

Landsat 9 Land Surface Temperature Retrieval Based on SE-ResUNet

Pan Huang-Fu-Yu^{1,2}, Tang Guo-Liang¹, Zhang Xu-Dong¹, Chen Hong-Yi^{1*}, Qi Hong-Xing^{1*}

(1. Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310016, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Accurate retrieval of Land Surface Temperature (LST) from satellite thermal infrared data remains challenging due to the reliance of physical models on real-time atmospheric profiles and the difficulty in characterizing surface emissivity over heterogeneous landscapes. To address these limitations, this study proposes SE-ResUNet, a deep learning framework for Landsat 9 thermal infrared images. To overcome the scarcity of large-scale in-situ measurements for training, we construct a high-quality synthetic dataset by coupling the MODTRAN 5 radiative transfer model with ERA5 atmospheric reanalysis data. The network adopts a U-Net encoder-decoder structure with a modified ResNet50 backbone to capture multi-scale features. Squeeze-and-Excitation (SE) attention modules are embedded in the residual blocks and physical prior knowledge is directly added to the input tensor. By integrating skip connections and an adaptive calibration mechanism for thermal signals under physical constraints, our method achieves precise pixel-by-pixel temperature reconstruction. Experiments show that SE-ResUNet effectively mitigates the overfitting problem linked to spatial autocorrelation. The model shows strong robustness against simulated noise and complicated terrain variability. Evaluations on multiple datasets show that it achieves a Root Mean Square Error (RMSE) of around 0.7 K and a Mean Absolute Error (MAE) of 0.5 K. These results confirm the effectiveness of SE-ResUNet as a high-precision, end-to-end solution for LST retrieval without real-time external atmospheric inputs at the inference stage.

Key words: Land Surface Temperature Retrieval, Landsat 9, Deep Learning, ResNet, U-Net

PACS:

Introduction

LST is a key variable in determining terrestrial energy and water exchanges, and is a widely used parameter for monitoring the global climate, modeling hydrology, and evaluating ecosystems^[1-2]. In the context of human-induced changes, LST further supports Urban Heat Island effects and evaluate the severity of agricultural droughts, driving the demand for thermal data with high spatiotemporal resolution^[3-4].

The Landsat satellite series provides one of the biggest datasets of medium-to-high-resolution thermal infrared data in remote sensing^[5]. In 2021, Landsat 9 was launched, ensuring observational continuity through the Operational Land Imager-2 and the Thermal Infrared Sensor-2. Compared with TIRS-1, TIRS-2 offers improved radiometric accuracy, reduced stray light contamination, and higher signal-to-noise ratios^[6]. These improvements help build a solid methodological framework for precise LST retrieval from Landsat data^[7].

Retrieving LST from Top-of-Atmosphere (TOA) ra-

diance is still an ill-posed and non-linear problem, even with advances in remote sensing instruments^[8]. The conventional physics-based approaches to retrieve LST from radiance data include the Split-Window Algorithm (SWA) and the Single-Channel Algorithm (SCA), both of which are derived from the Radiative Transfer Equation (RTE). A major limitation of physics-based approaches is the reliance on atmospheric parameters that are rarely available at pixel-level accuracy^[9]. Conventional machine learning methods do not use complex physical models to derive LST. In many cases, they work at pixel level, thereby failing to incorporate the contextual and textural information needed for generalization among different landscapes^[10].

To overcome these challenges, Deep Convolutional Neural Networks (DCNNs) have emerged as an alternative in quantitative remote sensing. Unlike shallow statistical models, DCNNs can approximate complex non-linear radiative transfer processes while preserving spatial structure^[11]. This capability helps build robust LST retrieval that minimizes explicit dependence on external at-

Received date: 2026-04-25,

收稿日期: 2026-04-25,

Foundation items: Supported by Intelligent Preprocessing Technology for Spaceborne Remote Sensing Spectral Data (B02006C021035), the Hangzhou Institute for Advanced Study, UCAS, China.

Biography: PAN Huangfuyu (1999-), male, ZheJiang, master. Research area involves remote sensing image processing and artificial intelligence. E-mail: panhuangfuyu23@mails.ucas.ac.cn.

* **Corresponding author:** E-mail: hongyichen@ucas.ac.cn

mospheric parameters.

In this paper, we propose SE-ResUNet, a deep learning model that incorporates physical constraints. The network learns an end-to-end non-linear mapping between a multi-source input tensor (multispectral reflectance, thermal radiance and physical priors) and pixel-wise LST. We further incorporate the SE attention mechanism that dynamically adjusts channel weights while suppressing atmospheric noise^[12]. The integration of high-level semantic and low-level spatial features enables robust, high-precision LST retrieval without dependence on real-time atmospheric data at the inference stage.

1 Related Works

1.1 Traditional Physical Retrieval Algorithms

The retrieval of LST from Landsat data has long relied on physical principles, especially the Radiative Transfer Equation (RTE) and Planck's Law of blackbody radiation. Variations in the physical formulation of atmospheric correction and the treatment of downwelling/upwelling radiances have given rise to three principal algorithms for Landsat-based LST retrieval: the direct RTE method, the SCA, and the SWA.

The Direct RTE utilizes a physically rigorous approach to retrieve LST. It requires precise synchronous atmospheric profiles to estimate transmittance, upwelling and downwelling radiance^[13]. Barsi et al. developed a web-based Atmospheric Correction Tool that combines NCEP global re-analysis data (including atmospheric parameters) with MODTRAN RTE codes to correct thermal imagery data^[14-15]. The accuracy of this method is directly correlated to the accuracy of the external atmospheric profile data (i. e. , water vapour, air temperature) used to generate the LST estimation (i. e. , RTE). As a result, the direct RTE method is limited by the availability of accurate atmospheric profiles synchronized with the satellite overpass.

For sensors that only have one thermal channel, the SCA provides a more practical methodology for retrieving LST. This algorithm either employs empirical approximations or linearizes the Planck function and therefore reduces the radiative transfer process to one band to derive LST from that band. Qin et al. linearize the Planck function and parameterize, which linearizes the Planck function and parameterizes atmospheric transmittance and mean effective temperature^[16]. Another example of a SCA method is the single-channel approach to retrieve LST based upon Water Vapour Content (WVC) without the requirement of completing the RTE simulation proposed by Jiménez-Muñoz and Sobrino^[17]. Similar to the Direct RTE approach, the accuracy of the SCA is dependent upon the accuracy of the near-surface meteorological data, which makes the SCA approach susceptible to error due to inaccuracies in atmospheric parameterization.

The SWA exploits the differential atmospheric absorption between two adjacent thermal bands to correct for water vapour effects without real-time radiosonde data. The generalized formulation of the SWA by Wan and Dozier serves as the basis for standard MODIS prod-

ucts^[9,18]. The SWA typically utilizes the refined coefficients of Jiménez-Muñoz et al.^[8] and Du et al.^[19] for LSE estimation. Although this method is theoretically sound, it still requires land surface emissivity as input to resolve the ill-posed inversion problem.

A common limitation of these physical approaches is their reliance on external data, such as water vapor content and land surface emissivity. Obtaining such data at the spatial and temporal resolution of the satellite remains difficult. Input parameter uncertainties propagate through the retrieval chain, limiting algorithm performance across diverse land covers and seasons.

1.2 Artificial Intelligence-Driven Retrieval Algorithms

Machine learning offers a data-driven alternative to physics-based LST retrieval. Traditional machine learning algorithms such as Random Forest (RF)^[20], Support Vector Regression (SVR)^[21], and Cubist models^[22] have been applied to simulate complex relationships between remote sensing data and LST estimates. For example, Hutengs et al.^[23] demonstrated the efficacy of RF in downscaling LST by effectively handling high-dimensional covariates. SVR handles noisy data through kernel mapping, which is useful when training samples are scarce^[24]. However, these methods are pixel-based and ignore spatial context. This limits their performance in heterogeneous landscapes where thermal continuity between pixels is low.

Recently, neural networks, with their universal approximation capability, have been applied as data-driven alternatives to physical radiative transfer models. Early research by Mao et al. utilized NNs to separate temperature and emissivity from ASTER data, effectively demonstrating the potential of NNs in solving non-linear radiative transfer problems^[25]. Shen et al. proposed a deep NN-based LST retrieval algorithm for Gaofen-5B, demonstrating that deep learning models can achieve robust performance without explicit atmospheric priors by learning interference patterns directly from data.^[26] However, prevailing NN-based approaches for LST retrieval, such as shallow Multi-Layer Perceptrons and basic CNNs, are fundamentally constrained by their limited receptive fields. By processing pixels in isolation or within confined local windows, these architectures fail to capture the long-range spatial dependencies essential for resolving complex thermal boundaries.

Combining residual networks with channel attention mechanisms offers a potential solution, as the residual structure supports deeper feature extraction while channel attention selectively emphasizes informative spectral bands.

2 Methods

2.1 Theoretical Reasoning

The input bands are selected based on their physical roles in land-atmosphere radiative transfer. The split-window technique uses the differential absorption between Band 10 and Band 11 to correct for atmospheric effects. Since it is difficult to measure surface emissivity

directly, we include OLI-2 bands (visible to SWIR) as proxy inputs for surface properties. These bands encode land-cover type and moisture conditions that correlate with emissivity.

Traditional SWA is derived from the linearization of the RTE. The TOA radiance L_i observed by sensor band i is expressed as:

$$L_i = \tau_i \varepsilon_i B_i(T_s) + L_i^\uparrow + \tau_i (1 - \varepsilon_i) L_i^\downarrow, \quad (1)$$

where τ_i is atmospheric transmittance, ε_i is surface emissivity, $B_i(T_s)$ is the Planck function for surface temperature T_s , and L_i^\uparrow , L_i^\downarrow represent upwelling and downwelling atmospheric radiance, respectively^[13].

To calculate T_s , standard SW algorithms simplify Eq. (1) into a linear combination of brightness temperatures (T_{10} , T_{11}):

$$T_s = c_0 + c_1 T_{10} + c_2 (T_{10} - T_{11}) + c_3 (T_{10} - T_{11})^2. \quad (2)$$

This formulation has two drawbacks: (1) the coefficients c_n are fixed or depend on external water vapor estimates, and (2) it requires accurate emissivity, which is difficult to obtain dynamically^[8].

For deep learning methods, the inverse problem was redefined as a non-linear mapping function in which the neural network approximated the inverse physics without explicitly parameterizing ε or atmospheric profiles.

$$\widehat{T}_s = F_\theta(L_{TIRS}, R_{OLI}, \mathbf{X}_{aux}). \quad (3)$$

Here, L_{TIRS} represents thermal radiance, and R_{OLI} denotes optical reflectance. Unlike the linear approximation in Eq. (2), the deep learning model F_θ (parameterized by weights θ) learns emissivity corrections from the optical bands and uses the spatial context of thermal bands to resolve atmospheric heterogeneity. This end-to-end formulation avoids the error propagation of stepwise physical inversions.

2.2 Data Preparation and Simulation Pipeline

2.2.1 Study Area and Representative Land Covers

To evaluate the model generalization, we selected experimental regions covering five major land cover types. Details are given in Table 1.

2.2.2 Data Sources and Preprocessing Pipeline

Primary satellite data were obtained from the Land-

sat 9 mission via the USGS official portal. We utilized both Level-1 (L1) products for raw radiometric calibration and Level-2 (L2) products for surface reflectance validation. The preprocessing workflow shown in Figure 1 follows a rigorous protocol to transform raw Digital Numbers (DN) into structured input tensors:

1. Radiometric Calibration: Gain and bias coefficients from the metadata are applied to convert DN to at-sensor radiance (L_λ) and TOA reflectance.

2. Geometric Correction and Resampling: The 100 m TIRS bands are resampled to 30 m using bilinear interpolation to match the OLI-2 spatial resolution. All data are reprojected to the Universal Transverse Mercator coordinate system.

3. Cloud Masking: Using the QA bands, we mask pixels contaminated by cloud or cloud shadow, ensuring that only clear-sky pixels for model input.

2.2.3 Physics-Informed Dataset Construction via MODTRAN 5

Given the scarcity of large-scale, pixel-aligned ground truth LST data, we used the MODTRAN 5 radiative transfer model to generate synthetic training samples through forward simulation, calculating atmospheric transmittance (τ), upwelling radiance (L_i^\uparrow), and downwelling radiance (L_i^\downarrow) at a spectral resolution of 0.2 cm^{-1} . This dataset contains approximately 40,000 "Atmosphere-Surface-Sensor" coupled samples.

We use specific atmospheric profile from the ERA5 reanalysis dataset rather than standard atmospheric profiles. ERA5 provides high spatio-temporal resolution ($0.25^\circ \times 0.25^\circ$, hourly). We applied stratified random sampling across various climate zones (tropical, temperate, frigid) and seasons to extract profiles that reflect real-world atmospheric variability.

These specific profiles include vertical distributions of temperature, humidity, and pressure, covering a wide range of Column Water Vapor (CWV) from 0.5 to 6.0 g/cm^2 , which is crucial for capturing non-linear atmospheric attenuation.

To ensure diverse surface and atmospheric conditions, we implemented the following strategies:

1. Surface Emissivity: As shown in Figure 2, we

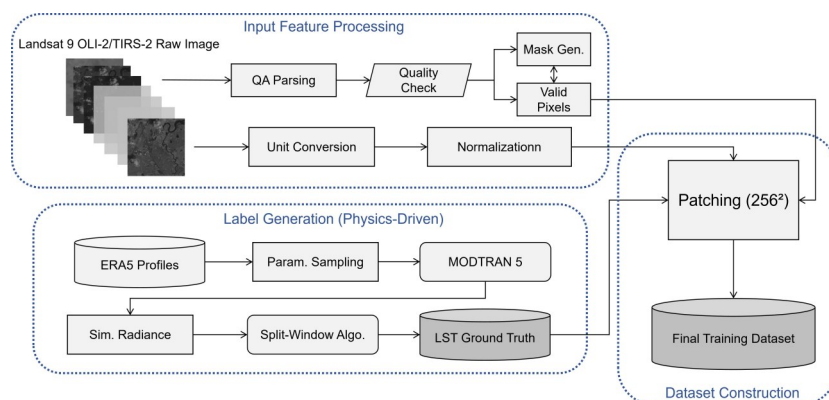
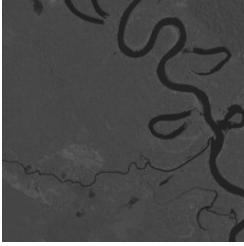
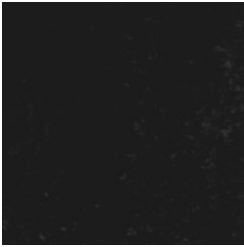

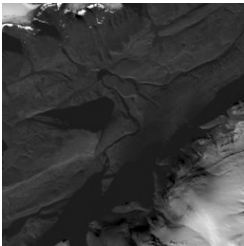
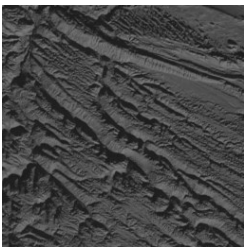


Figure 1 The data generation and processing workflow: from raw Landsat 9 and ERA5 data to MODTRAN simulation and final tensor construction

图 1 数据生成和处理工作流程: 从原始 Landsat 9 和 ERA5 数据到 MODTRAN 模拟和最终张量构建

Table 1 Overview of the validation dataset for heterogeneous complex land surfaces
表 1 异质性复杂地表验证数据集概况

Surface type	Sample	Description of Radiation	Range of Temperature	Sample Capacity
Forest		High vegetation coverage, high and stable emissivity ($\epsilon \approx 0.98$). The spectral characteristics are significantly affected by chlorophyll, and the thermal texture changes with vegetation density.	285 K–305 K	2272
Water Body		Extremely high emissivity ($\epsilon \approx 0.99$), large thermal inertia, and small temperature difference between day and night. Strong near-infrared absorption facilitates spectral separation from other surfaces.	278 K–295 K	2245
Built-up Area		Low emissivity and extremely heterogeneous (concrete, asphalt mixed). There is a significant urban heat island effect, with broken thermal texture and serious mixed pixels.	295 K–325 K	2236
Snow and Frozen Ground		High reflectivity, low surface temperature. Water vapor content is usually low, atmospheric transmission is high, but it is also susceptible to thin cloud disturbances.	250 K–273 K	1104
Desert		The emissivity of soil fluctuates greatly, which is related to soil moisture and surface roughness, and the surface temperature rises rapidly during the day and the thermal radiation signal is strong.	290 K–330 K	2267

utilized the ASTER Spectral Library, extracting emissivity curves for diverse surfaces, including water, sand, soil, vegetation, and urban man-made materials (concrete/asphalt).

2. **Thermodynamic State:** Instead of a fixed value, LST was simulated by adding a dynamic random perturbation (ΔT ranging from -10 K to +20 K) to the bottom-layer air temperature, accounting for both nocturnal cooling and daytime solar heating.

3. **Observation Geometry:** We accounted for varying view angles by setting the View Zenith Angle (VZA) from 0° to 7.5° for Landsat 9 (near-nadir) to simulate at-

mospheric path-length effects.

4. **Sensor Integration:** The high-resolution simulated radiation was integrated with the specific Spectral Response Functions (SRF) of Landsat 9 TIRS-2 to ensure the physical consistency of the simulated data with actual satellite observations.

2.3 Proposed Method: SE-ResUNet Architecture

The SE-ResUNet model illustrated in Figure 3 learns a non-linear mapping from multi-source inputs to pixel-wise LST, bypassing explicit atmospheric inversion.

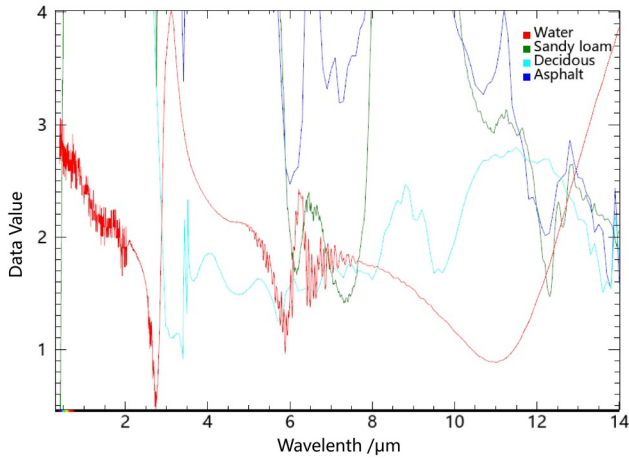


Figure 2 Surface emissivity curve of materials in the ASTER Spectral Library

图2 ASTER 光谱库材料表面发射率曲线图

1. 12 -Channel Input Tensor Design

The 12 input channels are grouped as follows: (1) Spectral & Thermal (B2 - B7, B10 - B11) provide the fundamental radiometric basis and land cover context. (2) ST_EMIS (NDVI-derived emissivity) provides a coarse physical baseline, while ST_CDIST (distance to

clouds) helps the model account for environmental neighborhood effects. (3) QA_PIXEL (quality assessment band) flags invalid pixels such as clouds and cloud shadows, enabling the network to learn to disregard contaminated observations. (4) The Split-Window Difference ($T_{B10} - T_{B11}$) is included as a channel to serve as a direct proxy for atmospheric water vapor absorption.

2 Encoder with Custom Bottleneck Internals

The encoder utilizes a modified ResNet50 backbone^[27]. Each Custom Bottleneck Block consists of: (1) 1×1 Convolution: Performs dimensionality reduction to minimize computational overhead. (2) 3×3 Convolution: Extracts spatial features and thermal gradients in the reduced dimensional space. (3) SE modules recalibrate channel dependencies allowing the network to emphasize informative channels and down-weight less relevant ones.

3 Decoder and Progressive Reconstruction

The decoder follows a U-Net style symmetry to restore spatial resolution. Skip Connections concatenate the encoder and decoder, fusing low-level spatial details

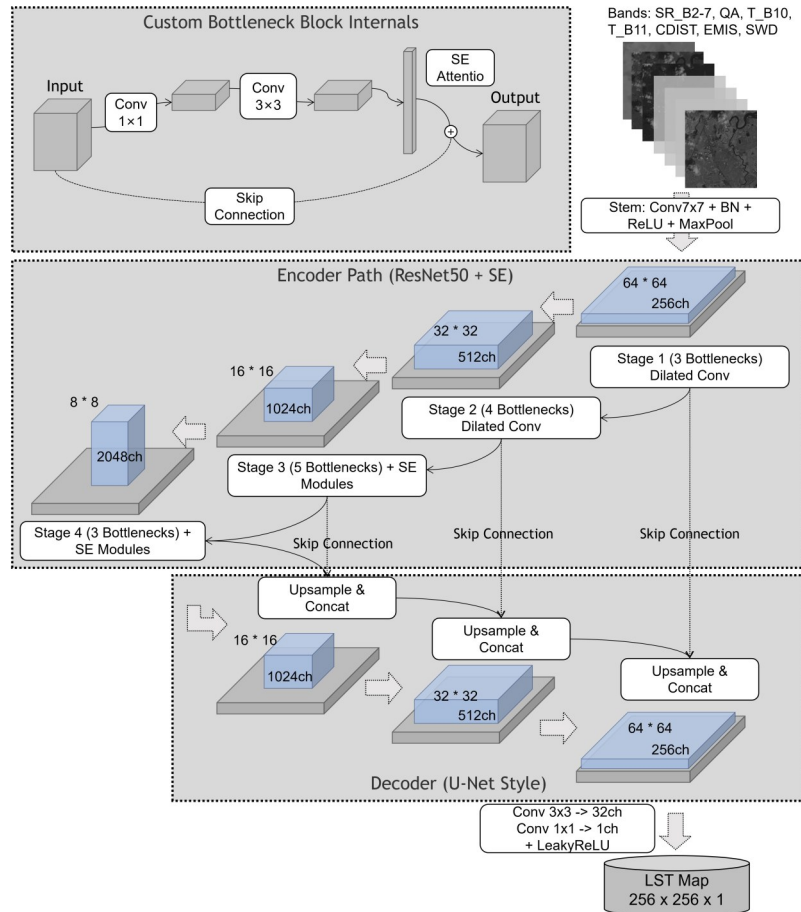


Figure 3 SE-ResUNet model overall architecture diagram

图3 SE-ResUNet 模型总体架构图

with high-level semantic features. Additionally, Leaky ReLU is used throughout the decoder to avoid dead neurons^[28].

4. Output Layer and Physical Anchoring

The final layer uses a 1×1 convolution to produce the LST map. To accelerate convergence and ensure physical feasibility, we incorporate a global learnable bias initialized at ~ 295 K.

2.4 Implementation and Evaluation Strategy

The model was implemented in PyTorch and trained on an NVIDIA GeForce RTX 4090 GPU.

We evaluate the model at four levels:

(a) Homogeneous Terrain Benchmark: A desert only subset to test whether spatial autocorrelation inflates accuracy in pixel-wise models.

(b) Heterogeneous Generalization Test: a dataset composed only of various types of land cover (vegetated, water, artificial, forest, and frozen soil) created to test how well the model generalizes.

(c) Radiometric Robustness Stress Test: Gaussian noise (at levels of 0-30%) and stripe artifacts are added to the test set to simulate sensor degradation and atmospheric effects on the observations.

(d) Ablation study: key components (SE module, skip connections, SWD channel, loss function) are removed individually to measure their contributions. An additional experiment disables ST_EMIS to check for data leakage.

A summary of the hyperparameters used throughout the training process is indicated in Table 2. The input image dimension is 256×256 pixels, with a total of 12 input channels. The dataset was randomly split into training and validation data sets at an 8:2 ratio.

The performance of the model was assessed through the RMSE, the MAE, and the Coefficient of Determination (R^2) and the calculation methods for these measures were specified below:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (5)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad (6)$$

Table 2 Detailed training hyperparameters

表 2 详细训练超参数表

Hyperparameter	Value	Description
Input Size	256×256×12	Height × Width × Channels
Batch Size	32	Adjusted based on GPU memory
Epochs	150	Maximum number of epochs
Optimizer	Adam	$\beta_1=0.9, \beta_2=0.999$
Initial Learning Rate	1×10^{-4}	Decays with factor 0.5
Weight Decay	1×10^{-5}	L2 regularization term
Dropout Rate	0.3	Applied in deep encoder layers
Bias Init	295.0	Initial output temperature baseline

where y_i is the ground truth value of LST from the simulation or in-situ measurement, \hat{y}_i is the predicted LST, and n is the total number of valid samples. RMSE and MAE measure the retrieval accuracy, while R^2 indicates the coefficient of determination.

3 Results And Discussion

3.1 Retrieval Performance To assess the performance of the proposed model under challenging radiometric conditions, we conducted a comparative study focused on the desert regions of Xinjiang, China, primarily within the Tarim Basin and the Taklamakan Desert.

These arid landscapes have extreme surface temperatures and highly variable spectral emissivity, which challenge conventional retrieval methods. Figure 4 compares the MODTRAN ground truth, SE-ResUNet predictions, and the official Landsat 9 ST_B10 product. The SE-ResUNet predictions closely match the ground truth and preserve fine-scale thermal patterns. By contrast, the standard ST_B10 product shows noticeable deviations in certain sub-regions. The difference arises because the official product relies on external emissivity priors which could potentially introduce biases across heterogeneous barren surfaces.

To quantify these differences, we compare scatter density plots, difference histograms, and 2-D spatial residual maps in Figure 5 and Figure 6.

Figure 5 shows two sub-regions (rows): predicted LST (left), MODTRAN ground truth (middle), and official ST_B10 (right). SE-ResUNet is spatially consistent

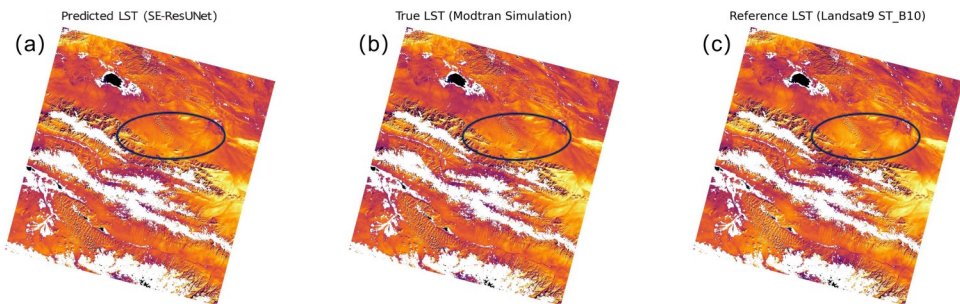


Figure 4 Global comparison of Predicted LST, MODTRAN simulation results, and ST_B10 data. : (a) SE-ResUNet inference results; (b) MODTRAN simulation result; (c) Landsat 9 ST_B10 thermodynamic map

图 4 SE-ResUNet 推理结果与 MODTRAN 模拟结果、ST_B10 数据的全局对比图: (a) SE-ResUNet 推理结果; (b) MODTRAN 模拟结果; (c) Landsat 9 ST_B10 热力图

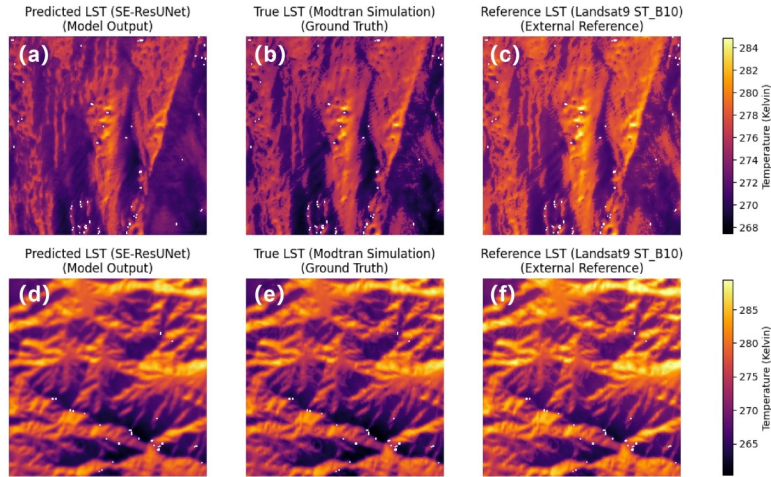


Figure 5 Local comparison of SE-ResUNet retrieval results, MODTRAN simulation results, and ST_B10 data: (a) SE-ResUNet inference results of subdomain 1; (b) MODTRAN simulation result of subdomain 1; (c) Landsat 9 ST_B10 thermodynamic map of subdomain 1; (d) SE-ResUNet Inference Results of subdomain 2; (e) MODTRAN simulation result of subdomain 2; (f) Landsat 9 ST_B10 thermodynamic map of subdomain 2

图5 SE-ResUNet推理结果与MODTRAN模拟结果、ST_B10数据的局部对比图:(a)子域1的SE-ResUNet推理结果;(b)子域1的MODTRAN模拟结果;(c)子域1的Landsat 9 ST_B10热力学图;(d)子域2的SE-ResUNet推理结果;(e)子域2的MODTRAN模拟结果;(f)子域2的Landsat 9 ST_B10热力学图

with ST_B10 but numerically closer to the MODTRAN truth.

The scatter plots in Figure 6(a) show R^2 of 0.83 - 0.99 between SE-ResUNet and ST_B10. The high-density point clusters are closely aligned with the 1:1 diagonal line, validating the macro-scale accuracy of the proposed model. However, a systematic offset is present across the temperature range, a trend further corroborated by the difference histogram in Figure 6(b). The histogram exhibits a near-normal distribution predominantly situated within the negative domain, indicating the presence of a systematic discrepancy between the two datasets under specific environmental conditions. The spatial residual maps in Figure 6(c) reveal that the bias is spatially structured between the SE-ResUNet predictions and the official values, with a systematic bias reaching -1.91 K. In regions characterized by sharp emissivity gradients—such as dune peripheries and the interfaces between bare rock and sandy terrain—pronounced red or blue clusters (representing $\Delta T > 1.5$ K) are evident. The primary driver of this spatial non-stationarity error is the official algorithm's heavy reliance on external emissivity priors. In arid regions such as deserts, the ASTER GED emissivity data utilized in the official products often suffers from insufficient spatio-temporal resolution or significant temporal latency. Consequently, it fails to accurately capture micro-scale variations in surface composition, thereby inducing characteristic "patch-like" temperature biases. In contrast, SE-ResUNet learns surface properties implicitly from multi-spectral inputs, reducing these emissivity-related biases.

Figure 7 shows the comparison of ground truth and predicted LST. The SE-ResUNet model has the best accuracy with $R^2=0.986$, $RMSE=0.55$ K and $MAE=0.37$ K. Both metrics are well below the 1 K benchmark for satellite LST retrieval.

3.2 Comparative Analysis and Robustness Evalua-

tion

We evaluate the model at three levels: (i) an evaluation on homogeneous terrain to assess potential overfitting, (ii) a generalization experiment on complex and diverse terrains, and (iii) a robustness test with simulated sensor noise and atmospheric disturbance. For comparison, five established regression methods were also employed to retrieve LST: Linear Regression (LR), RF, LightGBM, Multi-Layer Perceptron (MLP) and the traditional U-Net.

3.2.1 Multiple surface generalization tests

The initial training and assessment were conducted solely on a homogeneous desert dataset characterized by smooth spectral and thermal gradients. As shown in Figure 9, tree-based machine learning models, particularly RF, exhibited abnormally low error metrics ($MAE = 0.05$ K), apparently outperforming SE-ResUNet ($MAE = 0.37$ K).

An MAE of ~ 0.05 K is physically unrealistic for satellite LST retrieval. Standard sensor noise ($NE\Delta T$) for Landsat 9 is approximately 0.1-0.2 K, and uncertainties in atmospheric correction and surface emissivity typically limit the theoretical accuracy to 0.5-1.0 K. An MAE of 0.05 K suggests that the model is not "retrieving" LST based on physics, but rather "replicating" the labels through statistical memorization.

To test whether this reflects genuine learning or overfitting to the homogeneous training domain, we expanded the evaluation to complex, heterogeneous scenarios. All models were retrained on a diverse dataset encompassing vegetation, water bodies, built-up areas, forests, and frozen soil. As shown in Figure 9, the results diverge while SE-ResUNet maintained high stability, the performance degradation of MLP was particularly pronounced and the accuracy of RF and LightGBM deteriorated markedly.

We analyzed spatial surface temperature (LST)

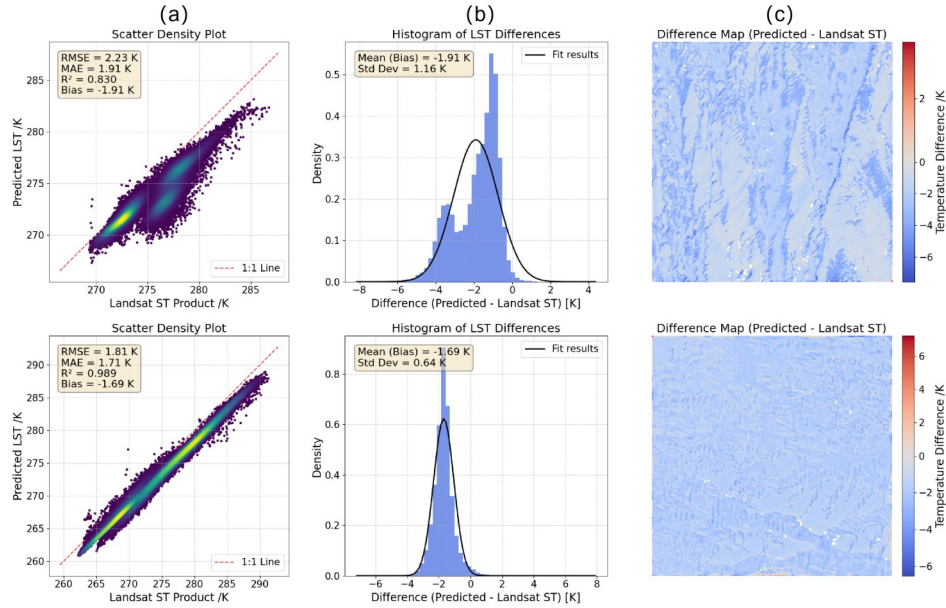


Figure 6 Quantitative discrepancy analysis between SE-ResUNet retrieval results and the official Landsat 9 Level-2 standard products: (a) Scatter plot of the two subdomains; (b) Difference histogram of the two subdomains; (c) Visual difference map of the two subdomains
图6 SE-ResUNet反演结果与Landsat 9 Level-2官方标准产品的定量差异分析图:(a)两个子域的散点图;(b)两个子域的差异直方图;(c)两个子域的视觉差值图

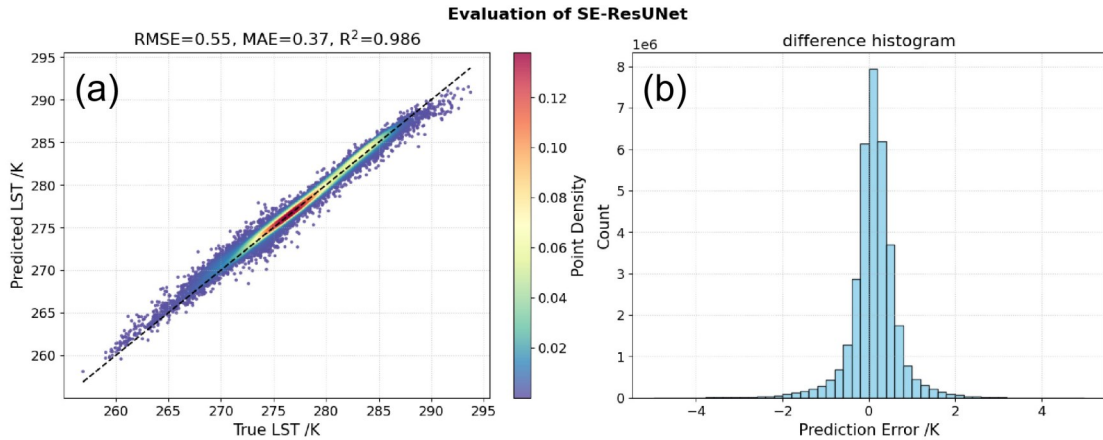


Figure 7 The scatter density plot and difference histogram between the predicted LST and the ground truth: (a) Scatter plot of SE-ResUNet; (b) Difference histogram of SE-ResUNet
图7 预测地表温度与真实地表温度的散点密度图和差异直方图:(a) SE-ResUNet的散点图;(b) SE-ResUNet的差异直方图

maps and evaluated the models' ability to resolve complex topographical features and land-cover boundaries. As shown in Figure 10, the sequential presentation includes:

(a) Linear Regression: As a baseline, LR fails to capture spatial variability. The output is significantly blurred, losing high-frequency textures and failing to reflect the temperature fluctuations associated with terrain relief.

(b) and (c) Traditional Machine Learning (RF & LightGBM): While these models capture the macroscopic temperature distribution, they exhibit noticeable "speckle" noise at the pixel level. As isolated pixel-wise regressors, they lack spatial context, leading to discontinuous temperature gradients at terrain transitions that appear statistically fitted rather than physically consis-

tent.

(d) and (e) CNN-based Models (Basic U-Net & SE-ResUNet): These models leverage regional receptive fields to perceive spatial textures. The SE-ResUNet performs resolves thermal contrasts between sunny and shady slopes and reproduce smooth, physically consistent gradients.

(f) Deep MLP: As a point-to-point non-linear regressor, the MLP yields the least competitive performance metrics among the deep learning candidates and even lags behind the ensemble tree models. Without spatial context, the MLP relies on single-pixel spectra and is sensitive to sensor artifacts.

(g) Ground Truth Reference: The baseline reference map.

The high accuracy of tree-based models on the ho-

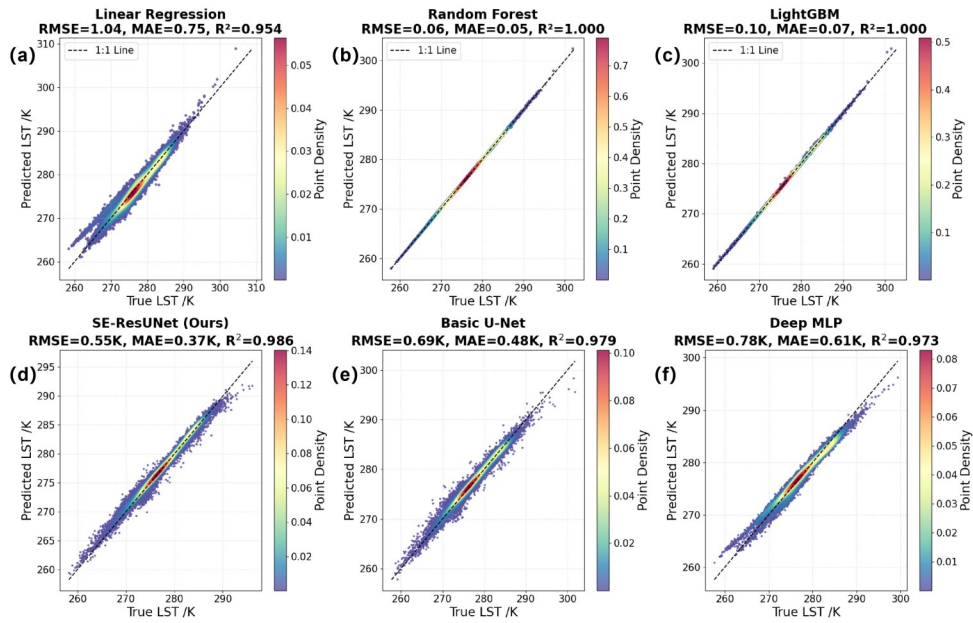


Figure 8 The scatter plots of the LR, RF, and LightGBM models together with the SE-ResUNet, Basic U-Net and MLP architectures: (a) Linear Regression; (b) Random Forest; (c) LightGBM; (d) SE-ResUNet; (e) Basic U-Net; (f) Deep MLP

图 8 线性回归、随机森林和 LightGBM 模型以及 ResNet-UNet, 基本 U-Net 和多层感知机架构的散点图: (a) 线性回归; (b) 随机森林; (c) LightGBM; (d) SE-ResUNet; (e) 基本 U-Net; (f) 多层感知机

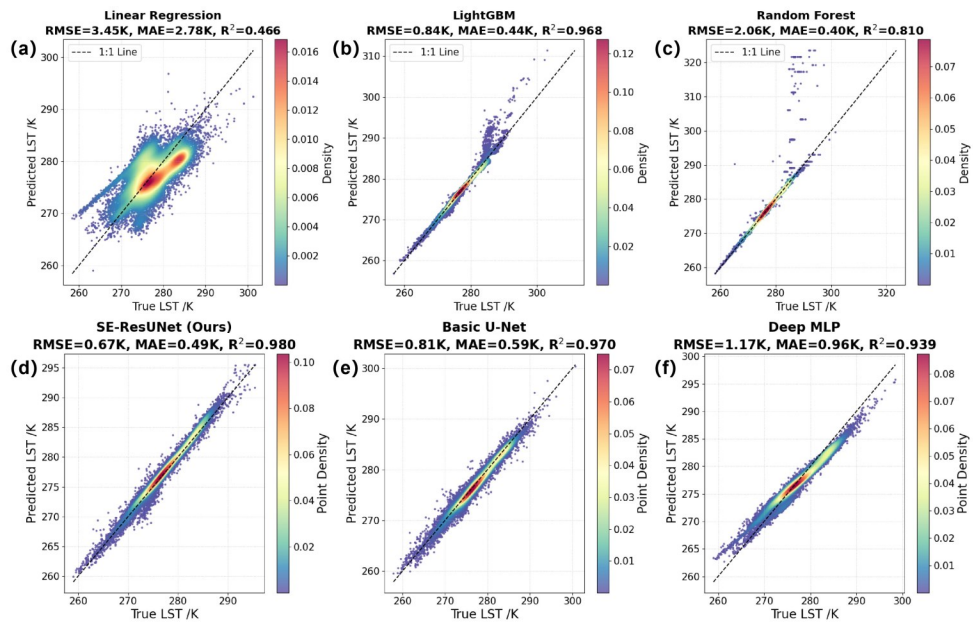


Figure 9 Performance comparison on Multi-Surface Terrains: (a) Linear Regression; (b) Random Forest; (c) LightGBM; (d) SE-ResUNet; (e) Basic U-Net; (f) Deep MLP

图 9 多种类表面地形上的性能比较: (a) 线性回归; (b) 随机森林; (c) LightGBM; (d) SE-ResUNet; (e) 基本 U-Net; (f) 多层感知机

homogeneous dataset is an artifact of spatial autocorrelation: random pixel-level splits place spectrally near-identical neighbors in both training and test sets, enabling memorization rather than physical learning. In heterogeneous environments, this strategy is ineffective because neighboring pixels often belong to different land cover types. The MLP suffers similarly: without spatial context, it produces patchy, inconsistent predictions. By contrast, SE-ResUNet extract regional features through convolutional receptive fields, trading marginal accuracy

on uniform terrain for robust generalization across diverse landscapes.

3.2.2 Noise Robustness Test

Real-world satellite images are often degraded by sensor disruptions, the presence of atmospheric particles, and the presence of light cloud cover. To understand the operational robustness of the model under degraded conditions, the test dataset was expanded by the addition of Gaussian noise (0% - 30%) and stripe masks.

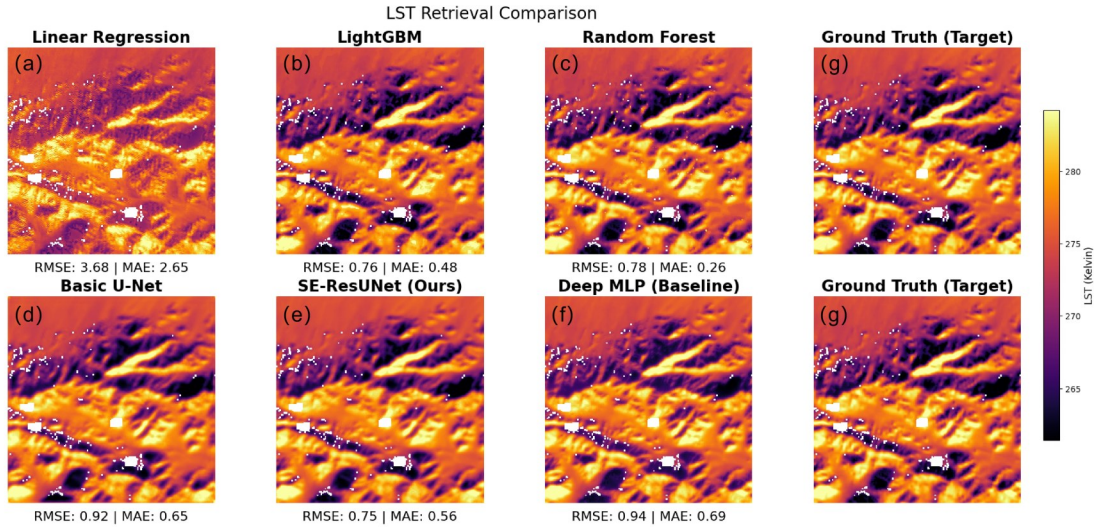


Figure 10 Comparison of actual retrieval results in heterogeneous land surface areas: (a) Inference results of Linear Regression; (b) Inference results of LightGBM; (c) Inference results of Random Forest; (d) Inference results of Basic U-Net; (e) Inference results of SE-ResUNet; (f) Inference results of Deep MLP; (g) Ground Truth reference

图 10 异质地表区域的实际推理结果对比图: (a) 线性回归推理结果; (b) LightGBM 推理结果; (c) 随机森林推理结果; (d) 基本 U-Net 推理结果; (e) SE-ResUNet 推理结果; (f) 多层感知机推理结果; (g) 真值参考数据

Table 3 presents the metrics of six models at 0%, 10%, 20%, and 30% noise.

Notably, both traditional models (LR and RF) actually perform better at the 10% level of noise than they do under noise-free conditions (e. g. , LR RMSE improved from 3.45 to 1.63 K). This improvement indicates that the raw input features contain a large number of high-frequency outliers and/or instrument noise to which traditional rigid pixel-based models are particularly sensitive, leading to overfitting. The presence of low-intensity noise that is present is likely acting as an implicit regularization term (similar to input jittering) thereby smoothing the irregularities present in the local features and preventing the model from fitting the trivial spectral spikes.

This performance benefit is limited to low-noise input. As the level of noise increases to 20% - 30%, the performance of the pixel-based models degrades dramatically. As seen in Figure 11, there is a considerable spread and systematic error in terms of their distance from the 45-degree line within the scatter plots. The Deep MLP follows a similar trend of rapid deterioration; its MAE rises from 0.964 K to 1.828 K, and its R^2 score drops from 0.973 to 0.764 as noise levels reach 30%. Without spatial context, input noise propagates directly to predictions.

Conversely, CNN-based models have better robustness. Figure 12 shows that the predictions of SE-ResUNet are highly consistent with the ground truth even at 30% noise intensity and demonstrate negligible degradation in accuracy as illustrated by the scatter plots. This robustness is explained by two major architectural mechanisms:

Spatial Filtering through Convolution: The convolutional architecture, unlike pixel-based models, sums the features of local receptive fields. This allows the network to make valid temperature estimates even when some pix-

Table 3 Performance statistics (RMSE/MAE/ R^2) of six algorithms under 0 - 30% noise injection

表 3 六种算法在 0%~30% 噪声注入下的性能统计数据 (RMSE/MAE/ R^2)

Models	Noise rate	RMSE (K)	MAE (K)	R^2 Score
LinearRegression	Original	3.453	2.775	0.466
	10% Noise	1.635	1.129	0.880
	20% Noise	1.800	1.379	0.855
	30% Noise	2.349	2.054	0.753
LightGBM	Original	0.840	0.444	0.968
	10% Noise	1.200	0.772	0.936
	20% Noise	1.634	1.276	0.881
	30% Noise	2.368	1.872	0.783
RandomForest	Original	2.062	0.401	0.810
	10% Noise	0.852	0.589	0.968
	20% Noise	1.554	1.202	0.892
	30% Noise	2.203	1.760	0.783
Multi-Layer Perceptron	Original	1.171	0.964	0.973
	10% Noise	1.454	1.187	0.925
	20% Noise	1.690	1.348	0.873
	30% Noise	2.304	1.828	0.764
Basic U-Net	Original	0.814	0.591	0.970
	10% Noise	0.836	0.604	0.968
	20% Noise	0.897	0.628	0.964
	30% Noise	0.933	0.678	0.962
SE-ResUNet (Ours)	Original	0.669	0.487	0.980
	10% Noise	0.671	0.493	0.981
	20% Noise	0.714	0.552	0.974
	30% Noise	0.790	0.613	0.972

els are corrupted since inference is based on synthesized features of the surrounding context as opposed to single data points.

Adaptive Recalibration through SE: SE mechanism enables adaptive recalibration by dynamically modulating channel-wise feature responses. In case optical bands are corrupted by atmospheric interference, the SE module down-weights the respective feature maps and thus reduces their contribution to the final prediction and shifts the model focus to more reliable thermal bands or auxiliary information (e. g., ST_CDIST). This soft-attention

ability is inherent in preserving the integrity of the model in adverse conditions enabling graceful degradation instead of catastrophic failure.

As shown in Figure 13, we also compared spatial LST maps under each noise levels.

The figure longitudinally displays thermal maps from 10% to 30% noise gradients (a-c), and ground truth surface temperature data. As noise increases, traditional pixel models like LR, RF, and LightGBM rapidly develop fragmented "salt-and-pepper noise" and patchy artifacts (Figure 14), disrupting the originally smooth

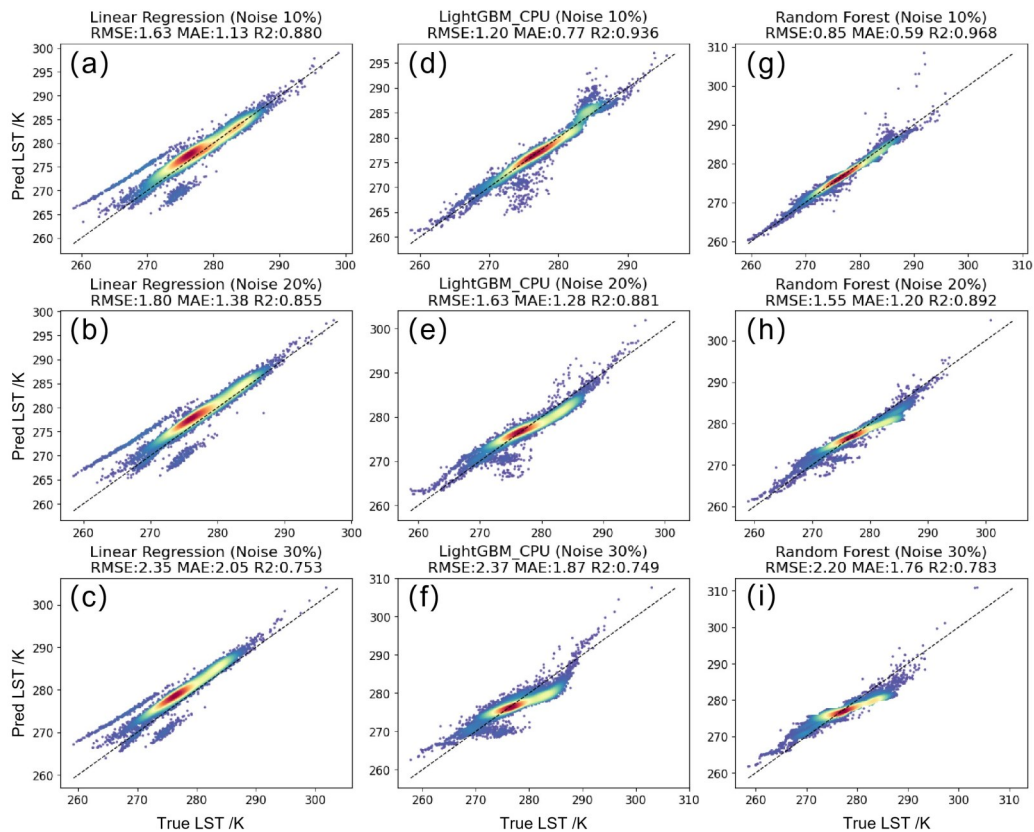


Figure 11 Scatter plots of LR, RF, and LightGBM under 10%-30% noise. Note the increasing dispersion and bias drift: (a-c) Linear Regression at noise level from 10% to 30%; (d-f) Random Forest at noise level from 10% to 30%; (g-i) LightGBM from at noise level from 10% to 30%

图 11 噪声水平为 10%-30% 时线性回归、RF 和 LightGBM 的散点图, 注意色散和偏差漂移的增加: (a) 噪声水平为 10%-30% 时的线性回归; (b) 噪声水平为 10%-30% 时的随机森林; (c) 噪声水平为 10%-30% 时的 LightGBM

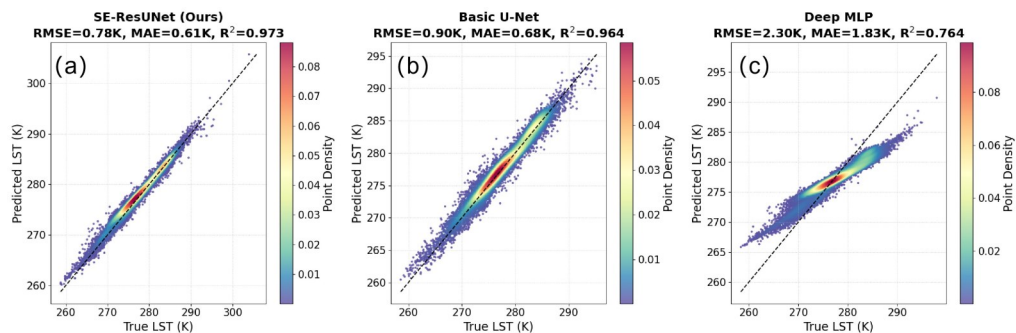


Figure 12 Scatter plots of Standard U-Net and SE-ResUNet under 30% noise. : (a) SE-ResUNet at noise level of 30%; (b) Basic U-Net at noise level of 30%; (c) Deep MLP at noise level of 30%

图 12 标准 U-Net 和 SE-ResUNet 在 30% 噪声下的散点图: (a) SE-ResUNet 在 30% 噪声水平下的表现; (b) 基本 U-Net 在 30% 噪声水平下的表现; (c) 多层感知机在 30% 噪声水平下的表现

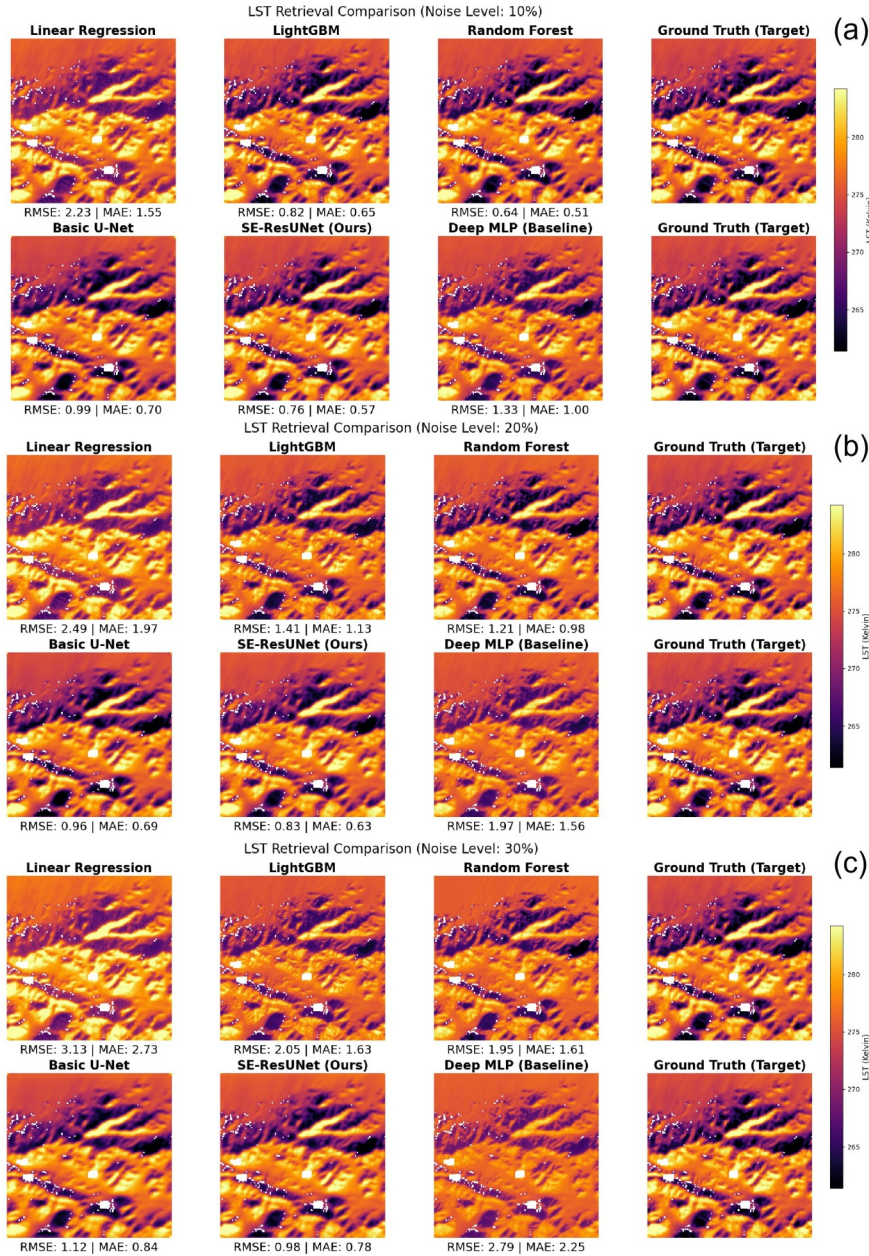


Figure 13 Comparison of retrieval results from six algorithms under different noise levels; (a) Inference results of the six algorithms at a 10% noise level; (b) Inference results of the six algorithms at a 20% noise level; (c) Inference results of the six algorithms at a 30% noise level

图 13 六种算法在不同噪声等级下的推理结果对比图:(a) 六种算法在 10% 噪声水平下的推理结果;(b) 20% 噪声水平下的推理结果;(c) 30% 噪声水平下的推理结果

surface temperature texture and causing overall color shift, and MLP even generates systematic overall bias in high-noise environments. In contrast, convolutional architectures such as U-Net and SE-ResUNet maintain intact spatial structures and clear texture edges even under 30% high-noise conditions without significant noise accumulation.

As the Figure 15 shows below, we selected the Yangtze River Delta urban agglomeration as a representative experimental region to further verify the SE-ResUNet's performance in heterogeneous scenarios. Due to its high building density, fragmented underlying surfaces, and significant mixed-pixel issues, the surface emissivity

and thermal states in this area exhibit intense spatial non-stationarity.

Experimental results show that in the noise environment with 0%-20% noise level, SE-ResUNet achieves the best accuracy, resolving street-level thermal patterns. At 30% noise, SE-ResUNet remains stable with its RMSE remaining below 1.3 K. In comparison, the RMSE values of all other models are nearly above 2 K.

3.3 Ablation Study and Sensitivity Analysis We conduct ablation experiments to isolate the contribution of each component. These tests compare the performance of the full-fledged SE-ResUNet model against several modified variants to isolate the impact of specific mod-

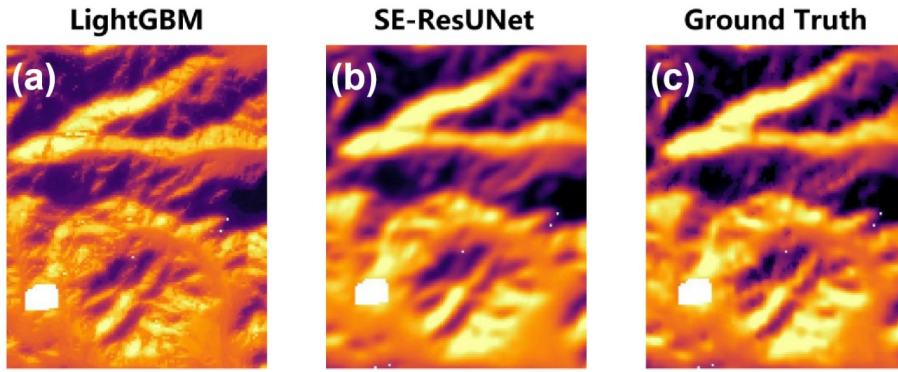


Figure 14 Comparison of retrieval details under 30% noise level: (a) LightGBM inference result details; (b) SE-ResUNet inference result details; (c) Details of Ground Truth reference

图 14 30% 噪声环境下推理细节对比图: (a) LightGBM 推理结果细节图; (b) SE-ResUNet 推理结果细节图; (c) 真值参考数据细节图

ules and features.

3.3.1 Architectural and Physics-Informed Components

In this stage, we evaluated four primary modifications:

- (1) Removing the SE attention block.
- (2) Excluding the SWD input.
- (3) Replacing the robust L1-based loss with MSE.
- (4) Severing the U-Net skip connections.

Table 4 summarizes the quantitative results. The comparative study between complete model and model without SE highlights the importance of adaptive feature adjustment as even though the global MAE showed a slight enhancement of around 0.04K, this overall measure hid the great influence that the module had in situations with complicated radiation.

After dividing the test data into two atmospheric groups based on the values of QA_Pixel and water vapor amount, Clear Sky and High Humidity, to examine how atmospheric conditions impacted model performance. The difference in model performance due to atmospheric conditions was greater under poor overall conditions than it was under clear sky conditions. Specifically, although the models had similar performance levels in clear sky conditions, the SE-ResUNet model had a lower mean er-

ror compared to the ResNet model in high humidity, resulting in an approximately 0.19 K difference in mean error.

This conditional performance gain confirms the efficiency of the SE module as a soft attention filter. In good atmospheric conditions, optical bands are used as good proxies to emissivity and get balanced weights on attention. On the other hand, when the optical data is impaired by atmospheric scattering (e. g., thin cloud cover), then the SE mechanism automatically suppresses these corrupted channels. As a result, the network changes its dependence to the more penetrating thermal infrared bands and the feature of ST_CDIST.

The Model without SWD Input shows that the lack of explicit SWD input leads to a substantial performance degradation, where RMSE increases from 0.73 K to 0.99 K. This confirms the value of injecting SWD as an explicit physical prior. Without SWD, the network must infer atmospheric water-vapour effects from raw radiance alone, which is a harder task.

The analysis of the structural ablation provides two major findings: (a) Significance of Feature Fusion: It was catastrophic to eliminate skip connections as it resulted in a breakdown of performance (RMSE > 2.7 K). This failure shows that successful pixel-wise LST retriev-

Table 4 Ablation study results on the test dataset

表 4 测试数据集的消融研究结果

Model Variant	Modification	RMSE (K)	MAE (K)	R ²
SE-ResUNet (Ours)	None (Full Model)	0.7327	0.5187	0.9784
w/o SE Block	Remove SE Attention	0.7923	0.5602	0.9734
w/o SWD Input	Remove SWD Feature	0.9963	0.7655	0.9580
w/ MSE Loss	Replace Loss with MSE	0.9805	0.7008	0.9593
w/o Skip Conn.	Remove Skip Connections	2.7398	1.9796	0.6823

Table 5 MAE under stratified atmospheric conditions

表 5 分层大气条件下的性能比较(MAE)

Atmospheric Condition	Metric	w/o SE Block	SE-ResUNet (Ours)	Improvement
Global Average	MAE (K)	0.5602	0.5187	-0.0415
Clear Sky	MAE (K)	0.3902	0.3746	-0.0156
High Humidity (> 3.5 g/cm ²)	MAE (K)	0.9348	0.7464	-0.1884

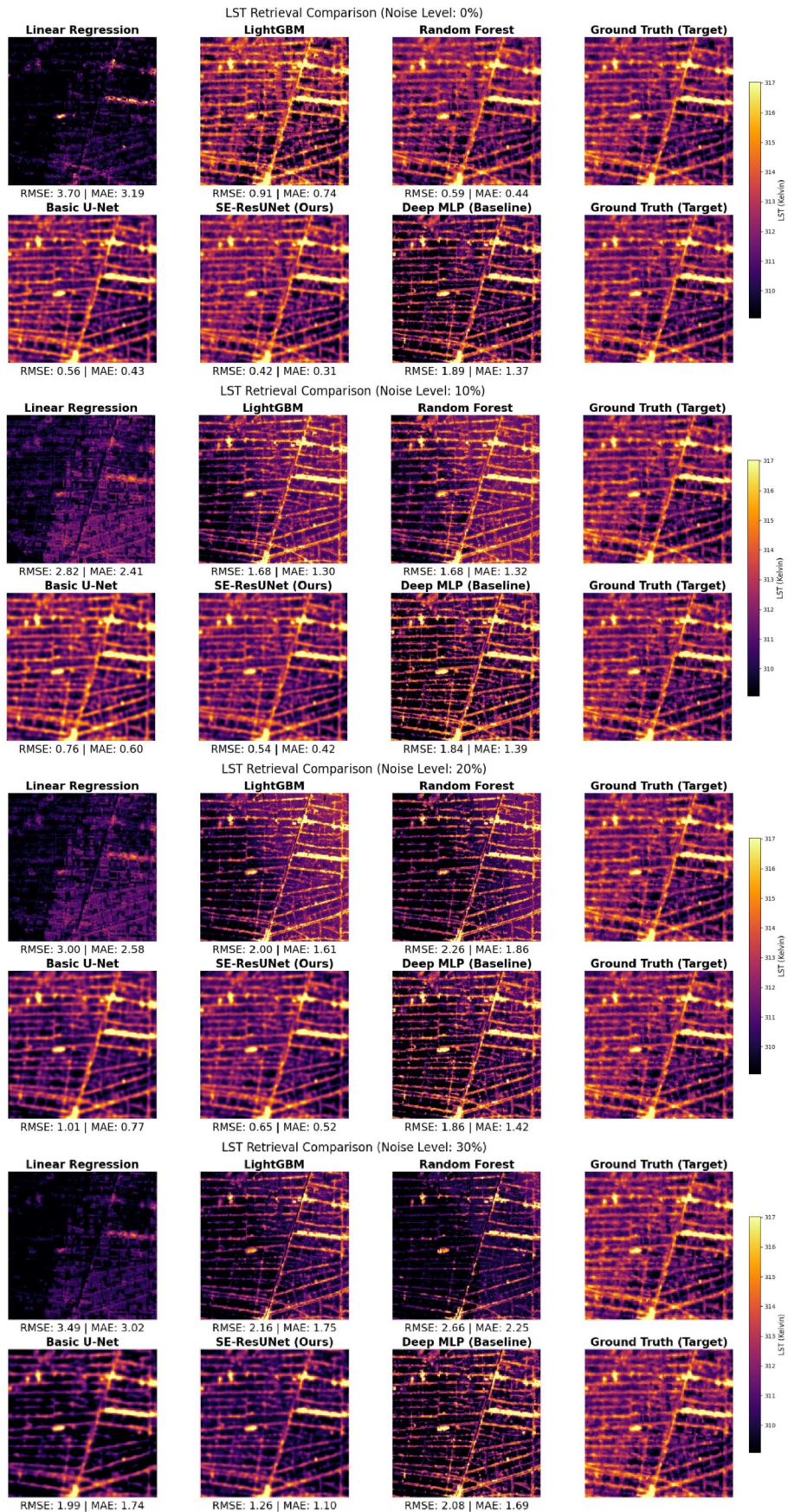


Figure 15 Comparison of retrieval results of City type from six algorithms under different noise levels
 图 15 六种算法在不同噪声等级下的城市类型推理结果对比图

al is only possible with semantic and spatial features fusion. (b) Loss Function Sensitivity: The standard MSE loss model had poorer performance than the full model trained using Charbonnier loss. This difference is explained by the fact that MSE is sensitive to outliers, which are common in thermal remote sensing data. Conversely, the robust Charbonnier loss counteracts this effect, thus stabilizing the optimization process. 3.3.2 Contribution of Surface Emissivity Prior

A specific concern in deep-learning-based LST retrieval is "information leakage," where the model might over-rely on the surface emissivity prior (ST_EMIS) rather than learning the underlying radiative transfer physics. To evaluate this, we conducted an additional ablation study by completely removing the ST_EMIS channel from the input tensor.

Table 6 Ablation Study Results: Impact of ST_EMIS on LST Retrieval Accuracy

表6 消融研究结果:ST_EMIS对LST检索准确率的影响

Model Configuration	RMSE (K)	MAE (K)	R2
Full SE-ResUNet (All 12 Channels)	0.7327	0.5187	0.9784
SE-ResUNet (w/o ST_EMIS)	0.8381	0.6041	0.9712
Standard U-Net (w/o ST_EMIS)	0.8842	0.6242	0.9675
Standard ResNet (w/o ST_EMIS)	0.9774	0.7525	0.9608

As shown in Table 6, even without the explicit ST_EMIS prior, the SE-ResUNet maintains high retrieval accuracy (MAE = 0.6242 K). This indicates that the model is capable of implicitly sensing surface emissivity variations by extracting deep features from the visible, NIR, and SWIR channels. While the inclusion of ST_EMIS further refines the results by providing a physical baseline, the model does not treat it as an exclusive "crutch."

This robustness is vital for operational applications where high-quality emissivity priors (e. g., ASTER GED) might be outdated or unavailable.

3.4 Validation with Independent Gridded Meteorological Data

To assess the operational performance of the SE-ResUNet model beyond synthetic simulations, we validate the retrieval results against independent, real-world observations. In this section, we evaluate the model's accuracy using the "Hourly Gridded Surface Meteorological Data of China" as a reference.

The validation reference used in this study is a high-quality product developed by the National Meteorological Information Center (NMIC) of the China Meteorological Administration. This dataset integrates the ERA5-Land reanalysis with observations from over 2,400 basic meteorological stations across China.

For this validation exercise, we selected a representative spatial window in the arid desert region of Xinjiang (88°E-92°E, 38°N-40°N). The analysis was conducted for December 26, 2024, at 04:00 (UTC), which coincides with the Landsat 9 satellite overpass time.

A comparison between the LST retrieved by the SE-ResUNet model and the gridded meteorological reference data is presented in Table 7.

Table 7 Comparison between SE-ResUNet predicted LST and China Gridded Meteorological Data (2024-12-26 04:00 UTC)

表7 SE-ResUNet预测地表温度与中国网格气象数据对比(2024年12月26日04:00 UTC)

Sample ID	Longitude (°E)	Latitude (°N)	SE-ResUNet LST (K)	Gridded Ref. LST (K)	Difference (K)
1	88.42	39.15	268.45	267.32	+1.13
2	89.10	38.85	265.12	266.05	-0.93
3	89.75	39.42	271.30	270.18	+1.12
4	90.22	38.25	263.88	264.40	-0.52
5	90.85	39.88	269.54	268.21	+1.33
6	91.30	38.60	266.75	267.55	-0.80
7	91.95	39.20	272.15	271.05	+1.10
8	91.12	39.45	270.40	269.58	+0.82

As shown in the table, the LST values retrieved by the proposed SE-ResUNet model demonstrate high agreement with the independent gridded meteorological data. The discrepancies are relatively small, with absolute differences generally remaining within ± 1.5 K. The Mean Absolute Error (MAE) for this validation set was calculated to be approximately 0.97 K.

The results indicate that the SE-ResUNet, trained on physics-based simulation data, can produce reasonable LST estimates for real satellite scenes.

4 Conclusion

This paper presents SE-ResUNet, a physics-guided deep learning architecture that is customized to the Landsat 9 OLI-2/TIRS-2 system. Our model incorporates domain physics SWD and spectral emissivity into a ResNet50 based encoder decoder with SE attention.

The model is validated against high-precision MODTRAN simulations in detail, which confirms its efficiency. SE-ResUNet reached RMSE of 0.67 K and MAE of 0.49 K ($R^2 = 0.98$) on independent test sets, which is a significant improvement compared to the official Level-2 product (ST_B10), especially in heterogeneous arid areas where classical algorithms tend to be biased due to emissivity. Comparative studies also show that although traditional machine learning models like RF and LightGBM are spatially overfitting and noise sensitive, and deep MLP models exhibit relatively weak performance while also being highly sensitive to noise, SE-ResUNet utilizes intrinsic spatial filtering and dynamic channel-wise recalibration to significantly enhance noise robustness and generalization. These mechanisms maintain retrieval accuracy under complex terrain and up to 30% noise sensor noise, without requiring real-time atmospheric data.

Lastly, the ablation studies highlight the need to in-

tegrate physical priors with deep semantic features; when major elements like SWD are removed, the performance deteriorates substantially and this confirms the hybrid design philosophy. SE-ResUNet removes reliance on real-time atmospheric data when applying by integrating strict physical modeling with data-driven intelligence. The proposed automated system of regional thermal monitoring is a framework. It is intended to advance the processing of satellite-based LST data, and its performance may serve as a useful reference for future studies.

Despite the significant retrieval precision achieved, the current study is primarily constrained by its specific optimization for the Landsat 9 sensor configuration, and the sample capacity of the synthetic dataset could be further expanded to enhance robustness under rare global meteorological anomalies.

Future research will focus on migrating this physics-guided architecture to multi-source platforms, including MODIS, FY-series satellites, and even uncooled thermal infrared sensors from commercial aerospace, to facilitate consistent cross-sensor Land Surface Temperature monitoring globally.

References

- [1] Li Zhao-Liang, Tang Bo-Hui, Wu Hua, et al. Satellite-derived Land Surface Temperature: Current Status and Perspectives [J]. *Remote Sensing of Environment*, 2013, 131: 14-37.
- [2] Karnieli A, Agam N, Pinker R T, et al. Use of NDVI and Land Surface Temperature for Drought Assessment: Merits and Limitations [J]. *Journal of Climate*, 2010, 23(3): 618-633.
- [3] Weng Qi-Hao. Thermal Infrared Remote Sensing for Urban Climate and Environmental Studies: Methods, Applications, and Trends [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2009, 64(4): 335-344.
- [4] Voogt J A, Oke T R. Thermal Remote Sensing of Urban Climates [J]. *Remote Sensing of Environment*, 2003, 86(3): 370-384.
- [5] Wulder M A, Loveland T R, Roy D P, et al. Current Status of Landsat Program, Science, and Applications [J]. *Remote Sensing of Environment*, 2019, 225: 127-147.
- [6] Masek J G, Wulder M A, Markham B, et al. Landsat 9: Empowering Open Science and Applications Through Continuity [J]. *Remote Sensing of Environment*, 2020, 248: 111968.
- [7] Montanaro M, McCorkel J, Tveekrem J, et al. Landsat 9 Thermal Infrared Sensor 2 (TIRS-2) Stray Light Mitigation and Assessment [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-8.
- [8] Jiménez-Muñoz J C, Sobrino J A, Skokovic D, et al. Land Surface Temperature Retrieval Methods From Landsat-8 Thermal Infrared Sensor Data [J]. *IEEE Geoscience and Remote Sensing Letters*, 2014, 11(10): 1840-1843.
- [9] Wan Zheng-Ming, Dozier J. A Generalized Split-window Algorithm for Retrieving Land-surface Temperature From Space [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1996, 34(4): 892-905.
- [10] Zhu Xiao-Xiang, Tuia D, Mou Li-Chao, et al. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources [J]. *IEEE Geoscience and Remote Sensing Magazine*, 2017, 5(4): 8-36.
- [11] Wang Han, Mao Ke-biao, Yuan Zi-jin, et al. A method for land surface temperature retrieval based on model-data-knowledge-driven and deep learning [J]. *Remote Sensing of Environment*, 2021, 265: 112665.
- [12] Hu Jie, Shen Li, Sun Gang. Squeeze-and-Excitation Networks [C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018: 7132-7141.
- [13] Sobrino J A, Jiménez-Muñoz J C, Paolini L. Land surface temperature retrieval from LANDSAT TM 5 [J]. *Remote Sensing of Environment*, 2004, 90(4): 434-440.
- [14] Barsi J A, Barker J L, Schott J R. An Atmospheric Correction Parameter Calculator for a Single Thermal Band Earth-Sensing Instrument [C]. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2003, 5: 3014-3016.
- [15] Barsi J A, Schott J R, Palluconi F D, et al. Validation of a Web-Based Atmospheric Correction Tool for Single Thermal Band Instruments [J]. *Proceedings of SPIE*, 2005, 5882: 58820E.
- [16] Qin Zhi-Hao, Karnieli A, Berliner P. A Mono-window Algorithm for Retrieving Land Surface Temperature From Landsat TM Data and Its Application to the Israel-Egypt Border Region [J]. *International Journal of Remote Sensing*, 2001, 22(18): 3719-3746.
- [17] Jiménez-Muñoz J C, Sobrino J A. A generalized single-channel method for retrieving land surface temperature from remote sensing data [J]. *Journal of Geophysical Research: Atmospheres*, 2003, 108(D22): 4688.
- [18] Wan Zheng-Ming. New refinements and validation of the collection-6 MODIS land-surface temperature products [J]. *Remote Sensing of Environment*, 2014, 140: 36-45.
- [19] Du Chen, Ren Hua-Zhong, Qin Qi-Ming, et al. A Practical Split-Window Algorithm for Estimating Land Surface Temperature from Landsat 8 Data [J]. *Remote Sensing*, 2015, 7(1): 647-665.
- [20] Breiman L. Random Forests [J]. *Machine Learning*, 2001, 45(1): 5-32.
- [21] Drucker H, Burges C J C, Kaufman L, et al. Support Vector Regression Machines [C]. *Advances in Neural Information Processing Systems*, 1997, 9: 155-161.
- [22] Quinlan J R. Learning with Continuous Classes [C]. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 1992: 343-348.
- [23] Hutengs C, Vohland M. Downscaling Land Surface Temperatures at Regional Scales with Random Forest Regression [J]. *Remote Sensing of Environment*, 2016, 178: 127-141.
- [24] Mountrakis G, Im J, Ogole C. Support Vector Machines in Remote Sensing: A Review [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2011, 66(3): 247-259.
- [25] Mao Ke-Biao, Qin Zhi-Hao, Shi Jian-Cheng, et al. A Neural Network Technique for Separating Land Surface Emissivity and Temperature From ASTER Imagery [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, 46(1): 200-208.
- [26] Shen Zheng-Yu, Huang Hong-Lian, Liu Xiao, et al. Neural Network-Based Land Surface Temperature Retrieval from Chinese GF-5B Satellite [J]. *Acta Optica Sinica*, 2025, 45(24): 2428008.
- [27] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep Residual Learning for Image Recognition [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 770-778.
- [28] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [C]. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015: 234-241.

基于SE-ResUNet的Landsat 9地表温度反演算法研究

潘皇甫钰^{1,2}, 唐国良¹, 张旭东¹, 陈弘毅^{1*}, 亓洪兴^{1*}

(1. 国科大杭州高等研究院, 浙江 杭州 310016;

2. 中国科学院大学, 北京 100049)

摘要: 传统地表温度反演技术高度依赖大气水汽等同步气象参数, 且既有机器学习模型在空间数据分析时泛化性能有限。为此, 本研究提出一种新的 Landsat9 地表温度反演方法, 该方法基于 SE-ResUNet 深度学习架构, 融合 MODTRAN5 辐射传输模型和 ERA5 再分析资料生成一套涵盖多种地表类型的高精度物理模拟数据集, 作为训练标签。模型采用 U-Net 基础架构, 其中编码器部分运用改进的 ResNet50 主干网络来提取多尺度空间特征信息, 并且在残差模块中引入通道注意力机制, 把劈窗差值等物理先验作为显式输入特征以增强模型对热红外信号的响应能力。通过跳跃连接融合深层语义特征和浅层空间细节, 最终实现像素级的精准温度反演。实验表明, SE-ResUNet 有效规避了传统方法由于空间自相关特性导致的精度虚高问题, 在模拟传感器噪声环境以及复杂地形场景下表现出良好的鲁棒性。在验证数据集中, 该模型取得了 RMSE 0.7 K 和 MAE 0.5 K 的反演精度, 表明该模型在推理阶段无需依赖实时大气参数, 即可达成高精度端到端的陆地表面温度反演效果。

关键词: 地表温度反演; 陆地卫星 9 号; 深度学习; 残差网络; U-Net