

MoCoLSK: Modality Conditioned High-Resolution Downscaling for Land Surface Temperature

Qun Dai, Chunyang Yuan, Yimian Dai, Yuxuan Li, Xiang Li, Kang Ni, Jianhui Xu, Xiangbo Shu, Jian Yang

Abstract—Land Surface Temperature (LST) is a critical parameter for environmental studies, but directly obtaining high spatial resolution LST data remains challenging due to the spatio-temporal trade-off in satellite remote sensing. Guided LST downscaling has emerged as an alternative solution to overcome these limitations, but current methods often neglect spatial non-stationarity, and there is a lack of an open-source ecosystem for deep learning methods. In this paper, we propose the Modality-Conditional Large Selective Kernel (MoCoLSK) Network, a novel architecture that dynamically fuses multi-modal data through modality-conditioned projections. MoCoLSK achieves a confluence of dynamic receptive field adjustment and multi-modal feature fusion, leading to enhanced LST prediction accuracy. Furthermore, we establish the GrokLST project, a comprehensive open-source ecosystem featuring the GrokLST dataset, a high-resolution benchmark, and the GrokLST toolkit, an open-source PyTorch-based toolkit encapsulating MoCoLSK alongside 40+ state-of-the-art approaches. Extensive experimental results validate MoCoLSK’s effectiveness in capturing complex dependencies and subtle variations within multispectral data, outperforming existing methods in LST downscaling. Our code, dataset, and toolkit are available at <https://github.com/GrokCV/GrokLST>.

Index Terms—Land surface temperature, guided image super-resolution, multi-modal fusion, receptive field, benchmark dataset

I. INTRODUCTION

Land Surface Temperature (LST) reflects the complex mass and energy exchanges between the Earth’s surface and the atmosphere [1]. It serves as a critical indicator for evaluating ecological and climatic dynamics across various scales and plays a vital role in environmental studies, including urban heat island analysis [2], forest fire monitoring, land surface evapotranspiration [3], soil moisture inversion [4], and geothermal anomaly detection. However, the inherent limitations of

This work was supported by the National Natural Science Foundation of China (62301261, 62206134, 62101280, 62222207, 62332010, 62427808, U24A20330, 62361166670), China Postdoctoral Science Foundation (2021M701727, 2023M731781), and the GDAS’ Project of Science and Technology Development (2023GDASZH-2023010101). The first two authors contributed equally to this work. (Corresponding author: Yimian Dai, Kang Ni, Jianhui Xu)

Qun Dai and Xiangbo Shu are with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. (e-mail: qun.dai.grokcv@gmail.com; shuxb@njust.edu.cn).

Chunyang Yuan and Kang Ni are with School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. (e-mail: chunyang.yuan.cs@gmail.com; tznikang@163.com).

Yimian Dai, Yuxuan Li, Xiang Li, and Jian Yang are with PCA Lab, VCIP, College of Computer Science, Nankai University. Xiang Li also holds a position at the NKIARI, Shenzhen Futian. (e-mail: yimian.dai@gmail.com; yuxuan.li.17@ucl.ac.uk; xiang.li.implus@nankai.edu.cn; csjyang@nankai.edu.cn)

Jianhui Xu is with Guangdong Provincial Key Laboratory of Utilization of Remote Sensing and Geographical Information System, Guangzhou Institute of Geography, Guangdong Academy of Sciences, Guangzhou, 510070, China. (e-mail: xujianhui306@gdas.ac.cn).

satellite remote sensing hinder the acquisition of high spatial resolution LST data, specifically the unavoidable trade-off between temporal and spatial resolutions [5]. For instance, Landsat 8 offers a spatial resolution of 100 meters but revisits the same region only once every 16 days [6]. In contrast, MODIS provides observations twice a day but at a coarser spatial resolution of 1 kilometer [7]. To address this challenge, one approach is to optimize sensing instruments and enhance satellite data transmission capabilities, but this is costly and time-consuming [8]. A more feasible alternative is to develop LST downscaling models.

Downscaling refers to transforming low-resolution (LR) images into high-resolution (HR) ones to enhance spatial detail information [5]. Over the past two decades, various LST downscaling techniques have emerged, primarily categorized into statistical regression models, machine learning-based models, fusion models, and physical models [5]. Classical linear statistical models, such as Disaggregation of Radiometric Surface Temperature (DisTrad) [9] and Thermal Sharpening (TsHARP) [10], rely on the scale-invariant relationship between the Normalized Difference Vegetation Index (NDVI) and LST, employing global regression for downscaling. However, since LST is influenced by multiple factors such as wind, terrain, and land cover types, using a single biophysical parameter, like NDVI, as a predictor is insufficient [8]. To address this limitation, machine learning-based models, such as Random Forest (RF) [11] and Extreme Gradient Boosting (XGBoost) [12], leverage multiple biophysical parameters to effectively achieve LST downscaling while mitigating the risk of overfitting [8]. However, these models primarily adopt global regression paradigms, which perform well in homogeneous areas but often fall short in highly heterogeneous regions, such as urban environments [8]. Methods like Geographically Weighted Regression (GWR) [13] and Multiscale Geographically Weighted Regression (MGWR) [14] effectively address the spatial heterogeneity of LST. Additionally, Geographically and Temporally Weighted Regression (GTWR) [15] models the spatiotemporal non-stationarity between LST and environmental factors in time-series datasets. On the physical modeling front, the DTsEB method [16], based on the Surface Energy Balance (SEB), explains the interactions between biophysical parameters and LST from a physical mechanism perspective.

Recently, deep learning has catalyzed a paradigm shift in computer vision and remote sensing, also markedly affecting LST downscaling [17–21]. These models leverage the ability of deep neural networks to learn complex spatial and temporal patterns from data, enabling them to effectively capture the relationships between LR and HR data.

In the context of HR climate data generation, Wang *et al.* pioneered the Super Resolution Deep Residual Network (SRDRN) to refine the downscaling of daily meteorological parameters like precipitation and temperature [22], vastly outstripping conventional methods. Building upon this, Mital *et al.* contributed a fine-scale (400 m) dataset, achieved via a data-driven downscaling model that discerned the impact of topography on climate variables [23]. Furthermore, Vaughan *et al.* introduced convolutional conditional neural processes, a versatile deep learning framework for multisite statistical downscaling, which enabled the generation of continuous stochastic forecasts for climate variables across any geographic location [24]. Unlike methods that focus solely on spatial downscaling, Geographically and Temporally Neural Network Weighted Autoregression (GTNNWAR) [25] uses a two-stage deep neural network and autoregressive model to downscale MODIS LST from 1 km to 100 m through spatiotemporal fusion. Yu *et al.* introduced a DisTrad-Super-Resolution Convolutional Network, integrating statistical methods with deep learning to significantly improve the spatial and temporal resolutions of remote sensing imagery, enabling more refined analysis of lake surface temperature dynamics [26]. Compared to methods that only consider spatial non-stationarity, they additionally leverage temporal information from time-series data and model its temporal non-stationarity, offering a greater advantage in LST downscaling. Besides, Mukherjee and Liu developed an encoder-decoder super-resolution architecture that incorporated a custom loss function and a self-attention mechanism, adeptly increasing the resolution of MODIS spectral bands while maintaining spatial and spectral fidelity without supplementary spatial inputs [27]. Although significant progress has been made in downscaling meteorological data using deep learning methods, there are still challenges, such as the lack of effective dynamic fusion architectures and a specialized open-source ecosystem for downscaling.

Shifting focus to the field of computer vision, super-resolution (SR) aligns closely with the concept of downscaling. SR can be categorized into single-image SR (SISR) [17–19, 28–31] and guided image SR (GISR) [20, 32–35]. GISR aims to restore HR images from LR ones by leveraging structural information from HR guidance images of the same scene, while SISR does not rely on these HR guidance data. Super-Resolution Convolutional Neural Network (SRCNN) [17], the first SISR method to learn the mapping between high-resolution and low-resolution images in an end-to-end manner, significantly boosted the development of the SR field. To fully exploit the hierarchical features of LR images, Residual Dense Network (RDN) [19] adopts a dense residual strategy, effectively leveraging the hierarchical information from the original LR images, achieving excellent performance. Additionally, Residual Channel Attention Networks (RCAN) [28] introduces channel attention mechanisms into deep residual networks, resulting in improved accuracy and enhanced visual quality. In recent years, many works have incorporated Transformer into SR tasks, such as SwinIR [29], DAT [36], and SRFormer [30], benefiting from the global receptive field of the self-attention mechanism. However, the quadratic complexity and the need for large-scale training data remain challenges. Zhang *et al.* [31]

proposed a general strategy to convert Transformer-based SR networks into Hierarchical Transformer (HiT-SR), enhancing SR performance through multi-scale features while maintaining an efficient design.

Recent advances in GISR can be broadly categorized into two main approaches: cross-modal feature fusion and shared-private feature separation. Cross-modal feature fusion methods focus on effectively combining information from the target and guidance images. For instance, Zhong *et al.* [37] introduced an attention-based hierarchical multi-modal fusion strategy that selected structurally consistent features. Building upon this, Shi *et al.* [38] proposed a symmetric uncertainty-aware transformation to filter out harmful information from the guidance image, ensuring more reliable feature fusion. Furthermore, Wang *et al.* [21] developed a structure-guided method that propagates high-frequency components from the guidance to the target image in both the frequency and gradient domains, enabling more comprehensive fusion of structural details. On the other hand, shared-private feature separation methods aim to disentangle the common and unique information between the target and guidance images. Deng *et al.* [39] employed convolutional sparse coding to split the shared and private information across different modalities, facilitating more targeted feature fusion. Building on this concept, He *et al.* [40] separated RGB features into high-frequency and low-frequency components using octave convolution, allowing for more fine-grained information integration. Additionally, Xiang *et al.* [34] introduced a detail injection fusion network to fully utilize the nonlinear complementary features of both the target and guidance images, achieving more effective detail restoration. Existing GISR methods typically rely on fixed receptive fields and vanilla multimodal fusion methods (e.g., addition, concatenation), which may fail to effectively capture the multi-scale dependencies in LST data and the complex interactions between different modalities.

Despite these advancements, LST downscaling has not kept pace with the rapid developments seen in SISR and GISR [41]. This stagnation may stem from a lack of a supportive ecosystem for deep learning innovation, with two primary obstacles identified:

- 1) **Absence of High-Resolution Benchmark Dataset:** Satellite data disparities in region, time, and sensor selection hinder methodological comparisons. The lack of uniformity in satellite data selection and the scarcity of *HR thermal infrared data* (≤ 30 m) pose significant challenges for LST SR research. Therefore, establishing a standardized HR benchmark dataset are crucial for advancing the field.
- 2) **Scarcity of Open-Source LST SR Toolkit:** The absence of a dedicated open-source toolkit for LST SR hinders the community's ability to replicate, refine, and challenge existing methods. Such a toolkit would be essential for fostering collaborative development and accelerating progress in the field.

Moreover, most current deep learning models for LST downscaling are straightforward adaptations from GISR models in computer vision, without fully considering the unique

characteristics of LST data and its associated challenges [41]. As the resolution of thermal infrared bands reaches high levels (≤ 30 m), small-scale local features, such as buildings and roads, emerge alongside large-scale land cover types like water bodies, deserts, and grasslands. These local features are prone to mixing with their surroundings, introducing additional complexity to the downscaling process. According to an analysis of the HR LST data, we identify two primary limitations in existing methodologies:

- 1) **Inability to Dynamically Adjust Receptive Fields:** The stark spatial heterogeneity of LST necessitates a model capable of adjusting its receptive field to the diverse scales of temperature fluctuations. This adaptability is crucial for accurately capturing the local contrasts within LST distributions over various spatial extents.
- 2) **Multi-modal Fusion in a Uni-dimensional Manner:** Existing approaches to integrating multi-modal auxiliary data with LST features have been restricted to simplistic, uni-dimensional operations, such as addition, multiplication, or concatenation. These approaches do not suffice to unravel the complex interdependencies within HR guidance data.

To address these challenges, we propose the **Modal-Conditioned Large Selective Kernel (MoCoLSK) Network**, a novel dynamic multimodal fusion framework. MoCoLSK builds upon our previous Large Selective Kernel Network (LSKNet) [42] by replacing the static convolution in the kernel selection mechanism with a dynamic modal-conditioned projection. This projection is determined jointly by coarse-resolution LST and fine-resolution guidance data, enabling dynamic receptive field adjustment. Consequently, MoCoLSK adaptively learns fine-grained, discriminative texture features, precisely modeling the mapping between coarse-resolution LST and fine-resolution guidance data, thereby enhancing the accuracy of LST downscaling.

Furthermore, to foster research and advancement in LST downscaling, we establish a **comprehensive open-source ecosystem termed the GrokLST project**. Our contributions include the **GrokLST dataset**, a benchmark featuring 641 pairs of LR and HR LST images from the SDGSAT-1 satellite data, along with corresponding auxiliary data of multiple modalities. Accompanying the dataset is **GrokLST toolkit**, an open-source PyTorch-based toolkit encapsulating our MoCoLSK model alongside other **40+** state-of-the-art approaches, empowering researchers to effortlessly leverage the GrokLST dataset and conduct standardized evaluations.

Through extensive experimental results, we validate the effectiveness of MoCoLSK, showcasing its ability to capture the complex dependencies and subtle variations within multispectral data, outperforming existing methods in LST downscaling. The proposed MoCoLSK architecture and the GrokLST ecosystem pave the way for advancing research and applications in HR LST retrieval, providing a solid foundation for future developments in this domain.

II. GROKLST: OPEN-SOURCE ECOSYSTEM

A. GrokLST Dataset

The recent proliferation of accessible satellite imagery has catalyzed the development of deep learning models in the

domain of thermal remote sensing. However, the field of LST downscaling currently lacks HR open-source datasets, which hinders the comprehensive evaluation and comparison of emerging models. Moreover, the disparate preprocessing practices and dataset structures across different research efforts further impede the uniform assessment of state-of-the-art techniques.

Recognizing the need for consistency in model evaluation and the importance of HR data, we introduce *GrokLST*, an open-source benchmark dataset specifically designed for LST downscaling. GrokLST fills a critical gap in the field by providing a HR dataset that enables researchers to evaluate and compare their models on a standardized platform, fostering the advancement of LST downscaling through rigorous and consistent algorithm assessments.

1) *Study Area:* As depicted in Fig. 1, the pivotal focus of this study is the Heihe River Basin, the second-largest inland river basin in Northwestern China. Geographically positioned between 98° to 101° E longitude and 38° to 42° N latitude, the basin is nestled within the Hexi Corridor, serving as the primary inland watershed in Western Gansu and Qinghai provinces.

The Heihe River Basin's unique positioning amidst the Eurasian landmass and its adjacency to towering mountain ranges bestow upon it a distinct continental climate. This climate is predominantly shaped by the mid-to-high latitude westerly wind circulation and periodic influxes of polar cold air masses. The basin is characterized by its arid conditions, punctuated by sparse and concentrated precipitation, frequent high winds, abundant sunshine, intense solar radiation, and significant diurnal temperature variations. Spanning 821 kilometers from its source to its terminus at Lake Juyan, the Heihe River carves its path through three distinct ecological environments, covering an area of approximately 142,900 square kilometers.

The intricate interplay of climatic factors and geographical diversity renders the Heihe River Basin a prime candidate for environmental remote sensing and land surface temperature downscaling studies. Its vast and varied land covers, which include impervious urban structures, verdant vegetation, and sprawling water bodies, provide a diverse palette for implementing advanced deep learning techniques and computational vision approaches. These methods are employed to super-resolve imagery, facilitating a granular environmental analysis, and thereby highlighting the unique value of this study area.

2) *Data Source and Preparation:* Our GrokLST dataset leverages the cutting-edge remote sensing capabilities of the Sustainable Development Goals Science Satellite 1 (SDGSAT-1), which was launched on November 5, 2021, to bolster the United Nations Sustainable Development Goals [43]. SDGSAT-1's Multispectral Imager for Inshore (MII) and Thermal Infrared Spectrometer (TIS) sensors synergistically contribute to this dataset, with their spectral characteristics and band designations detailed in Tab. I.

For the specific LST retrieval algorithm of SDGSAT-1, please refer to our latest work [44]. The validation of the LST retrieval accuracy against in-situ measurements from the HiWATER sites, available at the National Cryosphere Desert Data Center (<http://www.ncdc.ac.cn>), reveals an RMSE of 2.598 K and an R^2 of 0.977.

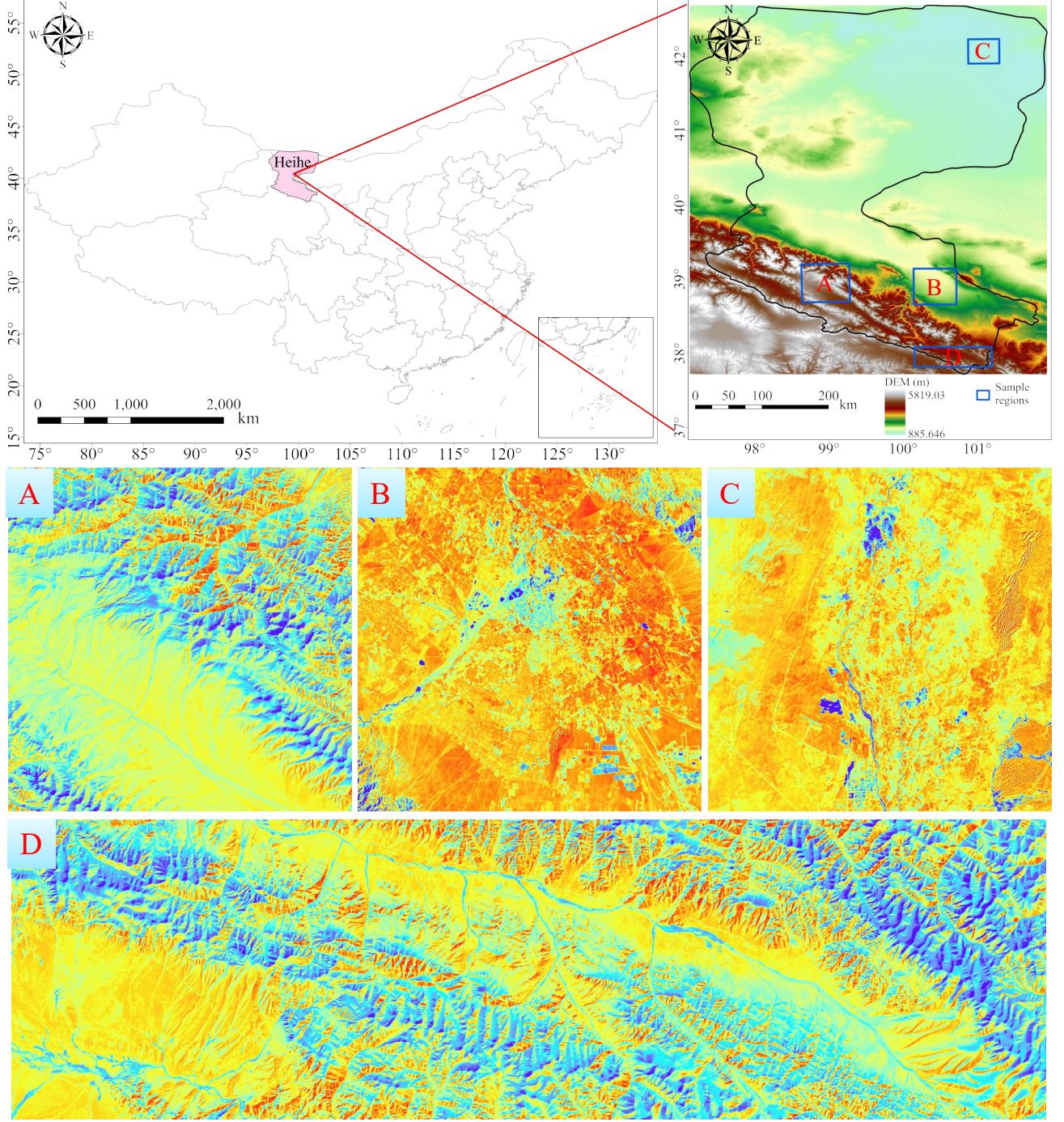


Fig. 1. Selection area and representative images of our GrokLST dataset. As demonstrated, upon reaching a resolution of 30 meters or lower, numerous local details emerge. This phenomenon underscores the critical need for models with a dynamic receptive field capable of capturing these intricate patterns.

3) Dataset Description: The GrokLST dataset includes 10 different types of HR (30m) auxiliary data. The selected bands include “B2 Deepblue”, “B3 Blue”, “B4 Green”, “B5 Red”, “B6 VRE”, and “B7 NIR”, while key indices feature the Digital Elevation Model (DEM), Normalized Difference Water Index (NDWI), NDVI, and the Normalized Difference Moisture Vegetation Index (NDMVI). These auxiliary data play a critical role in enriching the contextual understanding required for accurate LST modeling. Moreover, the LST data is provided at a resolution of 30 meters, offering detailed thermal spectral profiles, making it suitable for high-precision studies. Specifically, the GrokLST dataset consists of 641 pairs of

image data from the Heihe River Basin, covering four different scales (i.e., 30m, 60m, 120m, and 240m), including both LST data and HR guidance data, to address downscaling challenges across various scales. The detailed dataset dimensions and experimental setup can be found in IV-A1.

B. GrokLST Toolkit

The field of LST downscaling has long been hindered by a lack of accessible, open-source tools that foster innovation and reproducibility. To address this gap, we introduce GrokLST, a comprehensive deep learning toolkit designed specifically for LST downscaling tasks. Built on the PyTorch

TABLE I
SDGSAT-1 MULTISPECTRAL AND TIR BANDS USED IN CREATION OF OUR GROKLST DATASET.

Sensor	Band Name	Bandwidth (μm)	Resolution (m)	Note
MII	Band 2	0.410 ~ 0.467	10	Deep Blue
	Band 3	0.457 ~ 0.529	10	Blue
	Band 4	0.510 ~ 0.597	10	Green
	Band 5	0.618 ~ 0.696	10	Red
	Band 6	0.744 ~ 0.813	10	VRE
	Band 7	0.798 ~ 0.911	10	NIR
	Band 1	8.0 ~ 10.5	30	
TIS	Band 2	10.3 ~ 11.3	30	
	Band 3	11.5 ~ 12.5	30	

framework, GrokLST offers high flexibility and speed in model development and training, drawing inspiration from proven architectures in generic computer vision toolboxes like MMDetection and Detectron2.

GrokLST distinguishes itself through several key features that cater to the unique demands of LST downscaling.

- 1) **Comprehensive Model Support:** GrokLST provides out-of-the-box support for over 40 state-of-the-art super-resolution models. This extensive library not only facilitates easy comparison of different methods but also serves as a foundation for further research and development.
- 2) **Customizable Components:** Unlike general-purpose toolkits, GrokLST offers enhanced flexibility in model configuration. Users can choose from a variety of backbones, necks, and attention mechanisms, tailoring the architecture to specific LST downscaling needs.
- 3) **Specialized Tools and Metrics:** The toolkit includes specialized dataset loaders, data augmentation pipelines, and LST-specific evaluation metrics. These components are essential for accurately assessing model performance under diverse environmental conditions.

III. METHOD

The problem of guided LST downscaling can be formulated as follows: Given a LR LST map $T_{lr} \in \mathbb{R}^{1 \times H \times W}$ and HR guided data $G_{hr} \in \mathbb{R}^{K \times sH \times sW}$, the goal is to estimate an HR LST map $T_{sr} \in \mathbb{R}^{1 \times sH \times sW}$ that approximates the true HR LST map $T_{hr} \in \mathbb{R}^{1 \times sH \times sW}$. Here, H and W denote the height and width of the LR LST map, s is the scaling factor, and K represents the number of channels in the guided data.

In recent years, deep learning has emerged as a powerful tool for LST downscaling [45]. These methods leverage the ability of deep neural networks to learn intricate feature representations and model complex relationships between input data and the desired output. A typical deep learning-based LST downscaling model can be expressed as:

$$T_{sr} = \mathcal{F}(T_{lr}, G_{hr}; \theta), \quad (1)$$

where \mathcal{F} represents the deep neural network with learnable parameters θ . The network takes the LR LST map T_{lr} and the HR guided data G_{hr} as inputs and generates the HR LST map T_{sr} . The network is trained on a dataset of paired LR-HR LST

maps and HR guided data, with the objective of minimizing a L_1 loss function that measures the discrepancy between the predicted HR LST map and the ground truth, defined as:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \|\mathcal{F}(T_{lr}^i, G_{hr}^i) - T_{hr}^i\|_1. \quad (2)$$

A. MoCoLSK-Net Architecture

As depicted in Fig. 2, our proposed MoCoLSK-Net comprises four primary components: LST branch, guidance branch, MoCoLSK module, and reconstruction module, each designed to process and refine environmental data effectively.

LST and Guidance Branches: Apart from the different inputs, these two branches are almost completely homogeneous in structure. Each branch initiates with a convolutional stem that is responsible for extracting initial features from the input LR LST map or HR guidance image. Following the convolutional stem are N stages of Residual Groups, where each stage consists of multiple residual blocks [46] with channel attention [47]. Additionally, the two branches differ in one aspect: the LR LST map in the LST branch is processed through a bicubic upsample layer to match the desired output resolution, which serves as the preliminary step for further refinement.

MoCoLSK Module: The MoCoLSK module is the core component of our network, designed to perform dynamic multi-modal fusion. It takes as input the features from the corresponding stages of the LST and guidance branches, and performs dynamic multimodal fusion and refinement. Please refer to Section III-B for more details.

Reconstruction Module: The reconstruction module is responsible for aggregating the refined features from the MoCoLSK modules and generating final downscaled LST. Among a series of residual groups, this stage employs an up-projection unit [48] to generate HR features. Finally, projection head, consisting of two convolutional layers and a LeakyReLU activation layer, works together with the bicubic interpolation results to generate the final downscaled LST.

B. MoCoLSK Module

As illustrated in Fig. 3, the MoCoLSK module consists of two primary pathways: the Large Selective Kernel (LSK) pathway and the Modality-Conditioned Weight Generation (MCWG) pathway. Additionally, up and down projection layers [48] are introduced to perform upsampling and downsampling of LST features, enabling the framework to stack MoCoLSK multiple times for finer LST feature reconstruction. The complete MoCoLSK module can be formulated as:

$$\begin{aligned} T_{lr}^{(l)} &= \text{MoCoLSK}(T_{lr}^{(l-1)}, G_{hr}^{(l-1)}), \\ &= \text{Down}([\text{Up}(T_{lr}^{(l-1)}), Z]), \end{aligned} \quad (3)$$

where Up and Down refer to up-projection and down-projection layers [48], respectively. Z is output feature of LSK pathway, $[.]$ indicates channel concatenation, and $T_{lr}^{(l)}$ represents output of MoCoLSK at l -th stage.

Our MoCoLSK is based on LSK [42] and aims to dynamicize the key static convolutions, achieving multimodal

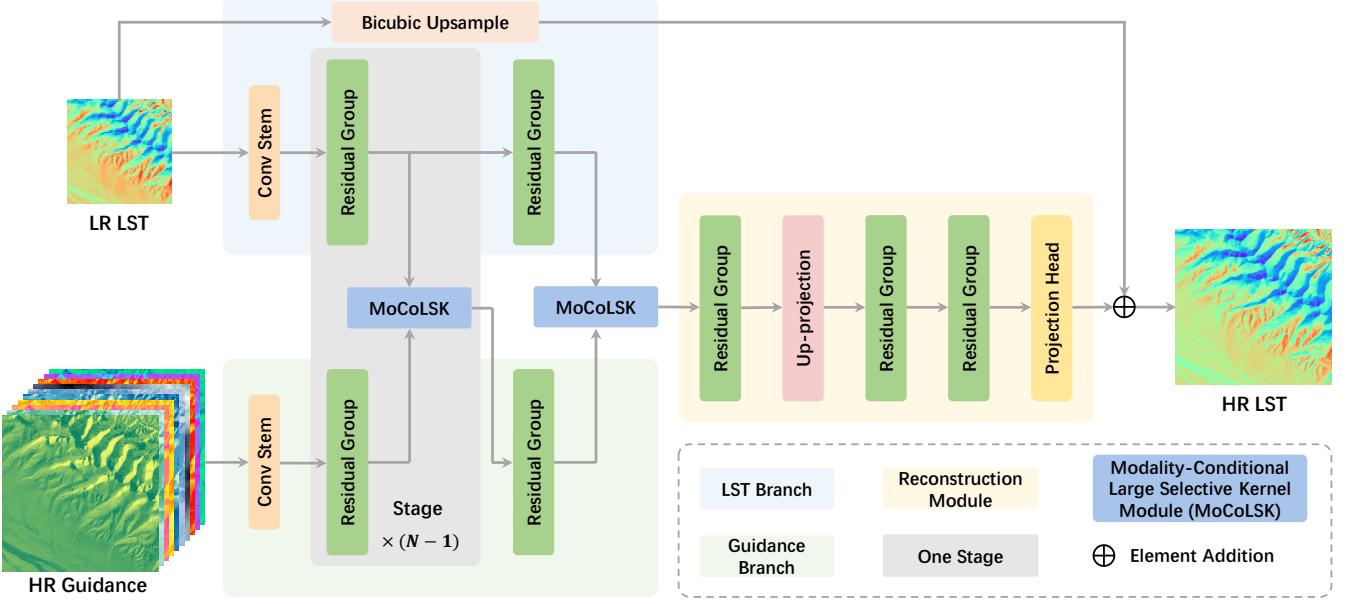


Fig. 2. The overall framework of MoCoLSK-Net primarily includes LST branch, guidance branch, MoCoLSK module, and reconstruction module. MoCoLSK-Net stacks N stages, with each stage comprising two residual groups and one MoCoLSK module. The output from the N -th stage is fed into the reconstruction module. The downscaled HR LST is obtained by adding output of the reconstruction module to bicubic upsampling result of the original LR LST data.

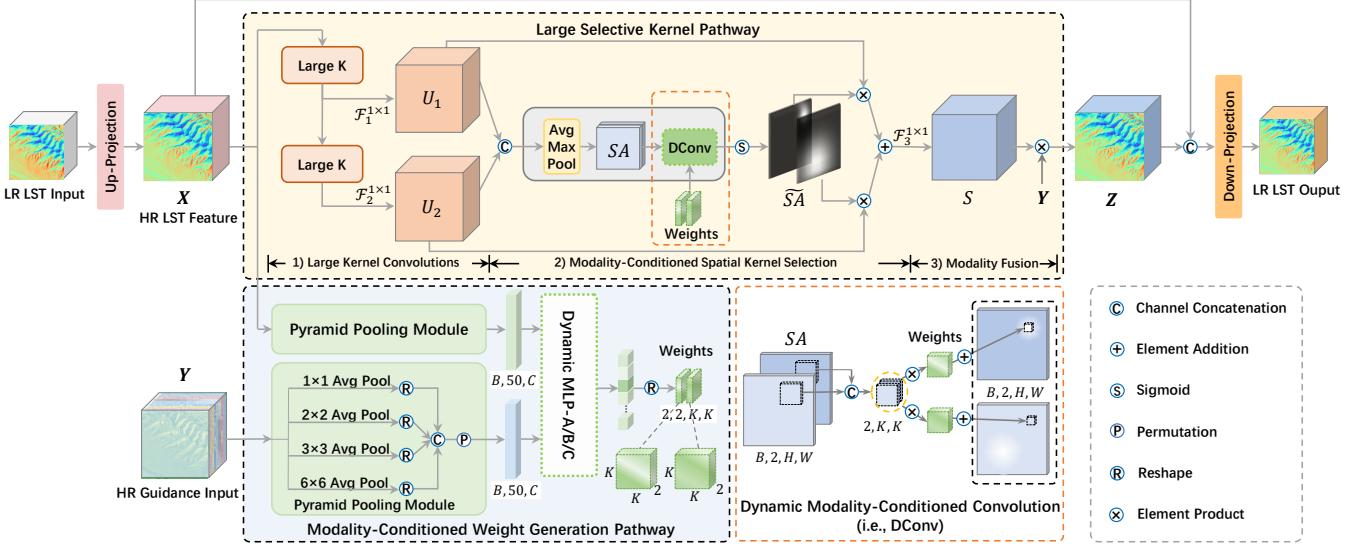


Fig. 3. Overview of our proposed MoCoLSK Module. The MoCoLSK module primarily consists of the large selective kernel pathway and the modality-conditioned weight generation pathway. For the LST pathway, it essentially follows the original LSK [42] module configuration but with two key differences: 1) the generation of the spatial selection mask $\bar{S}A$ is modulated by the modality-conditioned weights from the MCWG pathway; 2) the output feature Z is the result of fusing two modality features. For the MCWG pathway, the HR LST features and HR guidance features are deeply fused using the pyramid pooling module and the dynamic MLP [49] to generate modality-conditioned weights. Additionally, to facilitate modality fusion and the stacking of multiple MoCoLSK modules for more refined LST feature reconstruction, we utilize up-projection and down-projection units to upsample the LR LST features and downsample the concatenated result of the modality fusion output Z with the HR LST features X .

feature fusion through a dynamic receptive field driven by modality-conditioned weights. These weights are determined by dynamically fusing LST and guidance features through MCWG pathway and serve as convolution kernels for the dynamic modality-conditioned convolution (denoted as DConv), facilitating dynamic adjustments to the receptive field.

1) *Large Selective Kernel Pathway*: LSK pathway closely follows the original LSK, with two key distinctions: first, the static convolution used to generate the spatial selection masks is replaced by DConv; second, the feature S is multiplied by HR guidance feature Y instead of X .

Specifically, the LSK pathway takes HR LST features X and

HR guidance features Y as inputs, and outputs the refined HR LST Z through three steps: (1) large kernel decomposition; (2) modality-conditioned spatial kernel selection; and (3) modality fusion.

Large Kernel Decomposition: The large kernel decomposition leverages HR LST feature X to generate large kernel features U_1 and U_2 at different scales, defined as follows:

$$\begin{aligned} U_1 &= \mathcal{F}_{lk}^{1 \times 1}(\mathcal{F}_{lk}^{5 \times 5}(X)), \\ U_2 &= \mathcal{F}_{lk}^{1 \times 1}(\mathcal{F}_{lk}^{7 \times 7}(\mathcal{F}_{lk}^{5 \times 5}(X))), \end{aligned} \quad (4)$$

where $\{\mathcal{F}_i^{1 \times 1}, i = 1, 2\}$ are point-wise convolutions. $\mathcal{F}_{lk}^{5 \times 5}$

and $\mathcal{F}_{lk}^{7 \times 7}$ denote depth-wise convolutions with kernel size 5, dilation 1 and kernel size 7, dilation 3, respectively.

Modality-Conditioned Spatial Kernel Selection: This selection aims to dynamically select features from spatial kernels with different receptive fields (i.e., U_1 and U_2) that are effective for refining HR LST feature, assisted by modality-conditioned weights generated through MCWG pathway. Specifically, U_1 and U_2 are first concatenated along the channel dimension, followed by channel-wise average pooling \mathcal{P}_{avg} and maximum pooling \mathcal{P}_{max} , and then concatenated again to obtain preliminary spatial attention weights SA . This process is formulated as:

$$SA = [\mathcal{P}_{avg}([U_1, U_2]), \mathcal{P}_{max}([U_1, U_2])]. \quad (5)$$

To obtain modality-conditioned spatial selection masks \widetilde{SA} , we introduce a dynamic modality-conditioned convolution layer (denoted as \mathcal{F}_{dconv}) powered by modality-conditioned weights from the MCWG pathway (see III-B2), as detailed in the following:

$$\widetilde{SA} = \sigma(\mathcal{F}_{dconv}^{2 \rightarrow 2}(SA, \text{weights})), \quad (6)$$

where superscript $(\cdot)^{2 \rightarrow 2}$ indicates that the number of channels remains 2. σ is sigmoid activation function.

The large kernel features U_1 and U_2 are spatially weighted by their corresponding spatial masks (i.e., \widetilde{SA}_1 and \widetilde{SA}_2), then added and passed through a point-wise convolutional layer $\mathcal{F}_3^{1 \times 1}$ to obtain the dynamically selected features S :

$$S = \mathcal{F}_3^{1 \times 1}(\sum_{i=1}^2 (\widetilde{SA}_i \otimes U_i)), \quad (7)$$

where \otimes is the element-wise multiplication.

Modality Fusion: To obtain the final modality fusion feature Z , we treat S as the attention weights for HR guidance feature Y and perform element-wise multiplication, denoted as:

$$Z = Y \otimes S. \quad (8)$$

2) Modality-Conditioned Weight Generation Pathway: The MCWG pathway generates modality-conditioned weights for HR LST feature X under the guidance of modality Y using a pyramid pooling module (PPM) and a dynamic MLP (DMLP) [49], which can be formulated as (with reshaping and other operations omitted for clarity):

$$\begin{aligned} \text{weights} &= \text{MCWG}(X, Y) \\ &= \text{DMLP}(\text{PPM}(X), \text{PPM}(Y)), \end{aligned} \quad (9)$$

$$\text{PPM}(X/Y) = [\text{AvgPool}_i(X/Y)], i = 1, 2, 3, 4, \quad (10)$$

where $\text{AvgPool}_i(\cdot)$ represents a series of global average poolings with bin sizes $\{1, 2, 3, 6\}$, similar to pyramid scene parsing network (PSPNet) [50].

IV. EXPERIMENTS

A. Experimental Settings

1) Dataset: We utilize our GrokLST dataset for experiments. To address the challenge of downscaling across different resolutions, we adhere to the Wald's protocol, downsampling the 30m resolution data to three distinct resolutions of 60m, 120m, and 240m, thereby enabling $\times 2$, $\times 4$, and $\times 8$ downscaling tasks.

TABLE II
THE CORRESPONDING SIZES OF LST AND GUIDANCE AT DIFFERENT RESOLUTIONS IN THE GROKLST DATASET. H: HEIGHT, W: WIDTH, C: CHANNEL.

Resolution	Scale	LST Size (H×W×C)	Guidance Size (H×W×C)
30m	-	512×512×1	512×512×10
60m	×2	256×256×1	256×256×10
120m	×4	128×128×1	128×128×10
240m	×8	64×64×1	64×64×10

Specifically, the 30m resolution LST data is used as the ground truth (GT), while LST data at other resolutions is downsampled with the aid of 30m guidance data to reconstruct the predicted 30m resolution LST. The specific spatial resolutions of the GrokLST dataset are detailed in Table II. Fig. 4 provides visual representations of these bands and indices, highlighting the spectral characteristics and quality of the dataset. For effective model training and validation, the GrokLST dataset is carefully divided into three subsets in a 6:1:3 ratio: 384 samples for training, 64 for validation, and 193 for testing. All LST and guidance data are processed using the Z-score normalization strategy. For an in-depth analysis of different normalization strategies, please refer to V-F.

2) Evaluation Metrics: Some key statistical indicators such as root mean square error (RMSE), mean absolute error (MAE), bias (BIAS), correlation coefficient (CC), and ratio of standard deviations (RSD) are utilized to quantitatively evaluate reconstruction performance of one downscaling model.

RMSE is the square root of the average of the squared differences between the predicted HR LST T_{sr} and the ground truth T_{hr} :

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (T_{sr}^i - T_{hr}^i)^2}. \quad (11)$$

MAE represents the average of the absolute differences between T_{sr} and T_{hr} :

$$MAE = \frac{1}{N} \sum_{i=1}^N |T_{sr}^i - T_{hr}^i|. \quad (12)$$

BIAS shows the average of the differences between T_{sr} and T_{hr} :

$$BIAS = \frac{1}{N} \sum_{i=1}^N (T_{sr}^i - T_{hr}^i). \quad (13)$$

CC evaluates the correlation between T_{sr} and T_{hr} , with a value of 1 indicating perfect correlation:

$$CC = \frac{\frac{1}{N} \sum_{i=1}^N (\Delta T_{sr}^i)(\Delta T_{hr}^i)}{\sqrt{\frac{1}{N} \sum_{i=1}^N (\Delta T_{sr}^i)^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (\Delta T_{hr}^i)^2}}, \quad (14)$$

where

$$\begin{aligned} \Delta T_{sr}^i &= T_{sr}^i - \frac{1}{N} \sum_{i=1}^N T_{sr}^i, \\ \Delta T_{hr}^i &= T_{hr}^i - \frac{1}{N} \sum_{i=1}^N T_{hr}^i. \end{aligned} \quad (15)$$

RSD quantifies how closely the distribution of T_{sr} matches the distribution of T_{hr} :

$$RSD = \frac{|\sigma_{sr} - \sigma_{hr}|}{\sigma_{hr}}, \quad (16)$$

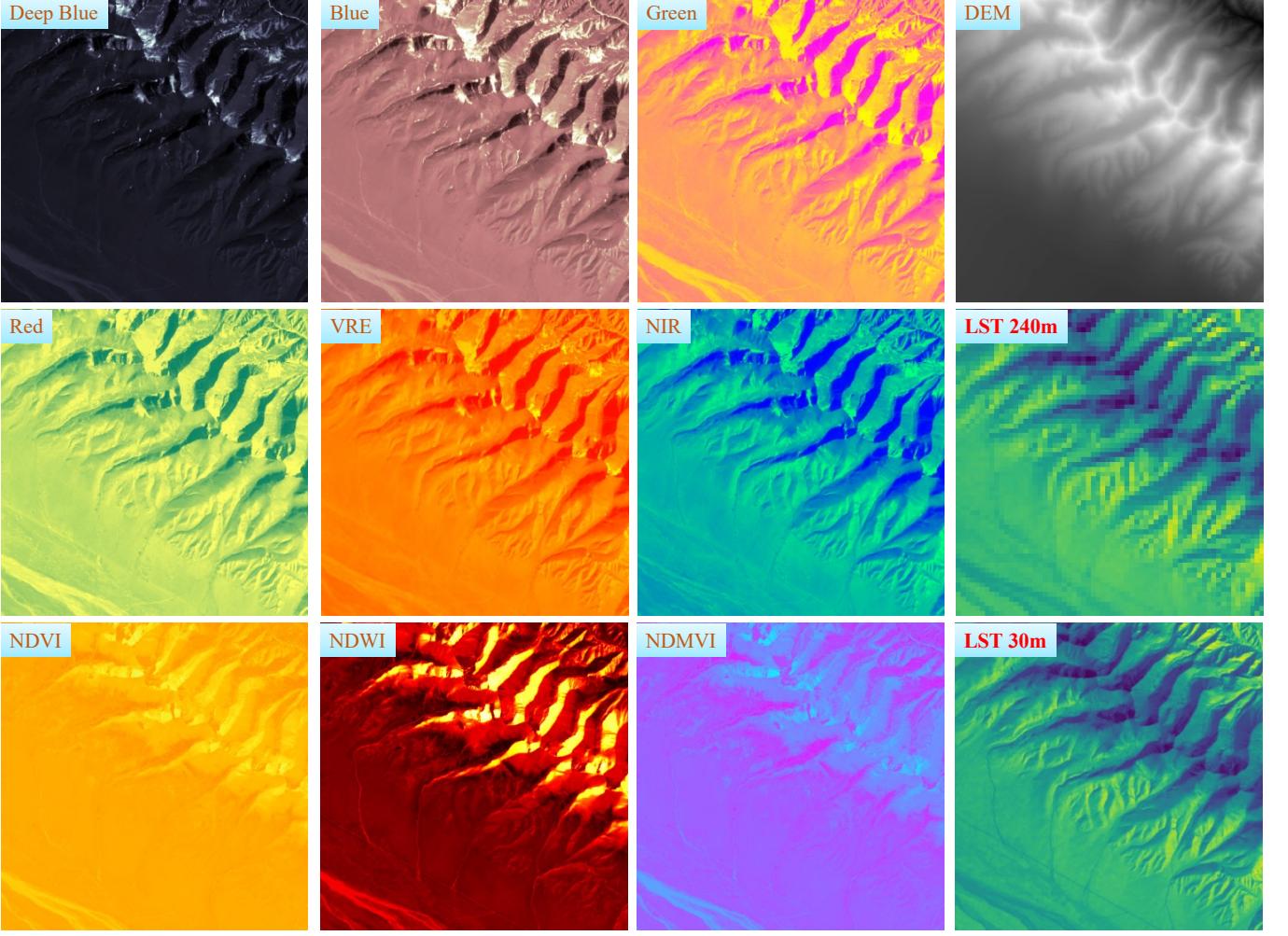


Fig. 4. Gallery of the GrokLST dataset, showcasing the comparison between LR (240 m) and HR (30 m) LST images, along with a suite of 10 auxiliary data.

where

$$\begin{aligned}\sigma_{sr} &= \sqrt{\frac{1}{N-1} \sum_{i=1}^N (T_{sr}^i - \bar{T}_{sr})^2}, \\ \sigma_{hr} &= \sqrt{\frac{1}{N-1} \sum_{i=1}^N (T_{hr}^i - \bar{T}_{hr})^2}.\end{aligned}\quad (17)$$

The closer the values of RMSE, MAE, BIAS, and RSD are to 0, and the closer CC is to 1, the better the downscaling method's reconstruction performance.

3) Implementation Details: We implemented our MoCoLSK-Net in our GrokLST toolkits and trained it on a platform equipped with four NVIDIA GeForce RTX 4090 GPUs using a distributed training approach. During training, we employ the AdamW optimizer [51] and a cosine annealing learning rate scheduler with warm restarts [52]. The initial learning rate is set to $1e-4$ and weight decay by a factor of $1e-5$ for $10k$ iterations. Each GPU is assigned one training sample, and the batch size is fixed to 4. All other deep learning-based methods in the GrokLST toolkits use the above experimental configuration.

Next, we detail the key hyperparameters in MoCoLSK-Net.

1) Base Feature Dimension: In the guidance branch, the feature dimension of all residual groups remains fixed at 32. In contrast, in the LST branch, the feature dimension of the residual groups increases by 32 in each stage compared

to the previous stage. Besides, the feature dimension of all submodules within the reconstruction module is maintained at $N \times 32$.

2) Number of Stages: MoCoLSK-Net by default has 4 stages (i.e., $N = 4$), with each stage containing two residual groups and one MoCoLSK module.

3) Number of Layers in DMLP: The DMLP contains multiple linear layers to enhance the dynamic fitting capability of the module. In MoCoLSK-Net, the default number of DMLP layers is 1.

4) DMLP Versions: There are three versions of standard DMLP, namely A, B, and C. For details, please refer to [49].

5) Size of Weights: The weights dynamically generated by MCWG pathway are used in DConv in LSK pathway to obtain modality-conditioned spatial selection masks. The default size of the weights is 3×3 .

B. Comparison with State-of-the-Arts

We benchmarked the reconstruction performance of MoCoLSK-Net against current state-of-the-art downscaling methods, including four machine learning methods, nineteen single-image downscaling methods, and thirteen guided image downscaling methods. The benchmarking covers three scales: $\times 2$, $\times 4$, and $\times 8$, with evaluation conducted on our GrokLST

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE **GROKLST** DATASET. THE SYMBOL “-” INDICATES INSUFFICIENT MEMORY TO EXECUTE THE ALGORITHM, WHILE “X” DENOTES THAT THE ALGORITHM DOES NOT SUPPORT THE CORRESPONDING DOWNSCALING FACTOR.

Method	$\times 2$					$\times 4$					$\times 8$				
	RMSE \downarrow	MAE \downarrow	BIAS	CC \uparrow	RSD \downarrow	RMSE \downarrow	MAE \downarrow	BIAS	CC \uparrow	RSD \downarrow	RMSE \downarrow	MAE \downarrow	BIAS	CC \uparrow	RSD \downarrow
<i>Machine Learning</i>															
Random Forest [53]	-	-	-	-	-	1.3900	0.9494	-0.0317	0.9055	0.0712	1.7367	1.2264	-0.0456	0.8477	0.1330
XGBoost [54]	-	-	-	-	-	1.8209	1.3000	-0.0108	0.8244	0.1228	1.9825	1.4342	-0.0156	0.7846	0.1755
LightGBM [55]	-	-	-	-	-	1.7826	1.2680	-0.0202	0.8340	0.1016	2.0205	1.4632	-0.0267	0.7772	0.1425
CatBoost [56]	-	-	-	-	-	1.5350	1.0639	-0.0175	0.8787	0.0866	1.9089	1.3683	-0.0074	0.8030	0.1537
<i>Single Image Downscaling</i>															
EDSR [18]	0.4010	0.2605	0.0018	0.9889	0.0114	0.8921	0.6042	0.0061	0.9559	0.0441	1.4855	1.0397	0.0112	0.8933	0.1024
RDN [19]	0.3802	0.2478	0.0018	0.9898	0.0104	0.8227	0.5598	0.0066	0.9595	0.0400	1.2497	0.8742	0.0133	0.9110	0.0856
RCAN [28]	0.4046	0.2644	0.0017	0.9887	0.0117	0.8826	0.6009	0.0065	0.9562	0.0440	1.4446	1.0147	0.0114	0.8958	0.1017
DBPN [48]	0.4257	0.2803	0.0008	0.9879	0.0120	0.8865	0.6008	0.0072	0.9564	0.0431	1.4303	0.9982	0.0146	0.8975	0.0991
CTNet [57]	0.4012	0.2627	0.0021	0.9889	0.0118	0.8954	0.6064	0.0065	0.9561	0.0442	X	X	X	X	X
FeNet [58]	0.3995	0.2616	0.0017	0.9890	0.0116	0.9009	0.6118	0.0064	0.9557	0.0457	1.5193	1.0644	0.0113	0.8917	0.1074
FENet [59]	0.4018	0.2632	0.0020	0.9889	0.0117	0.8879	0.6017	0.0070	0.9564	0.0440	1.4844	1.0366	0.0135	0.8940	0.1030
SRFBN [60]	0.3962	0.2586	0.0020	0.9891	0.0113	0.8969	0.6067	0.0073	0.9560	0.0446	1.5212	1.0589	0.0131	0.8919	0.1060
CFGN [61]	0.4283	0.2814	0.0014	0.9877	0.0127	0.9386	0.6395	0.0053	0.9529	0.0485	1.5724	1.1052	0.0108	0.8867	0.1134
SwinIR [29]	-	-	-	-	-	0.9259	0.6297	0.0059	0.9537	0.0474	1.5549	1.0906	0.0104	0.8884	0.1112
DAT [36]	-	-	-	-	-	-	-	-	-	-	1.2605	0.8816	0.0153	0.9131	0.0842
SRFormer [30]	-	-	-	-	-	0.9203	0.6252	0.0065	0.9544	0.0473	1.5584	1.0929	0.0123	0.8882	0.1122
DLGSANet [62]	-	-	-	-	-	0.9559	0.6537	0.0059	0.9514	0.0483	1.5848	1.1159	0.0097	0.8852	0.1128
ACT [63]	-	-	-	-	-	0.8908	0.6054	0.0057	0.9560	0.0444	1.4592	1.0265	0.0117	0.8952	0.1021
NGSwin [64]	0.4902	0.2897	0.0006	0.9831	0.0113	1.3168	0.6786	0.0017	0.8904	0.0694	X	X	X	X	X
DCTLSA [65]	0.4616	0.2981	-0.0056	0.9855	0.0226	0.9837	0.6720	0.0212	0.9456	0.0590	1.6282	1.1488	-0.0107	0.8708	0.1156
HiT-SIR [31]	0.4078	0.2675	0.0023	0.9886	0.0117	0.9009	0.6121	0.0067	0.9555	0.0444	1.5301	1.0735	0.0132	0.8904	0.1066
HiT-SNG [31]	0.4084	0.2677	0.0014	0.9886	0.0118	0.9079	0.6176	0.0056	0.9551	0.0453	1.5207	1.0669	0.0123	0.8912	0.1057
HiT-SRF [31]	0.4080	0.2674	0.0018	0.9886	0.0116	0.9028	0.6136	0.0064	0.9554	0.0448	1.5142	1.0622	0.0092	0.8916	0.1049
<i>Guided Image Downscaling</i>															
MSG-Net [66]	0.4294	0.2829	0.0021	0.9877	0.0120	0.8651	0.5914	0.0043	0.9578	0.0412	1.3418	0.9442	0.0070	0.9048	0.0893
SVLRLM [67]	0.4612	0.3085	0.0045	0.9863	0.0232	0.8611	0.5974	0.0017	0.9567	0.0513	1.2357	0.8815	-0.0101	0.9135	0.0863
DJFR [68]	0.3933	0.2603	0.0013	0.9891	0.0112	0.7784	0.5382	0.0031	0.9642	0.0390	1.1892	0.8436	0.0010	0.9181	0.0809
P2P [69]	0.4788	0.3200	-0.0041	0.9860	0.0227	1.0003	0.6898	-0.0298	0.9497	0.0602	1.5409	1.0952	-0.0964	0.8914	0.1062
DSRN [70]	0.4480	0.2956	0.0574	0.9875	0.0243	0.9562	0.6543	0.1458	0.9525	0.0724	1.5587	1.1023	0.2025	0.8891	0.1477
FDSR [40]	0.4065	0.2698	0.0013	0.9885	0.0125	0.7779	0.5371	0.0047	0.9619	0.0396	1.1395	0.8096	0.0047	0.9210	0.0783
DKN [71]	0.4071	0.2695	0.0041	0.9884	0.0135	0.8388	0.5727	0.0026	0.9574	0.0416	1.3719	0.9589	0.0036	0.8976	0.1042
FDKN [71]	0.3717	0.2449	0.0032	0.9901	0.0099	0.7946	0.5456	0.0032	0.9612	0.0387	1.3312	0.9335	0.0038	0.9061	0.0978
AHMF [37]	0.3557	0.2348	0.0017	0.9908	0.0097	0.7224	0.4996	0.0028	0.9655	0.0352	1.1246	0.7959	0.0019	0.9229	0.0764
CODON [72]	X	X	X	X	X	0.9617	0.6642	0.0210	0.9492	0.0531	1.6690	1.1799	-0.0001	0.8789	0.1608
SUFT [38]	0.3130	0.2093	0.0011	0.9927	0.0075	0.6046	0.4207	0.0021	0.9737	0.0265	0.8598	0.6061	0.0025	0.9468	0.0489
DAGF [73]	0.3917	0.2589	0.0005	0.9892	0.0110	0.7910	0.5451	0.0063	0.9613	0.0391	1.1935	0.8469	0.0070	0.9170	0.0879
RSAG [20]	X	X	X	X	X	0.7223	0.4990	0.0026	0.9654	0.0350	1.0118	0.7154	0.0012	0.9330	0.0638
* MoCoLSK	0.2902	0.1951	0.0009	0.9937	0.0062	0.5590	0.3883	0.0020	0.9771	0.0218	0.8031	0.5642	0.0027	0.9514	0.0456

dataset using five metrics: RMSE, MAE, BIAS, CC, and RSD. All comparison methods and experimental results are listed in Table III, and the following key conclusions can be drawn:

- 1) Our MoCoLSK-Net achieves state-of-the-art performance across nearly all metrics (except BIAS) in downscaling challenges at various scales, underscoring the effectiveness of MoCoLSK, which utilizes modality-conditioned dynamic receptive fields for multimodal fusion. The MoCoLSK module integrates the LSK and MCWG pathways, working synergistically to enable deep dynamic fusion of LST and guidance features, continuously refining the discriminative LST features. As a result, MoCoLSK-Net outperforms all other SOTA methods, delivering the most accurate LST downscaling results.
- 2) All deep learning-based downscaling methods significantly outperformed the four traditional machine learning methods (Random Forest [53], XGBoost [54], LightGBM [55], and

CatBoost [56]) across all reconstruction scales. This indicates that the four traditional machine learning methods struggle to accurately capture the fine-grained mapping relationships between multiple guidance data and LST. This may be due to the fact that these models learn the global mapping between LST and the multiple guidance data, and then predict LST on a pixel-by-pixel basis [8]. Although LST is highly correlated with surface attributes in the guidance data, this relationship may vary across different regions, and the global mapping may not fully satisfy the local downscaling of LST. Furthermore, the pixel-wise reconstruction process can disrupt the spatial texture of LST, resulting in noticeable temperature biases, either very high or very low, in the downscaled LST.

- 3) Among single-image downscaling methods, we did not observe Transformer-based downscaling algorithms (italicized in Table III) outperforming those based on CNNs, spatial

attention, or channel attention across various metrics. Instead, Transformer-based methods introduced greater computational complexity, as evidenced by the significant number of “-” entries in Table III. This suggests that algorithms leveraging CNNs, spatial, or channel attention mechanisms are not inferior to Transformer methods with global attention. It is worth noting, however, that the relatively small size of our GrokLST dataset might limit the reconstruction performance of Transformer-based methods.

4) Overall, guided downscaling methods exhibited significantly better performance across reconstruction metrics compared to single-image downscaling methods. Examples include FDKN, AHMF, SUFT, and RSAG, with our MoCoLSK significantly outperforming RDN (the most effective single-image downscaling algorithm). This highlights not only the higher reconstruction quality ceiling of guided downscaling methods compared to single-image methods but also underscores the importance and necessity of HR guided data.

C. Visual Analysis

Fig. 5 provides intuitive visualizations of different downscaling methods on GrokLST dataset at $\times 8$ downscaling challenge, with each method showing its downscaled result and difference map compared to GT, complementing quantitative analysis from a qualitative perspective. From these visualizations, we can intuitively draw the following crucial insights:

1) From downscaled result maps, MoCoLSK-Net demonstrates the clearest land surface textures and most accurate temperature predictions. From difference maps, it is evident that MoCoLSK-Net’s difference map aligns most closely with label differences, tending towards white (whiter difference maps indicate better reconstruction performance). This once again offers a comprehensive and intuitive qualitative validation of MoCoLSK-Net’s superior LST downscaling performance.

2) Downscaling results of four traditional machine learning algorithms perform the worst visually. These methods fail to reconstruct land surface textures and structures, appearing disordered and lacking smoothness, while showing significant temperature bias. This suggests difficulty in accurately capturing fine-grained mapping between guide data and LST, visually reinforcing significant limitations of paradigms that learn global mappings between LST and multiple guide data for point-by-point LST prediction.

3) In single-image downscaling methods, most algorithms produce blurred LST results with unclear textures and noticeable temperature biases, whereas RDN and DAT yield relatively more accurate LST downscaling. Furthermore, most Transformer-based algorithms (excluding DAT), despite global receptive fields, do not outperform CNN-based or spatial/channel attention-based methods and introduce higher computational complexity. This suggests that for downscaling tasks, CNN-based or attention mechanism-based methods perform similarly to Transformer methods with global receptive fields, further confirmed from a visual perspective.

4) Most guided downscaling methods (e.g., AHMF, SUFT, RSAG) exhibit significantly better reconstruction results than single-image downscaling methods, especially MoCoLSK. This

TABLE IV
VALIDATION OF KEY COMPONENTS IN MoCoLSK MODULE.

Case	LSK Pathway			MCWG pathway		Metrics	
	LKD	DConv	MF	PPM	AvgMax	RMSE \downarrow	CC \uparrow
1	-	-	-	-	-	0.7405	0.9605
2		✓	✓	✓		0.7154	0.9612
3	✓		✓	✓		0.7267	0.9603
4	✓	✓		✓		0.9407	0.9414
5	✓	✓	✓		✓	0.7153	0.9613
6	✓	✓	✓	✓		0.7133	0.9613

TABLE V
ABLATION STUDY ON THE EFFECTIVENESS OF PPM IN MCWG.

Avg.	Pooling Max.	PPM	Metrics	
			RMSE \downarrow	CC \uparrow
✓			0.7175	0.9610
	✓		0.7237	0.9608
✓	✓		0.7153	0.9613
		✓	0.7133	0.9613

not only highlights superior reconstruction potential of guide-based methods but also further validates the critical role of HR guide data in improving LST downscaling performance.

V. DISCUSSION

A. Ablation Study

This section presents ablation study results for key components of MoCoLSK module on GrokLST dataset with $\times 8$ downscaling and 20k iterations to ensure more reliable results, including large kernel decomposition (LKD), DConv, modality fusion (MF) in LSK pathway, and PPM in MCWG pathway.

Table IV presents the results of the ablation experiments. Case 1 is the baseline, which only uses up and down projection layers in MoCoLSK without utilizing the LSK and MCWG pathways, as shown in Fig. 8(a). “LKD” indicates whether large kernel decomposition is utilized; if unchecked, it signifies the use of one large kernel depth-wise convolution with same receptive field (i.e., 23) to replace two decomposed large kernels. “DConv” refers to dynamic modality-conditioned convolution; if unchecked, original static depth-wise convolution in LSK is employed. “MF” denotes modal fusion, as represented in Equation (8); if unchecked, it implements $Z = X \otimes S$, similar to the original LSK.

Cases 2 and 6 validate that large kernel decomposition is superior to a single larger kernel. Cases 3 and 6 show that DConv driven by the modality-conditional weights generated by the MCWG pathway performs better than the original static convolution. Cases 4 and 6 demonstrate the necessity of further modality fusion. Cases 5 and 6 validate the effectiveness of proposed PPM.

B. Hyperparameters Analysis

1) *Pooling in MCWG Pathway*: We conduct an in-depth exploration of different poolings in MCWG pathway. Table V presents the comparative results of different poolings. It can be observed that: 1) Average pooling performs better than max pooling; 2) Using both average and max pooling together is

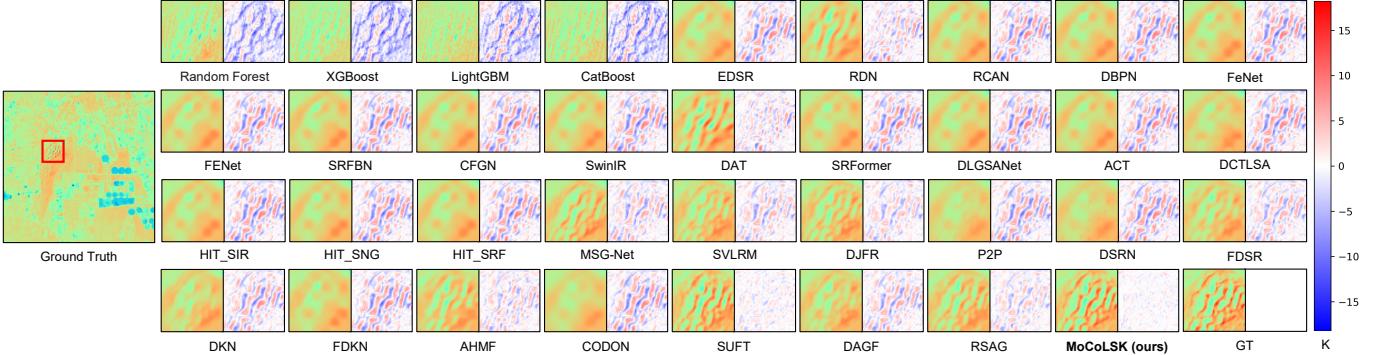


Fig. 5. Visual comparison of $\times 8$ reconstruction images from different methods. Each method is represented by two images: the left image shows the reconstruction result, while the right image illustrates the difference between the reconstruction result and the GT. A Kelvin (K) temperature color bar is shown on the far right, where pixels with values larger than the GT are displayed in red, smaller values are displayed in blue, and identical values are shown in white.

TABLE VI
STUDY ON THE IMPACT OF BASE FEATURE DIMENSION ON RECONSTRUCTION PERFORMANCE IN MoCoLSK-NET.

Dim.	RMSE \downarrow	MAE \downarrow	BIAS	CC \uparrow	RSD \downarrow
16	0.8719	0.6175	0.0030	0.9446	0.0512
24	0.7932	0.5549	0.0030	0.9536	0.0407
32	0.7133	0.4849	0.0038	0.9613	0.0312
40	0.6697	0.4445	0.0037	0.9663	0.0259

TABLE VII
STUDY ON THE IMPACT OF STAGE COUNT ON RECONSTRUCTION PERFORMANCE IN MoCoLSK-NET.

Stages	RMSE \downarrow	MAE \downarrow	BIAS	CC \uparrow	RSD \downarrow
1	0.9076	0.6449	0.0017	0.9413	0.0553
2	0.8146	0.5738	0.0030	0.9506	0.0445
3	0.7511	0.5200	0.0041	0.9575	0.0360
4	0.7133	0.4849	0.0038	0.9613	0.0312
5	0.7441	0.4957	0.0042	0.9605	0.0306

more effective than using average or max pooling alone; 3) Our proposed PPM outperforms the other three pooling strategies.

2) *Base Feature Dimension*: In MoCoLSK-Net, the base feature dimension is also a critical hyperparameter. As the overall channel dimension of MoCoLSK-Net increases, its reconstruction performance naturally improves, as shown in Table VI. Due to memory limitations, we increased the base dimension only up to 40, but we believe that further appropriate increases in dimension would lead to even better reconstruction performance.

3) *Number of Stages*: Table VII presents the impact of stacking different numbers of residual groups and the MoCoLSK module on LST reconstruction performance. Notably, a stage number of 4 achieves optimal reconstruction performance, whereas a number of 5 leads to a decline in performance.

4) *Number of Layers in Different Versions of DMLP*: The depth of the linear layers in different versions of DMLP determines the quality of modality-conditioned weights, which subsequently affects the LST downscaling performance of MoCoLSK-Net. Fig. 6 illustrates the reconstruction performance, as reflected by RMSE, MAE, and CC metrics, for three DMLP versions (A, B, and C) with varying numbers of linear layers. It can be observed that version A of DMLP achieves

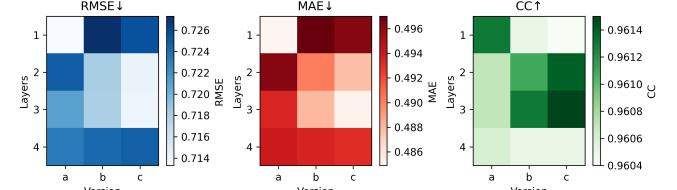


Fig. 6. Heatmap of the impact of different layer numbers for the three versions of dynamic MLP (i.e., A, B, and C) on reconstruction performance in the MCWG pathway.

optimal performance with 1 layer, whereas versions B and C achieve the best reconstruction with 3 layers.

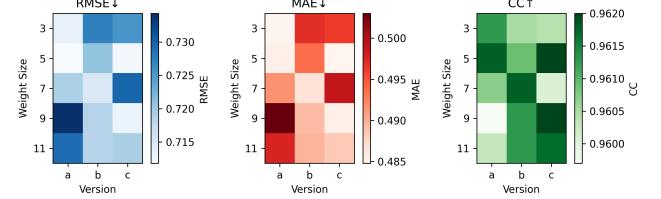


Fig. 7. Heatmap of experimental results on reconstruction performance with different weight sizes from three versions of dynamic MLP, i.e., A, B, and C.

5) *Size of Weights*: We conducted a study of how different sizes of modality-conditioned weights, generated by three versions of DMLP, affect LST reconstruction performance. Fig. 7 provides detailed experimental results, revealing the following: the optimal weight size varies across DMLP versions but remains broadly consistent. For version A, the best size is 5×5 (instead of the default 3×3); for version B, it is 7×7 ; and for version C, 5×5 , slightly outperforming 9×9 . Moreover, the optimal weight sizes, typically around 5×5 or 7×7 , indicate that larger weights do not necessarily lead to more accurate LST downscaling predictions.

C. Larger Kernel, Better Performance?

We conducted a deeper investigation into large-kernel decomposition to determine whether larger kernels result in better reconstruction performance. Table VIII shows the impact

TABLE VIII

RESEARCH ON LARGE KERNEL CONVOLUTIONS WITH DIFFERENT RECEPTIVE FIELDS. NOTE THAT THE #P AND FLOPs COLUMNS ONLY FOCUS ON THE SINGLE MoCoLSK MODULE (IGNORING THE UP AND DOWN PROJECTION UNITS). K: KERNEL, D: DILATION, RF: RECEPTIVE FIELD.

(K, D) Sequences	RF	#P	FLOPs	RMSE↓	CC↑
(23, 1)	23	0.02M	4.44G	0.7154	0.9612
(3, 1) → (3, 2)	7	0.04M	0.56G	0.7233	0.9609
(3, 1) → (5, 2)	11	0.04M	0.69G	0.7183	0.9613
(5, 1) → (7, 3)	23	0.04M	1.03G	0.7133	0.9613
(7, 1) → (9, 4)	39	0.04M	1.49G	0.7091	0.9620
(9, 1) → (11, 5)	59	0.05M	2.10G	0.7092	0.9622

of a single large kernel and a series of two consecutive large kernels with varying receptive fields on LST downscaling performance. The results reveal the following: a single large kernel convolution with same receptive field 23 is less effective than two consecutive decomposed large-kernel convolutions. Furthermore, the LST downscaling performance of MoCoLSK-Net improves as the receptive field of the large-kernel convolution group increases.

TABLE IX

THE IMPACT OF DIFFERENT CONFIGURATION SELECTION MECHANISMS ON THE DOWNSCALING PERFORMANCE OF MoCoLSK-NET. S: MoCoLSK-SS, C: MoCoLSK-CS. FOR EXAMPLE, (C, S, C, C) IN SECOND COLUMN MEANS THAT SECOND STAGE OF MoCoLSK-NET USES MoCoLSK AND OTHER STAGES USE MoCoLSK-CS.

Fusion Modules	S/C Sequences	RMSE↓	CC↑
MoCoLSK-SS only	(S, S, S, S)	0.7133	0.9613
MoCoLSK-CS only	(C, C, C, C)	0.7486	0.9599
Interleaved MoCoLSK-SS & MoCoLSK-CS	(S, C, S, C)	0.7044	0.9627
	(C, S, C, S)	0.7040	0.9627
	(C, C, S, S)	0.7181	0.9609
	(S, S, C, C)	0.7077	0.9621
	(C, S, C, C)	0.7000	0.9633
	(S, C, S, S)	0.7205	0.9606

D. Spatial Selection or Channel Selection?

LSKNet [42] introduces both the large spatially selective kernel module (LSK-SS) and the large channel-selective kernel module (LSK-CS). Based on this, we develop MoCoLSK-CS module, integrating modality-conditioned weights generated from MCWG pathway with features obtained from LSK-CS after global average pooling, using element-wise addition.

The default configuration of MoCoLSK-Net comprises four stages, each containing a MoCoLSK module (referred to as MoCoLSK-SS). To investigate which selection mechanism is more effective, we configure different selection mechanisms for the four stages: MoCoLSK-SS for spatial selection or MoCoLSK-CS for channel selection. Table IX presents all selection mechanism configurations and their corresponding experimental results, leading to the following conclusions: reconstruction performance of MoCoLSK-Net configured exclusively with MoCoLSK-SS is superior to that of the network configured solely with MoCoLSK-CS modules. This indicates that spatial selection is significantly more critical than channel selection for LST downscaling tasks. Moreover, interleaving MoCoLSK-SS and MoCoLSK-CS modules within MoCoLSK-

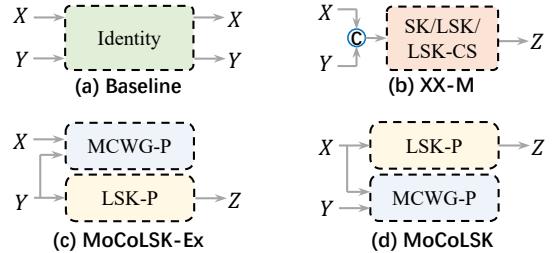


Fig. 8. Thumbnails of different multimodal selection mechanisms (up/down-projection layers are ignored). \mathbf{X} : HR LST, \mathbf{Y} : HR guidance, P: pathway, M: multimodal. (a) Baseline, meaning the output is the same as the input. (b) XX-M represents three modules: SK-M, LSK-M, and LSK-CS-M, which concatenate \mathbf{X} and \mathbf{Y} before feeding them into the original SK [74], LSK [42], and LSK-CS [42] modules. (c) MoCoLSK-Ex (pathway exchange) means the guidance feature \mathbf{Y} enters the LSK pathway, and the output of modality fusion is denoted as $\mathbf{Z} = \mathbf{X} \otimes \mathbf{S}$. (d) Our MoCoLSK module.

TABLE X
COMPARATIVE STUDY OF DIFFERENT MULTIMODAL SELECTIVE MECHANISMS.

No.	Fusion Modules	RMSE↓	CC↑
1	Baseline	0.7405	0.9605
2	SK-M	1.0700	0.9181
3	LSK-M	0.7193	0.9612
4	LSK-CS-M	0.7314	0.9610
5	MoCoLSK-Ex	0.7461	0.9587
6	MoCoLSK	0.7133	0.9613

Net achieves relatively better reconstruction performance compared to configurations using only MoCoLSK-SS or MoCoLSK-CS. For instance, configurations such as (S, C, S, C), (C, S, C, S), (S, S, C, C), and especially (C, S, C, C), support this observation.

E. Comparison of Different Multimodal Selective Mechanisms.

To further explore the effectiveness of our MoCoLSK module, we compare it with several other multimodal variants. Fig. 8 presents the schematic diagrams and configurations of all the variants. The conclusions from Table X are as follows:

1) The SK-M and LSK-CS-M modules demonstrate poor reconstruction performance, reaffirming that a multimodal fusion strategy relying solely on channel selection mechanisms may not be suitable for LST downscaling tasks.

2) The LSK-M module exhibits excellent reconstruction performance, nearly matching that of our MoCoLSK module. This highlights the importance of the spatial selection mechanism, especially for the challenges posed by LST downscaling.

3) The MoCoLSK-Ex module performs worse than our MoCoLSK module and even falls below the baseline. This indicates that feeding LST features, rather than guidance features, into the LSK pathway is critical for reconstruction.

F. Which Normalization Performs Best?

We delve into the impact of three normalization strategies on the performance of a downscaling method: no normalization (denoted as None), Z-score, and Min-max. The definitions of Z-score and Min-max are as follows:

$$\text{Z-score}(X) = \frac{X - \bar{\mu}}{\sigma}, \quad (18)$$

TABLE XI
COMPARISON OF VARIOUS LOSS FUNCTIONS ON MoCoLSK RECONSTRUCTION PERFORMANCE.

Loss	$\times 2$					$\times 4$					$\times 8$				
	RMSE↓	MAE↓	BIAS	CC↑	RSD↓	RMSE↓	MAE↓	BIAS	CC↑	RSD↓	RMSE↓	MAE↓	BIAS	CC↑	RSD↓
L1	0.2902	0.1951	0.0009	0.9937	0.0062	0.5590	0.3883	0.0020	0.9771	0.0218	0.8031	0.5642	0.0027	0.9514	0.0456
SSIM	0.2932	0.2011	0.0112	0.9938	0.0052	0.5668	0.4015	0.0096	0.9770	0.0184	0.8388	0.6052	0.0138	0.9479	0.0422
MS-SSIM	0.2935	0.2009	0.0000	0.9937	0.0059	0.5745	0.4058	0.0003	0.9763	0.0215	0.8229	0.5908	0.0075	0.9504	0.0429
0.3 SSIM + 0.7 L1	0.2906	0.1956	0.0011	0.9937	0.0063	0.5606	0.3899	0.0017	0.9771	0.0219	0.8030	0.5647	0.0024	0.9518	0.0451
0.5 SSIM + 0.5 L1	0.2898	0.1950	0.0010	0.9938	0.0062	0.5642	0.3925	0.0020	0.9769	0.0221	0.7982	0.5615	0.0034	0.9521	0.0446
0.7 SSIM + 0.3 L1	0.2886	0.1943	0.0009	0.9938	0.0061	0.5587	0.3889	0.0020	0.9771	0.0217	0.8028	0.5658	0.5658	0.9514	0.0456
0.84 SSIM + 0.16 L1	0.2890	0.1948	0.0009	0.9938	0.0061	0.5679	0.3956	0.0019	0.9765	0.0225	0.8033	0.5673	0.0031	0.9514	0.0453
0.84 MS-SSIM + 0.16 L1	0.2890	0.1945	0.0011	0.9938	0.0061	0.5614	0.3907	0.0017	0.9770	0.0221	0.8104	0.5700	0.0026	0.9506	0.0465

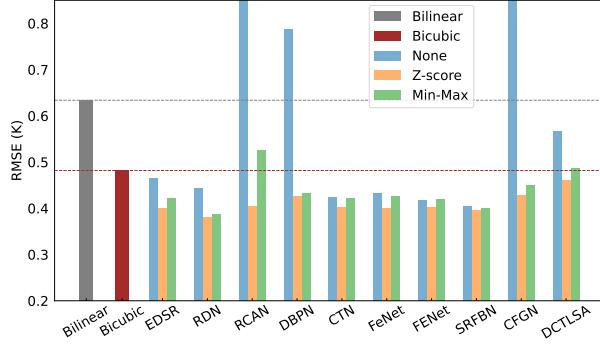


Fig. 9. RMSE comparison between existing SOTA SISR methods ($\times 2$) under different normalization strategies. The experiments follow the default configuration.

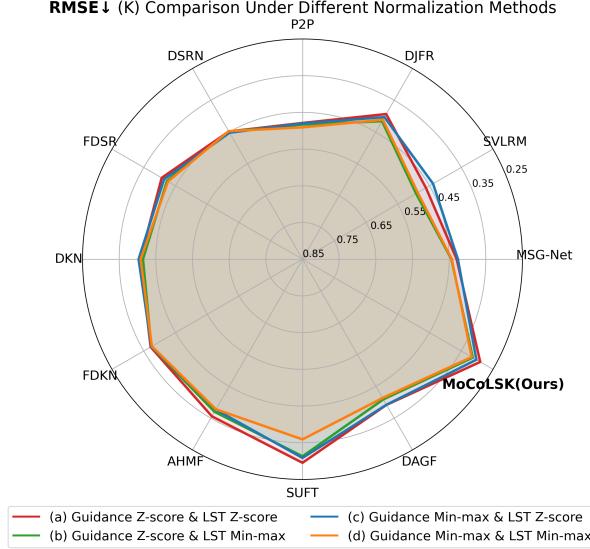


Fig. 10. RMSE comparison between our MoCoLSK-Net and existing SOTA GDSR methods ($\times 2$) under different normalization strategies. The experiments follow the default configuration.

$$\text{Min-max}(X) = \frac{X - \bar{X}_{\min}}{\bar{X}_{\max} - \bar{X}_{\min}}, \quad (19)$$

where X is LST data, $\bar{\mu}$, $\bar{\sigma}$, \bar{X}_{\min} and \bar{X}_{\max} are the mean, standard deviation, minimum value, and maximum value of all LST data in GrokLST dataset, respectively.

Fig. 9 compares the RMSE of ten single-image super-resolution methods applied to LST data using different normalization strategies. It is evident that the None strategy performs significantly worse than the Z-score and Min-max strategies, with Z-score achieving the best results. This strongly highlights

the necessity of normalization. Fig. 10 presents the RMSE results of twelve guided downscaling methods applied to LST data and guidance data using different normalization strategies. It can be observed that the Z-score strategy is particularly suitable for LST data. For example, models employing strategies (a) and (c) exhibit significantly better reconstruction performance compared to those using strategies (b) and (d). Furthermore, models utilizing strategies (a) and (c) achieve commendable reconstruction performance, with strategy (a) yielding the most superior results. This indicates that the Z-score strategy is effective not only for LST data but also for guidance data. In conclusion, we believe that normalization strategies are essential for downscaling tasks. Whether for single-image downscaling or guided downscaling methods, the Z-score strategy is worth considering.

G. Which Loss Works Best?

To explore the impact of different loss functions on MoCoLSK reconstruction performance, we conduct a detailed comparison using several mainstream loss functions, including L1 loss, structural similarity index measure (SSIM) loss, multi scale structural similarity index measure (MS-SSIM) loss, and their combinations. Key insights from Table XI are as follows:

1) Using SSIM or MS-SSIM alone as the loss function to supervise MoCoLSK-Net learning yielded noticeably worse performance than using L1 loss alone, which was observed across all three reconstruction scales.

2) When combining SSIM and L1 loss, MoCoLSK-Net achieved the best reconstruction metrics across all three scales. Furthermore, increasing the weight of the SSIM loss slightly improved reconstruction results, though the improvements were minimal.

3) When combining MS-SSIM and L1 loss, we observed a slight improvement in the $\times 2$ reconstruction task compared to L1 loss alone. However, at larger scales (i.e., $\times 4$, $\times 8$), we saw the opposite trend, with performance even worse than when using MS-SSIM alone.

Overall, we recommend using a combination of SSIM loss with a higher weight and L1 loss with a lower weight for LST downscaling research.

VI. CONCLUSION

To promote the thriving development of LST downscaling, we contribute a comprehensive open-source ecosystem, GrokLST project, which includes GrokLST dataset, a high-resolution benchmark dataset specifically designed for LST

downscaling, and a toolkit featuring over 40 advanced downscaling methods along with various downscaling metrics. Additionally, we propose a novel and effective modality-conditioned multimodal fusion network, MoCoLSK-Net, to address guided LST downscaling challenges. Through extensive quantitative and qualitative comparisons on GrokLST dataset with four machine learning methods, nineteen single-image downscaling methods, and thirteen guided image downscaling methods, MoCoLSK-Net demonstrates superior reconstruction performance, achieving the most accurate LST predictions.

ACKNOWLEDGMENT

The authors would like to thank the International Research Center of Big Data for Sustainable Development Goals (CBAS) for kindly providing the SDGSAT-1 data. We acknowledge the Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP) for their essential resources. Computation is partially supported by the Supercomputing Center of Nankai University (NKSC).

REFERENCES

- [1] Y. Su, C. Zhang, P. Ciais, Z. Zeng, A. Cescatti, J. Shang, J. M. Chen, J. Liu, Y.-P. Wang, W. Yuan *et al.*, “Asymmetric influence of forest cover gain and loss on land surface temperature,” *Nature Climate Change*, vol. 13, no. 8, pp. 823–831, 2023.
- [2] D. Wang, Y. Chen, L. Hu, J. A. Voogt, J.-P. Gastellu-Etchegorry, and E. S. Krayenhoff, “Modeling the angular effect of MODIS LST in urban areas: A case study of Toulouse, France,” *Remote Sensing of Environment*, vol. 257, p. 112361, 2021.
- [3] Y. Bai, N. Bhattacharai, K. Mallick, S. Zhang, T. Hu, and J. Zhang, “Thermally derived evapotranspiration from the surface temperature initiated closure (STIC) model improves cropland GPP estimates under dry conditions,” *Remote Sensing of Environment*, vol. 271, p. 112901, 2022.
- [4] R. Tang, Z. Peng, M. Liu, Z.-L. Li, Y. Jiang, Y. Hu, L. Huang, Y. Wang, J. Wang, L. Jia, C. Zheng, Y. Zhang, K. Zhang, Y. Yao, X. Chen, Y. Xiong, Z. Zeng, and J. B. Fisher, “Spatial-temporal patterns of land surface evapotranspiration from global products,” *Remote Sensing of Environment*, vol. 304, p. 114066, 2024.
- [5] D. Hu, F. Guo, Q. Meng, U. Schlink, S. Wang, D. Hertel, and J. Gao, “A novel dual-layer composite framework for downscaling urban land surface temperature coupled with spatial autocorrelation and spatial heterogeneity,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, p. 103900, 2024.
- [6] X. Ye, J. Hui, P. Wang, J. Zhu, and B. Yang, “A modified transfer-learning-based approach for retrieving land surface temperature from Landsat-8 TIRS data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.
- [7] W. Tang, J. Zhou, J. Ma, Z. Wang, L. Ding, X. Zhang, and X. Zhang, “TRIMS LST: a daily 1 km all-weather land surface temperature dataset for China’s landmass and surrounding areas (2000–2022),” *Earth System Science Data*, vol. 16, no. 1, pp. 387–419, 2024.
- [8] F. Guo, D. Hu, and U. Schlink, “A new nonlinear method for downscaling land surface temperature by integrating guided and Gaussian filtering,” *Remote Sensing of Environment*, vol. 271, p. 112915, 2022.
- [9] 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.
- [10] N. Agam, W. P. Kustas, M. C. Anderson, F. Li, and C. M. Neale, “A vegetation index based technique for spatial sharpening of thermal imagery,” *Remote Sensing of Environment*, vol. 107, no. 4, pp. 545–558, 2007.
- [11] 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.
- [12] P. Sismanidis, B. Bechtel, I. Keramitsoglou, F. Goetsche, and C. T. Kiranoudis, “Satellite-derived quantification of the diurnal and annual dynamics of land surface temperature,” *Remote Sensing of Environment*, vol. 265, p. 112642, 2021.
- [13] S.-B. Duan and Z.-L. Li, “Spatial downscaling of MODIS land surface temperatures using geographically weighted regression: Case study in northern China,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6458–6469, 2016.
- [14] A. S. Fotheringham, W. Yang, and W. Kang, “Multiscale geographically weighted regression (MGWR),” *Annals of the American Association of Geographers*, vol. 107, no. 6, pp. 1247–1265, 2017.
- [15] Y. Peng, W. Li, X. Luo, and H. Li, “A geographically and temporally weighted regression model for spatial downscaling of MODIS land surface temperatures over urban heterogeneous regions,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5012–5027, 2019.
- [16] Y. Hu, R. Tang, X. Jiang, Z.-L. Li, Y. Jiang, M. Liu, C. Gao, and X. Zhou, “A physical method for downscaling land surface temperatures using surface energy balance theory,” *Remote Sensing of Environment*, vol. 286, p. 113421, 2023.
- [17] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *European Conference on Computer Vision*, 2014, pp. 184–199.
- [18] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.
- [19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.
- [20] J. Yuan, H. Jiang, X. Li, J. Qian, J. Li, and J. Yang,

- “Recurrent structure attention guidance for depth super-resolution,” in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3331–3339.
- [21] Z. Wang, Z. Yan, and J. Yang, “SGNet: Structure guided network via gradient-frequency awareness for depth map super-resolution,” in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5823–5831.
- [22] F. Wang, D. Tian, L. Lowe, L. Kalin, and J. Lehrter, “Deep learning for daily precipitation and temperature downscaling,” *Water Resources Research*, vol. 57, no. 4, p. e2020WR029308, 2021.
- [23] U. Mital, D. Dwivedi, J. B. Brown, and C. I. Steefel, “Downscaled hyper-resolution (400 m) gridded datasets of daily precipitation and temperature (2008–2019) for the East–Taylor subbasin (western United States),” *Earth System Science Data*, vol. 14, no. 11, pp. 4949–4966, 2022.
- [24] A. Vaughan, W. Tebbutt, J. S. Hosking, and R. E. Turner, “Convolutional conditional neural processes for local climate downscaling,” *Geoscientific Model Development*, vol. 15, no. 1, pp. 251–268, 2022.
- [25] J. Wu, L. Xia, T. O. Chan, J. Awange, and B. Zhong, “Downscaling land surface temperature: A framework based on geographically and temporally neural network weighted autoregressive model with spatio-temporal fused scaling factors,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, pp. 259–272, 2022.
- [26] Z. Yu, K. Yang, Y. Luo, P. Wang, and Z. Yang, “Research on the lake surface water temperature downscaling based on deep learning,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5550–5558, 2021.
- [27] R. Mukherjee and D. Liu, “Downscaling MODIS spectral bands using deep learning,” *GIScience & Remote Sensing*, vol. 58, no. 8, pp. 1300–1315, 2021.
- [28] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *European Conference on Computer Vision*, 2018, pp. 286–301.
- [29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin Transformer,” in *IEEE International Conference on Computer Vision Workshops*, 2021, pp. 1833–1844.
- [30] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, “SRFormer: Permuted self-attention for single image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 791.
- [31] X. Zhang, Y. Zhang, and F. Yu, “HiT-SR: Hierarchical transformer for efficient image super-resolution,” in *European Conference on Computer Vision*, 2025, pp. 483–500.
- [32] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister, “Discrete cosine transform network for guided depth map super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5697–5707.
- [33] N. Metzger, R. C. Daudt, and K. Schindler, “Guided depth super-resolution by deep anisotropic diffusion,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 237–18 246.
- [34] Z. Xiang, L. Xiao, J. Yang, W. Liao, and W. Philips, “Detail-injection-model-inspired deep fusion network for pansharpening,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
- [35] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, and X. Cao, “Memory-augmented deep unfolding network for guided image super-resolution,” *International Journal of Computer Vision*, vol. 131, no. 1, pp. 215–242, 2023.
- [36] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, “Dual aggregation transformer for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 312–12 321.
- [37] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji, “High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion,” *IEEE Transactions on Image Processing*, vol. 31, pp. 648–663, 2021.
- [38] W. Shi, M. Ye, and B. Du, “Symmetric uncertainty-aware feature transmission for depth super-resolution,” in *ACM International Conference on Multimedia*, 2022, pp. 3867–3876.
- [39] X. Deng and P. L. Dragotti, “Deep convolutional neural network for multi-modal image restoration and fusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3333–3348, 2020.
- [40] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. Zhao, “Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9229–9238.
- [41] Y. Sun, K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, “Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 14–38, 2024.
- [42] Y. Li, X. Li, Y. Dai, Q. Hou, L. Liu, Y. Liu, M.-M. Cheng, and J. Yang, “LSKNet: A foundation lightweight backbone for remote sensing,” *International Journal of Computer Vision*, 2024.
- [43] H. Guo, C. Dou, H. Chen, J. Liu, B. Fu, X. Li, Z. Zou, and D. Liang, “SDGSAT-1: the world’s first scientific satellite for sustainable development goals,” *Science Bulletin*, vol. 68, no. 1, pp. 34–38, 2023.
- [44] N. Li, J. Xu, X. Li, B. Qin, Y. Wang, D. Fu, K. Zhong, and Z. Qin, “A novel land surface temperature retrieval algorithm for SDGSAT-1 images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [45] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, “Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [47] S. Woo, J. Park, J. Lee, and I. S. Kweon, “CBAM: convolutional block attention module,” in *European Conference on Computer Vision*, 2018, pp. 3–19.

- [48] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-projection networks for super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1664–1673.
- [49] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, and J. Yang, “Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information,” in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 10 945–10 954.
- [50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6230–6239.
- [51] I. Loshchilov, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [52] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.
- [53] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.
- [54] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, 2016, pp. 785–794.
- [55] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [56] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [57] S. Wang, T. Zhou, Y. Lu, and H. Di, “Contextual transformation network for lightweight remote-sensing image super-resolution,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [58] Z. Wang, L. Li, Y. Xue, C. Jiang, J. Wang, K. Sun, and H. Ma, “FeNet: Feature enhancement network for lightweight remote-sensing image super-resolution,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [59] P. Behjati, P. Rodriguez, C. F. Tena, A. Mehri, F. X. Roca, S. Ozawa, and J. González, “Frequency-based enhancement network for efficient super-resolution,” *IEEE Access*, vol. 10, pp. 57 383–57 397, 2022.
- [60] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, “Feedback network for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3867–3876.
- [61] T. Dai, M. Ya, J. Li, X. Zhang, S.-T. Xia, and Z. Zhu, “CFGNet: A lightweight context feature guided network for image super-resolution,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 855–865, 2024.
- [62] X. Li, J. Dong, J. Tang, and J. Pan, “DLGSANet: Lightweight dynamic local and global self-attention network for image super-resolution,” in *IEEE International Conference on Computer Vision*, 2023, pp. 12 792–12 801.
- [63] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, “Enriched CNN-Transformer feature aggregation networks for super-resolution,” in *IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 4945–4954.
- [64] H. Choi, J. Lee, and J. Yang, “N-gram in swin transformers for efficient lightweight image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2071–2081.
- [65] K. Zeng, H. Lin, Z. Yan, and J. Fang, “Densely connected transformer with linear self-attention for lightweight image super-resolution,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.
- [66] T.-W. Hui, C. C. Loy, , and X. Tang, “Depth map super-resolution by deep multi-scale guidance,” in *European Conference on Computer Vision*, 2016, pp. 353–369.
- [67] J. Pan, J. Dong, J. S. Ren, L. Lin, J. Tang, and M.-H. Yang, “Spatially variant linear representation models for joint filtering,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1702–1711.
- [68] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Joint image filtering with deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1909–1923, 2019.
- [69] R. D. Lutio, S. D’aronco, J. D. Wegner, and K. Schindler, “Guided super-resolution as pixel-to-pixel transformation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8829–8837.
- [70] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, “Image super-resolution via dual-state recurrent networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1654–1663.
- [71] B. Kim, J. Ponce, and B. Ham, “Deformable kernel networks for joint image filtering,” *International Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.
- [72] Y. Yang, Q. Cao, J. Zhang, and D. Tao, “CODON: on orchestrating cross-domain attentions for depth super-resolution,” *International Journal of Computer Vision*, vol. 130, no. 2, pp. 267–284, 2022.
- [73] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, “Deep attentional guided image filtering,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.
- [74] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.



Qun Dai received her M.S. degree in Avionics Engineering in 2017. Subsequently, she worked as an Avionics Engineer at China Eastern Airlines Technology Co., Ltd., where she gained extensive experience in the field of aviation electronics and systems integration. Currently, she is a Ph.D. candidate at Nanjing University of Science and Technology, where she focuses on advancing the state-of-the-art in image restoration techniques for enhanced object detection performance.



Chunyang Yuan received the B.S degree in energy and power engineering from Henan Polytechnic University, Henan, China, in 2021. He is currently pursuing a master's degree in Computer Science and Technology at Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include machine learning, SAR image processing, and computer vision.



Kang Ni (Member, IEEE) received the M.S. degrees from Changchun University of Technology, Jilin, China, in 2016, and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Jiangsu, China, in 2020. He is an Associate Professor with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, and also a member with the Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing. His research interests include machine learning, SAR image processing, and computer vision.



Yimian Dai (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013 and 2020, respectively. From 2021 to 2024, he was a Postdoctoral Researcher with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He is currently an Associate Professor with the College of Computer Science, Nankai University, Tianjin, China. His research interests include computer vision, deep learning, and their applications in remote sensing. For more information, please visit the link (<https://yimian.grokcv.ai/>).



Jianhui Xu received the Ph.D. degree in multi-source remote sensing data assimilation from Wuhan University, Wuhan, China, in 2015. He is currently an Associate Professor with the Guangzhou Institute of Geography, Guangdong Academy of Sciences. He has published numerous SCI papers in prestigious journals, such as *The Science of The Total Environment*, *Journal of Geophysical Research Atmospheres and Building and Environment*, in the area of data fusion and urban remote sensing. His research interests include data fusion and assimilation, land surface temperature and urban remote sensing.



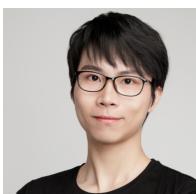
Xuxuan Li is currently a Ph.D. student at the Department of Computer Science, Nankai University, China. He graduated from University College London (UCL) with a first-class degree in Computer Science. He was the champion of the Second Jittor Artificial Intelligence Challenge in 2022, was awarded 2nd place in Facebook Hack-a-Project in 2019 and was awarded 2nd place in the Greater Bay Area International Algorithm Competition in 2022. His research interests include neural architecture design, and remote sensing object detection.



Xiangbo Shu (Senior Member, IEEE) received the Ph.D. degree from Nanjing University of Science and Technology, Nanjing, China, in 2016.

From 2014 to 2015, he worked as a Visiting Scholar at the National University of Singapore, Singapore. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include computer vision, and multimedia. He has authored over 100 journals and conference papers in these areas.

Dr. Shu is a member of ACM and a Senior Member of CCF. He has received the Best Student Paper Award in International Conference on MultiMedia Modeling (MMM) 2016 and the Best Paper Runner-Up in ACM MM 2015. He has served on an editorial boards for IEEE TNNLS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO (TCSVT), and Information Sciences.



Xiang Li is an Associate Professor in College of Computer Science, Nankai University. He received the PhD degree from the Department of Computer Science and Technology, Nanjing University of Science and Technology (NJUST) in 2020. There, he started the postdoctoral career in NJUST as a candidate for the 2020 Postdoctoral Innovative Talent Program. In 2016, he spent 8 months as a research intern in Microsoft Research Asia, supervised by Prof. Tao Qin and Prof. Tie-Yan Liu. He was a visiting scholar at Momenta, mainly focusing on monocular perception algorithm. His recent works are mainly on: neural architecture design, CNN/Transformer, object detection/recognition, unsupervised learning, and knowledge distillation. He has published 20+ papers in top journals and conferences such as TPAMI, CVPR, NeurIPS, etc.



Jian Yang received the PhD degree from Nanjing University of Science and Technology (NJUST) in 2002, majoring in pattern recognition and intelligence systems. From 2003 to 2007, he was a Postdoctoral Fellow at the University of Zaragoza, Hong Kong Polytechnic University and New Jersey Institute of Technology, respectively. From 2007 to present, he is a professor in the School of Computer Science and Technology of NJUST. His papers have been cited over 50000 times in the Scholar Google. His research interests include pattern recognition and computer vision. Currently, he is/was an associate editor of Pattern Recognition, Pattern Recognition Letters, IEEE Trans. Neural Networks and Learning Systems, and Neurocomputing. He is a Fellow of IAPR.