Residual-Guided and Diversity-Aware Multi-Source Selection for Spatio-Temporal Transfer Forecasting

Zhiwen Liang[1], Yuanbo Xu[1,*], Hangtong Xu[1], Mijia Zhang[1], Dongming Luan[1], Lu Jiang[2], Pengyang Wang[3] and Pengfei Wang[4]

[1] *College of Computer Science and Technology, Jilin University, Changchun, China*

[2] *Dalian Maritime University, Dalian, China*

[3] *University of Macau, Macau, China*

[4] *Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*

{liangzw25,xuht24}@mails.jlu.edu.cn, {yuanbox,zhangmijia,luandm}@jlu.edu.cn, jiangl761@dlmu.edu.cn, pywang@um.edu.mo, wpf2106@gmail.com

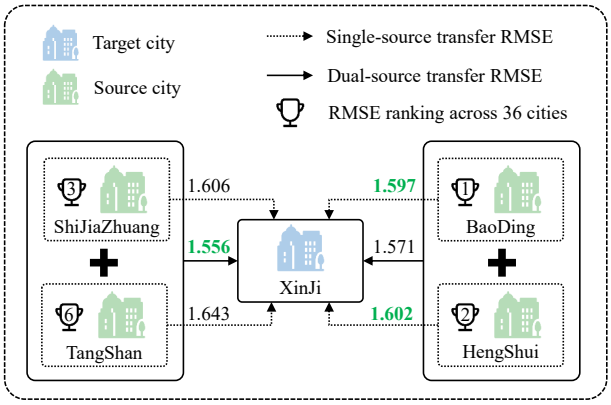
***Abstract*—Multi-source transfer learning is an effective way to improve prediction in data-scarce target domains by leveraging knowledge from related source domains. It is especially valuable in spatio-temporal forecasting, where a target city often has limited historical observations but many neighboring or structurally similar cities may provide transferable patterns. However, selecting a reliable source subset is challenging because accurate transferability estimation usually requires repeated downstream training, redundant sources may hurt the final combination, and the number of possible subsets grows exponentially with the number of candidates.To address these issues, we propose a residual-guided and diversity-aware multi-source selection framework that combines lightweight transferability estimation, Determinantal Point Process (DPP)-based subset scoring, and a multi-stage filtering strategy. Specifically, the framework estimates cross-domain transferability using residual distributions, temporal dynamics, and frequency-domain characteristics, while the DPP formulation jointly rewards target relevance and penalizes inter-source redundancy. In addition, the multi-stage filtering strategy progressively narrows the search space, reducing computational complexity and focusing the selection process on more promising source combinations. Experiments on a real-world air quality forecasting benchmark with 4 target cities and 36 candidate source cities show that the proposed method achieves consistently lower RMSE and MAE across most target cities and forecasting architectures. It also substantially reduces the computational time required for transferability estimation, reducing runtime by approximately one order of magnitude on average.**

***Index Terms*—Transfer Learning, Multi-source Selection, Transferability Estimation, Determinantal Point Process (DPP), Air Quality Forecasting**

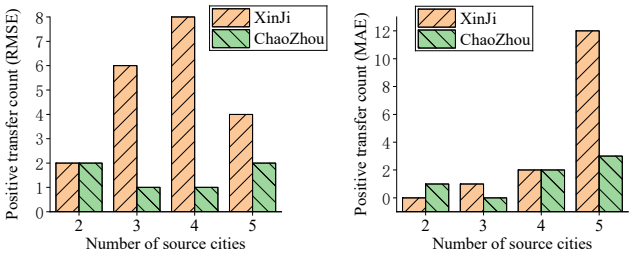
I. Introduction

Transfer learning improves prediction performance in data-scarce target domains by leveraging supervision from related source domains [1]. In spatio-temporal forecasting tasks such as air quality prediction, newly deployed monitoring systems often contain limited historical observations, while other cities provide abundant multivariate time series for transferable knowledge [2], [3]. Compared with single-source transfer learning, multi-source transfer learning can provide stronger robustness and broader coverage of heterogeneous temporal patterns [4]. However, directly aggregating all source domains may introduce redundancy and domain mismatch, making effective multi-source selection a critical problem [5]–[7].

Despite recent progress, multi-source selection remains challenging for three reasons. First, accurately estimating transferability usually requires repeatedly training downstream forecasting models, resulting in substantial computational cost. Second, existing methods mainly evaluate source-target similarity independently while neglecting redundancy and com-



(a) Single-source transferability does not necessarily imply optimal multi-source transferability.



(b) For different subset sizes k, randomly sampled source subsets (100 trials per k) can outperform Top-K source selection based on individual transferability.

Fig. 1. Motivation examples for multi-source selection. Existing methods mainly evaluate source-target similarity independently while ignoring redundancy and complementarity among source domains.

* Corresponding author.
Code: https://github.com/zhiwenCode/RDMSS

plementarity among source domains [8]–[10]. Figures I and I show that the top single-source pair (BaoDing and HengShui) underperforms a weaker pair (TangShan and ShiJiaZhuang), and many random source subsets outperform Top-K selections based on individual transferability across subset sizes. These findings imply that individually strong sources do not necessarily form the optimal subset. Third, the number of possible source subsets grows exponentially with the number of candidate domains, making exhaustive search impractical for realistic applications [11]–[13].

To address these challenges, we propose a residual-guided and diversity-aware multi-source selection framework that combines lightweight transferability estimation, DPP-based subset scoring, and progressive candidate filtering. First, we design a lightweight proxy model to estimate single-source transferability without repeatedly training downstream forecasting models. The proxy model extracts residual representations and characterizes domain discrepancies from distributional, temporal-dynamic, and frequency-domain perspectives. Second, based on the resulting single-source transferability scores, we construct a Determinantal Point Process (DPP) kernel for multi-source evaluation, where quality and similarity terms are defined by source-target and source-source transferability, respectively. This formulation jointly captures target relevance and inter-source diversity. Third, we develop a multi-stage filtering strategy that progressively focuses computation on promising regions of the combinatorial search space, thereby avoiding exhaustive search over exponentially many source subsets.

The main contributions of this paper are summarized as follows:

- We propose a residual-guided and diversity-aware multi-source selection framework that combines lightweight transferability estimation with Determinantal Point Process (DPP)-based subset scoring for spatio-temporal transfer forecasting.
- We develop a multi-stage filtering strategy that progressively narrows the search space, reducing computational complexity while focusing the selection process on more promising source combinations.
- We conduct extensive experiments on a Chinese air quality forecasting benchmark with 4 target and 36 source cities using Transformer and TCN models. Results show that the proposed method achieves lower RMSE and MAE across most cities and models, while reducing transferability estimation time by approximately one order of magnitude.

II. Related Work

A. Cross-city Transfer and Multi-source Domain Adaptation

Urban computing tasks increasingly rely on spatio-temporal data from multiple correlated cities with uneven supervision [2], [3]. Cross-city transfer learning transfers knowledge from data-rich cities to lightly labeled targets [14], [15], yet many methods assume a fixed source pool. In practice, source availability, sensor coverage, and inter-city correlations evolve over time, making source selection a key challenge [16], [17].

Deep domain adaptation mitigates domain gaps through feature alignment, adversarial learning, and discrepancy minimization [18]–[20]. Multi-source variants employ mixture-of-experts routing and collaborative representations [4], [21], but mainly focus on *how* to combine sources rather than *which* sources should be excluded to prevent negative transfer.

For long-horizon regression tasks such as PM2.5 forecasting, distributional similarity alone is often insufficient to predict transfer quality [22]. Domain adaptation theory and recent multi-source studies also emphasize interaction effects among domains [6], [23], motivating source selection strategies beyond simple similarity matching.

B. Transferability Estimation and Divergence-based Screening

Maximum mean discrepancy (MMD) [24] and related kernel methods measure covariate shift without retraining target models [25]. Metrics including H-score, LEEP, and LogME estimate transferability efficiently [8]–[10], but are primarily designed for classification and capture marginal similarity rather than temporal compatibility in forecasting.

Submodular maximization offers efficient guarantees for diversity-aware subset selection [26], while determinantal point processes (DPP) encourage diversity through log-determinant objectives that suppress redundant sources [27]. However, multi-city forecasting also depends on nonlinear temporal interactions and target-specific utility.

Data Shapley and influence-function approaches estimate the contribution of samples or domains to downstream performance [11], [12], but repeated retraining is computationally expensive for Transformer-based forecasting [13]. Surrogate-assisted optimization alleviates this cost using lightweight approximations [28].

III. Preliminary

A. Basic Definition

Suppose that there are N candidate source domains, $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, and one target domain T. Each domain contains multivariate temporal sequences $\mathbf{V} = \{\mathbf{v}_t\}_{t=1}^{L}$, where $\mathbf{v}_t \in \mathbb{R}^d$ denotes the feature vector at time step t, d is the number of features, and L is the total number of time steps.

We use $\mathcal{T}(S_i, T)$ to denote the transferability from a source domain S_i to the target domain T. For a subset of source domains $\mathcal{A} \subseteq \mathcal{S}$, its multi-source transferability to T is denoted by $\mathcal{T}(\mathcal{A}, T)$.

B. Problem statement

Given source domains $\mathcal{S}$ and target domain T, the goal of multi-source selection is to identify an optimal subset:

$$\mathcal{A}^* = \arg\max_{\mathcal{A} \subseteq \mathcal{S}} \mathcal{T}(\mathcal{A}, T). \tag{1}$$

In practice, source selection should balance target relevance and source diversity to maximize transferability while avoiding redundancy and conflicts. However, the number of candidate

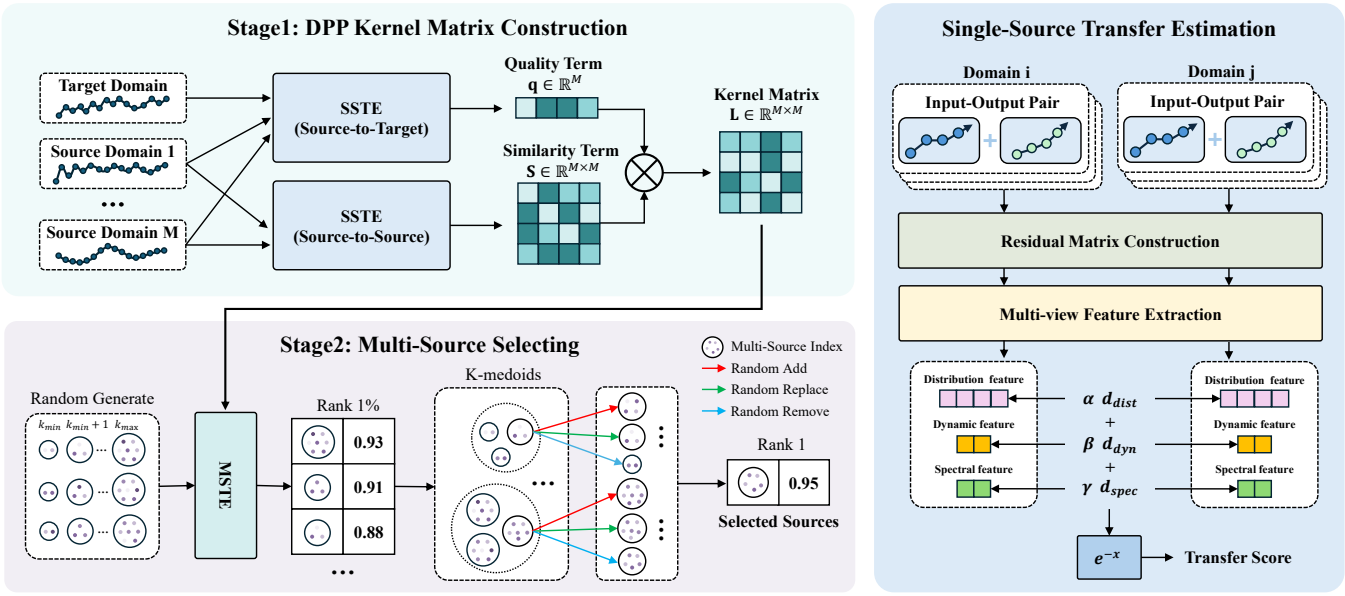


Fig. 2. Overall framework of the proposed residual-guided and diversity-aware multi-source selection

subsets grows exponentially with N, making exhaustive evaluation impractical due to the cost of downstream model training and validation.

IV. Methodology

The proposed framework estimates multi-source transferability without repeatedly training downstream forecasting models on all source combinations. As illustrated in Figure 2, the framework contains two stages. Stage 1, a single-source transferability estimation method is used to estimate transferability between arbitrary domain pairs, based on which a DPP kernel matrix is constructed to model source quality and diversity jointly. Stage 2, a multi-stage source selection algorithm is performed on the DPP kernel to identify the optimal source subset while avoiding exhaustive search.

A. Stage 1: DPP Kernel Matrix Construction

The proposed transferability estimation method is first used to compute:

- transferability scores between source domains;
- transferability scores between each source domain and the target domain.

The inter-source transferability scores form a similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, where $\mathbf{S}_{ij} = \mathrm{Sim}(S_i, S_j)$. The source-target transferability scores form a diagonal quality matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$, where $\mathbf{Q} = \mathrm{diag}(q_1, \ldots, q_N)$ and $q_i = \mathrm{Sim}(S_i, T)$, where $\mathrm{Sim}(\cdot,\cdot)$ denotes the estimated transferability score between two domains.

The final DPP kernel matrix is constructed as

$$\mathbf{L} = \mathbf{QSQ}. \tag{2}$$

This formulation favors source subsets with both high target relevance and low redundancy. Diagonal entries encode source quality, while off-diagonal entries encode the similarity between source domains.

1) Single-Source Transferability Estimation (SSTE): The proposed method estimates transferability between arbitrary domains without distinguishing source and target domains.

For a domain D, its data contains multiple Input-Output Pairs:

$$D = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N_D}, \tag{3}$$

where $X_i \in \mathbb{R}^{H\times d}$ denotes the historical observation sequence, and $Y_i \in \mathbb{R}^{T\times 1}$ denotes the corresponding future sequence. This representation is shared by both source and target domain, so the same scoring pipeline can be applied to all domain pairs.

For two domains D_a and D_b, their data are first processed by a residual matrix construction module to obtain residual matrices. The residual matrices are then fed into a multi-view feature extraction module, which outputs distribution, dynamic, and spectral features.

Distances are computed independently in the three feature spaces: $d_{dist}, d_{dyn}, d_{spec}$. The detailed formulations are given in Eqs. (16)–(18).

The total distance is defined as

$$d_{total} = \alpha d_{dist} + \beta d_{dyn} + \gamma d_{spec}, \tag{4}$$

where $\alpha + \beta + \gamma = 1$.

The final transferability score is obtained by

$$\mathrm{Sim}(D_a, D_b) = \exp(-d_{total}). \tag{5}$$

This exponential mapping converts feature-space distances into normalized transferability scores. A smaller feature discrepancy yields a score closer to one, whereas a larger discrepancy yields a score closer to zero, which makes the transferability values easy to compare across domains.

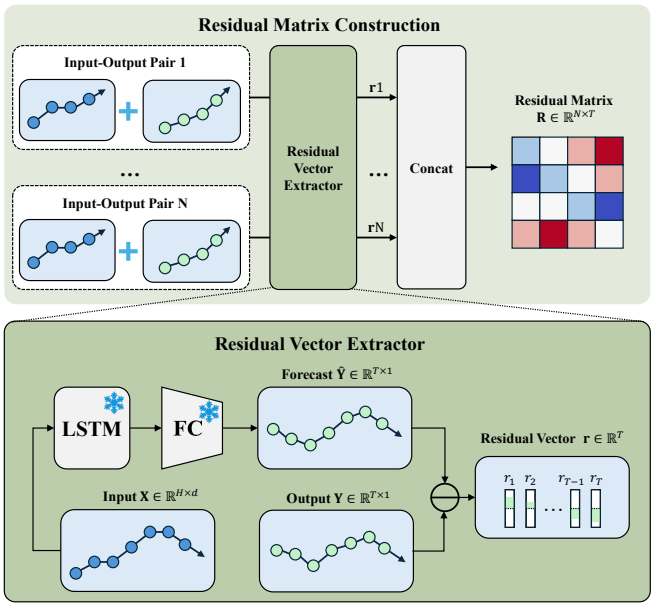


Fig. 3. Structure of the residual matrix construction module.

2) Residual Matrix Construction Module: The residual matrix construction module converts multiple Input-Output Pairs into a structured residual representation, as illustrated in Figure 3. This module is designed to retain the error behavior of a lightweight forecasting proxy, because residuals are often more informative than raw predictions when measuring cross-domain compatibility. In our implementation, the proxy model is trained once and then frozen, so all residual matrices remain directly comparable across domains.

Given a domain data $D = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N_D}$, each pair is processed by a residual vector extractor. Specifically, $\mathbf{X}_i$ is first fed into an LSTM encoder and a fully connected layer to produce the prediction $\hat{\mathbf{Y}}_i$:

$$\hat{\mathbf{Y}}_i = \mathrm{FC}(\mathrm{LSTM}(\mathbf{X}_i)). \tag{6}$$

The residual vector is computed as

$$\mathbf{r}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i. \tag{7}$$

The LSTM encoder and FC layers are jointly pre-trained on all source domains and then kept fixed during transferability estimation, allowing the extractor to capture shared spatio-temporal forecasting patterns across diverse source cities. By using a unified pre-trained model, residual patterns from different domains are represented in a common feature space, enabling more consistent and comparable cross-domain transferability analysis.

All residual vectors are concatenated row-wise to form the residual matrix:

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}1^\top & \mathbf{r}2^\top & \cdots & \mathbf{r}N_D{}^\top \end{bmatrix}^\top \in \mathbb{R}^{N_D \times T}, \tag{8}$$

where T is the forecasting horizon.

3) Multi-View Feature Extraction Module: The multi-view feature extraction module extracts complementary characteristics from the residual matrix. As shown in Figure 4, the residual matrix is simultaneously fed into three parallel encoders. Each encoder focuses on a different aspect of transferability, so the final score reflects not only how large the residual error is, but also how that error changes over time and across frequencies.

a) Distribution encoder: For each forecasting step, the distribution encoder computes the mean μ_t, variance σ_t, skewness γ_t, and extreme-value statistic e_t, where e_t is defined as the mean absolute value of the top 5% residuals. These statistics are concatenated as: $\mathbf{z}_t = [\mu_t, \sigma_t, \gamma_t, e_t]$.

The final distribution feature is $\mathbf{f}_{dist} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \ldots, \mathbf{z}_T^\top]^\top$. This representation captures both regular error distributions and severe forecasting failures.

b) Dynamic encoder: The dynamic encoder models temporal dependency patterns in residual sequences. First, the mean residual sequence $\bar{\mathbf{r}} = [\bar{\mathbf{R}}^{\mathbf{1}}, \ldots, \bar{\mathbf{R}}^{\mathbf{T}}]$ is computed. The Pearson Correlation between $\bar{\mathbf{r}}$ and forecasting steps is used as the first dynamic feature:

$$c = \mathrm{PearsonCorr}([1, \ldots, T], \bar{\mathbf{r}}). \tag{9}$$

Next, the Lag-1 Correlation of each residual vector is calculated and averaged:

$$\rho = \frac{1}{N_D} \sum_{i=1}^{N_D} \mathrm{PearsonCorr}(\mathbf{R}_i^{1:T-1}, \mathbf{R}_i^{2:T}). \tag{10}$$

The final dynamic feature is $\mathbf{f}_{dyn} = [c, \rho]$. This feature captures whether residual errors drift with forecast horizon and whether adjacent horizons exhibit similar correction patterns.

c) Spectral encoder: The spectral encoder analyzes residual patterns in the frequency domain. For a residual vector $\mathbf{r} = [r_1, r_2, \ldots, r_T]$, Fast Fourier Transform (FFT) is first applied:

$$\hat{\mathrm{R}}_i(k) = \sum_{n=1}^{T} r_n e^{-j2\pi k(n-1)/T}, \quad k = 0, 1, \ldots, T-1, \tag{11}$$

where $\hat{\mathrm{R}}_i(k)$ denotes the frequency component at frequency index k.

The corresponding power spectrum is then computed as

$$\mathrm{P}_i(k) = |\hat{\mathrm{R}}_i(k)|^2. \tag{12}$$

Low-frequency Energy Ratio (0–2) and High-frequency Energy Ratio (8–12) are subsequently computed from the normalized power spectrum:

$$LF = \frac{\sum_{k=0}^{2} \mathrm{P}_i(k)}{\sum_{k=0}^{L-1} \mathrm{P}_i(k)}, \tag{13}$$

$$HF = \frac{\sum_{k=8}^{12} \mathrm{P}_i(k)}{\sum_{k=0}^{L-1} \mathrm{P}_i(k)}. \tag{14}$$

The bins $k = 0$–2 mainly capture slowly varying components, while the bins $k = 8$–12 characterize relatively rapid oscillatory variations. These bands are used as empirical

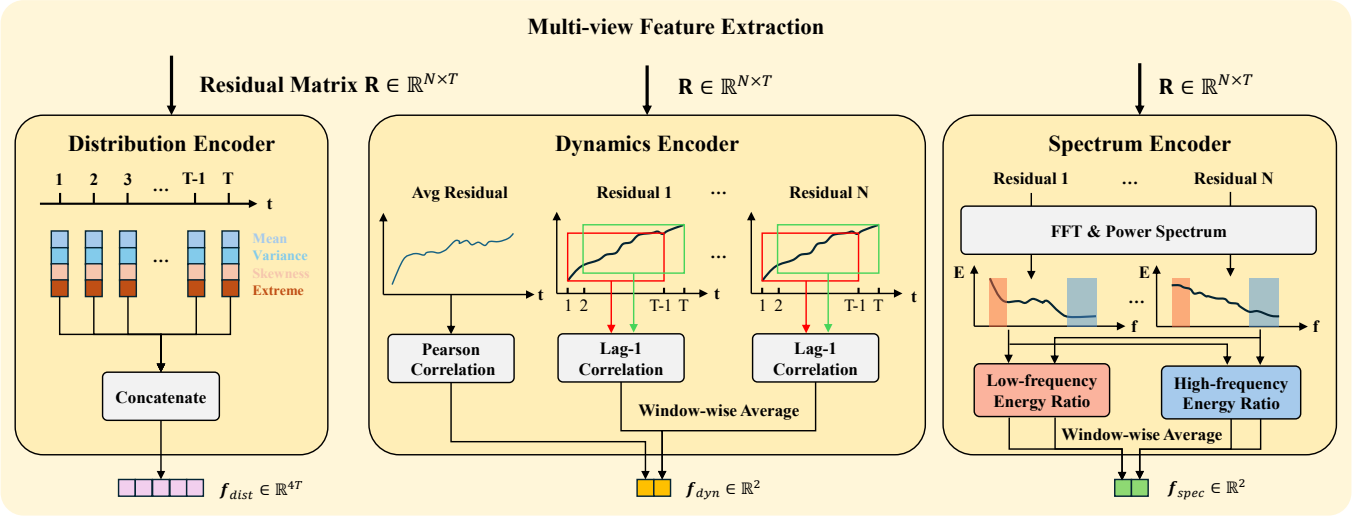


Fig. 4. Structure of the multi-view feature extraction module.

spectral descriptors rather than universal physical frequency intervals.

The final frequency-domain feature is obtained by averaging these statistics across all residual vectors:

$$\mathbf{f}_{spec} = [LF, HF]. \tag{15}$$

Frequency-domain analysis complements time-domain statistics by characterizing how residual energy is distributed across different frequency bands, thereby capturing both long-term trend discrepancies and short-term fluctuation patterns between domains.

For two domains D_a and D_b, distances in the three feature spaces are computed as:

$$d_{dist}(D_a, D_b) = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{z}_t^{D_a} - \mathbf{z}_t^{D_b}\|_2, \tag{16}$$

$$d_{dyn}(D_a, D_b) = \sqrt{(c^{D_a} - c^{D_b})^2 + (\rho^{D_a} - \rho^{D_b})^2}, \tag{17}$$

$$d_{freq}(D_a, D_b) = \sqrt{(LF^{D_a} - LF^{D_b})^2 + (HF^{D_a} - HF^{D_b})^2}, \tag{18}$$

B. Stage 2: Multi-Source Selecting

After constructing the DPP kernel matrix, the proposed algorithm searches for source subsets with high transferability and low redundancy.

A total of M candidate subsets are randomly generated, with subset sizes uniformly distributed between $k_{\min}$ and $k_{\max}$. Denote the candidate subset collection as

$$\mathcal{C} = \{C_1, C_2, \ldots, C_M\}, \tag{19}$$

where each subset $C_m \subseteq \mathcal{S}$. For each subset $C_m \in \mathcal{C}$, perform the multi-source transfer estimation (MSTE) process

$$\mathrm{E}(C_m) = \frac{\log\det(\mathbf{I} + \mathbf{L}_{C_m})}{|C_m|^a}, \quad a \geq 0, \tag{20}$$

where $\mathbf{L}_{C_m}$ denotes the principal submatrix indexed by C_m. Its determinant favors relevant, low-redundancy source combinations, while a penalizes subset size. We tuned a over $[0, 1]$ by ranking subsets via $\mathrm{E}(C_m)$ and analyzing the size distribution of the top 1%. The optimal $a = 0.78$ maximizes the normalized entropy of that distribution.

After scoring all candidate subsets, only the top 1% are retained as elite candidates, since lower-ranked subsets provide limited additional benefit while substantially increasing the search space. K-medoids clustering is then performed using Jaccard distance, the highest-scoring subset in each cluster is selected as the representative subset:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{21}$$

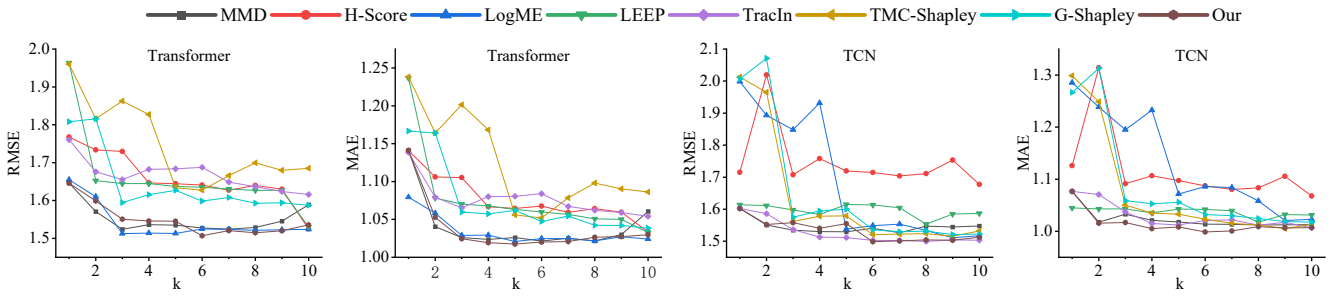
Next, stochastic neighborhood refinement is performed through random add, remove, and replace operations to further explore promising regions of the search space. Each operation perturbs the current subset by a small amount, which helps the algorithm escape local optima while preserving strong candidates.

Finally, all refined subsets are rescored using Eq. (20), and the subset with the highest score is selected as the final source-domain combination for downstream transfer learning.

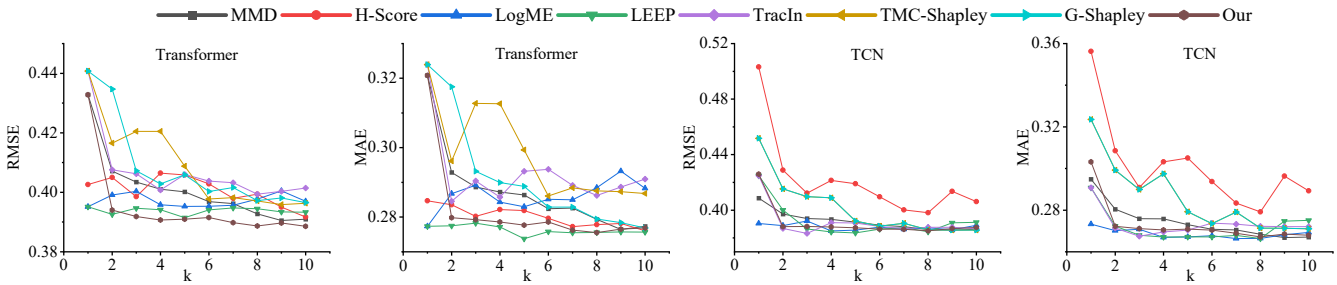
V. Experiments

A. Dataset

We use a spatio-temporal air quality dataset covering 43 cities in China from May 2014 to April 2015, including Beijing, Tianjin, Guangzhou, Shenzhen, and surrounding regions. The dataset contains hourly air quality measurements (PM2.5, PM10, NO2, CO, O3, SO2) collected from 437 monitoring stations (2.89M records), hourly meteorological observations (temperature, pressure, humidity, wind; 1.90M records), and 2-day weather forecasts issued every 3/6/12 hours (0.91M records) [29]–[31].



(a) XinJi



(b) ChaoZhou

Fig. 5. Forecasting performance of different source selection methods under varying Top-K settings.

TABLE I
STATISTICS OF THE AIR QUALITY DATASET.

Item	Value
Time Span	May 2014 – Apr 2015
Air Quality Records	2,891,393
Meteorological Records	1,898,453
Air Pollutants	PM2.5, PM10, NO2, CO, O3, SO2
Meteorological Features	Temperature, pressure, humidity, wind
Target Cities	XinJi, ChaoZhou, DaTong, DeZhou
Source Cities	Beijing, Shenzhen, Tianjin, etc. (36 cities)

After preprocessing, three cities with extremely limited data volume (ChenZhou, WuZhou, and HeZhou) were removed, leaving 40 cities for experiments. To construct geographically diverse target domains, we selected one representative city with relatively limited data from 4 provinces and finally chose XinJi, ChaoZhou, DaTong and DeZhou as the target cities. Detailed dataset statistics are summarized in Table I. For source cities, the data are chronologically divided into training, validation, and test sets using an 8-month / 2-month / 2-month split. For target cities, the first 2 months are treated as the visible set, while the remaining 10 months are used as the unseen set (i.e., the test set).

B. Baseline

We compare the proposed framework with representative transferability estimation methods for multi-source transfer learning, including **MMD** (Maximum Mean Discrepancy) [24], which measures feature-distribution discrepancy between source and target domains in reproducing kernel Hilbert space (RKHS); **H-Score** [8], which evaluates feature–label dependence through covariance statistics; **LogME** (Logarithm of Maximum Evidence) [10], which estimates transferability using Bayesian evidence; **LEEP** (Log Expected Empirical Prediction) [9], which measures conditional likelihood transferability based on source predictions; **TracIn** [32], which estimates sample influence through accumulated gradient similarity along optimization trajectories; **TMC-Shapley** (Truncated Monte Carlo Shapley) [12], which approximates source-domain Shapley values through truncated random permutations; and **G-Shapley** (Gradient Shapley) [12], which evaluates source importance in gradient space.

We further compare against several commonly used strategies, including **All-Sources**, which directly uses all available source cities; **Best-Single**, which selects the best-performing individual source city; **Random**, which uniformly samples source subsets of fixed size; and **Top-K**, which ranks sources according to a surrogate transferability score and selects the top-ranked candidates. The **Top-K** strategy is applied to all compared transferability estimation methods.

C. Experimental Settings and Evaluation Metrics

To evaluate the effectiveness of the proposed framework, experiments were conducted using two representative downstream forecasting architectures: Transformer and Temporal Convolutional Network (TCN). For all experiments, both the input sequence length and forecasting horizon were set to 24 time steps. The proxy model was trained using the Adam optimizer with a learning rate of 1×10^{-5}, a batch size of 128, and 50 training epochs. For multi-source subset generation, the number of randomly sampled candidate subsets was set to $M = 1000000$, $k_{min} = 2$ and $k_{max} = 10$. All experiments were conducted on a workstation equipped with an

TABLE II

FORECASTING PERFORMANCE OF DIFFERENT TRANSFERABILITY ESTIMATION METHODS FOR SOURCE DOMAIN SELECTION. LOWER VALUES INDICATE BETTER PERFORMANCE; BEST RESULTS ARE SHOWN IN **BOLD** AND SECOND-BEST RESULTS ARE UNDERLINED.

Target City	Model	Metric	All-Sources	Best-Single	Random	MMD	H-Score	LogME	LEEP	TracIn	TMC-Shapley	G-Shapley	Ours
XinJi	Transformer	RMSE	1.5284	1.5870	1.5562	1.5276	1.6408	1.5246	1.6350	1.6878	1.6271	1.5985	**1.5196**
		MAE	1.0207	1.0947	1.0473	1.0214	1.0677	1.0246	1.0598	1.0839	1.0518	1.0474	**1.0179**
	TCN	RMSE	1.5424	1.5440	1.5266	1.5405	1.7150	1.5488	1.6136	**1.5026**	1.5210	1.5357	1.5116
		MAE	1.0040	1.0570	1.0287	1.0138	1.0872	1.0859	1.0419	1.0205	1.0231	1.0320	**1.0029**
ChaoZhou	Transformer	RMSE	0.4094	0.3952	0.4106	0.3969	0.4028	0.3953	0.3941	0.4038	0.3978	0.4003	**0.3892**
		MAE	0.2938	0.2773	0.2960	0.2824	0.2796	0.2851	**0.2758**	0.2938	0.2861	0.2829	0.2760
	TCN	RMSE	0.4029	0.3905	0.4095	0.3892	0.4096	0.3877	0.3864	0.3882	0.3885	0.3885	**0.3854**
		MAE	0.2873	0.2733	0.2952	0.2708	0.2938	0.2677	0.2671	0.2737	0.2736	0.2736	**0.2667**
DaTong	Transformer	RMSE	0.4419	0.4553	0.4913	0.4422	0.4423	0.4403	0.4594	0.4586	0.4632	0.4668	**0.4343**
		MAE	0.2884	0.2990	0.3526	0.2885	0.2909	0.2893	0.3037	0.3044	0.3184	0.3146	**0.2861**
	TCN	RMSE	0.4341	0.4446	0.4849	0.4478	0.4238	0.4615	0.4431	0.5841	0.4973	0.5036	**0.4234**
		MAE	0.2850	0.2921	0.3500	0.2983	0.2816	0.3195	0.2904	0.4687	0.3031	0.3002	**0.2808**
DeZhou	Transformer	RMSE	**1.0285**	1.0522	1.0956	1.0497	1.0669	1.0417	1.0419	1.0619	1.0613	1.0713	1.0388
		MAE	**0.7038**	0.7160	0.7590	0.7093	0.7194	0.6987	0.7051	0.7186	0.7347	0.7248	0.7162
	TCN	RMSE	1.0242	1.0402	1.0721	1.0429	1.0268	1.0592	1.0645	1.0531	1.0663	1.0524	**1.0198**
		MAE	**0.6845**	0.6899	0.7319	0.6993	0.6853	0.6893	0.6987	0.6964	0.7276	0.7133	0.6856

NVIDIA RTX 3090 GPU with 32GB memory. Performance was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for forecasting accuracy, Pearson Correlation (P.) and Spearman Correlation (S.) for measuring the consistency between estimated transferability scores and actual transfer performance, and Source Selection Time for evaluating computational efficiency.

D. Impact of Transferability Estimation on Source Selection

Table II reports the best forecasting performance of each transferability estimation method among settings with the number of selected source cities ranging from 2 to 10, with the exception of the All-Sources and Best-Single strategies. Since all compared methods share the same Top-K strategy and downstream forecasting backbones, the performance differences mainly reflect the quality of transferability estimation. Overall, the proposed method achieves the most consistent performance across different target cities and architectures, obtaining 11 best results and 3 second-best results among all RMSE/MAE evaluations. It consistently performs well on XinJi, ChaoZhou, and DaTong under both Transformer and TCN, achieving the best RMSE and MAE in most cases. For DeZhou, the performance gap between methods becomes smaller, and the All-Sources strategy achieves the best Transformer results. Compared with existing transferability estimators, H-Score and Shapley-based methods generally show weaker and less stable performance, while MMD and LogME achieve relatively competitive results in several cases.

Figure 5 further illustrates the performance trends under varying Top-K settings. As k changes, our method consistently achieves superior or comparable performance in most cases, demonstrating stronger robustness to the choice of source city number.

E. Computational Efficiency of Transferability Estimation

Table III summarizes the computational overhead of different transferability estimation methods prior to downstream adaptation. Traditional distribution-based metrics such as MMD achieve the lowest computational cost, but their transferability ranking quality remains limited according to Table II. Although H-Score, LogME, and LEEP are theoretically training-free once feature representations are available, they still require pre-training a feature extractor in the spatio-temporal forecasting setting considered in this work, resulting in runtimes of several thousand seconds for each city split. TracIn and Shapley-based approaches further introduce repeated gradient accumulation or Monte Carlo subset evaluation, leading to substantially higher computational overhead and making iterative source selection impractical. These results indicate that transferability estimation quality and computational efficiency are often difficult to balance simultaneously. Our method avoids repeated downstream forecasting training by introducing a lightweight proxy model for transferability estimation. After a one-time proxy-model initialization stage, the complete source ranking and subset evaluation procedure can be completed within only a few hundred seconds. This efficiency advantage is particularly important in dynamic urban sensing scenarios, where spatial correlations and candidate source relationships may change frequently and source domains must be re-evaluated with low latency [33].

F. Correlation Between Transferability Estimation and Forecasting Performance

Table IV reports Pearson and Spearman correlations between transferability scores and downstream errors on Transformer and TCN models. Our method achieves the strongest correlations across both cities and architectures. It consistently obtains the highest Spearman correlations on Transformer targets for both RMSE and MAE, indicating better ranking of high-transferability source domains. On TCN, LogME is competitive in several settings, especially ChaoZhou, while our method still achieves best or second-best in most cases.

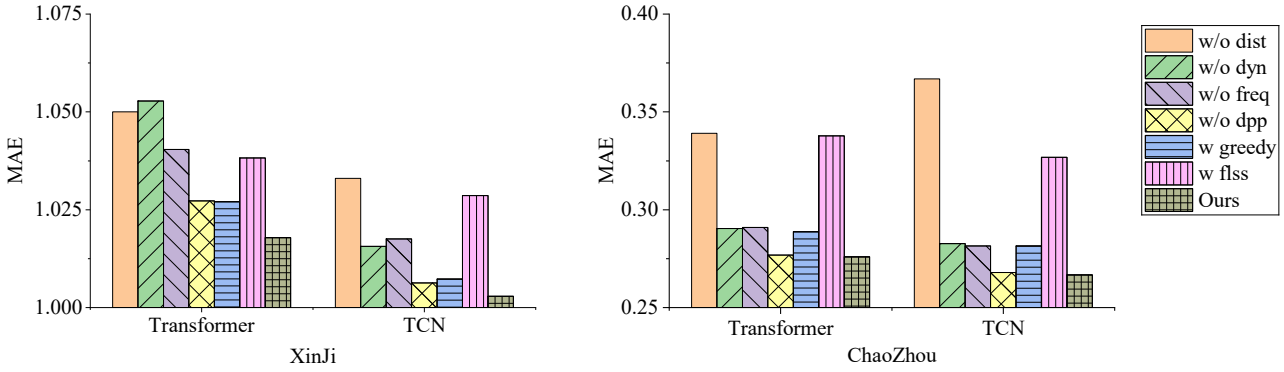


Fig. 6. Results of ablation study

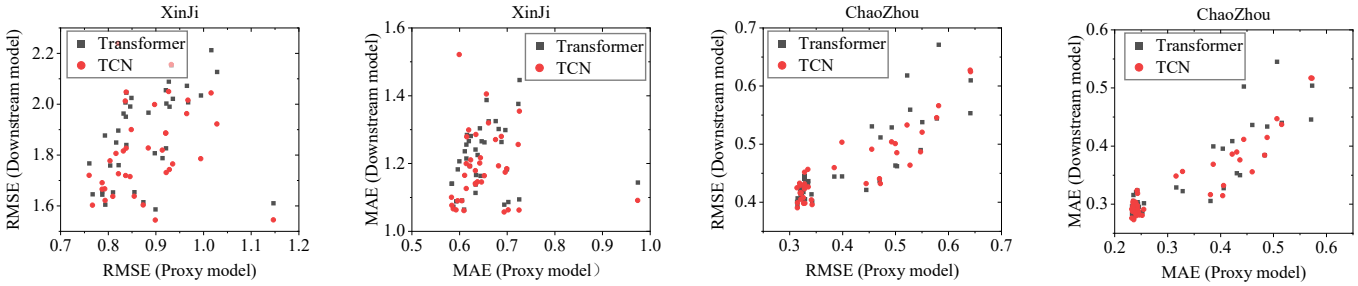


Fig. 7. Approximate linear correlation between proxy-model performance on the visible target set and downstream transfer performance on the unseen target set across different source cities.

TABLE III
RUNTIME COMPLEXITY AND WALL-CLOCK COST OF TRANSFERABILITY ESTIMATION METHODS (SECONDS).

Method	Complexity	XinJi		ChaoZhou	
		Transf.	TCN	Transf.	TCN
MMD	$O(n^2 d)$	26	26	45	45
H-Score	$O(f_{\text{pre}} + nd_f^2)$	3992	1545	3992	1545
LogME	$O(f_{\text{pre}} + d_f^3 + nd_f^2)$	3990	1544	3991	1544
LEEP	$O(f_{\text{pre}} + nC)$	3990	1543	3991	1544
TracIn	$O(f_{\text{train}} + TnmP)$	4350	1897	4313	1861
TMC-Shapley	$O(Nnf_{\text{train}})$	107013	30622	106974	30685
G-Shapley	$O(f_{\text{train}} + NnP)$	61373	30663	61382	30415
Ours	$O(f_{\text{train}}^{\text{prox}} + MK^3)$	389	389	352	352

Notation: n: sample count per domain, d: input dimension, d_f: feature dimension, C: class number, P: model parameter number, m: evaluation sample number, T: saved checkpoint number, N: Monte Carlo permutation number, M: sampled source subset number, and K: average subset cardinality, f_{train} / f_{pre} / $f_{\text{train}}^{\text{prox}}$: training cost of the downstream model / feature extractor / proxy-model. Transf. is the abbreviation of Transformer. The wall-clock cost of MMD and Ours is independent of the downstream model.

In contrast, TracIn and Shapley-based methods show weak correlations, suggesting limited effectiveness for transferability estimation in cross-domain time-series forecasting.

G. Ablation Study

We conduct ablation studies on two target cities, XinJi and ChaoZhou, using Transformer and TCN as downstream models, MAE as evaluation metrics. For the similarity score, we consider three variants: **w/o dist**, **w/o dyn**, and **w/o freq**, which remove the distributional, dynamic, and spectral components in Eq. (4), respectively. For source selection, we further compare several alternatives, including **w/o dpp** (Top-K selection), **w greedy**, and **w flss**. Greedy marginal gain adds sources with largest validation-error reduction until saturation. FLSS (Facility location source selection) [34] greedily maximizes incremental facility-location coverage via pairwise transferability, promoting diversity and representativeness.

The results in Figure 6 show that removing the distributional component causes the largest performance drop across both cities and downstream models, indicating that distributional residual statistics are the most important factor in similarity modeling. Removing the dynamic or spectral component also degrades performance, demonstrating their complementary contributions. For Eq. (4), we set $\alpha = 0.7$, $\beta = 0.2$, and $\gamma = 0.1$. Top-K and greedy selection achieve comparable performance and both outperform FLSS. This is mainly because FLSS focuses on source coverage while ignoring target-oriented transferability. Overall, the results verify the effectiveness of both the proposed evaluation score and source selection strategy.

H. Sensitivity Analysis

We further examine the effectiveness of the lightweight proxy model and the sensitivity of the size-penalty exponent a in Eq. (20). Figure 7 shows that the source-city performances produced by the proxy and downstream models align

TABLE IV

P. (PEARSON) AND S. (SPEARMAN) CORRELATIONS BETWEEN TRANSFERABILITY SCORE AND DOWNSTREAM ERRORS ON DIFFERENT CITIES. BEST IN **BOLD**; SECOND-BEST UNDERLINED (* INDICATES THAT THE P-VALUE FOR THE CORRELATION COEFFICIENT IS GREATER THAN 1×10^{-4}).

Target City	Model	Metric	MMD		H-Score		LogME		LEEP		TracIn		TMC-Shapley		G-Shapley		Ours	
			P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.	S.
XinJi	Transformer	RMSE	0.7661	0.7681	0.3405	0.2965	0.7632	0.7115	0.2140	0.1210*	0.1564	0.1122*	0.1918	0.1926	0.2540	0.2838	**0.7768**	**0.7940**
		MAE	0.7301	0.7293	0.3780	0.3508	**0.7522**	0.7191	0.2987	0.2147	0.1982	0.2425	0.1618	0.0529*	0.1667	0.1558	0.7482	**0.8003**
	TCN	RMSE	0.6074	0.6227	0.3300	0.4331	-0.0812*	-0.0268*	0.0571*	0.0480*	0.3817	0.4703	0.2320	0.3331	0.2205	0.4395	**0.7333**	**0.7638**
		MAE	0.5979	0.6596	0.2380	0.3937	-0.1064*	-0.0995*	0.1073*	0.0842*	0.3286	0.3824	0.2307	0.3392	0.1841	0.3239	**0.7350**	**0.8220**
ChaoZhou	Transformer	RMSE	0.5119	0.5764	0.4341	0.4247	0.4261	0.3885	0.5174	0.6148	0.0418*	0.0838*	0.1185	0.3912	0.5269	0.5439	**0.6156**	**0.7151**
		MAE	0.5269	0.5969	0.4489	0.4668	0.4114	0.3894	0.5421	0.6510	0.0026*	0.0236*	0.1134	0.3946	0.5632	0.6110	**0.6362**	**0.7457**
	TCN	RMSE	0.5564	0.6183	0.3140	0.3782	0.7429	0.7831	0.5645	0.6623	0.6182	0.6810	0.1653	0.3481	0.5027	0.5598	**0.7707**	**0.8029**
		MAE	0.5685	0.6421	0.3539	0.4348	0.7652	**0.8379**	0.5955	0.7106	0.6141	0.6877	0.1673	0.3324	0.5324	0.5996	**0.7772**	0.8015

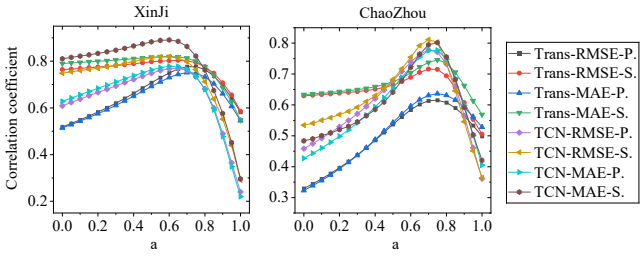


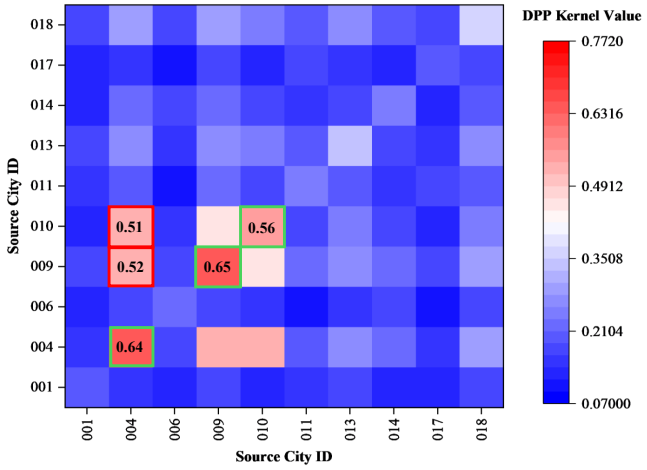
Fig. 8. Effect of the subset-size penalty exponent a on the Pearson/Spearman correlation with downstream performance.

approximately along a linear trend, despite being evaluated on different settings and scales. Specifically, the x-axis represents the proxy-model performance on the visible set of the target city, whereas the y-axis represents the downstream forecasting performance on the unseen set. The consistent ranking trend across source cities suggests that the proxy model serves as an effective indicator of transferability.

Figure 8 illustrates the influence of the exponent a on the correlation between the transferability score and downstream performance, measured by both Pearson and Spearman coefficients. In both cases, the correlation coefficients consistently peak around $a = 0.75$, suggesting that this weighting achieves the best balance in source selection. This observation is also consistent with the optimal value $a = 0.78$ derived from the normalized entropy analysis in Section IV-B.

I. Visualization of the DPP Kernel Matrix

Figure 9 visualizes a partial DPP kernel matrix for the target city XinJi, constructed from the remaining 36 candidate source cities. For clarity, only a sub-matrix containing 10 source-city IDs is shown. In the upper heatmap, diagonal entries denote source-quality scores with respect to the target city, while off-diagonal entries represent pairwise source similarities. The green boxes highlight the quality scores of source cities 004, 009, and 010, and the red boxes indicate the similarities between source city 004 and source cities 009 and 010. The two similarity values are both around 0.52, indicating comparable redundancy levels. The lower part compares the multi-source transfer scores (MSTE), RMSE, and MAE of source sets $\{004, 009\}$ and $\{004, 010\}$. Since the redundancy levels are similar, the performance difference is mainly determined by



Sources	MSTE↑	RMSE↓	MAE↓
004+009	**0.53**	**1.97**	**1.24**
004+010	0.49	2.18	1.41

Fig. 9. Visualization of a partial DPP kernel matrix for target city XinJi. Diagonal entries represent source-domain quality scores, while off-diagonal entries represent inter-source similarities. The lower table further compares the transfer scores and forecasting performance of different source combinations.

source quality. Source set $\{004, 009\}$ achieves better transfer performance due to the higher quality score of source city 009. This result shows that the proposed method effectively balances source quality and inter-source redundancy during source selection.

VI. CONCLUSION

In this paper, we showed that effective source selection is crucial for spatio-temporal transfer forecasting. The proposed residual-guided scoring strategy estimates transferability from distributional, temporal, and spectral perspectives without repeated downstream retraining, and the DPP-based subset scoring mechanism further balances target relevance and source diversity. Experiments on a real-world PM2.5 forecasting benchmark show that the proposed method achieves competitive forecasting accuracy while reducing source-evaluation cost by approximately one order of magnitude. These results indicate that lightweight residual representations and diversity-aware selection are practical for cross-city transfer forecasting.

Future work will focus on robustness under sparse observations and sensor drift, as well as uncertainty-aware and continual source updating for larger urban sensing networks.

VII. ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grant No. 92567204, No. 62472196 and No. 62502177, and in part by the Postdoctoral Fellowship Program of the CPSF under Grant No. GZC20251096.

REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[2] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–41, 2015.

[3] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "Geoman: Multi-level attention networks for geo-sensory time series prediction." in *Ijcai*, vol. 2018, 2018, pp. 3428–3434.

[4] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.

[5] Y. Tang, M. Rahmani Dehaghani, P. Sajadi, and G. G. Wang, "Selecting subsets of source data for transfer learning with applications in metal additive manufacturing," *Journal of Intelligent Manufacturing*, pp. 1–22, 2024.

[6] Q. Zhang, H. Fu, G. Huang, Y. Liang, C. Chu, T. Peng, Y. Wu, Q. Li, Y. Li, and S.-L. Huang, "A high-dimensional statistical method for optimizing transfer quantities in multi-source transfer learning," *Advances in Neural Information Processing Systems*, vol. 38, pp. 25 528–25 563, 2026.

[7] Y. Hu, H. Jin, X. Chu, and R. Cao, "Selective multi-source transfer learning and ensemble learning for piezoelectric actuator feedforward control." in *Actuators*, vol. 15, no. 1, 2026.

[8] Y. Bao, Y. Li, S.-L. Huang, L. Zhang, L. Zheng, A. Zamir, and L. Guibas, "An information-theoretic approach to transferability in task transfer learning," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 2309–2313.

[9] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau, "Leep: A new measure to evaluate transferability of learned representations," in *International conference on machine learning*. PMLR, 2020, pp. 7294–7305.

[10] K. You, Y. Liu, J. Wang, and M. Long, "Logme: Practical assessment of pre-trained models for transfer learning," in *International conference on machine learning*. PMLR, 2021, pp. 12 133–12 143.

[11] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[12] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *International conference on machine learning*. PMLR, 2019, pp. 2242–2251.

[13] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 1167–1176.

[14] L. Wang, X. Geng, X. Ma, F. Liu, and Q. Yang, "Cross-city transfer learning for deep spatio-temporal prediction," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 1893–1899.

[15] X. Wei, T. Guo, H. Yu, Z. Li, H. Guo, and X. Li, "Areatransfer: A cross-city crowd flow prediction framework based on transfer learning," in *International Conference on Smart Computing and Communication*. Springer, 2021, pp. 238–253.

[16] Z. Fang, D. Wu, L. Pan, L. Chen, and Y. Gao, "When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters." in *IJCAI*, vol. 22, 2022, pp. 2030–2036.

[17] Y. Zhang, X. Wang, X. Yu, Z. Sun, K. Wang, and Y. Wang, "Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, 2025, pp. 1147–1155.

[18] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[20] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.

[21] J. Guo, D. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 4694–4703.

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[23] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[24] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[26] K. Wei, R. Iyer, and J. Bilmes, "Submodularity in data subset selection and active learning," in *International conference on machine learning*. PMLR, 2015, pp. 1954–1963.

[27] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2-3, pp. 123–286, 2012.

[28] X. Xue, Y. Hu, L. Feng, K. Zhang, L. Song, and K. C. Tan, "Surrogate-assisted search with competitive knowledge transfer for expensive optimization," *IEEE Transactions on Evolutionary Computation*, 2024.

[29] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 2267–2276.

[30] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, pp. 1–55, 2014.

[31] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1436–1444.

[32] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 920–19 930, 2020.

[33] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[34] H. Zhao, M. Li, L. Sun, and T. Zhou, "Bento: Benchmark reduction with in-context transferability," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 29 086–29 102.