Statistically-Guided Deep Network Transformation to Harness Heterogeneity in Space (Extended Abstract)[†]

**Yiqun Xie[‡][1*] , Erhu He[2*] , Xiaowei Jia[2] , Han Bao[3] , Xun Zhou[3] ,
Rahul Ghosh[4] , Praveen Ravirathinam[4]**
[1]University of Maryland
[2]University of Pittsburgh
[3]University of Iowa
[4]University of Minnesota
xie@umd.edu, {erh108,xiaowei}@pitt.edu, {han-bao,xun-zhou}@uiowa.edu,
{ghosh128,pravirat}@umn.edu

Abstract

Spatial data are ubiquitous and have transformed decision-making in many critical domains, including public health, agriculture, transportation, etc. While recent advances in machine learning offer promising ways to harness massive spatial datasets (e.g., satellite imagery), spatial heterogeneity – a fundamental property of spatial data – poses a major challenge as data distributions or generative processes often vary over space. Recent studies targeting this difficult problem either require a known space-partitioning as the input, or can only support limited special cases (e.g., binary classification). Moreover, heterogeneity-pattern learned by these methods are locked to the locations of the training samples, and cannot be applied to new locations. We propose a statistically-guided framework to adaptively partition data in space during training using distribution-driven optimization and transform a deep learning model (of user's choice) into a heterogeneity-aware architecture. We also propose a spatial moderator to generalize learned patterns to new test regions. Experiment results on real-world datasets show that the framework can effectively capture footprints of heterogeneity and substantially improve prediction performances.

1 Introduction

Spatial datasets are ubiquitous and collected at ever-growing scale, resolution, frequency and variety. Common types of spatial data include satellite/UAV imagery, points-of-interest (POI), GPS locations/trajectories, geo-tagged tweets, census data, maps (e.g., land cover, crimes, traffic accidents, COVID statistics), and many more. These spatial datasets are critical in a variety of societal applications, such as Earth observation (e.g., crop monitoring [GEOGLAM, 2021]), public health (e.g., COVID-19 mobility analysis [Kraemer and others, 2020]), public safety, transportation, etc.

While spatial datasets are both important and widely used, their intrinsic properties – spatial autocorrelation and heterogeneity – often undermine the independent and identical distribution (i.i.d.) assumption [Atluri *et al.*, 2018; Shekhar *et al.*, 2015]. Spatial autocorrelation means data samples are not independent as nearby ones tend to share higher similarity (e.g., landcover, temperature, mobility). Spatial heterogeneity, on the other hand, violates the identical distribution assumption as data are often generated by different processes over space. Even more challenging, such differences in distributions may not be reflected by variations in observed features [Goodchild and Li, 2021], and the spatial footprints of different generative processes could be arbitrary in shape due to complex social and physical contexts.

While the wide adoption of convolutional kernels [Krizhevsky *et al.*, 2012] has explicitly filled the missing representation of spatial autocorrelation by local connections and maintained spatial relationships (i.e., non-vectorized), the complex spatial heterogeneity challenge has not been sufficiently addressed. In a recent study, a spatial-variability-aware neural network (SVANN) was developed [Gupta and others, 2020; Gupta and others, 2021]. SVANN demonstrates the benefit (e.g., increase in accuracy) of separating out training data subsets belonging to known different distributions, but it requires the spatial footprints of heterogeneous processes to be known as an input, which is often unavailable in practice. Explicit spatial ensemble approaches aim to adaptively partition a dataset [Jiang *et al.*, 2019], but the algorithm and its variation are specifically designed for binary classification problems and only allow two partitions. Outside recent literature on deep learning, a traditional approach to handle spatial heterogeneity is geographically-weighted regression (GWR) [Brunsdon *et al.*, 1999; Fotheringham *et al.*, 2017]. However, GWR is mainly designed for linear inference, and cannot handle complex prediction tasks commonly addressed by deep learning. Most existing methods also require dense training data across space to train models for individual partitions or locations, and cannot be applied to other regions outside the spatial extent of the training samples. Additional

[†]The original version of this paper [Xie *et al.*, 2021] was published at IEEE ICDM 2021, and received the Best Paper Award.

[‡]Contact Author

related work is discussed in [Xie *et al.*, 2021a].

We propose a model-agnostic Spatial Transformation And modeRation (STAR) framework to automatically learn arbitrarily-shaped spatial footprints of different data distributions during network training, and generalize the learned partitioning scheme and model weights to other spatial regions.

Through experiments on real world tasks, i.e., satellite-based crop monitoring and COVID-19 human mobility projection, we show that the STAR framework can substantially improve model performance, capture flexibly-shaped spatial footprints of different processes, and can be effectively applied to prediction tasks in new test regions.

2 Deep Network Transformation

2.1 Heterogeneity in Space

Definition 1. ***Spatial process*** Φ***:*** *A function* $\Phi : \boldsymbol{X} \mapsto \boldsymbol{y}$ *governing data generation in a spatial region, which may involve observed and unobserved (or unknown) features as variables.*

Definition 2. ***Spatial heterogeneity:*** *An intrinsic property of spatial data stating that data are generated by different spatial processes* $\{\Phi\}$ *in different regions [Atluri* et al.*, 2018; Shekhar* et al.*, 2015; Goodchild and Li, 2021].*

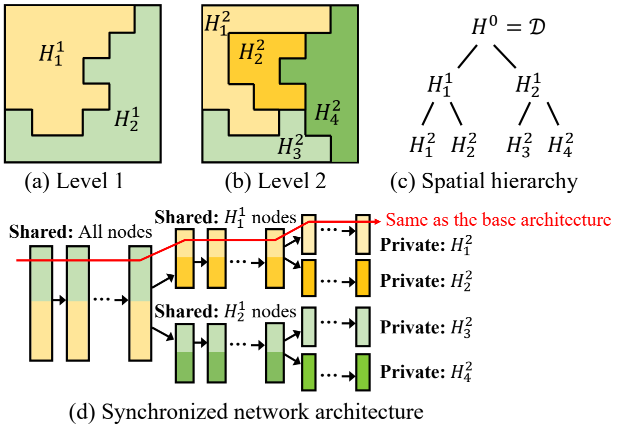
While deep networks can function as universal approximators for data following identical distributions [Kratsios and Bilokopytov, 2020], spatial heterogeneity commonly existed in spatial data violates this assumption (e.g., spatial data generated by two simple scalar functions $y = x$ and $y = -x$ across space cannot be approximated by a single network). As a result, the heterogeneous processes $\{\Phi\}$ will cause confusion on data distribution during training, and hamper prediction performance and stability.

Moreover, another complicating factor we need to consider is the hierarchy of spatial processes across scales and their corresponding heterogeneity. For example, higher-level heterogeneity in the hierarchy may be caused by policies at larger scales, climate zones, major geographical barriers (e.g., mountains), whereas lower-level processes may vary by local policies, demographics, social/cultural contexts, and personal decisions. In addition, the spatial footprints of these different processes may be arbitrary in shape. Fig. 1 (a) and (b) show an example of mixtures of spatial processes at two different scales/levels, and this hierarchy is formally defined in Def. 3.

Definition 3. ***Spatial hierarchy of processes*** $\mathcal{H}$***:*** *A multi-scale representation of spatial heterogeneity [Xie* et al.*, 2021b].* $\mathcal{H}$ *represents the input spatial domain* $\mathcal{D}$ *as a tree; each node* $\mathcal{H}_j^i \in \mathcal{H}$ *is a partition of* $\mathcal{D}$*, where* i *is the level in the hierarchy, and* j *is the unique ID for each partition at level-*i*. Children of a partition* $\mathcal{H}_j^i$ *share the same lower-level processes (i.e.,* $\{\Phi\}$ *at levels* $i' < i$*).* Φ *is homogeneous within a leaf-node and heterogeneous across leaf-nodes.*

2.2 Statistically-Guided Transformation

Based on Def. 3, we separate data samples belonging to different spatial processes Φ using a hierarchical structure. Specifically, at each step (Fig. 2), we identify an optimal space-bi-partitioning that maximizes the discrepancy of Φ between the partitions, and verify if the impact of separation is statistically significant for learning enhancement. The partitioning continues hierarchically until no significant heterogeneity can be recognized in new partitions. This transformation framework is a Dynamic and Learning-engaged generalization of the Multivariate Scan Statistic (DL-MSS) [Kulldorff and others, 2007; Neill and others, 2013; Xie *et al.*, 2021c]. DL-MSS has three major components:



(a) Level 1

(b) Level 2

(c) Spatial hierarchy

(d) Synchronized network architecture

Figure 1: Spatial processes, hierarchy and network architecture.

- *Space-partitioning optimization with prediction error distribution:* As spatial processes $\Phi : \mathbf{X} \to \mathbf{y}$ are not directly observable, we leverage the function approximation power of a deep network $\mathcal{F}$ and use its prediction error distribution for a partition $H_j^i \in \mathcal{H}$ as a proxy of Φ_j^i. Intuitively, if all data belonging to a partition H_j^i are generated by a homogeneous Φ_j^i, we expect the error distribution for each target (e.g., class) predicted by a single model for H_j^i to follow a homogeneous distribution over space; otherwise, heterogeneous. To obtain statistics on error rates, locations are first aggregated into unit local groups using a grid, where all samples in a grid cell form a group; a user may also choose a different grouping strategy. Using classification as an example, denote $\hat{\mathbf{y}}_{k,m}$ as the predicted labels for samples with class m (i.e., true labels are m) at a cell s_k. The number of misclassified samples of class m at s_k is then $e_{k,m} = |\hat{\mathbf{y}}_{k,m} \neq m|$. Further, denote $n_{k,m}$ as the number of samples of class m at location s_k; and E_m and N_m as the number of misclassified and all samples of class m in the entire space. We identify an arbitrary set of cells $S = \{s_k\}$ that maximizes the error rate distribution discrepancy using Poisson-based likelihood ratio [Neill and others, 2013]:

$$S^* = \arg\max_{S} \frac{\text{Likelihood}(H_1, S)}{\text{Likelihood}(H_0)}$$

$$= \arg\max_{S} \prod_{s_k \in S} \prod_{m=1}^{M} \frac{\Pr(e_{k,m} \sim \text{Poisson}(q_m E(e_{k,m})))}{\Pr(e_{k,m} \sim \text{Poisson}(E(e_{k,m})))}$$

where the null hypothesis H_0 states that Φ_j^i is homogeneous, and H_1 states that there exists a set S where the rate of having an error is q_m times the expected rate under H_0; and $E(e_{k,m}) = E_m \cdot \frac{n_{k,m}}{N_m}$ is the expected number of misclassified samples at location s_k under H_0. S^* can be

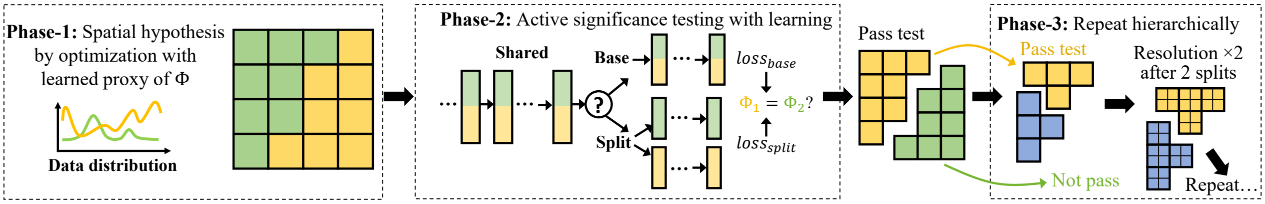

Phase-1: Spatial hypothesis by optimization with learned proxy of Φ
Data distribution
Phase-2: Active significance testing with learning
Shared
Base
Split
?
$loss_{base}$
$\Phi_1 = \Phi_2$?
$loss_{split}$
Pass test
Phase-3: Repeat hierarchically
Pass test
Resolution ×2 after 2 splits
Not pass
Repeat…

Figure 2: Illustrative example of the spatial transformation framework with dynamic and learning-engaged MSS.

efficiently solved using the Linear-Time Subset Scanning (LTSS) property [Neill, 2012; Xie *et al.*, 2021c] combined with coordinate ascent. More solution details and results on regression are available in [Xie *et al.*, 2021a].

- *Active significance testing with learning:* Once the optimal S^* is identified, the current node $\mathcal{H}_j^i$ will be temporarily split into two children $\mathcal{H}_{j1}^{i+1}$ and $\mathcal{H}_{j2}^{i+1}$, where one child corresponds to S^* and the other for the rest of the space in $\mathcal{H}_j^i$. To validate if the space-partitioning can lead to a statistically significant improvement on learning, we drop the traditional Monte-Carlo based descriptive test [Xie *et al.*, 2021c] and use a learning-engaged active test. Specifically, we carry out two training scenarios with and without the split (Fig. 2) and perform an upper-tailed dependent T-test on the losses from the two scenarios [Xie *et al.*, 2021a]. For the split scenario a new network branch will be created (e.g., adding a copy of the last L layers to the network) to allow private parameters (Fig. 1(d)). If the impact of S^* is significant, we approve the new partitioning and network branch; otherwise, the transformation on $\mathcal{H}_j^i$ is terminated.
- *Spatial transformation via a dynamic and learning-engaged spatial hierarchy $\mathcal{H}$:* As both the space-partitioning and network parameters may be constantly updated during training, DL-MSS performs the first two key components as sub-routines for new partitions added to $\mathcal{H}$ to dynamically converge to the final $\mathcal{H}$ and heterogeneity-aware network $\mathcal{F}_{\mathcal{H}}$.

2.3 Spatial Moderation

The spatial hierarchy $\mathcal{H}$ and "spatialized" deep network $\mathcal{F}_{\mathcal{H}}$ learned and trained from the transformation step aim to capture heterogeneity in the spatial extent of the input $\mathbf{X}$ and $\mathbf{y}$. However, the partitions cannot be directly applied to a new spatial region. To bridge this gap, we propose a spatial moderator, which translates the learned network branches in $\mathcal{F}_{\mathcal{H}}$ to prediction tasks in a new region. The key idea of the spatial moderator is to learn and predict a weight matrix $\mathbf{W}$ for all branches in $\mathcal{F}_{\mathcal{H}}$ (corresponding to all leaf-nodes in the spatial hierarchy $\mathcal{H}$), and then use the weights to ensemble the branches' predictions to get the final result (Fig. 3).

3 Beyond Spatial Data

The spatial transformation framework is designed for but not limited to spatial data. The model-agnostic approach can be applied to create explicit heterogeneity-awareness for general types of data, with the assumption that data samples can be combined into unit local groups (e.g., grid cells in this paper), as needed by the first component of DL-MSS in Sec. 2.2.

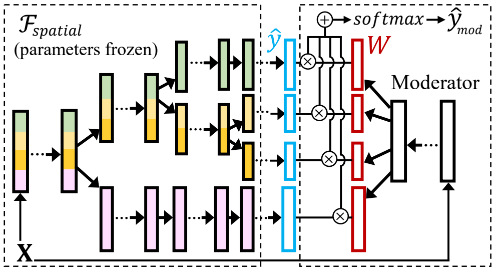


Figure 3: Illustrative example of the spatial moderator.

4 Experiments

4.1 Datasets

California land-cover classification. We use multi-spectral data from Sentinel-2 satellites in two regions in Central Valley, California. Each region has a size of 4096×4096 (~6711 km^2 in 20m resolution). We first learn the spatial partitioning using the data from Region $\mathcal{D}_A$ and then use the moderator to transfer it to Region $\mathcal{D}_B$. We use composite image series from May to October in 2018 (2 images/month) for time-series models, and one snapshot from August, 2018 for snapshot-based models. The labels are from the USDA Crop Data Layer (CDL) [CDL, 2021]. The training (and validation) set has 20% data at sampled locations in $\mathcal{D}_A$, and 1% data in $\mathcal{D}_B$ is used for fine-tuning.

Boston COVID-19 human mobility prediction. Human mobility provides critical information to COVID-19 transmission dynamics models. We use the Boston COVID-19 mobility dataset from [Bao *et al.*, 2020], which includes data from US census, CDC COVID statistics, and SafeGraph patterns. Here human mobility $\mathbf{y}$ is represented by the number of visits to points-of-interest (POIs; e.g., grocery stores, restaurants) and the counting is based on smartphone trajectories. The features include population, weekly COVID-19 cases and deaths, number of POIs, week ID and income. The dataset contains 12 weeks of data, and according to [Bao *et al.*, 2020], we use the first 11 weeks for training/validation and the final week for testing.

4.2 Candidate Methods

For California land-cover classification, we have 12 candidate methods: (1) base $\mathcal{F}$, spatially transformed $\mathcal{F}_{\mathcal{H}}$, and $\mathcal{F}_M$ ($\mathcal{F}_{\mathcal{H}}$ with moderator), each for three base architectures: DNN (snapshot-based and fully-connected), LSTM and LSTM+Attention [Jia *et al.*, 2019]; and (2) $\mathcal{F}$ integrated with the model-agnostic meta-learning (denoted as $meta$)

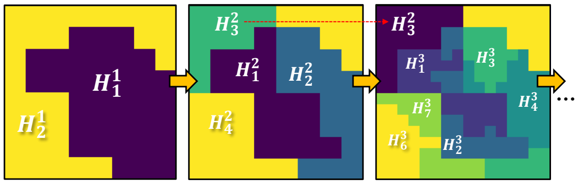


Figure 4: Spatial hierarchy learned in region $\mathcal{D}_A$ (first 3 levels).

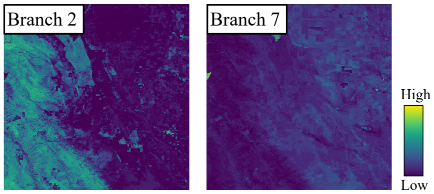


Figure 5: Learned branch weights across space (across samples).

[Finn *et al.*, 2017; Yao *et al.*, 2019] for fast adaptation in region $\mathcal{D}_B$ with the available 1% of data ($\mathcal{D}_A$'s data are clustered into tasks), for the three architectures.

For Boston COVID-19 human mobility regression, we have 5 candidate methods for comparison, i.e., ridge regression, geographically-weighted regression (GWR), COVID-GAN [Bao *et al.*, 2020], DNN, and DNN$_\mathcal{H}$ (spatially transformed version). As time-series is not used to construct additional features in [Bao *et al.*, 2020] and some features are aggregated to week-levels, we follow the same strategy and only use week IDs as features, which also allows a more direct comparison with COVID-GAN. In addition, as training and test samples are from the same set of spatial locations (timestamps are different), we directly use spatial transformation and skipped the moderator in this comparison.

4.3 Results

Land-cover classification. As shown in Fig. 6, the "spatialized" network architectures overall achieved the highest F1-scores for different types of base models in both regions. For region $\mathcal{D}_A$, the general trend is that the results of a base model $\mathcal{F}$ gradually improve with the addition of spatial transformation $\mathcal{F}_\mathcal{H}$, and the spatial moderator $\mathcal{F}_M$ (e.g., the weighted F1 score increases from 0.49 to 0.59, and finally to 0.67 for DNN). In addition, Fig. 4 shows the hierarchical process of space-partitioning during spatial transformation for the first 3 levels. In the first level (largest scale), for example, $\mathcal{H}_1^1$ is a mix of urban and suburban areas, whereas $\mathcal{H}_2^1$ contains more rural and mountainous areas. Note that some partitions (e.g., $\mathcal{H}_3^2$) are not further split, as determined by significance testing. Finally, Fig. 5 visualizes the weights predicted by the moderator for two example network branches in $\mathcal{F}_\mathcal{H}$ for DNN for all locations in region $\mathcal{D}_B$. For each branch, the weight is averaged over all classes in the predicted $\mathbf{W}$ at each location. As we can see, in the new region $\mathcal{D}_B$, branch-2 is given higher weights for the left-side of the region, which is a mountainous area, whereas the weights for branch-7 shows the opposite spatial pattern.

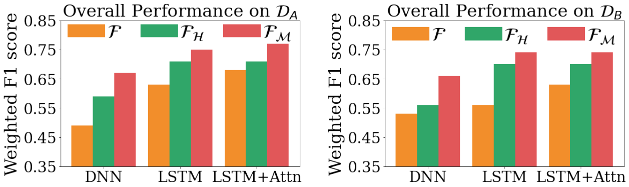


Figure 6: Weighted F1 scores for regions $\mathcal{D}_A$ and $\mathcal{D}_B$.

	Ridge	GWR	COVID-GAN	DNN	DNN$_\mathcal{H}$
MAE	220	160	178	159	**139**
RMSE	440	388	388	405	**341**
Diff.	-87K	-176K	-20K	-209K	**18K**

Table 1: COVID-19 human mobility projection

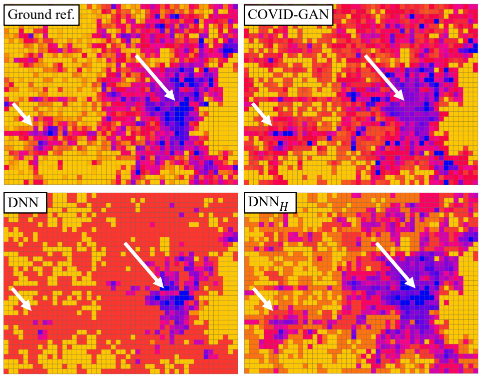


Figure 7: Visualization of human mobility maps (blue: high).

COVID-19 mobility regression. Table **??** and Fig. 7 show the results of the candidate methods, where "Diff." refers to the difference between the total predicted and true mobility values (POI visits). Several potential causes of the spatial heterogeneity here include different mobility patterns in the more populous downtown area versus the suburban regions, and several "hotspot" areas of POI visits that are a bit abnormal compared to the rest. As we can see, overall DNN$_\mathcal{H}$ achieved better results for the three measures. One interesting observation is that DNN (the base model $\mathcal{F}$ used for DNN$_\mathcal{H}$), while obtained better MAE than ridge regression and COVID-GAN, substantially underestimates the total mobility, which could be a result of incorrect predictions on several mobility hotspots, whose patterns do not follow the global pattern. Similarly, GWR shows a similar trend. Although GWR performs spatially-localized regression, it can only handle simple linear relationships using input variables and apply the same spatial neighborhood for all locations, which cannot well capture non-stationary mobility hotspots and variation in the data. Finally, DNN$_\mathcal{H}$ automatically identified three heterogeneous partitions (other splits are statistically insignificant) and branched out downtown, suburban and several mobility hotspots (our method allows a single partition to contain multiple disjoint large-footprints), greatly improving the prediction performance.

Acknowledgments

Yiqun Xie is supported in part by NSF awards 2105133, 2126474 and 2147195, Google's AI for Social Good Impact Scholars program, and the DRI award at the University of Maryland; Erhu He and Xiaowei Jia are supported in part by NSF award 2147195, USGS award G21AC10207, Pitt Momentum Funds award, and CRC at the University of Pittsburgh; Han Bao and Xun Zhou are supported in part by the ISSSF grant from the University of Iowa, and SAFER-SIM funded by US-DOT award 69A3551747131.

References

[Atluri *et al.*, 2018] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41, 2018.

[Bao *et al.*, 2020] Han Bao, Xun Zhou, Yingxue Zhang, Yanhua Li, and Yiqun Xie. Covid-gan: Estimating human mobility responses to covid-19 pandemic through spatio-temporal conditional generative adversarial networks. In *Proceedings of the 28th international conference on advances in geographic information systems*, pages 273–282, 2020.

[Brunsdon *et al.*, 1999] Chris Brunsdon, A Stewart Fotheringham, and Martin Charlton. Some notes on parametric significance tests for geographically weighted regression. *Journal of regional science*, 39(3):497–524, 1999.

[CDL, 2021] USDA cropland data layer. https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php, 2021. Accessed: 2022-06-03.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[Fotheringham *et al.*, 2017] A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265, 2017.

[GEOGLAM, 2021] Group on earth observations global agricultural monitoring initiative. https://earthobservations.org/geoglam.php, 2021. Accessed: 2022-06-03.

[Goodchild and Li, 2021] Michael F Goodchild and Wenwen Li. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35), 2021.

[Gupta and others, 2020] Jayant Gupta et al. Towards spatial variability aware deep neural networks (svann): A summary of results. In *ACM SIGKDD workshop on deep learning for spatiotemporal data, app. & sys.*, 2020.

[Gupta and others, 2021] Jayant Gupta et al. Spatial variability aware deep neural networks (svann): A general approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(6):1–21, 2021.

[Jia *et al.*, 2019] Xiaowei Jia, Sheng Li, Ankush Khandelwal, Guruprasad Nayak, Anuj Karpatne, and Vipin Kumar. Spatial context-aware networks for mining temporal discriminative period in land cover detection. In *SDM*, pages 513–521. SIAM, 2019.

[Jiang *et al.*, 2019] Zhe Jiang, Arpan Man Sainju, Yan Li, Shashi Shekhar, and Joseph Knight. Spatial ensemble learning for heterogeneous geographic data with class ambiguity. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4):1–25, 2019.

[Kraemer and others, 2020] Moritz UG Kraemer et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[Kratsios and Bilokopytov, 2020] Anastasis Kratsios and Ievgen Bilokopytov. Non-euclidean universal approximation. *NeurIPS*, 33, 2020.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[Kulldorff and others, 2007] Martin Kulldorff et al. Multivariate scan statistics for disease surveillance. *Statistics in medicine*, 26(8):1824–1833, 2007.

[Neill and others, 2013] Daniel B Neill et al. Fast subset scan for multivariate event detection. *Statistics in medicine*, 32(13):2185–2208, 2013.

[Neill, 2012] Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society*, 74(2):337–360, 2012.

[Shekhar *et al.*, 2015] Shashi Shekhar, Steven K Feiner, and Walid G Aref. Spatial computing. *Communications of the ACM*, 59(1):72–81, 2015.

[Xie *et al.*, 2021a] Yiqun Xie, Erhu He, Xiaowei Jia, Han Bao, Xun Zhou, Rahul Ghosh, and Praveen Ravirathinam. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 767–776. IEEE, 2021.

[Xie *et al.*, 2021b] Yiqun Xie, Xiaowei Jia, Han Bao, Xun Zhou, Jia Yu, Rahul Ghosh, and Praveen Ravirathinam. Spatial-net: A self-adaptive and model-agnostic deep learning framework for spatially heterogeneous datasets. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, pages 313–323, 2021.

[Xie *et al.*, 2021c] Yiqun Xie, Shashi Shekhar, and Yan Li. Statistically-robust clustering techniques for mapping spatial hotspots: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[Yao *et al.*, 2019] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *The World Wide Web Conference*, pages 2181–2191, 2019.