Pilot Analysis for: Learning to Encode Multi-level Dynamics in Effect Heterogeneity Estimation

Fucheng Warren Zhu* **Connor T. Jerzak †** **Adel Daoud‡**

Abstract

Earth Observation (EO) data are increasingly used in policy analysis by enabling granular estimation of treatment effects. However, a challenge in EO-based causal inference lies in balancing the trade-off between capturing fine-grained individual heterogeneity and broader contextual information. This paper introduces algorithms that address this challenge by combining satellite images of different sizes to estimate Conditional Average Treatment Effects (CATEs). This multi-scale approach employs Vision Transformer (ViT) models fine-tuned on satellite images, then applied to images of varying patch sizes to capture both household- and neighborhood-level information. We first perform simulation studies, showing how a multi-scale approach captures multi-level dynamics that single-scale ViT models fail to capture. We then apply the multi-scale method to two randomized controlled trials (RCTs) conducted in Peru and Uganda using Landsat satellite imagery. The Rank Average Treatment Effect (RATE) measure is employed in this pilot analysis to assess performance without ground truth individual treatment effects. The results of this analysis indicate that our dual-size inference technique improves the performance of deep learning models in EO-based CATE estimation.

Introduction

Earth Observation (EO) data play an increasing role in policy analysis because these data provide researchers with contextual information to estimate treatment effects at a more granular level, with that contextual information relating to environmental conditions, land use patterns, economic development, urban design, and climate variables (Anderson et al., 2017). A growing body of work therefore focuses on estimating household or neighborhood-specific Conditional Average Treatment Effects (CATE) (Sakamoto et al., 2024; Jerzak, Fredrik Johansson, et al., 2023; Serdavaa, 2023; Giannarakis et al., 2023; Go et al., 2022). A characteristic of EO-based representation learning is the inherent multi-level setting of each observational unit, where information about heterogeneity is encoded in the local area around a household and also the broader neighborhood around which the household is situated (Xiong et al., 2022). To our knowledge, there is yet no established methodology for incorporating these dynamics in EO-based causal inference.

Causal Representation Learning Workshop at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

*Undergraduate, Department of Statistics, Harvard University. ORCID: 0009-0001-5692-7572. Email: wzhu@college.harvard.edu

†Assistant Professor, Department of Government, University of Texas at Austin. ORCID: 0000-0003-1914-8905. Email: connor.jerzak@austin.utexas.edu URL: ConnorJerzak.com

‡Linköping University. ORCID: 0000-0001-7478-8345. Email: adel.daoud@liu.se URL: AdelDaoud.se AI & Global Development Lab: global-lab.ai

Multi-level dynamics call for more careful examination into the size of images used to estimate causal effects for households or neighborhoods, as well as the representation functions used to generate useful features from the satellite image information for downstream causal analysis. Larger images, while providing extensive contextual information, often lead to significant image overlap within villages (see Figure 5), making it difficult for models to capture individual-level effect heterogeneity. Conversely, smaller image sizes, although better suited for detecting localized effects, may lack sufficient neighborhood-level context, resulting in higher variance and less robust estimates. When conducting inference with images of a single scale, there is an inherent trade-off between individual-specific and broader contextual information. This household-neighborhood tradeoff, however, could be overcome if images of different scales are used for inference. Designing algorithms to learn multi-level dynamics from multi-scale imagery thus presents an important methodological challenge.

We address this methodological challenge by proposing algorithms that combine satellite images of different sizes to capture both fine-grained individual-level details and broader contextual information. We employ a remote-sensing fine-tuned image model, CLIP-RSICD (Lu et al., 2017; Arutiunian et al., 2021; Radford et al., 2021), for multi-scale inference, generating combined numerical representations used to estimate CATEs using the approach developed in Jerzak, F. Johansson, et al. (2023).

To quantify performance in real randomized controlled trials (RCTs) by evaluating the importance of household-level and neighborhood-level information for heterogeneity, we examine Rank Average Treatment Effect (RATE) values (Yadlowsky et al., 2021). By quantifying the gains from our multi-scale approach, we can explore if we can obtain meaningful improvements in effect estimation accuracy and targeting efficiency, potentially increasing the impact of poverty alleviation programs without additional resource expenditure.

Due to the difficulty in evaluating model performance on RCT data where ground-truth CATE is unknown, simulation studies are also used to validate results from the RCT analysis. The simulations show that multi-scale inference significantly outperforms single-scale approaches when heterogeneity information exists at multiple levels.

Finally, although in this paper we have only investigated the effect of multi-level modeling with varying image sizes, we note that this technique could be applied to the selection of temporal information or with images of varying resolutions. Multi-level dynamics can be encoded into a model by combining high temporal or spatial resolution satellite imagery data centered at a household with larger, lower-resolution satellite imagery of the household neighborhood to improve CATE estimation.

1 Background and Contributions

An enduring question in EO-based causal inference is how to best estimate $\tau(\mathbf{m}) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{M}_i = \mathbf{m}]$, i.e., the CATE for pre-treatment image array, $\mathbf{M}_i$ (Jerzak, F. Johansson, et al., 2023; Athey et al., 2018). In theory, to obtain the most accurate causal estimate, one would provide all available covariates to an oracle function generating estimated $\tau(\mathbf{m})$'s. For EO-based causal inference, this would ideally involve using the largest, highest-resolution satellite image available.

However, complexities are introduced in practice when we estimate $\tau(\mathbf{m})$ using a model $g_\theta(\mathbf{m})$ with parameters θ. The performance of $g_\theta(\mathbf{m})$ can degrade if the parameter estimation process is not probabilistically principled (e.g., lacking proper likelihood-based or Bayesian methods), or when there is a distribution shift between the training and inference data (e.g. of lower resolution) (Meng et al., 2014; Wu, 2021). For example, a pre-trained image model may have difficulty capturing both household-specific and neighborhood-level contextual information. The addition of extra information could in this case even degrade performance as the model has trouble distinguishing relevant from non-relevant information.

When leveraging the representation-extraction power of state-of-the-art pre-trained image models for EO-based causal inference, it is difficult for the image models, pre-trained on classification or self-supervised learning tasks, to disentangle the multi-level dynamics in causal inference. At the same time, due to the scarcity of RCT data, it is infeasible to capture these multi-level dynamics by training a model on RCT data alone. Therefore, techniques

must be developed to encode multi-level dynamics in effect heterogeneity modeling, both with representation concatenation and eventually with model fine-tuning.

Our main contribution is to develop a way of encoding multi-level dynamics in effect heterogeneity modeling by systematically varying the image size and combining image representations at different scales. We examine this inference pipeline in the context of randomized controlled trials (RCTs) conducted in Peru and Uganda.

Figure 1 provides some intuition for the contribution. We see in the left panel a satellite image of a household (in this case, the Washington-Longfellow National Historic Site in Cambridge, MA, USA). The right panel contains the same information about this household but also broader information about the context surrounding this household. In analyzing how someone living in this household may respond differentially to an intervention (e.g., voter turnout or economic intervention), both scales might be relevant. However, there is little methodological guidance as to how—motivating the core work of this paper.

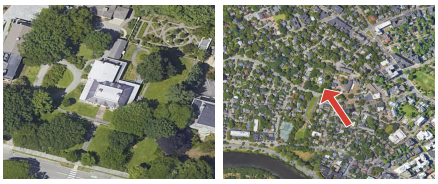

Figure 1: *The Washington-Longfellow National Historic Site* (LEFT)*, with context* (RIGHT)*.*

2 Methodology

Let i index the experimental units in the study. Each i has a geo-location denoted by $\mathbf{x}_i \in \mathbb{R}^2$, representing spatial coordinates (e.g., latitude and longitude), and a binary treatment indicator W_i encoding whether the unit received treatment. The observed outcome from the RCT for unit i is $Y_i \in \mathbb{R}$. Let $Y_i(1)$ and $Y_i(0)$ denote potential outcomes under treatment and control. Identification will be performed assuming unconfoundedness and SUTVA.

Define a function $f_I : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathcal{M}$, where $\mathcal{M}$ is the space of images. This function generates an image centered at a given location with a specified size:

$$\mathbf{M}_{i,s} = f_I(\mathbf{x}_i, s) \in \mathcal{M},$$

where $\mathbf{M}_{i,s}$ is the image of size $s > 0$ centered at location $\mathbf{x}_i$.

Next, we introduce a causal representation extraction function $\phi : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^d$, which takes two images as input and outputs a d-dimensional feature vector:

$$\boldsymbol{\phi}_{i,s_1,s_2} = \phi\left(\mathbf{M}_{i,s_1}, \mathbf{M}_{i,s_2}\right) \in \mathbb{R}^d,$$

where s_1 and s_2 are two different image sizes. With representation concatenation, $\phi\left(\mathbf{M}_{i,s_1}, \mathbf{M}_{i,s_2}\right) = (\phi'(\mathbf{M}_{i,s_1}), \phi'(\mathbf{M}_{i,s_2}))$ for a single image encoder ϕ'.

Given SUTVA, the CATE given representation $\boldsymbol{\phi}_{i,s_1,s_2}$ is defined as:

$$\tau\left(\boldsymbol{\phi}_{i,s_1,s_2}\right) \coloneqq \mathbb{E}\left[Y_i(1) - Y_i(0) \mid \boldsymbol{\phi}_{i,s_1,s_2}\right].$$

We estimate the CATE using a function, $g : \mathbb{R}^d \to \mathbb{R}$: $\hat{\tau}_i = g_\theta\left(\boldsymbol{\phi}_{i,s_1,s_2}\right)$, where $\hat{\tau}_i$ is the estimated CATE for unit i based on the extracted features.

To estimate the CATEs, we need to estimate both the representation extraction function ϕ and the estimation function g_θ. Our goal is to design a function ϕ that extracts the most causally relevant covariates from the image at multiple levels of representation (e.g., household and village). Having estimated ϕ, we draw upon the Causal Forest approach, a well-established method for estimating g_θ under unconfoundedness (Athey and Wager, 2019).

With this algorithm for estimating CATEs, we then seek to maximize the heterogeneity signal of different multi-scale representations using a metric, $\mu(\cdot)$, designed to quantify the extent of effect heterogeneity detected given input features. In our case, $\mu(\cdot)$ will denote the RATE ratio as a principled metric that allows one to evaluate model performance without

ground truth individual treatment effects, overcoming the unobservability of true CATE values (Yadlowsky et al., 2021). For a set of image conditioning variables, $\mathbf{M}_i$, the RATE is:

$$\text{RATE} = \int_0^1 \alpha(q) \Bigg(\underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid F(\tau(\mathbf{M}_i)) \geq 1 - q]}_{\text{ATE among top } q\text{-th percentile under rule } \tau} - \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{\text{Baseline ATE}} \Bigg) \, \mathrm{d}q, \quad (1)$$

See §7.1 for details. The RATE ratio (our measure μ) is defined as $\frac{\text{RATE}}{\text{se(RATE)}}$. Motivation for using the RATE ratio lies in the fact that it can be used as an asymptotic t-statistic of the existence of treatment heterogeneity in the population given the conditioning data (here, pre-treatment satellite image arrays).

With this way of generating representations from images, CATEs from representations, and heterogeneity measures from CATEs, we now formalize our optimization when performing multi-scale inference.[4] Our optimization proceeds by comparing a multi-scale heterogeneity signal against a baseline of comparison involving the optimal single-scale-only input. Specifically, we have the following optimization:

$$\textit{Goal: } \text{maximize}_{s_1, s_2} \left\{ \mathbb{E}\Big[\mu \left(g \left(\phi \left(\mathbf{M}_{i,s_1}, \mathbf{M}_{i,s_2} \right) \right) \right) \Big] - \max_s \mathbb{E}\Big[\mu \left(g \left(\phi \left(\mathbf{M}_{i,s}, \mathbf{M}_{i,s} \right) \right) \right) \Big] \right\}, \quad (2)$$

Expectation are taken over population variability. While optimization over s_1 and s_2 does not depend on s, we include the term involving s to establish a baseline—if the loss is below 0, we have evidence in favor of using single-scale-only representations. This optimization task codifies the task of quantifying multi-level dynamics in effect heterogeneity estimation.

For this pilot analysis, we use grid search to optimize Equation 2 (see §7.3 for details).

3 Simulation

While the RATE ratio is a useful tool, large-scale evaluation in a causal context is difficult due to the lack of ground-truth CATE data. To overcome this evaluation challenge, we use a simulation to supplement findings from the later RCT analysis. Here, we employ a prediction-oriented framework, with the 5-fold cross-validated out-of-sample R^2 as our quantity of interest to measure how well we identify features driving the outcome of interest.

In our simulation, we employ images of size 32×32 and 256×256 pixels drawn from the Peru RCT. We design three image perturbations corresponding to three scales of causal features (see Figure 2). First, we perturb the images with *household-level* features by masking the center of the image, with *neighborhood-level* features by adding an image fading to the edge of the larger scale image, and with *global context features* by applying image contrast.

In each experiment, a set of perturbations are chosen; for each perturbation, half of the satellite images over RCT participants are independently sampled to be perturbed; synthetic outcomes are constructed by adding a deterministic signal with Gaussian noise

$$Y_i \sim \begin{cases} \mathcal{N}(\mu = 0, \sigma^2 = 0^2) & \text{if } i \text{ is not perturbed,} \\ \mathcal{N}(100, 100^2) & \text{if } i \text{ is Masked,} \\ \mathcal{N}(-100, 100^2) & \text{if } i \text{ is Edge Faded,} \\ \mathcal{N}(100, 100^2) & \text{if } i \text{ is Contrasted,} \\ \mathcal{N}(0, 200^2) & \text{if } i \text{ is Masked and Edge Faded.} \end{cases}$$

With the outcome and image data defined, we then train a Multi-Layer Perceptron (MLP) on top of representations generated by the CLIP model using the perturbed and, separately, unperturbed images, computing the resulting R^2 in both cases.

For a single-scale approach, the input to the MLP is the representation generated by the CLIP image encoder of an image of a fixed size. We employ a simple multi-scale approach for our simulation, concatenating representations generated by the two image sizes.

Three sets of experiments are performed on the perturbed dataset to explore performance of a multi-scale modeling approach when (a) only household or neighborhood level information are present in the dataset, (b) when both levels of information are present in the dataset, and

[4]For simplicity, we leave out discussion of test set evaluation, although in practice this is desirable.

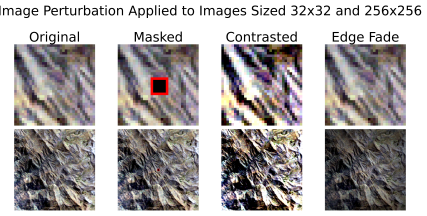

Figure 2: *Image perturbations are visualized for an image of size 32×32 and 256×256. Images are not centered around RCT participants due to privacy considerations. Masking used in the simulation experiments is 2×2 pixels, here enlarged for visibility.*

(c) when the global feature is present in the dataset. We found that our simple concatenation-based multi-scale modeling can recover most of the information present on one level of resolution and capture signal simultaneously from both scales if present. Perhaps more surprisingly, multi-scale modeling also better captures global features (see Figure 3).

We observe that models are fragile to seemingly innocent perturbations to the image to detect the causal features. By applying contrast perturbation, for example, our model's ability to detect Edge Fade and Centered Mask decreased. This motivates further research into a robust fine-tuning pipeline for robust causal feature detection.

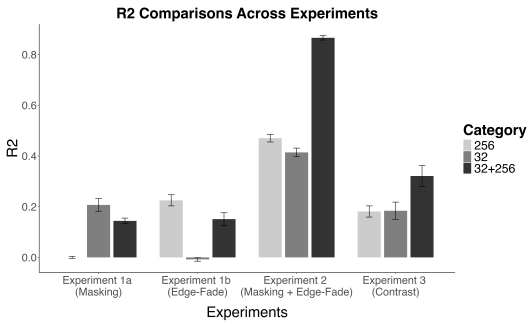

Figure 3: *Experiments 1a and 1b apply household/neighborhood-specific perturbation. Experiment 2 applies household and neighborhood-specific perturbation. Experiment 3 applies global perturbation that has a uniform effect on the whole image. Experiment 1 shows that even if there are no cross-scale effects, the simple multi-scale inferential procedure can recover much of the signal at one scale. Experiment 2 validates that the procedure can capture multi-level signals if present and outperform a single scale approach. Experiment 3 suggests that a multi-scale architecture would allow the model to capture predictive signals better even when the signal is recoverable from any one scale.*

4 Application to Anti-Poverty RCTs

While the simulation results are suggestive of the potential benefits of multi-scale analysis, we now turn to quantify its benefits in the context of real RCTs. Our analysis here draws upon unique experimental datasets from diverse country contexts: Peru and Uganda. For both, there is evidence that SUTVA is satisfied as spillover effects are determined to be unlikely and treatment implementation is standardized. These datasets therefore provide a rich foundation for exploring the impacts of scale across different societal settings.

The Peru dataset is drawn from the Multifaceted Graduation Program (Banerjee et al., 2015), with treatment occurring from 2007-2014. We use Landsat 5 and 7 Satellite Imagery between 2000 to 2003 due to image availability, applying a cloud mask and median filter over the satellite images, visualized in Figure 4. Images have 30 by 30 m pixels, with experiments performed on images of varying width and height around each household. In this application, we study household poverty as it responds to a multi-faceted intervention that combines short-term aid and long-term support for the very poor.

The Uganda RCT was also designed to reduce poverty, here, by giving young people business grants (Blattman et al., 2020); data are again drawn from Landsat. Whereas, for Peru, we have geolocations at the household level, for Uganda, geolocations are made for villages.

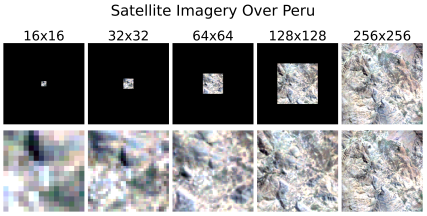

Figure 4: *Peru images size 16×16 pixels to 256 by 256 pixels from* 2007*. Image resolution is held at* 30*m the highest resolution for the Landsat* 5 *and* 7 *satellites. Images are not centered around RCT participants for privacy considerations. The top row shows images without resizing across changing dimensions; bottom panel shows resized images.*

Because the Peru geolocations are at the household level, we can analyze the average pairwise image overlap of individuals in the same village in our data is shown below, showing significant difference in percentage overlap as image size varies (see left panel of Figure 5). The image representation distance similarly decreases as the image size increases (see right panel), emphasizing the need to use large and small images to obtain heterogeneous representations between individuals inside the same village.

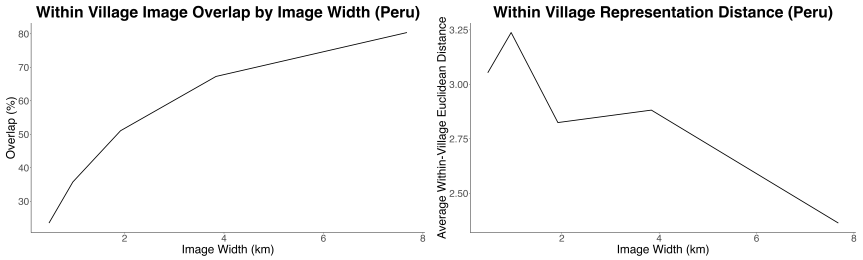

Figure 5: Left. *Average pairwise overlap of input images for individuals in a Peru village when image width is varied.* Right. *For CLIP-RSICD, the mean Euclidean distance of representations inside the same village decreases with increasing image width.*

We next conduct a multi-scale RCT analysis using the CLIP-RSICD image model. Because CLIP-RSICD is not designed to extract representations at varying scales, we will concatenate the first 50 principal components of household image representations obtained from images of widths {16, 32, 64, 128, 256, 349} (i.e., 0.5-10 km).

In Figure 6, we see that the optimal RATE ratio occurs when we concatenate representations from a large with those from a more localized context. Interestingly, the optimal choice in Uganda (where we have village-only geolocations) involves a larger localized context than that in Peru—where we have exact household geolocations and where precise image information from around a household most improves CATE signal when combined with broader context.

Overall, we see that dual-size inference techniques may significantly improve the performance of deep learning models in EO-based CATE estimation. By integrating small and large images, we better capture both individual heterogeneity and neighborhood-level effects, which allows for more informative causal inferences.

5 Towards Individualized Multi-scale Causal Representations

Despite grid search's promising performance, it is computationally infeasible when extending to individualized multi-scale analysis or even to non-individual multi-scale analysis involving more than two resolutions. Grid search is also architecturally unprincipled (as we use architectures designed for a single resolution across multiple resolutions).

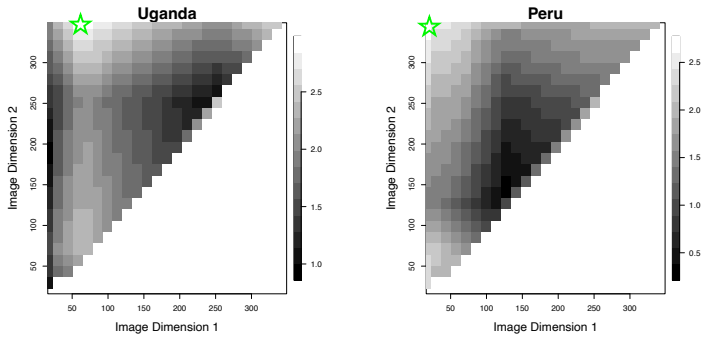

Figure 6: *Analysis of RATE ratios for Uganda and Peru RCTs across a range of s_1 and s_2 values. We see that the maximum heterogeneity signal in both RCTs is detected with small/medium-sized s_1 ($\sim$64) and large s_2 ($\sim$350). $\star$ indicates optimal selection.*

We aim to allow for individualized multi-scale analysis to allow the multi-resolution choice for each unit to be selected in a data-dependent manner. This adaptivity may be particularly relevant for analyses involving individuals in both rural and urban areas, as the notion of scale can differ markedly across those contexts. We present the optimization problem with just two scales, but we could employ additional values based on CATE estimation context.

The optimization problem for adaptively tuning s_1 and s_2 for different images is then:

$$\text{maximize}_{\{s_{1(i)}, s_{2(i)}\}_{i=1}^n} \left\{ \mathbb{E}\left[\mu\left(g\left(\phi\left(\mathbf{M}_{i,s_{1(i)}}, \mathbf{M}_{i,s_{2(i)}}\right)\right)\right)\right] - \max_s \mathbb{E}\left[\mu\left(g\left(\phi'\left(\mathbf{M}_{i,s}, \mathbf{M}_{i,s}\right)\right)\right)\right] \right\},$$

where $\{s_{1(i)}, s_{2(i)}\}_{i=1}^n$ represents the set of image sizes adaptively chosen for each unit based on image information. In this equation, we seek to find the set of individualized sizes that maximizes the expected improvement in the heterogeneity measure, μ, over a baseline.

We present as an open problem whether it might be feasible to remove the need for two image sizes altogether—setting a single s as the largest patch size and training a single model to optimize the image regions attended. This approach faces significant computational challenges when imagery is of high resolution (Bakhtiarnia et al., 2022). Moreover, many vision foundational models, especially those using Transformer backbones, are trained with a fixed patch dimension (e.g., 224×224, as in the case of CLIP). This could be overcome through sub-sampling imagery over a large scale and sampling specific locations to zoom into to extract fine-grained local signals. RL-based approaches (where an agent navigates sub-samples of high-resolution satellite imagery to identify regions with heterogeneity information) could also allow for individualized multi-scale inference (Rocamonde et al., 2024).

6 Limitations & Conclusion

In this paper, we have addressed the methodological challenge of capturing multi-level dynamics in EO-based causal inference by leveraging multi-scale image representations. By combining representations across scales, our approach effectively captures both fine-grained individual-level details and broader contextual information, enhancing the estimation of CATEs. Simulation studies and analysis of two RCTs demonstrate the promise of multi-scale inference in outperforming single-scale-only methods when effect heterogeneity information exists at multiple levels. This offers a promising solution to the inherent trade-off between individual heterogeneity and neighborhood-level context in causal inference and contributes to the growing literature of deep learning architectures for EO-based causal inference.

Limitations, however, remain. First, the approach here assumes SUTVA and unconfoundedness (i.e., RCT data) for identification. Second, images used here have low resolution; hence, many causal signals may be undetectable. Third, using high-resolution images in a multi-scale approach raises significant privacy concerns. Further research could be done to address these limitations of the proposed multi-scale effect heterogeneity algorithm. □

References

Anderson, Katherine et al. (2017). "Earth observation in service of the 2030 Agenda for Sustainable Development". In: *Geo-spatial Information Science* 20.2, pp. 77–96.

Arutiunian, Artashes et al. (2021). *CLIP-RSICD v2*. https://huggingface.co/flax-community/clip-rsicd-v2.

Athey, Susan et al. (2018). "The impact of machine learning on economics". In: *The economics of artificial intelligence: An agenda*, pp. 507–547.

Athey, Susan and Stefan Wager (2019). *Estimating Treatment Effects with Causal Forests: An Application*. arXiv: 1902.07409 [stat.ME]. URL: https://arxiv.org/abs/1902.07409.

Bakhtiarnia, Arian et al. (2022). *Efficient High-Resolution Deep Learning: A Survey*. arXiv: 2207.13050 [cs.CV]. URL: https://arxiv.org/abs/2207.13050.

Banerjee, Abhijit et al. (2015). "A multifaceted program causes lasting progress for the very poor: Evidence from six countries". In: *Science* 348.6236, p. 1260799. DOI: 10.1126/science.1260799. eprint: https://www.science.org/doi/pdf/10.1126/science.1260799. URL: https://www.science.org/doi/abs/10.1126/science.1260799.

Blattman, Christopher et al. (2020). "The long-term impacts of grants on poverty: Nine-year evidence from Uganda's Youth Opportunities Program". In: *American Economic Review: Insights* 2.3, pp. 287–304.

Giannarakis, Georgios et al. (July 2023). *Understanding the Impacts of Crop Diversification in the Context of Climate Change: A Machine Learning Approach*. arXiv:2307.08617 [cs, q-bio] version: 1. URL: http://arxiv.org/abs/2307.08617 (visited on 04/16/2024).

Go, Eugenia et al. (Mar. 2022). *On the Use of Satellite-Based Vehicle Flows Data to Assess Local Economic Activity: The Case of Philippine Cities*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4057690. URL: https://papers.ssrn.com/abstract=4057690 (visited on 04/16/2024).

Jerzak, Connor T., F. Johansson, et al. (2023). "Image-based Treatment Effect Heterogeneity". In: vol. 213, pp. 531–552. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147679914&partnerID=40&md5=7866b08ec210422443d7ad823941845f.

Jerzak, Connor T., Fredrik Johansson, et al. (2023). "Image-based Treatment Effect Heterogeneity". In: *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR), Proceedings of Machine Learning Research (PMLR)* 213, pp. 531–552.

Lu, Xiaoqiang et al. (2017). "Exploring Models and Data for Remote Sensing Image Caption Generation". In: *IEEE Transactions on Geoscience and Remote Sensing* 56.4, pp. 2183–2195. DOI: 10.1109/TGRS.2017.2776321.

Meng, Xiao-Li et al. (Feb. 2014). "I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb?" In: *Econometric Reviews* 33. DOI: 10.1080/07474938.2013.808567.

Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Rocamonde, Juan et al. (2024). *Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning*. arXiv: 2310.12921 [cs.LG]. URL: https://arxiv.org/abs/2310.12921.

Sakamoto, Kazuki et al. (2024). *A Scoping Review of Earth Observation and Machine Learning for Causal Inference: Implications for the Geography of Poverty*. arXiv: 2406.02584 [cs.LG]. URL: https://arxiv.org/abs/2406.02584.

Serdavaa, Batkhurel (Dec. 2023). *A Satellite Image Analysis on Housing Conditions and the Effectiveness of the Affordable Housing Mortgage Program in Mongolia: A Deep Learning Approach*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4664966. URL: https://papers.ssrn.com/abstract=4664966 (visited on 04/16/2024).

Wu, Yifan (Sept. 2021). "Learning to Predict and Make Decisions under Distribution Shift". Thesis Committee: Zachary Lipton (Chair), Andrej Risteski, Sivaraman Balakrishnan, Alexander Smola (Amazon). Ph.D. Thesis. Pittsburgh, PA: Carnegie Mellon University.

Xiong, Wei et al. (2022). "A Confounder-Free Fusion Network for Aerial Image Scene Feature Representation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 5440–5454. ISSN: 2151-1535. DOI: 10.1109/JSTARS.2022.3189052. URL: https://ieeexplore.ieee.org/document/9817622 (visited on 04/16/2024).

Yadlowsky, Steve et al. (2021). "Evaluating treatment prioritization rules via rank-weighted average treatment effects". In: *arXiv preprint arXiv:2111.07966*.

7 Supplementary Information

7.1 RATE Ratio Details

The RATE ratio is calculated through sample splitting, with $\hat{\tau}$ estimated from half of the samples, and $\hat{\mathbb{E}}$ from the other half (Yadlowsky et al., 2021). In Equation 1, $\hat{\tau}(\mathbf{M}_i)$ is used as a prioritization rule, with $\hat{\mathbb{E}}[Y_i(1) - Y_i(0) \mid F(\hat{\tau}(\mathbf{M}_i)) \geq 1 - q]$ being the Average Treatment Effect (ATE) among the top q-th percentile of treatment respondents as estimated by $\mathbf{M}_i$. $\mathbb{E}[Y_i(1) - Y_i(0)]$ is the baseline ATE, and the difference $\mathbb{E}[Y_i(1) - Y_i(0) \mid F(\hat{\tau}(\mathbf{M}_i)) \geq 1 - q] - \mathbb{E}[Y_i(1) - Y_i(0)]$ represents the gain in ATE in the respondents in the top q-th percentile of estimated CATE over the general population. Finally, this difference is weighed through $\alpha(q)$ and integrated to produce a scalar output. There are at least two weighting functions under which the RATE ratio ($\frac{\text{RATE mean}}{\text{RATE variance}}$) has hypothesis testing guarantees in detecting heterogeneity in a population—i.e., $\alpha_{AUTOC}(q) = 1$ and $\alpha_{QINI}(q) = q$. In our context, we report the AUTOC weighting function, which gives more weight to individuals with high response in the integration. This weighting is more relevant for policy analysis when not all individuals will be treated.

7.2 Simulation Details

Fold and perturbation indices are generated from simple random samples from available indices. We found that making outcomes Gaussian (rather than deterministic) had no statistically significant effect on model performance. We also found that the representations learned by the model are not always robust to small perturbations. By applying contrast image perturbation on top of other image perturbations to experiment 1 without changing the outcome, R^2 decreased by 55%. This provides motivation to develop model architectures or fine-tuning procedures that are more robust to noise in multi-scale features.

We used a Multi-Layer Perceptron (MLP) with three linear layers, each followed by ReLU activation. The input layer connects to a hidden layer of 128 neurons, followed by a second bottleneck layer of 32 neurons, and an output layer providing a scalar prediction. The architecture is mathematically described as follows:

$$f(\mathbf{x}) = \sigma(W_3 \cdot \sigma(W_2 \cdot \sigma(W_1\mathbf{x} + b_1) + b_2) + b_3)$$

where W_i and b_i represent layer weights and biases, and σ is the ReLU activation.

For the cross-scale model, input dimensionality is doubled, as two image representations are concatenated. We experimented with Principal Component Analysis (PCA) for dimensionality reduction; using only subsets of principal components often led to large performance degradations.

The Edge Fade image perturbation is implemented using a radial distance-based mask:

$$\texttt{mask} = \texttt{Clip}(1 - \texttt{distance} \times \texttt{fade\_size}, 0, 1),$$

and the Contrast image perturbation is applied by transforming each pixel values using,

$$M^{\text{New}}_{i,w,h,b} = \overline{M}_b + c \times (M_{i,w,n,b} - \overline{M}_b),$$

where $M_{i,w,h,b}$ is the original pixel value, $\overline{M}_b$ is the mean band intensity, and c scales the contrast.

7.3 Algorithms

Algorithm 1 Grid search optimizing multi-scale representations in CATE estimation.

Input:

- $i \in \{1, \cdots, n\}$ denote the index for observational units.
- $\{\mathbf{x}_i\}_{i=1}^n$ the sets of locations of those units.
- Sets of image sizes $\mathcal{S} = \{s_1, s_2, ..., s_k\}$.
- Image fetcher, $f_I(\mathbf{x}_i, s)$, that obtains an image centered at a given location $\mathbf{x}_i$ with a size s.
- Image encoder, ϕ', that extract representations from $\mathbf{M}_{i,s}$.
- Trainable CATE estimation function $g_\theta()$ parametrized by θ.
- Observed outcome of interest $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$.
- Binary treatment indicator $\mathbf{W} = (W_1, W_2, \cdots, W_n), W_i \in \{0, 1\}$

Output: Optimal image sizes s_1^*, s_2^*, with optimal RATE ratio*

Grid Search over s_1 and s_2:
Initialize $s_1^* \leftarrow 0$, $s_2^* \leftarrow 0$
foreach $s_1 \in S_1$ **do**
  foreach $s_2 \in S_2$ **do**
    Set `MaxRATERatio` to $-\infty$
    foreach $i, \mathbf{x}_i \in$ *enumerate*$(\{\mathbf{x}_i\}_{i=1}^n)$ **do**
      $\mathbf{M}_{i,s_1} = f_I(\mathbf{x}_i, s_1)$; $\mathbf{M}_{i,s_2} = f_I(\mathbf{x}_i, s_2)$
      $\phi_{i,s_1,s_2} = (\phi'(\mathbf{M}_{i,s_1}), \phi'(\mathbf{M}_{i,s_2}))$
      end
      Call the RATE ratio calculation function, R:
        $\widehat{\text{RATE Ratio}} = R(\mathbf{W}, \mathbf{Y}, \{\phi_{i,s_1,s_2}\}_{i=1}^n)$
      if $\widehat{RATE\ Ratio} >$ *MaxRATERatio* **then**
        `MaxRATERatio` $\leftarrow \widehat{\text{RATE Ratio}}$ $s_1^* \leftarrow s_1$ $s_2^* \leftarrow s_2$
      end
    end
  end
  return *Optimal sizes* s_1^*, s_2^*, *MaxRateRatio*

Data Availability Statement

Simulation code is to be made available on GitHub:

`GitHub.com/AIandGlobalDevelopmentLab/MultiScaler`

Uganda replication data are available at:

`doi.org/10.7910/DVN/O8XOSF`

For privacy reasons given household-level geolocations in Peru, we cannot make that RCT data available at this time.