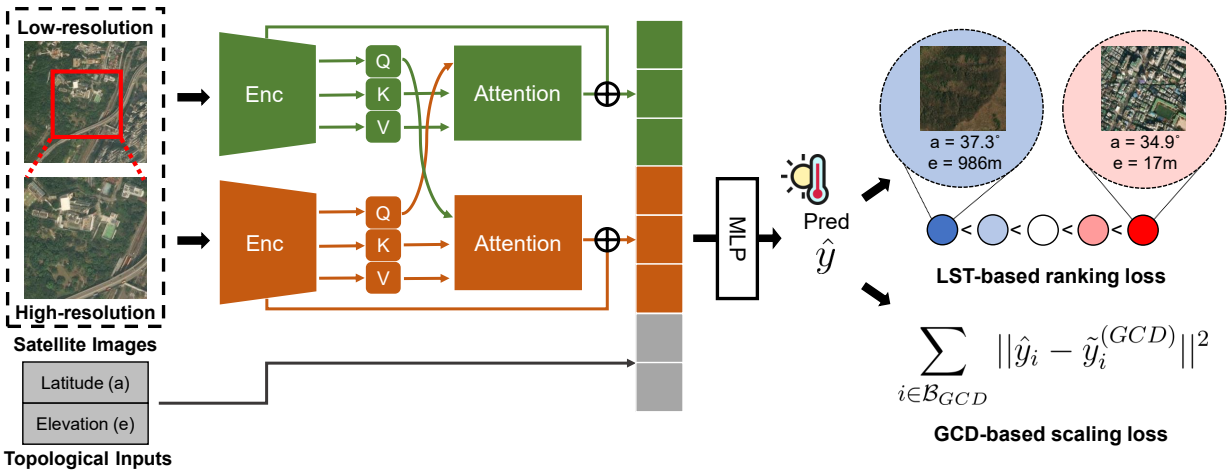


Figure 2: Overview of the proposed framework. With multi-resolution satellite images and topological information (i.e., latitude, elevation) as an input, the model predicts fine-grained urban air temperature $\hat{y}$, by passing a visual encoder, a cross-attention module, and an MLP layer. We train our model using two loss functions, by using proxy labels from LST and GCD.

urban thermosystems and provide insufficient resolution, thereby introducing errors in predictions at a granular level. Consequently, training a model solely with GCD labels inherits these shortcomings. To address this issue, it is necessary to incorporate an additional information source that can estimate temperature variations within granular areas.

Land Surface Temperature LST data are a relevant alternative to overcome the limitations inherent in GCD values. LST data, which are derived from a satellite's thermal band, reflect the temporal characteristics of regions with a relatively high spatial resolution (e.g., 0.1 km). Using LST data as a proxy label enables the model to learn the relationships between regional heat levels and the landscape information captured in optical satellite images. However, using raw LST values as ground-truth labels for model training also poses challenges. Firstly, while LST data effectively captures the ordinal temperature relationship between regions, its absolute scale significantly differs from real air temperature. Secondly, raw LST values are often noisy due to temporal cloud conditions. To address these issues, we do not use absolute LST values for training; instead, we group points based on LST values, allowing the model to learn the ordinal information among groups by *LST-based ranking loss*. In addition, we employ pruning based on the yearly record of LST values, which are more noise-tolerant.

We aggregate data points by the LST values of the target period into n groups in ascending order, where the set of groups is defined as $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$. All groups contain the same number of points. Then, we again group the points using yearly average LST values into the set of groups, denoted as $\hat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_n\}$. We denote the group index of the point c_i within the set of group $\mathcal{G}$ as $\mathcal{I}_{\mathcal{G}}(c_i)$. For each point c_i, if the group index difference between $\mathcal{G}$ and $\hat{\mathcal{G}}$ is larger than 1 (i.e., $\left|\mathcal{I}_{\mathcal{G}}(c_i) - \mathcal{I}_{\hat{\mathcal{G}}}(c_i)\right| > 1$), we considered this as a noisy LST value and pruned the point from the training dataset. This prevents points with erroneous satellite data, such as those obscured by clouds, from being included in the dataset. Here, we set the threshold to 1 to account for the seasonal variation in LST.

By using grouped LST labels, we ensure that the model's predictions align with the ordinal relationships in the LST data. We employ Spearman correlation as the loss function, which measures the degree of agreement between two ranked variables. In each training step, the given batch $\mathcal{B}_{LST}$ includes K ordered sets of points, each sampled from the groups $G \in \mathcal{G}$. Specifically, for a k-th ordered set, we sample one point from each group (i.e., $c_1^{(k)} \in G_1$, $c_2^{(k)} \in G_2$, $\ldots$, $c_n^{(k)} \in G_n$). Then we compute the model output of each points, resulting in list of predictions $P^{(k)} = [\hat{y}_1^{(k)}, \hat{y}_2^{(k)}, \ldots, \hat{y}_n^{(k)}]$ and their LST group indices $I^{(k)} = [\mathcal{I}_{\mathcal{G}}(c_1^{(k)}), \mathcal{I}_{\mathcal{G}}(c_2^{(k)}), \ldots, \mathcal{I}_{\mathcal{G}}(c_n^{(k)})]$. UrbanHeat maximizes the Spearman correlation of the following equation:

$$1 - \frac{6||\texttt{rank}(P^{(k)}) - \texttt{rank}(I^{(k)})||^2}{n(n^2-1)} \tag{7}$$

Here, the `rank` function returns the ranking of elements in a list. Because we sample points from each group in $\mathcal{G}$, which is already sorted in ascending order, $\texttt{rank}(I^{(k)})$ is always $[1, 2, \ldots, n]$.

However, since the `rank` function is not differentiable, Eq. 7 is infeasible for a loss function to train our model. Therefore, we adopt a method to approximate the ranking algorithm by using a differentiable sorter (Engilberge et al. 2019; Han et al. 2020b). We first define a function σ that compares two scalar values a and b:

$$\sigma(a, b) = \frac{1}{1 + e^{-\lambda(b-a)}} \tag{8}$$

The output of the function would be close to 0 (or 1) when a is larger (or smaller) than b. This behavior allows σ to approximate the relative rank between two input scalars, where the hyperparameter λ determines its sensitivity. Utilizing the

function σ, we define $R_{P^{(k)}}(\hat{y}_i^{(k)})$ as the differentiable ranking of $\hat{y}_i^{(k)}$ within $P^{(k)}$, calculated by summing the results of its comparisons with all other elements in $P^{(k)}$:

$$R_{P^{(k)}}(\hat{y}_i^{(k)}) = \sum_{j=1;\, j\neq i}^{n} \sigma(\hat{y}_i^{(k)}, \hat{y}_j^{(k)}). \tag{9}$$

We then define $\texttt{rank}_d$ as the function that returns the differentiable ranking of elements in a given list (e.g., $\texttt{rank}_d(P^{(k)}) = [R_{P^{(k)}}(\hat{y}_1), R_{P^{(k)}}(\hat{y}_2), \ldots, R_{P^{(k)}}(\hat{y}_n)]$). Here, we denote $\hat{y}_i^{(k)}$ as $\hat{y}_i$ for simplicity. Utilizing function $\texttt{rank}_d$, we formulate the loss objective to maximize the spearman correlation in Eq. 7 as following:

$$\begin{aligned} L_{LST} &= \frac{1}{|\mathcal{B}_{LST}|} \sum_{k\in\mathcal{B}_{LST}} ||\texttt{rank}_d(P^{(k)}) - \texttt{rank}(I^{(k)})||^2 \\ &= \frac{1}{|\mathcal{B}_{LST}|} \sum_{k\in\mathcal{B}_{LST}} \sum_{i=1}^{n} ||R_{P^{(k)}}(\hat{y}_i^{(k)}) - i||^2. \end{aligned} \tag{10}$$

Here, $|\mathcal{B}_{LST}| = K \times n$.

Final loss objective With the adjusting parameter α, UrbanHeat is trained to optimize the following loss objective:

$$L = L_{LST} + \alpha \times L_{GCD} \tag{11}$$

Experiments

Datasets

We collected satellite imagery corresponding to the longitudes and latitudes of Hong Kong and Seoul from World Imagery data of Esri between 2018 and 2019. Elevation data was gathered from the SRTM 1 Arc-Second Global dataset, using the EarthExplorer platform. We used the Google Earth Engine platform to collect the LST data using the Landsat 8 Collection 2 bands, which have a resolution of 0.1 km for the thermal band. For GCD, we used the CHELSA (Karger et al. 2017) dataset. For evaluation, we obtained real air temperature measurements from 37 AWS in Hong Kong through the Hong Kong CSDI Portal[2], and data from 62 stations in Seoul from the Korea Meteorological Administration's official website[3]. All temperature-related variables were collected for the months of July and August, the two hottest months in both cities, during the year 2019.

Experimental Setup

We used ImageNet-pretrained Resnet-18 (He et al. 2016) backbones for encoder $g^{(h)}$ and $g^{(l)}$, excluding the last average pooling layer. Latitude and elevation data are used with min-max normalization. The number of LST groups n was fixed to 5. The hyperparameter λ in Eq. 8 was fixed to 30, and α in Eq. 11 to 100.

Performance Evaluation

We evaluate the relationship between model predictions and real air temperatures using Spearman (ρ_s) and Pearson (ρ_p) correlations. We also provide the mean absolute error (MAE) between two values to measure the difference in absolute scale. Experiments were repeated three times with different seeds, and the average results were reported.

	Method	Hong Kong			Seoul		
		ρ_s	ρ_p	MAE	ρ_s	ρ_p	MAE
A	ERA-5	0.080	0.002	0.742	0.011	0.092	1.704
	SRGAN	0.404	0.925	0.534	0.317	0.367	0.745
	Top. Reg	0.476	0.941	<u>0.478</u>	0.449	0.495	0.580
	ResNet	0.407	0.924	0.511	0.439	0.506	0.574
	Tile2Vec	0.468	0.820	0.581	0.417	0.498	0.568
	LST	0.246	0.248	10.179	0.046	-0.008	10.223
	CHELSA	0.510	0.941	0.525	0.398	0.535	0.560
	Proxy Ens.	0.533	0.941	0.513	0.387	0.534	0.651
B	**UrbanHeat**	**0.625**	<u>0.958</u>	**0.469**	**0.533**	**0.579**	**0.512**
	w/o L_{LST}	0.518	0.949	0.493	0.438	0.533	0.596
	w/o L_{GCD}	0.492	0.650	3.707	0.472	0.525	2.004
	w/o cross-att	<u>0.607</u>	0.956	0.492	0.472	0.539	<u>0.541</u>
	One img: $\mathbf{x}^{(h)}$	0.603	**0.959**	0.503	<u>0.485</u>	<u>0.556</u>	0.544
	One img: $\mathbf{x}^{(l)}$	0.572	0.958	0.509	0.477	0.554	0.550

A : Baselines, B : Ours with ablation

Table 1: Comparison of performance evaluated on the Hong Kong and Seoul summer temperature data. We marked the best results in a bold text, and underlined the second-best.

Total eight baselines are compared (division A in Table 1): (1) ERA-5 is the most recent generation of ECMWF reanalysis for global climate and weather, with a resolution of $0.25°\times0.25°$. (2) SRGAN is a deep learning method for image super-resolution, which we used to downscale the CHELSA value to 4x the original scale. We used the DIV2K dataset pretrained model due to the small set of training samples resulting from the limited study size. (3) Top. Reg utilize topological inputs (i.e., latitude and elevation) and employ a Linear regressor by using CHELSA data as a training label. (4) ResNet and (5) Tile2Vec (Jean et al. 2019) additionally utilize satellite image embeddings with an ImageNet-pretrained and a Tile2Vec-pretrained ResNet-18 encoder, respectively. (6) LST refers to the land surface temperature value of the given point. Missing values were imputed with the average LST value of the points within the training set. (7) CHELSA refers to the value presented in the CHELSA dataset for the given point. (8) Proxy Ens. is a weighted average of the two proxy labels which were calculated as $\text{LST}\times\boldsymbol{\gamma}+\ \text{CHELSA}\times(\mathbf{1}-\boldsymbol{\gamma})$. Both values were min-max normalized before addition, and the γ value was grid-searched between 0.1 and 0.9. Top results are reported.

ERA-5 exhibited the lowest correlation, mainly due to its limited resolution. Applying SRGAN led to a performance drop compared to the raw CHELSA results, indicating that simply using SR models from computer vision domain is not a viable solution. Utilizing only topological information (e.g., Top. Reg) proved ineffective for producing finer-scale results without an understanding of the complex dynamics driven by human activities. Additionally utilizing embeddings from satellite images with conventional methodologies also proved to be ineffective. LST performed

[2]https://portal.csdi.gov.hk/csdi-webpage/

[3]https://data.kma.go.kr/cmmn/main.do

(a) Pred = 26.70°C (b) Pred = 26.56°C (c) Pred = 26.20°C

Figure 3: Qualitative analysis for changing input images. Both changes in high-resolution and low-resolution image result in a change in prediction, though to different extents

poorly due to the noise and presence of missing values in the raw LST data. CHELSA exhibits some deficiencies, as the model lacks detailed regional information. The results on the ensemble of LST and CHELSA verifies that each information complement each other, but simple weighted average could not fully utilize the benefit. UrbanHeat demonstrated the best result for both Hong Kong and Seoul datasets, showcasing the effectiveness of integrating both relative ranking and absolute scale along with satellite images.

Ablation Studies

We conducted an ablation study to confirm the contribution of each model component (division B in Table 1): (1) w/o L_{LST} and (2) w/o L_{GCD} denote the results after removing each loss, respectively. (3) w/o cross-att simply concatenates the two image embeddings from each encoder instead of using a cross-attention module. For (4) One img: $\mathbf{x}^{(h)}$ and (5) One img: $\mathbf{x}^{(l)}$, we used a single encoder and a single resolution satellite image to get the final image embedding.

Findings from the ablation study demonstrate that any modification or removal of individual components leads to decreased performance. Ablations (1) and (2) show that the loss functions are both essential for performance. Furthermore, the improved performance of ablation (2) over the raw LST values demonstrates that the grouping and filtering process for LST brings an additional benefit. Ablations (4) and (5) show that even using a single resolution satellite image can greatly enhance the model's performance, compared to the baseline results. In ablation (3), we can observe that the simple concatenation of these embeddings does not necessarily improve the performance compared to results with a single image, verifying the effectiveness of the cross-attention module.

Qualitative Analysis

Urban forests are known to be effective in mitigating the effects of the UHI by providing cooling effects, which can help decrease urban air temperatures (Livesley, McPherson, and Calfapietra 2016). We analyze whether the model can capture the impact of urban forests by changing the input satellite images. Figure 3 illustrates how the temperature prediction changes in response to modifications in the input images, while other topological inputs remain fixed. Figure 3a depicts the extensive urban region, where both high- and low-resolution satellite images present numerous artificial structures. The temperature prediction decreased when the images of surrounding regions were replaced with forest areas, as shown in Figure 3b, indicating that the construction of urban forests can positively influence the nearby temperatures. Figure 3c illustrates that the urban forest itself exhibits significantly lower heat levels, potentially providing a shelter from urban heat for city residents.

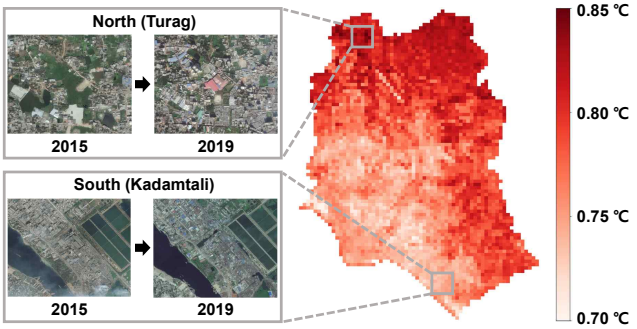


Figure 4: Visualization of the increase in predicted air temperature in Dhaka from 2015 and 2019. Satellite images reveal noticeable development in the northern region with large temperature changes, whereas the southern region with minor temperature changes shows little variation.

Case Study

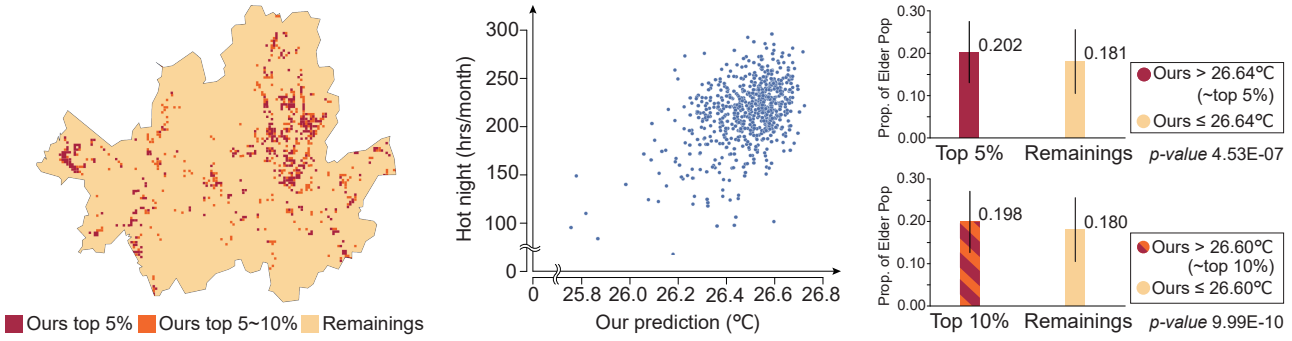
Unequal Temperature Rises in Dhaka, Bangladesh

We first investigate temperature changes within the capital city of Dhaka, one of the areas most vulnerable places to heat-related risks. We want to observe whether an AI model can measure spatial disparities in heat exposure increases within a city over time. Figure 4 shows the differences in predicted air temperatures by UrbanHeat between 2015 and 2019. The predictions were made by training separate models, each using a proxy label corresponding to its respective year. The average MAE evaluated in Dhaka for 2015, was 0.31°C (CHELSA: 0.34°C, LST: 16.80°C), indicating that UrbanHeat offers reliable indicators for further analysis.

We make two observations. First, Dhaka's overall temperature has significantly increased. Second, the temperature rise was not equal within the city. Higher rises are observed in the northern suburban areas, where urban expansion has been greater over the past decades, consistent with previous research (Uddin et al. 2022). This suggests that urban expansion exacerbates temperature rises, so potential risks from the UHI effect should be considered in urban planning and policymaking. Fine-grained temperature predictions provided by UrbanHeat, which uncover the complex and unequal patterns of global warming and human activity over time, can be of great value in such considerations.

Identifying Vulnerable Population in Seoul, Korea

We next apply UrbanHeat to the capital city Seoul to demonstrate how it can identify vulnerable local neighborhoods and populations to heat risks within a city. Figure 5a shows



(a) Heat risk areas in Seoul (b) Correlation with hot night hours (c) t-test for elderly proportion

Figure 5: (a) Heat risk area of the top 5% and 10% in a fine-grained level, defined by UrbanHeat. Each cell measures 0.25km by 0.25km. (b) Correlation between UrbanHeat predictions and hot night hours. (c) Comparison of the proportion of elderly individuals between the heat risk area and the remaining areas.

the regions with the top 5% and 10% of the predicted air temperatures. These precise predictions enable more efficient allocation of resources and protective measures to areas needing additional heat-related support.

Next, Figure 5b depicts the correlation between predictions made by UrbanHeat and the occurrence of extremely hot nights. A hot night hour is defined as an hour going over 25°C between 8PM and 8AM, following the definition of 'a tropical night' of the Korea Meteorological Administration and calculated from real air temperatures in the S-DoT dataset. The *tropical night phenomenon* is one of the noticeable problems caused by the UHI because artificial structures absorb heat, keeping city temperatures uncomfortably high even at night. It causes public health problems and increases energy consumption. This analysis suggests that heat risk areas identified by UrbanHeat are more likely to experience the tropical night phenomenon.

Lastly, Figure 5c shows that heat risk areas (Top 5% and Top 10% in our temperature predictions) are more likely to have a higher composition of the elderly population, who are more vulnerable to extreme heat. We calculated the elderly (aged 65 and above) proportion from the grid-level population data with a resolution of 0.25 km from the National Geographic Information Institute (NGII) of South Korea. We then conducted a t-test to compare the elderly proportions between heat risk areas and other regions. This analysis using measurements of urban microclimate reveals an interesting fact, especially for policymakers and urban planners, that the elderly population is at greater risk compared to the total population, making them more susceptible to heat-related issues such as the tropical night phenomenon. In sum, fine-grained measurements allow for the precise identification of populations exposed to extreme heat, enabling more targeted and effective support.

Conclusion

As more people migrate to cities, securing high-resolution climate data to understand urban microclimate has become essential for enhancing public health and well-being. This paper introduced a novel framework for predicting fine-grained urban air temperatures at neighborhood level (i.e., 0.25km×0.25km). We utilized satellite imagery at multiple resolutions, combined with the cross-attention module, to effectively capture regional characteristics (e.g., high-rise buildlings, river, park) that could influence the urban thermal environment. By employing multi-task learning with GCD and LST as proxy labels, our method tackles the limited availability of station-based real air temperature data through a weakly-supervised approach. Along with the extensive coverage and increasing temporal resolution of satellite imagery, this approach emphasizes its potential for usage in various climate downscaling tasks. In particular, our model employed a cross-attention module to extract meaningful relationships between local and broad land cover details in multi-resolution satellite imagery. Our evaluation featured case studies of the air temperature prediction model, tracking the urban heat island effect in Dhaka, Bangladesh, and Seoul, South Korea. We have made our model details and implementation codes available for research purposes.

Limitations The evaluation of our model was conducted with a limited number of data points due to the scarcity of ground-truth air temperature measurements, a situation that itself motivated this study. Additionally, the proposed model does not directly utilize established physical principles related to weather. Nevertheless, we believe that the model can indirectly learn these principles from the GCD, which already incorporates them. Despite these limitations, this research highlights the potential to monitor rapidly changing urban microclimates, offering benefits to vulnerable populations such as the elderly and people in developing countries.

Acknowledgments

J. Kim and M. Cha are co-corresponding authors. Authors affiliated with KAIST are part of the School of Computing, with J. Kim also affiliated with the College of Business. This work was supported by the National Research Foundation of Korea (NRF) grant (RS-2022-00165347).

References

Amit, S. N. K. B.; and Aoki, Y. 2017. Disaster detection from aerial imagery with convolutional neural network. In *IEEE International Electronics Symposium on Knowledge Creation and Intelligent Computing*, 239–245.

Baño-Medina, J.; Manzanas, R.; Cimadevilla, E.; Fernández, J.; González-Abad, J.; Cofiño, A. S.; and Gutiérrez, J. M. 2022. Downscaling multi-model climate projection ensembles with deep learning (DeepESD): Contribution to CORDEX EUR-44. *Geoscientific Model Development Discussions*, 2022: 1–14.

Engilberge, M.; Chevallier, L.; Pérez, P.; and Cord, M. 2019. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proceedings of the IEEE CVPR*, 10792–10801.

Estoque, R. C.; Ooba, M.; Seposo, X. T.; Togawa, T.; Hijioka, Y.; Takahashi, K.; and Nakamura, S. 2020. Heat health risk assessment in Philippine cities using remotely sensed data and social-ecological indicators. *Nature Communications*, 11(1): 1581.

Eyring, V.; Bony, S.; Meehl, G. A.; Senior, C. A.; Stevens, B.; Stouffer, R. J.; and Taylor, K. E. 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958.

Han, S.; Ahn, D.; Cha, H.; Yang, J.; Park, S.; and Cha, M. 2020a. Lightweight and robust representation of economic scales from satellite imagery. In *Proceedings of the AAAI*, volume 34, 428–436.

Han, S.; Ahn, D.; Park, S.; Yang, J.; Lee, S.; Kim, J.; Yang, H.; Park, S.; and Cha, M. 2020b. Learning to score economic development from satellite imagery. In *Proceedings of the ACM SIGKDD*, 2970–2979.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*, 770–778.

Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3967–3974.

Karger, D. N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R. W.; Zimmermann, N. E.; Linder, H. P.; and Kessler, M. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4(1): 1–20.

Livesley, S.; McPherson, E. G.; and Calfapietra, C. 2016. The urban forest and ecosystem services: impacts on urban water, heat, and pollution cycles at the tree, street, and city scale. *Journal of Environmental Quality*, 45(1): 119–124.

Menne, M. J.; Durre, I.; Vose, R. S.; Gleason, B. E.; and Houston, T. G. 2012. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7): 897–910.

Mistry, M. N.; Schneider, R.; Masselot, P.; Royé, D.; Armstrong, B.; Kyselỳ, J.; Orru, H.; Sera, F.; Tong, S.; Lavigne, É.; et al. 2022. Comparison of weather station and climate reanalysis data for modelling temperature-related mortality. *Scientific Reports*, 12(1): 5178.

Muller, C. L.; Chapman, L.; Grimmond, C.; Young, D. T.; and Cai, X. 2013. Sensors and the city: a review of urban meteorological networks. *Int. J. Climatol*, 33(7): 1585–1600.

Oukawa, G. Y.; Krecl, P.; and Targino, A. C. 2022. Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. *Science of the Total Environment*, 815: 152836.

Park, S.; Singh, K.; Nellikkattil, A.; Zeller, E.; Mai, T. D.; and Cha, M. 2022. Downscaling earth system models with deep learning. In *Proceedings of the ACM SIGKDD*, 3733–3742.

Portela, C. I.; Massi, K. G.; Rodrigues, T.; and Alcântara, E. 2020. Impact of urban and industrial features on land surface temperature: Evidences from satellite thermal indices. *Sustainable Cities and Society*, 56: 102100.

Rizwan, A. M.; Dennis, L. Y.; and Chunho, L. 2008. A review on the generation, determination and mitigation of Urban Heat Island. *Journal of Environmental Sciences*, 20(1): 120–128.

Sunyer, M.; Madsen, H.; and Ang, P. 2012. A comparison of different regional climate models and statistical downscaling methods for extreme rainfall estimation under climate change. *Atmospheric Research*, 103: 119–128.

Uddin, A. S.; Khan, N.; Islam, A. R. M. T.; Kamruzzaman, M.; and Shahid, S. 2022. Changes in urbanization and urban heat island effect in Dhaka city. *Theoretical and Applied Climatology*, 147(3): 891–907.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wei, X.; Guan, F.; Zhang, X.; Van de Weghe, N.; and Huang, H. 2023. Integrating planar and vertical environmental features for modelling land surface temperature based on street view images and land cover data. *Building and Environment*, 235: 110231.

Yoo, C.; Im, J.; Weng, Q.; Cho, D.; Kang, E.; and Shin, Y. 2023. Diurnal Urban Heat Risk Assessment: Using Extreme Air Temperatures and Real-Time Population Data in Seoul. *Iscience*, 108123.

Yoo, J.; Ahn, N.; and Sohn, K.-A. 2020. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE CVPR*, 8375–8384.