Proceedings of Machine Learning Research 275:1–24, 2025 4th Conference on Causal Learning and Reasoning

Optimizing Multi-Scale Representations to Detect Effect Heterogeneity Using Earth Observation and Computer Vision: Applications to Two Anti-Poverty RCTs

Fucheng Warren Zhu WZHU@COLLEGE.HARVARD.EDU
Department of Statistics, Harvard University
Department of Computer Science, Harvard University

Connor T. Jerzak CONNOR.JERZAK@AUSTIN.UTEXAS.EDU
Department of Government, University of Texas at Austin

Adel Daoud ADEL.DAOUD@LIU.SE
Institute for Analytical Sociology, Linköping University
Chalmers University of Technology

Editors: Biwei Huang and Mathias Drton

Abstract

Earth Observation (EO) data are increasingly used in policy analysis by enabling granular estimation of conditional average treatment effects (CATE). However, a challenge in EO-based causal inference is determining the scale of the input satellite imagery—balancing the trade-off between capturing fine-grained individual heterogeneity in smaller images and broader contextual information in larger ones. This paper introduces Multi-Scale Representation Concatenation, a set of composable procedures that transform arbitrary single-scale EO-based CATE estimation algorithms into multi-scale ones. We benchmark the performance of Multi-Scale Representation Concatenation on a CATE estimation pipeline that combines Vision Transformer (ViT) models (which encode images) with Causal Forests (CFs) to obtain CATE estimates from those encodings. We first perform simulation studies where the causal mechanism is known, showing that our multi-scale approach captures information relevant to effect heterogeneity that single-scale ViT models fail to capture as measured by R^2. We then apply the multi-scale method to two randomized controlled trials (RCTs) conducted in Peru and Uganda using Landsat satellite imagery. As we do not have access to ground truth CATEs in the RCT analysis, the Rank Average Treatment Effect Ratio (RATE Ratio) measure is employed to assess performance. Results indicate that Multi-Scale Representation Concatenation improves the performance of deep learning models in EO-based CATE estimation without the complexity of designing new multi-scale architectures for a specific use case. The application of Multi-Scale Representation Concatenation could have meaningful policy benefits—e.g., potentially increasing the impact of poverty alleviation programs without additional resource expenditure.

Keywords: Causal inference; Treatment effect heterogeneity; Earth observation; Image data; Multi-scale Inference; Probabilistic reasoning

1. Introduction

Earth Observation (EO) data play an increasingly important role in policy analysis by providing researchers with contextual information to estimate treatment effects at a more granular level, revealing characteristics of environmental conditions, land use patterns, economic development, and climate variables (Anderson et al., 2017; Daoud et al., 2023; Burke et al., 2021b; Pettersson et al., 2023; Kino et al., 2021). A growing body of work

© 2025 F.W. Zhu, C.T. Jerzak & A. Daoud.

therefore focuses on estimating household or neighborhood-specific Conditional Average Treatment Effects (CATE) (Sakamoto et al., 2024; Jerzak et al., 2023; Serdavaa, 2023; Giannarakis et al., 2023; Go et al., 2022; Daoud and Johansson, 2024; Shiba et al., 2022).

An important application of EO-based policy analysis lies in anti-poverty Randomized Controlled Trials (RCTs), where the outcome $\mathbf{Y}$ is the household income for the unit that the satellite imagery is centered on, and the treatment $\mathbf{W}$ is cash transfer to the household. The researcher's goal is to estimate effects conditioned on available satellite imagery $\mathbf{M}$. We denote this estimate for a given image, $\hat{\tau}(\mathbf{m})$. In order to do so, we encode $\mathbf{M}$ into a sufficiently low-dimensional representation, ϕ, that preserves maximal information about $\mathbf{Z}$, the set of unobserved background features that are relevant to treatment (e.g., the village geography/baseline wealth). See the white nodes of Figure 2 for illustration.

A key characteristic of EO-based images in the context of effect heterogeneity estimation is the inherent multi-scale nature of the relevant background features $\mathbf{Z}$ for observational units, a phenomenon termed "multi-scale dynamics" (Xiong et al., 2022; Reed et al., 2023). Figure 1 provides some intuition for such multi-scale dynamics in EO-based inference, where information relevant to heterogeneity is encoded in both the local area around a household and the broader neighborhood around which the household is situated.

Figure 1: *The Washington-Longfellow National Historic Site (Left), with context (Right). Using sentences in the RSICD dataset, CLIP's nearest-neighbor text embedding to the images are "the square with some halls and lawns is in a rectangular region which (sic.) is in the middle of the forest" and "here is a zone of apartment buildings between a river and a road' respectively, showing that the model extracts separate, novel information from the two images.*

Along with Figure 1, we provide a visualization in Figure 2 that provides a more formal illustration of our setting. Parts highlighted in blue refer to multi-scaled components of the DAG that were previously not considered in effect heterogeneity estimation methodologies.

Our contribution is to propose Multi-Scale Representation Concatenation, a procedure that transforms single-scale EO-based CATE estimation algorithms into multi-scale ones by concatenating the image representation tensors. We benchmark using a single-scale CATE estimation algorithm previously used in anti-poverty RCTs (Jerzak et al., 2023). A remote-sensing fine-tuned image model, CLIP-RSICD, is used to generate lower-dimensional representations of single-scaled satellite imagery (Lu et al., 2017; Arutiunian et al., 2021; Radford et al., 2021). A causal forest (CF) model then produces the final CATE estimates.[1]

Beyond building on top of previous work in Causal Inference, there is a rich body of work investigating multi-scale phenomena in Computer Vision which we have drawn inspiration from to develop Multi-Scale Representation Concatenation. We survey them in §8.4.

[1]We also report results for SWIN and Clay model architectures, but use CLIP results as baseline due to the possibility of text-based interpretability analyses and the wide use of this model class.

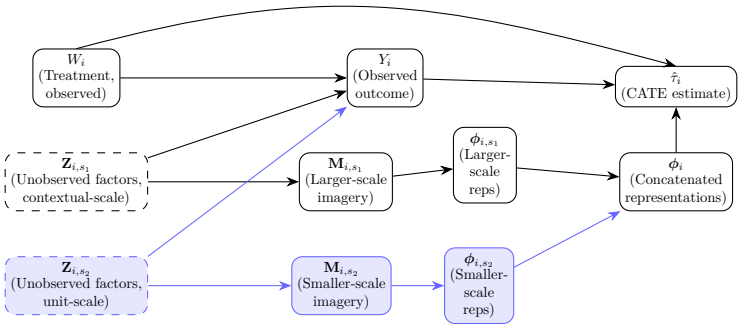

Figure 2: *A DAG showing the relationship between the various components involved in CATE estimation. The i subscript is the observation unit index; the subscript s (s_1 or s_2) is the scale of the variable. The blue components are the components of the figure that reflect multi-scale dynamics. Multi-Scale Representation Concatenation allows algorithms that previously were only a function of $\mathbf{M}_{i,s}$ for a single s to utilize both $\mathbf{M}_{i,s_1}$ and $\mathbf{M}_{i,s_2}$. Note that this diagram could be extended to additional scales. Image resolution is assumed fixed.*

To quantify performance in real randomized controlled trials (RCTs) lacking ground truth CATE data, we use a measure of the degree of effect heterogeneity detected by a model, Rank Average Treatment Effect Ratio (RATE Ratio), as our metric (Yadlowsky et al., 2021). We find evidence that applying Multi-Scale Representation Concatenation improves the RATE Ratio, and, thus, potentially improves accuracy of CATE estimates and policy targeting.

As extensions, simulation studies are used to validate results from the RCT analysis due to the difficulty in evaluating model performance on RCT data where ground-truth CATE is unknown (§4); a scaling analysis demonstrates that Multi-Scale Representation Concatenation can significantly boost the performance of CATE estimation models in EO-based causal inference beyond just two scales (§5.3); and an analysis where we assume the CATE analyst has limited prior information on the geographic distribution of household units finds that multi-scale inference improves the RATE metric under this weaker assumption (§5.2.2). These suggest promising further areas of exploration.

2. Background and Contributions

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

In this context, because there are well-understood methods to estimate CATEs from $\mathbf{W}$, $\mathbf{Y}$, and pre-treatment covariate vectors Chernozhukov et al. (2024), a natural approach in image-based CATE estimation would be to introduce an encoder ϕ that maps high-

dimensional imagery into a lower-dimensional representation. However, in an EO-based causal inference context, a pre-trained image encoder may struggle to capture both household-specific and neighborhood-level contextual information. In some cases, adding additional contextual information can even degrade performance if the model fails to distinguish relevant from non-relevant signals (D'Amario et al., 2022). Model fine-tuning could address this issue, yet it remains difficult due to the small size of most RCTs.

Our main contribution therefore is to develop Multi-Scale Representation Concatenation, a family of procedures that transform any previous single-scale procedure into a multi-scale one through tensor concatenation. It is especially suitable in data-constrained settings (e.g. RCTs), as it requires minimal additional data requirements. The strong interpretability of the procedure also makes it simple to explain to relevant stakeholders. Further, it requires limited model experimentation: The optimal concatenation procedure can be found by systematically varying the image size and combining image representations from different scales. Multi-Scale Representation Concatenation thus transforms the problem of designing architectures specific to the multi-scale dynamics of the problem at hand to a less complex inference-time computational search (Snell et al., 2024). We examine the improvements in model performance when incorporating Multi-Scale Representation Concatenation in the context of randomized controlled trials (RCTs) conducted in Peru and Uganda.

3. Methodology

Let i index the experimental units in the study. Each i has a geolocation denoted by $\mathbf{x}_i \in \mathbb{R}^2$, representing spatial coordinates (e.g., latitude/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 (Imbens and Rubin, 2016; Chernozhukov et al., 2024). Now define an image fetcher that generates an image from a coordinate $\mathbf{x}_i \in \mathbb{R}^2$ with a specified size $s \in \mathbb{N}_+$: $f_I : \mathbb{R}^2 \times \mathbb{N}_+ \to \mathcal{M}$, where $\mathcal{M}$ is the space of possible imagery:

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

where $\mathbf{M}_{i,s}$ is the image of size $s > 0$ centered at $\mathbf{x}_i$. Here, size refers to the width of the raw satellite imagery in terms of pixel number, with pixel resolution fixed.

Next, because of the high-dimensionality of $\mathbf{M}_{i,s}$, we need to introduce a causal representation extraction function $\phi : \mathcal{M}^c \to \mathbb{R}^d$, which takes $c \in \mathbb{N}_+$ images as input and outputs a d-dimensional feature vector. The projection of the image to a lower dimensionality makes estimation tractable in finite samples. We develop a Multi-Scale Representation Concatenation procedure that constructs ϕ from base image encoders, $f_{\phi'_{s_k}} : \mathcal{M} \to \mathbb{R}^{d'}$, to easily adapt single-scale representation extraction functions into a multi-scale setting.

Although our algorithm is generalizable to arbitrary c, for concreteness, our exposition will be limited to the case where $c = 2$, with $f_\phi : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^{2d'}$:

$$\boldsymbol{\phi}_{i,s_1,s_2} = f_\phi\left(\mathbf{M}_{i,s_1}, \mathbf{M}_{i,s_2}\right) = \left(f'_{\phi_{s_1}}(\mathbf{M}_{i,s_1}), f'_{\phi_{s_2}}(\mathbf{M}_{i,s_2})\right),$$

where s_1 and s_2 are two different image sizes. In this paper, f_ϕ is constructed through representation concatenation of the outputs of base image encoders $f'_{\phi_{s_k}}$ along with (optionally) a

dimensionality reduction function r, so that $f_\phi\left(\mathbf{M}_{i,s_1}, \mathbf{M}_{i,s_2}\right) = r((f'_{\phi_{s_1}}(\mathbf{M}_{i,s_1}), f'_{\phi_{s_2}}(\mathbf{M}_{i,s_2})))$. Here, $r : \mathbb{R}^{2d'} \to \mathbb{R}^l$, $l < 2d'$.

Given SUTVA and unconfoundedness, with W_i being the binary treatment indicator for unit i, the CATE given representation $\boldsymbol{\phi}_{i,s_1,s_2}$ is defined as:

$$\begin{aligned}\tau\left(\boldsymbol{\phi}_{i,s_1,s_2}\right) &:= \mathbb{E}\left[Y_i(1) - Y_i(0) \mid \boldsymbol{\phi}_{i,s_1,s_2}\right] \quad \textit{(SUTVA)} \\ &= \mathbb{E}\left[Y_i \mid \boldsymbol{\phi}_{i,s_1,s_2}, W_i = 1\right] - \mathbb{E}\left[Y_i \mid \boldsymbol{\phi}_{i,s_1,s_2}, W_i = 0\right] \quad \textit{(Unconfoundedness)} \qquad (1)\end{aligned}$$

Given $\boldsymbol{\phi}_{i,s_1,s_2}$ and unconfoundedness, CATE can be estimated using a function, $h_\theta : \mathbb{R}^{2d'} \to \mathbb{R}$: $\hat{\tau}_i = h_\theta\left(\boldsymbol{\phi}_{i,s_1,s_2}\right)$, where $\hat{\tau}_i$ is the estimated CATE for unit i based on the extracted features.

To estimate CATEs, we need to estimate both the representation extraction function ϕ and the estimation function h_θ. Our goal is to learn a function f_ϕ that extracts the most causally relevant covariates from the image at multiple levels of representation (e.g., household and village). Having found f_ϕ, we then draw upon the CF approach, a well-established method for estimating h_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 (Holland, 1986; Yadlowsky et al., 2021). For a set of image conditioning variables, $\mathbf{M}_i$, the RATE is:

$$\text{RATE} = \int_0^1 \alpha(q)\Bigg(\underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid F(\tau(\mathbf{M}_i)) \geq 1 - q]}_{\text{ATE among top } q\text{-th percentile under rule } \tau} - \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{\text{Baseline ATE}}\Bigg)\, \mathrm{d}q, \qquad (2)$$

A more detailed exposition is in §8.3. The RATE Ratio (the measure μ) is defined as $\frac{\text{RATE}}{\text{sd(RATE)}}$. Motivation for using the RATE Ratio lies in the fact that (1) it is interpretable and comparable across contexts as a scale-free quantification of heterogeneity signal and (2) 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). The detailed algorithm for RATE Ratio estimation is presented in §8.3.

The RATE Ratio is chosen as the primary metric for model evaluation due to its comparability across RCTs and outcomes. However, RATE Ratio is not as interpretable as the more policy-oriented metrics of RATE and Qini Curves (Yadlowsky et al., 2021; Sverdrup et al., 2024). RATE alone is a summary measure of heterogeneity in the data captured by the model and downstream policy relevance of the current heterogeneity model, and Qini Curves plot the gain over randomization of assigning treatment according to CATE estimates from a model. We validate the RATE Ratio metric with the values produced by these two alternatives. We find a high correlation between the three metrics, indicating that in our context RATE Ratio also quantifies the policy benefits of a specific CATE estimation model. Due to the high correlation, we leave RATE and Qini figures for the Appendix (Figure 12, Figure 13, Figure 14). For a discussion of Qini Curves, see §8.5.

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. 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:

$$\textit{Goal: } \max_{s_1,s_2} \left\{ \mathbb{E}\Big[\mu\left(h_\theta\left(f_\phi\left(\mathbf{M}_{i,s_1},\, \mathbf{M}_{i,s_2}\right)\right)\right)\Big] - \max_{s} \mathbb{E}\Big[\mu\left(h_\theta\left(f'_{\phi_s}\left(\mathbf{M}_{i,s}\right)\right)\right)\Big] \right\}, \tag{3}$$

Expectations are taken over population variability. The resulting optimized value is G, the Multi-scale Gain. While optimization over s_1 and s_2 does not depend on s, we include the term involving s to establish a baseline—if G is negative, we have evidence in favor of using single-scale-only representations. We use grid search for optimization (see §8.7).

4. Simulation

While the RATE Ratio is a useful theoretical metric, contextual-scale evaluation in an applied context is difficult due to the lack of ground-truth CATE data. In developing our methods, we thus use a simulation to supplement findings from the later RCT analysis. Our simulation employs 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 3). We perturb images' *household-level* features by masking the center of the image, *neighborhood-level* features by adding an image fading to the edge of the larger scale image, and *global context features* by applying image contrast. For each of our experiments, we choose a subset of these perturbations, each applied independently to half of the images. The different experiments help us discern how our model will perform when differently scaled covariates are present. Synthetic outcomes are then constructed by adding a deterministic signal with Gaussian noise (see Figure 3).

In our simulation, we design a setup that mimics an RCT. Each unit has an associated image, and we apply various perturbations—such as masking or edge fading—to these images. The specific type of perturbation applied to a unit's image determines its outcome (and associated CATE). We systematically test the ability of different modeling strategies to detect and leverage these image-based signals in estimation. We measure how well we identify features driving the outcome of interest with the 5-fold cross-validated out-of-sample R^2.

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

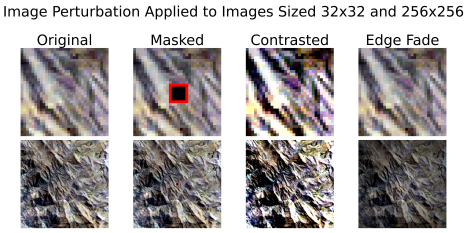

Figure 3: *Visualization of image perturbations (Right) and their corresponding distributions (Left). The equation defines the distribution of Y_i based on the perturbations applied, while the images illustrate these perturbations for different image sizes. Masking used in the simulation experiments is 2×2 pixels, enlarged here for visibility.*

With the outcome and image data defined, we then train a Multi-Layer Perceptron (MLP) on top of representations generated by the CLIP model to predict the outcome of

the units. We defer further details to the Appendix (§8.6). 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 then apply Multi-Scale Representation Concatenation with no dimensionality reduction to generate the multi-scale analog.

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 is present in the dataset, (b) when both levels of information are present in the dataset, and (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 captures signal simultaneously from both scales if present. Perhaps more surprisingly, multi-scale modeling also better captures global features (see Figure 4).

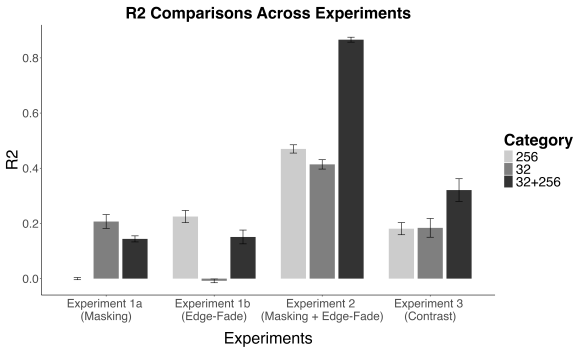

Figure 4: *Experiments 1a and 1b apply household/neighborhood-specific perturbation. Experiment 2 applies household and neighborhood-specific perturbation. Experiment 3 applies global perturbation that has a uniform effect on the whole image. Experiment 1 shows that even if there are no cross-scale effects, Multi-Scale Concatenated 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 Multi-Scale Representation Concatenation could improve a model's ability to capture predictive signals even when the signal is recoverable from any one scale.*

While the investigations present here isolate the influence of specific local, intermediate, and global perturbations, we investigate all possible combinations in the Appendix (Table 2), where an additional global perturbation of 90-degree image rotation is included. The results accord with those presented in Figure 4.

5. Application to Anti-Poverty RCTs

While the simulation results are suggestive of potential benefits of multi-scale analysis, we now turn to quantify its benefits in the context of real RCTs. Our analysis here draws on 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 solid foundation for exploring the impacts of scale across different societal settings.

5.1. Data: Treatment, Imagery, and Outcome

The Peru data are drawn from a program designed to alleviate poverty (Banerjee et al., 2015), with treatment occurring from 2007-2014. The outcome is household poverty; treatment is a multi-faceted intervention that combines short-term aid and long-term support. Due to image availability, we use Landsat 5 satellite imagery between 2000 to 2003 with 30 $\times$ 30 m pixels centered around each household, applying a cloud mask and median filter over images, visualized in the Appendix (Figure 10).

The Uganda RCT was also designed to reduce poverty, here, by giving young people business grants to improve human capital (Blattman et al., 2020). Treatment occurs from 2008-2012, and data are drawn from the Landsat 7 ETM+ Mosaics from 2000 for their high quality and image resolution (14.5$\times$14.5 m pixels). For Peru, geolocations at the household level are available, whilst for Uganda, geolocations are only available at the village level.

Because the Peru geolocations are at the household level, we can analyze the average pairwise image overlap of individuals in the same village. The analysis shows an increase in percentage overlap with image size (left panel of Figure 5). We further find that image representation distance decreases as the image size increases (right panel), emphasizing the need to use large and small images to obtain heterogeneous representations between individuals inside the same village.

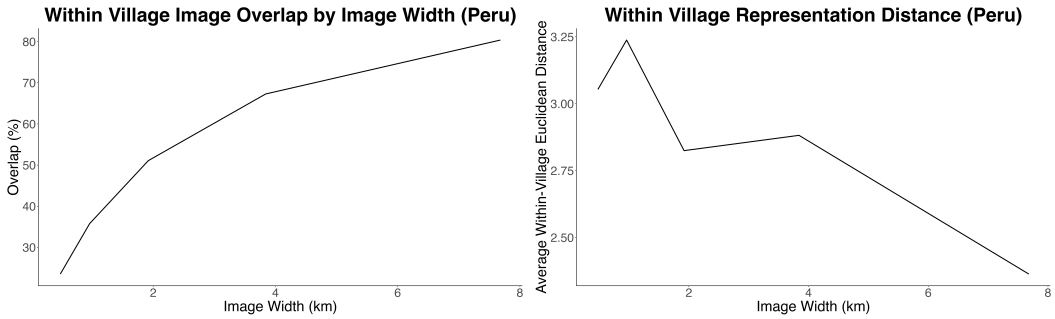

Figure 5: *Left. Average pairwise overlap of input images for individuals in a Peru village when image width is varied. Right. For CLIP-RSICD, the mean Euclidean distance of representations inside the same village decreases with increasing image width.*

5.2. Multi-scale RCT Analysis

We next conduct multi-scale RCT analysis on the two datasets using the CLIP-RSICD image model. Results from the SWIN and Clay model are reported in the Appendix (§8.9). We concatenate CLIP household image representations from images of widths {16, 32, 64, 128, 256, 349} (i.e., 0.5-10 km) and use a CF as CATE estimator. RATE Ratio is used as our evaluation metric, with results for RATE and Qini Curves in the Appendix (§8.9). Due to the composability of our multi-scale procedure, different base image encoders can be chosen for images at different scales. For simplicity, we use the same base encoder throughout.

Overall, we find that Multi-Scale Representation Concatenation improves or does no worse than single-scale-only analyses. Using a suitable dimensionality reduction function r further increases the performance of a multi-scale procedure.

5.2.1. Improvement Over Single-Scale Baseline

We first analyze the improvement of the multi-scale procedure over the baseline via Equation 3, where we subtract the optimal single-scale from the optimal multi-scale RATE Ratio to quantify the relative increase in heterogeneity signal. The Multi-scale Gain metric is reported in Table 1. With benchmark CLIP-RSICD-based analysis, strong evidence of multi-scale heterogeneity is found in Uganda ($G = 0.95$ (s.e.=0.10) and 0.41 (s.e.=0.09) for PC and raw representations, respectively); some evidence is found in Peru ($G = 0.68$ (s.e.=0.09) and 0.00 (s.e.=0.08) for PC and raw representations, respectively).

Table 1: *RATE ratio differences from Equation 3. Standard errors in parentheses. "clip-rsicd" an EO fine-tune of CLIP. PC denotes principal component representations.* s^*, s_1^*, s_2^* *denote optimal image dimensions in the single- and multi-scale cases using raw (uncompressed) representations.*

PC: Multi-scale Gain	Multi-scale Gain	PC: $\{s^*\}/\{s_1^*, s_2^*\}$	$\{s^*\}/\{s_1^*, s_2^*\}$	Case	Model
0.68 (0.09)	0.00 (0.08)	{32}/{32, 64}	{64}/{64, 64}	Peru	clip-rsicd
0.95 (0.10)	0.41 (0.09)	{349}/{32, 128}	{16}/{16, 349}	Uganda	clip-rsicd

Table 1 also provides insight into the optimal image dimensions selected in each data-model-analysis combination. Strikingly, although the largest image context (349 pixels) is the best for single-scale Uganda analysis, when compared with the Multi-scale Concatenated procedures, in no case is the largest image context best—i.e., using the largest image array does not maximize the heterogeneity signal in either of the RCTs. Moreover, smaller images generally generate higher heterogeneity signals than larger images. This hints at some of the challenges in multi-scale analysis, where features from larger images are packed into a fixed-dimension embedding space, leading to the possible degradation of information related to important household- or village-level dynamics.

5.2.2. Analysis of RCT Results

With further analysis, we find more evidence for the use of multi-scale information by the heterogeneity model. In Figure 6, we see that the optimal RATE Ratio occurs when we concatenate representations from a large image with those from a more localized context. This difference is larger in Uganda than in Peru in terms of raw pixel numbers, but may be partially explained by the higher-resolution of the Uganda images.

Because Multi-Scale Representation Concatenation increases the dimensionality of the input covariates to the CATE estimation model, dimensionality reduction techniques could improve model performance. Figure 7 displays the RATE Ratio heat maps from PC compressions of the neural features (50 dimensions). Using PC representations, we see even stronger evidence for the role of multi-scale representations in improving heterogeneity signal and stronger model performance as evaluated by the RATE Ratios, and the optimality point is shifted towards the center of the heat map. We did not do an exhaustive search for the optimal PC procedure but anticipate that the performance of multi-scale would improve further with further optimization of the dimensionality reduction technique employed.

Evidence of Multi-scale Use by Heterogeneity Model. Due to the interpretability of the concatenation-based multi-scale procedure, we can directly investigate whether there

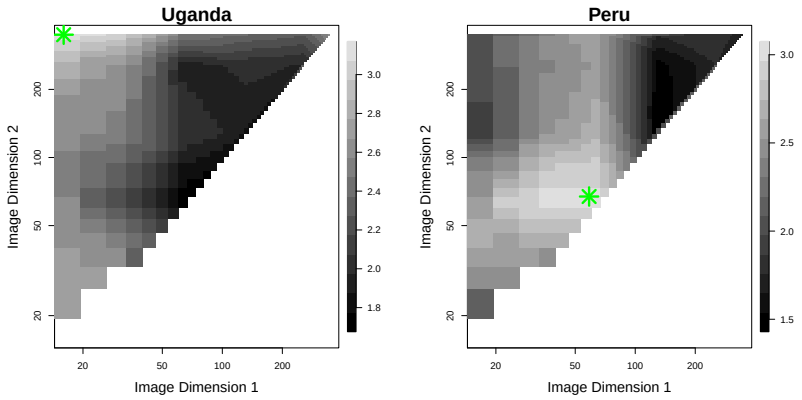

Figure 6: *Analysis of RATE Ratios for Uganda and Peru RCTs across a range of s_1, s_2 values (higher/lighter values are better). We see that the maximum heterogeneity signal in both RCTs is detected with small/medium-sized s_1 ($\sim$64) and large s_2 ($\sim$350). $*$ indicates optimal selection. See Figure 12 for analysis using RATE values (instead of ratios).*

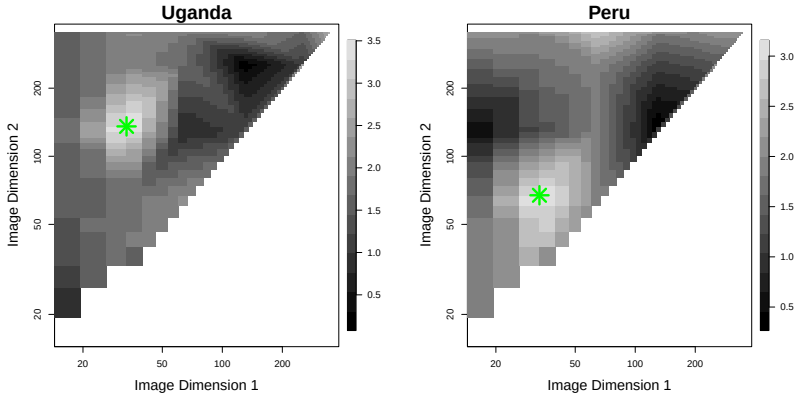

Figure 7: *Analysis of RATE Ratios across a range of s_1, s_2 values with PC representations. Higher/lighter values are better. $*$ indicates optimal selection.*

is evidence that the heterogeneity models are incorporating information from multiple scales in forming effect estimates. We probe the utilization of multi-scale information in Figure 11. Here, we find that the effect heterogeneity model, in general, uses information from both scales in modeling treatment effects.

We measure the CF model's information utilization by examining the fraction of the top 10 most informative features (via CF) derived from representations at the smaller scale (i.e., ϕ_{i,s_1}). Values near 0.5 indicate balanced use of smaller and larger scales; values near 1 indicate dominance of unit-scale features; those near 0 indicate dominance of larger-scale ones. According to this measure, gain from multi-scale information is stronger in Uganda imagery than in Peru imagery, a finding congruent with the analysis of RATE Ratios. A confounding factor here could be the lower resolution of the Peru Imagery (which decreases the amount of information detectable with smaller scales).

Multi-scale Dynamics With Weak Prior Information About Unit Locations. Although we observe an improvement in RATE Ratio in using Multi-Scale Representation Concatenation, the procedure so far described requires strong prior information on the geographic location of households. We therefore investigate performance when we randomly sample image centers so that they are no longer centered around villages/households. To investigate this, we displace household and village locations so that small images no longer contain information from the target units' location and re-run the same analysis as above. We find that there is a decrease in multi-scale dynamics for raw representations (for Uganda, a weak-prior-information G of 0.19 (s.e.=0.08), for Peru, a difference of 0.00 (s.e.=0.06)), but not for PC representations (for Uganda, a G of 1.35 (s.e.=0.10); for Peru, 0.68 (s.e.=0.07) and for PC and raw representations, respectively).

More broadly, Figure 8 shows that, even with only weak information about unit locations, multi-scale analysis can improve signals about heterogeneity; scale information is not specific to household-level features but also other features about localized contexts. We also find that heterogeneity signals are actually higher when using displaced images than the actual images around units (mean RATE Ratio of 2.24 vs. 2.13), averaging across all method and data choices. In other words, even without household-specific knowledge, heterogeneity signals can still propagate through the integration of multi-scale features into effect heterogeneity modeling. We investigated this in further simulations and found that multi-scale analysis improves signals about heterogeneity, although less than the case with strong prior information (see Table 3 in the Appendix).

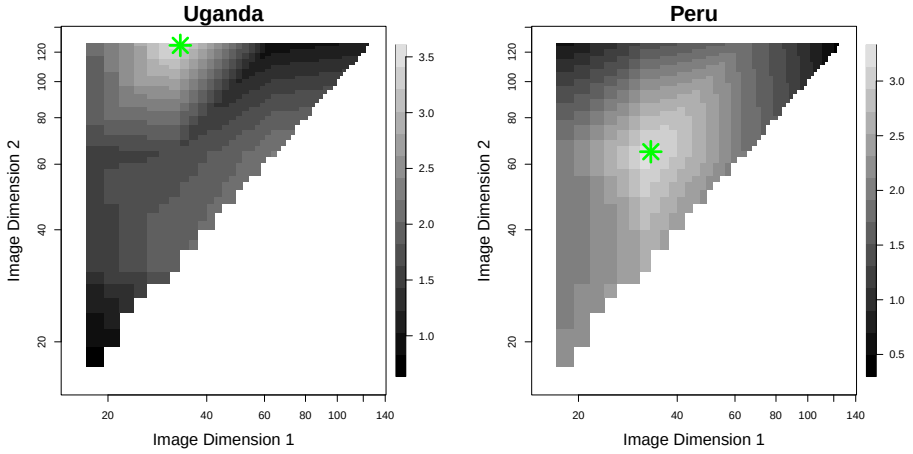

Figure 8: *Analysis of RATE Ratios for Uganda and Peru with displaced image centers across a range of s_1, s_2 values (higher/lighter values are better). $*$ indicates optimal selection.*

5.3. Scaling Scales

We next consider whether any scaling regularities emerge as the number of concatenated scales grows. Figure 9 investigates dynamics related to the number of scales employed. The average RATE Ratio increases with the number of concatenated scales, indicating that incorporating information from multiple scales enhances the model's ability to detect treatment effect heterogeneity. The RATE Ratio rises from approximately 2 for a single scale

to a peak of around 2.5 when five scales are concatenated. Beyond this point, performance stabilizes, as evidenced by the RATE Ratio remaining at 2.5 for six scales. This suggests that concatenating up to five scales captures increasingly diverse information related to household-specific features and broader contextual dynamics to improve the heterogeneity signal, though additional scales beyond this threshold yield diminishing returns due to potential redundancy. This pattern is consistent across Peru and Uganda RCTs.

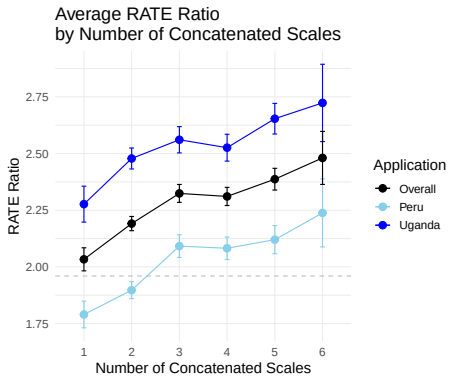

Figure 9: *As the number of scales grows, so too does average RATE ratio performance. The dotted grey line marks 1.96, which represents the 5% significance threshold for detecting heterogeneity (assuming asymptotic normality).*

Overall, we see that Multi-Scale Representation Concatenation improves the performance of deep learning models in EO-based CATE estimation, whether information is strong or weak about the exact location of observational units. Moreover, Multi-Scale Representation Concatenation is interpretable in that investigators can readily probe the relative performance of different multi-scale choices on heterogeneity signal—noting that, in comparison, many of the multi-scale inference strategies described in the computer vision literature up to this point are, to a point, black box, and do not offer a ready interpretation of which scales matter, and which do not. Multi-Scale Representation Concatenation thus introduces a composable procedure that allows any single-scale procedure to integrate heterogeneity information from small and large images to better capture both individual heterogeneity and neighborhood-level effects, allowing for more informative causal inferences.

6. Discussion

Larger Images Are Not Always Better. A counterintuitive finding from our analyses is that using the largest possible satellite image context (349 pixels) never maximizes the heterogeneity signal in any of our experimental conditions. This challenges the intuitive assumption that incorporating more contextual information necessarily improves causal inference. We find this "bigger is not better" principle manifests in three ways.

First, in single-scale analyses with unperturbed geolocations, smaller images often generate elevated heterogeneity signals, likely because they better capture household-specific characteristics without dilution from broader contexts. Second, even in multi-scale analyses, the optimal combinations typically pair a small or medium-sized image (around 64 pixels) with a large but not maximum-sized image, rather than using the largest available scale.

Third, multi-scale analysis can provide for robust inferences even when only approximate location information is available (as evidenced by the displaced locations analysis).

These findings have implications for EO-based causal inference. Rather than defaulting to the largest possible image context, researchers should carefully consider the trade-off between capturing relevant local heterogeneity and broader contextual information, either employing Multi-Scale Representation Concatenation or incorporating multi-scale dynamics into the inferential procedure in other ways to be robust to the intricacies of EO data.

Towards Adaptive Multi-Scale Representation Concatenation. Despite grid search's promising performance, its computational requirement is exponential in C, the number of scales considered for combined analysis. Further, the current Multi-Scale Representation Concatenation technique limits itself to a fixed image-scale combination for each observational unit, whilst in practice the optimal scale for each observational unit may differ.

Our finding that Multi-Scale Representation Concatenation outperforms single-scale analysis suggests that there may be a way to optimize Multi-Scale Representation Concatenation further by estimating an additional function $l : \mathcal{M} \to \mathbb{R}^2$ that takes in a large image over an entire region around an experimental unit and outputs locations $\mathbf{x_i}$ where an image at each scale is centered. This would involve evaluating each geographic location for suitability of analysis, where dense prediction techniques could prove helpful (Zuo et al., 2022). We present the estimation of l as an open problem.

Multi-scale dynamics beyond EO. Multi-scale dynamics are present in diverse geospatial, biological, and public health contexts. Beyond CATE estimation, these dynamics are relevant for environmental cost-benefit analysis (Druckenmiller et al., 2024), EO-based poverty imputation (Burke et al., 2021a), and confounder control—where the latent geospatial confounding structure may operate with multi-scale dynamics (Jerzak et al., 2022). We invite further work to incorporate multi-scale dynamics in these diverse settings.

7. Limitations & Conclusion

This paper addressed the methodological challenge of capturing multi-scale dynamics in EO-based causal inference. By combining representations across scales, our approach 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 finding a trade-off between individual heterogeneity and neighborhood-level context, and contributes to the growing literature on deep learning for causal inference (Daoud and Dubhashi, 2023).

Limitations remain. First, the approach here assumes SUTVA and unconfoundedness for identification, and thus, an extension to observational inference requires additional assumptions. Second, images used in our study have low resolution, and resolution likely interacts with the gain from doing multi-scale analysis. Third, using high-resolution images in a multi-scale approach raises privacy concerns. Fourth, the approach taken to assess the validity of Multi-Scale Representation Concatenation is empirical; theoretical results would strengthen our understanding of the method. Fifth, the approach is currently only validated on anti-poverty RCT datasets. Further research could be done to address these limitations of the general Multi-Scale Representation Concatenation approach. □

Acknowledgments

We thank members of the AI & Global Development Lab for helpful feedback, resources, and inspiration. We thank Dean Karlan and Andre Nickow for assistance in accessing data for the Peru experiment. We thank Hannah Druckenmiller, Antonio Linero, and SayedMorteza Malaekeh for helpful discussions. All remaining limitations are our own.

References

Katherine Anderson, Barbara Ryan, William Sonntag, Argyro Kavvada, and Lawrence Friedl. Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science*, 20(2):77–96, 2017.

Artashes Arutiunian, Dev Vidhani, and Goutham Venkatesh Mayank Bhaskar Ritobrata Ghosh Sujit Pal. Clip-rsicd v2. https://huggingface.co/flax-community/clip-rsicd-v2, 2021.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application, 2019. URL https://arxiv.org/abs/1902.07409.

Susan Athey et al. The impact of machine learning on economics. *The economics of artificial intelligence: An agenda*, pages 507–547, 2018.

Abhijit Banerjee, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236):1260799, 2015. doi: 10.1126/science.1260799. URL https://www.science.org/doi/abs/10.1126/science.1260799.

Christopher Blattman, Nathan Fiala, and Sebastian Martinez. The long-term impacts of grants on poverty: Nine-year evidence from uganda's youth opportunities program. *American Economic Review: Insights*, 2(3):287–304, 2020.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021a.

Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. Using Satellite Imagery to Understand and Promote Sustainable Development. *Science*, 371(6535), March 2021b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abe8628. URL https://www.sciencemag.org/lookup/doi/10.1126/science.abe8628.

Nick Chater and Gordon D.A. Brown. Scale-invariance as a unifying psychological principle. *Cognition*, 69(3):B17–B24, 1999. doi: 10.1016/s0010-0277(98)00066-3.

Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation, 2016. URL https://arxiv.org/abs/1511.03339.

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ml and ai, 2024. URL https://arxiv.org/abs/2403.02467.

Vanessa D'Amario, Sanjana Srivastava, Tomotake Sasaki, and Xavier Boix. The data efficiency of deep learning is degraded by unnecessary input dimensions. *Frontiers in Computational Neuroscience*, 16, 2022. ISSN 1662-5188. doi: 10.3389/fncom.2022.760085. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2022.760085.

Adel Daoud and Devdatt Dubhashi. Statistical Modeling: The Three Cultures. *Harvard Data Science Review*, 5(1), January 2023. ISSN 2644-2353, 2688-8513. doi: 10.1162/99608f92.89f6fe66.

Adel Daoud and Fredrik D. Johansson. The impact of austerity on children: Uncovering effect heterogeneity by political, economic, and family factors in low- and middle-income countries. *Social Science Research*, 118:102973, February 2024. ISSN 0049-089X. doi: 10.1016/j.ssresearch.2023.102973.

Adel Daoud, Felipe Jordán, Makkunda Sharma, Fredrik Johansson, Devdatt Dubhashi, Sourabh Paul, and Subhashis Banerjee. Using Satellite Images and Deep Learning to Measure Health and Living Standards in India. *Social Indicators Research*, 167(1): 475–505, June 2023. ISSN 1573-0921. doi: 10.1007/s11205-023-03112-x. URL https://doi.org/10.1007/s11205-023-03112-x.

Hannah Druckenmiller, Georgia Gkioxari, Connor T. Jerzak, and SayedMorteza Malaekeh. A multimodal causal framework for large-scale ecosystem valuation: Application to wetland benefits for flood mitigation. Presented at the NeurIPS Workshop - Tackling Climate Change with Machine Learning, 2024. Workshop Paper.

Georgios Giannarakis, Ilias Tsoumas, Stelios Neophytides, Christiana Papoutsa, Charalampos Kontoes, and Diofantos Hadjimitsis. Understanding the Impacts of Crop Diversification in the Context of Climate Change: A Machine Learning Approach, July 2023. URL http://arxiv.org/abs/2307.08617. arXiv:2307.08617 [cs, q-bio] version: 1.

Ross Girshick. Fast r-cnn, 2015. URL https://arxiv.org/abs/1504.08083.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014. URL https://arxiv.org/abs/1311.2524.

Eugenia Go, Kentaro Nakajima, Yasuyuki Sawada, and Kiyoshi Taniguchi. On the Use of Satellite-Based Vehicle Flows Data to Assess Local Economic Activity: The Case of Philippine Cities, March 2022. URL https://papers.ssrn.com/abstract=4057690.

Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features, 2014. URL https://arxiv.org/abs/1403.1840.

Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1411, January 2020. ISSN 2045-2322. doi: 10.1038/s41598-019-57261-6. URL https://doi.org/10.1038/s41598-019-57261-6.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Geoffrey Hinton. How to represent part-whole hierarchies in a neural network, 2021. URL https://arxiv.org/abs/2102.12627.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2289064.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Taylor & Francis, 2016.

Ylva Jansson and Tony Lindeberg. Scale-Invariant Scale-Channel Networks: Deep Networks That Generalise to Previously Unseen Scales. *Journal of Mathematical Imaging and Vision*, 64(5):506–536, June 2022. ISSN 1573-7683. doi: 10.1007/s10851-022-01082-2. URL https://doi.org/10.1007/s10851-022-01082-2.

Connor T Jerzak, Fredrik Johansson, and Adel Daoud. Estimating causal effects under image confounding bias with an application to poverty in africa. *arXiv preprint arXiv:2206.06410*, 2022.

Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity. *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR), Proceedings of Machine Learning Research (PMLR)*, 213:531–552, 2023.

Licheng Jiao, Jie Gao, Xu Liu, Fang Liu, Shuyuan Yang, and Biao Hou. Multiscale representation learning for image classification: A survey. *IEEE Transactions on Artificial Intelligence*, 4(1):23–43, 2023. doi: 10.1109/TAI.2021.3135248.

Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15: 100836, September 2021. ISSN 2352-8273. doi: 10.1016/j.ssmph.2021.100836. URL https://www.sciencedirect.com/science/article/pii/S2352827321001117.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. doi: 10.1109/TGRS.2017.2776321.

Markus B. Pettersson, Mohammad Kakooei, Julia Ortheden, Fredrik D. Johansson, and Adel Daoud. Time Series of Satellite Imagery Improve Deep Learning Estimates of Neighborhood-level Poverty in Africa. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6165–6173, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023. URL https://arxiv.org/abs/2212.14532.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. URL https://arxiv.org/abs/1506.01497.

Kazuki Sakamoto, Connor T. Jerzak, and Adel Daoud. A scoping review of earth observation and machine learning for causal inference: Implications for the geography of poverty, 2024. URL https://arxiv.org/abs/2406.02584.

Batkhurel Serdavaa. A Satellite Image Analysis on Housing Conditions and the Effectiveness of the Affordable Housing Mortgage Program in Mongolia: A Deep Learning Approach, December 2023. URL https://papers.ssrn.com/abstract=4664966.

Koichiro Shiba, Adel Daoud, Hiroyuki Hikichi, Aki Yazawa, Jun Aida, Katsunori Kondo, and Ichiro Kawachi. Uncovering Heterogeneous Associations Between Disaster-Related Trauma and Subsequent Functional Limitations: A Machine-Learning Approach. *American Journal of Epidemiology*, page kwac187, October 2022. ISSN 0002-9262. doi: 10.1093/aje/kwac187.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Erik Sverdrup, Han Wu, Susan Athey, and Stefan Wager. Qini curves for multi-armed treatment rules, 2024. URL https://arxiv.org/abs/2306.11979.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. URL https://arxiv.org/abs/1409.4842.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario

Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021. URL https://arxiv.org/abs/2105.01601.

Wei Xiong, Zhenyu Xiong, and Yaqi Cui. A Confounder-Free Fusion Network for Aerial Image Scene Feature Representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5440–5454, 2022. ISSN 2151-1535. doi: 10.1109/JSTARS.2022.3189052. URL https://ieeexplore.ieee.org/document/9817622. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey, 2023. URL https://arxiv.org/abs/2204.08610.

Shuangquan Zuo, Yun Xiao, Xiaojun Chang, and Xuanhong Wang. Vision transformers for dense prediction: A survey. *Knowledge-Based Systems*, 253:109552, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109552. URL https://www.sciencedirect.com/science/article/pii/S0950705122007821.

8. Appendix

8.1. Data Availability Statement

Simulation code is 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.

8.2. Satellite Imagery Processing Details

Direct image cropping is used to obtain satellite imagery at different scales. The effect of this cropping is shown in Figure 10.

8.3. 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). The value of sd(RATE) is estimated through half-sample bootstrap. In Equation 2, $\hat{\tau}(\mathbf{M}_i)$ 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

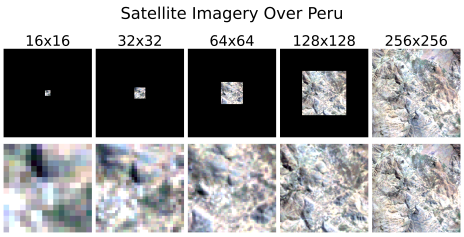

Figure 10: *Peru images size 16×16 pixels to 256×256 pixels from* 2007*. Image resolution is held at* 30*m the highest resolution for the Landsat* 5 *and* 7 *satellites. Visualized imagery are not centered around RCT participants for privacy considerations. The top row shows images without resizing across changing dimensions; bottom panel shows resized images. Images of different scales are obtained in identical fashion in Uganda.*

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 ($\frac{\text{RATE mean}}{\text{RATE SD}}$) 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. Indeed, in our study, where the focus is on anti-poverty interventions like cash transfers, identifying households with the largest treatment effects is often a priority for efficient resource allocation, hence the choice of AUTOC weighting over the equi-weighting approach taken by QINI (where all respondents contribute uniformly to the RATE calculation).

8.4. Related Work

There is a tradition of research into multi-scale phenomena in Computer Vision that has informed our approach to developing methodologies for EO-based Causal Inference. We survey them below.

Multi-scale Strategies in Deep Vision Networks. The success of deep vision networks has been attributed to their ability to learn hierarchical levels of abstractions present in variously-scaled images (Szegedy et al., 2014; LeCun et al., 2015). Deep residual networks enabled by skip connections are multi-scale feature extractors (Jiao et al., 2023), and interpretable architectures that take advantage of part-whole hierarchies have been proposed (Hinton, 2021). The SWIN Transformer and MLP-mixer are recent architectures that place the hierarchical multi-scale nature of images at the center of their architecture, achieving efficiency and interoperability, respectively (Liu et al., 2021; Tolstikhin et al., 2021).

Explicit Multi-scale Representation Learning in Computer Vision. A drawback to the above multi-scale approach is that they do not explicitly preserve representations at multiple scales, whilst in EO-contexts such information may be needed (e.g. exact details of the rooftop of a household's home that take up a small number of pixels but encode important information about their socio-economic status). Thus, more relevant to our application scenario are strategies that explicitly preserve lower-scale information. Region Proposal Networks achieve efficient object detection by processing an image globally

before identifying local sub-regions (Girshick et al., 2014; Girshick, 2015; Ren et al., 2016). Multi-scale residual networks combine residual connections with convolution kernels of varying sizes to adaptively detect features at different scales (Li et al., 2018). In EO-based Deep Learning, the Scale-MAE architecture tackles the scale dependence of geospatial representations by combining Laplacian Pyramids and Ground Sample Distance Encoding with a Transformers-based Masked Auto-encoder framework (Reed et al., 2023). Our work adds to this existing literature by introducing a composable and interpretable procedure that adapts arbitrary EO-based causal effect estimation algorithms to incorporate multi-scale dynamics in effect heterogeneity estimation, enabling practitioners to improve their pre-existing effect heterogeneity estimation pipeline with limited additional data requirement and model design effort.

Scale Invariant Vision Network. Scale-invariance has been recognized as an important characteristic of the natural world and a desirable property for vision networks (Chater and Brown, 1999; Han et al., 2020). Image augmentation through cropping and resizing are commonly used to encourage scale-invariance in models (He et al., 2016; Yang et al., 2023), and architectures with inductive biases for scale-invariance have been proposed (Chen et al., 2016; Gong et al., 2014; Jansson and Lindeberg, 2022). In the EO setting, where input scale and resolution differ markedly depending on modeling choices, designing architectures adaptable to different scales grows in importance (Reed et al., 2023). Our work employs Multi-Scale Representation Concatenation to extract features amenable to detection at different scales, providing a general procedure to enable extraction of arbitrarily scaled features in satellite imagery.

8.5. Qini Curve Details

Our exposition deals with Qini Curves with a single binary treatment and constant costs for each unit. There are generalizations of Qini Curves with varying costs and multiple treatment arms (Sverdrup et al., 2024), but they are not applicable to our RCT data.

Qini Curves employ the identical sample splitting strategy to obtain the estimated CATEs as with RATE. With the obtained CATEs, Qini Curves construct a policy function $\pi_B(M_i)$, which assigns treatment to a unit M_i if $\hat{\tau}(\mathbf{M}_i)$ greater than 0 and in the top $B\%$ of the CATE predictions. It then plots a curve with B as the X-axis and the expected gain in the outcome of interest when one targets the most responsive units according to $\pi_B(M_i)$ as the Y-axis. It is often plotted with a baseline ATE curve which shows the expected gain when the treatment is random assignment.

Both Qini Curves and RATE measure the degree of treatment effect heterogeneity that a model is able to capture from the input covariates. Qini curves are useful under the setting of cost-benefit analysis as its y-axis directly measures the expected gain.

We implement Qini Curves through the `maq` package. Doubly robust scores obtained through Causal Forests using the `grf` package are used as CATE estimators, the cost parameter is set as the constant 1, and 200 bootstrap replications are used for standard error calculation,

8.6. 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 for developing 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 Multi-scale Concatenated 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 value using,

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

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

Here is the full results for the simulations when taking all possible combinations of image perturbations. We additionally tested Rotation as a global context feature.

Table 2: *Simulation R^2 for All Combinations of Perturbation Parameters, Values are presented as mean (lower CI, upper CI). R is rotation, M is mask, C is contrast, E is edge.*

R	M	C	E	256+32	256	32
0	0	0	1	0.152 (0.139, 0.165)	0.198 (0.183, 0.212)	-0.007 (-0.011, -0.003)
0	0	1	0	0.295 (0.270, 0.320)	0.169 (0.153, 0.184)	0.198 (0.193, 0.202)
0	0	1	1	0.134 (0.124, 0.144)	0.097 (0.094, 0.099)	0.064 (0.058, 0.069)
0	1	0	0	0.146 (0.133, 0.160)	0.001 (-0.001, 0.003)	0.201 (0.191, 0.210)
0	1	0	1	0.136 (0.122, 0.150)	0.086 (0.078, 0.094)	0.078 (0.066, 0.091)
0	1	1	0	0.103 (0.068, 0.137)	0.048 (0.041, 0.056)	0.088 (0.082, 0.094)
0	1	1	1	0.082 (0.074, 0.090)	0.035 (0.027, 0.044)	0.038 (0.031, 0.045)
1	0	0	0	0.014 (0.009, 0.019)	0.010 (0.008, 0.013)	0.001 (-0.000, 0.003)
1	0	0	1	0.646 (0.636, 0.656)	0.565 (0.555, 0.576)	0.115 (0.098, 0.131)
1	0	1	0	0.558 (0.550, 0.566)	0.511 (0.498, 0.524)	0.440 (0.424, 0.455)
1	0	1	1	0.142 (0.125, 0.159)	0.131 (0.118, 0.144)	0.071 (0.062, 0.081)
1	1	0	0	0.593 (0.588, 0.598)	0.264 (0.248, 0.280)	0.430 (0.422, 0.437)
1	1	0	1	0.110 (0.102, 0.119)	0.074 (0.065, 0.083)	0.070 (0.064, 0.076)
1	1	1	0	0.137 (0.127, 0.148)	0.064 (0.059, 0.068)	0.097 (0.081, 0.113)
1	1	1	1	0.072 (0.065, 0.078)	0.049 (0.043, 0.055)	0.047 (0.041, 0.053)

The results for the case where there is weak prior information about unit locations is also presented. Under weak prior information, we combine a 256x256 image representation with a 32x32 image drawn uniformly from the image.

Table 3: *Simulation R^2 for All Combinations of Perturbation Parameters, Values are presented as mean (lower CI, upper CI). R is rotation, M is mask, C is contrast, E is edge. Weak Prior column combines a 256x256 image representation with a 32x32 image drawn uniformly from the image*

R	*M*	*C*	*E*	Weak Prior	256	32
0	0	0	1	0.243 (0.197, 0.289)	0.195 (0.193, 0.197)	-0.006 (-0.008, -0.004)
0	0	1	0	0.218 (0.188, 0.249)	0.193 (0.159, 0.228)	0.190 (0.178, 0.203)
0	0	1	1	0.119 (0.101, 0.136)	0.102 (0.094, 0.111)	0.065 (0.064, 0.066)
0	1	0	0	0.002 (0.001, 0.002)	0.001 (0.001, 0.001)	0.222 (0.175, 0.268)
0	1	0	1	0.084 (0.072, 0.096)	0.073 (0.071, 0.074)	0.086 (0.071, 0.100)
0	1	1	0	0.066 (0.056, 0.076)	0.058 (0.052, 0.064)	0.093 (0.078, 0.108)
0	1	1	1	0.053 (0.042, 0.065)	0.041 (0.027, 0.056)	0.024 (-0.006, 0.054)
1	0	0	0	0.013 (0.010, 0.016)	0.008 (0.005, 0.011)	0.000 (-0.001, 0.002)
1	0	0	1	0.578 (0.559, 0.596)	0.550 (0.525, 0.575)	0.100 (0.097, 0.103)
1	0	1	0	0.544 (0.522, 0.566)	0.522 (0.501, 0.543)	0.428 (0.415, 0.441)
1	0	1	1	0.126 (0.099, 0.154)	0.122 (0.106, 0.138)	0.073 (0.059, 0.087)
1	1	0	0	0.249 (0.225, 0.273)	0.276 (0.268, 0.283)	0.425 (0.417, 0.432)
1	1	0	1	0.079 (0.063, 0.094)	0.070 (0.064, 0.076)	0.071 (0.064, 0.078)
1	1	1	0	0.065 (0.054, 0.076)	0.062 (0.028, 0.096)	0.107 (0.074, 0.141)
1	1	1	1	0.054 (0.046, 0.061)	0.051 (0.046, 0.056)	0.046 (0.036, 0.055)

The simulation replication code is in the Sim folder of the code repository.

8.7. Algorithms

Input:

- $i \in \{1, \cdots, n\}$ denotes the index for observational units.
- $\{\mathbf{x}_i\}_{i=1}^n$ the set of locations of those units.
- Sets of image sizes $\mathcal{S} = \{s_1, s_2, \ldots, s_{\text{Max}}\}$.
- Image fetcher, $f_I(\mathbf{x}_i, s)$, that obtains an image centered at a given location $\mathbf{x}_i$ with size s.
- Image encoders, $f'_{\phi_{s_1}}$ and $f'_{\phi_{s_2}}$, that extracts representations from $\mathbf{M}_{i,s_1}$, $\mathbf{M}_{i,s_2}$.
- Dimensionality reduction function r that reduces the dimension of the concatenated representations.
- Trainable CATE estimation function $h_\theta(\cdot)$ parametrized by θ.
- Observed outcome of interest $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$.
- Binary treatment indicator $\mathbf{W} = (W_1, W_2, \ldots, W_n), W_i \in \{0, 1\}$.

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 $s_1 \in \mathcal{S}$ **do**
  foreach $s_2 \in \mathcal{S}$ **do**
    Set `MaxRATERatio` to $-\infty$
    foreach $i, \mathbf{x}_i \in$ *enumerate*$(\{\mathbf{x}_i\}_{i=1}^n)$ **do**
      $\mathbf{M}_{i,s_1} = f_I(\mathbf{x}_i, s_1)$;
      $\mathbf{M}_{i,s_2} = f_I(\mathbf{x}_i, s_2)$
      $\boldsymbol{\phi}_{i,s_1,s_2} = r((f_{\phi'_{s_1}}(\mathbf{M}_{i,s_1}), f_{\phi'_{s_2}}(\mathbf{M}_{i,s_2})))$
    end
    Compute $\widehat{\text{RATE Ratio}} = R(\mathbf{W}, \mathbf{Y}, \{\boldsymbol{\phi}_{i,s_1,s_2}\}_{i=1}^n)$
    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*

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

8.8. Causal Forest Implementation Detail

We use the `grf` package in R to implement our Causal Forest. The minimum node size is 5. The nuisance models are random forests. The number of trees grown is 2000.

8.9. Additional Empirical Results

Table 4: *RATE ratio differences from Equation 3. Standard errors in parentheses. "clay" denotes the Clay EO foundation model; "clip-rsicd" an EO fine-tune of CLIP; "swin" the SWIN Transformer. PC denotes principal component representations.* s^*, s_1^*, s_2^* *denote optimal image dimensions in the single- and multi-scale cases using raw (uncompressed) representations.*

PC: Multi-scale Gain	Multi-scale Gain	PC: $\{s^*\}/\{s_1^*, s_2^*\}$	$\{s^*\}/\{s_1^*, s_2^*\}$	Case	Model	Displaced?
0.00 (0.07)	0.00 (0.07)	{64}/{64, 64}	{64}/{64, 64}	Peru	clay	Yes
0.37 (0.09)	0.00 (0.09)	{16}/{32, 64}	{16}/{16, 16}	Uganda	clay	Yes
0.00 (0.09)	0.11 (0.08)	{349}/{349, 349}	{64}/{16, 64}	Peru	clay	No
0.58 (0.10)	0.01 (0.09)	{16}/{32, 349}	{16}/{16, 32}	Uganda	clay	No
0.68 (0.07)	0.00 (0.06)	{32}/{32, 64}	{64}/{64, 64}	Peru	clip-rsicd	Yes
1.35 (0.10)	0.19 (0.08)	{64}/{32, 128}	{16}/{16, 32}	Uganda	clip-rsicd	Yes
0.68 (0.09)	0.00 (0.08)	{32}/{32, 64}	{64}/{64, 64}	Peru	clip-rsicd	No
0.95 (0.10)	0.41 (0.09)	{349}/{32, 128}	{16}/{16, 349}	Uganda	clip-rsicd	No
0.06 (0.07)	0.13 (0.06)	{32}/{32, 64}	{16}/{16, 128}	Peru	swin	Yes
0.09 (0.11)	0.00 (0.10)	{128}/{32, 128}	{128}/{128, 128}	Uganda	swin	Yes
0.83 (0.09)	0.29 (0.09)	{32}/{32, 64}	{16}/{16, 32}	Peru	swin	No
0.24 (0.11)	0.00 (0.11)	{64}/{16, 32}	{128}/{128, 128}	Uganda	swin	No

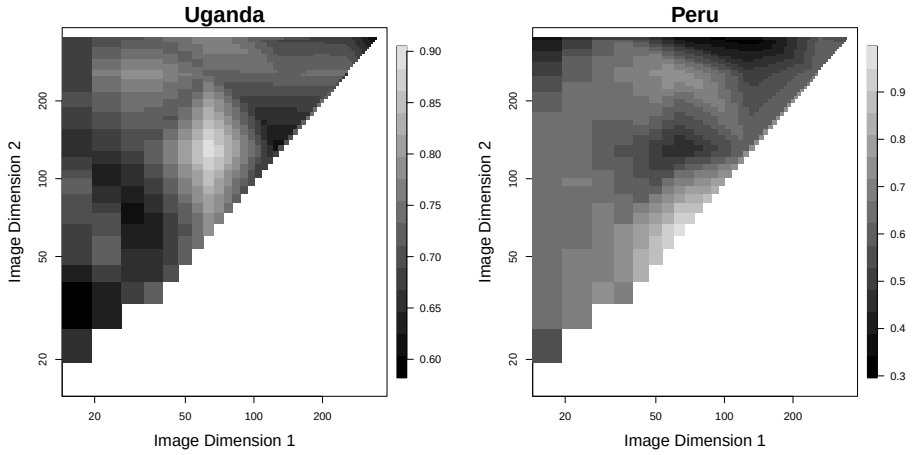

Figure 11: *Analysis of features predictive of treatment effects. Colors indicate the fraction of the top 10 features (via Causal Forest) predictive of treatment effects from the smaller image scale* (s_1)*. We see that, in general, information from both scales is used across the range of* s_1, s_2 *values.*

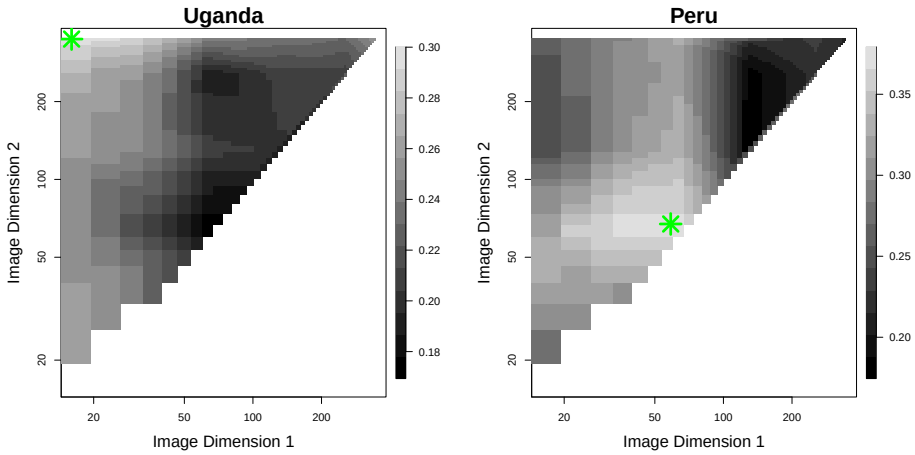

Figure 12: *Analysis of RATE values for Uganda and Peru RCTs across a range of s_1, s_2 values (higher/lighter values are better).*

Table 5: *Displaced RATE ratio differences from Equation 3. Standard errors in parentheses. "clip-rsicd" an EO fine-tune of CLIP. PC denotes principal component representations. s^*, s_1^*, s_2^* denote optimal image dimensions in the single- and multi-scale cases using raw (uncompressed) representations.*

PC: Multi-scale Gain	Multi-scale Gain	PC: $\{s^*\}/\{s_1^*, s_2^*\}$	$\{s^*\}/\{s_1^*, s_2^*\}$	Case	Model
0.68 (0.07)	0.00 (0.06)	{32}/{32, 64}	{64}/{64, 64}	Peru	clip-rsicd
1.35 (0.10)	0.19 (0.08)	{64}/{32, 128}	{16}/{16, 32}	Uganda	clip-rsicd

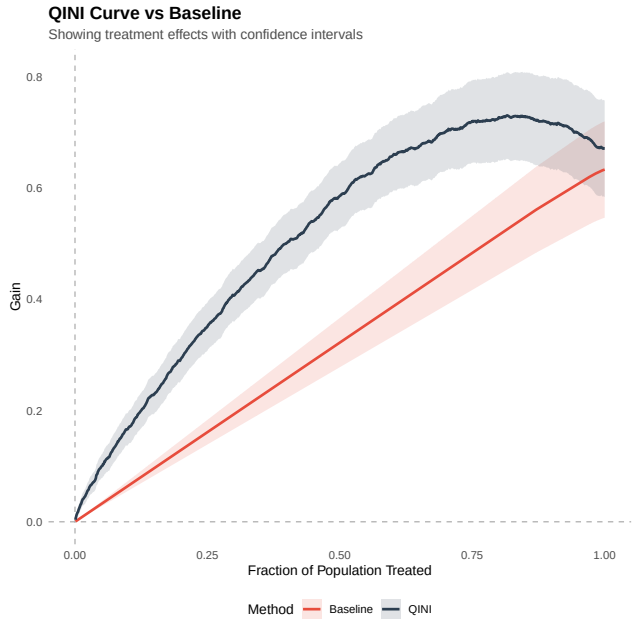

Figure 13: *Uganda Qini results with CLIP-RSICD representations ($s_1^* = 16$, $s_2^* = 349$).*

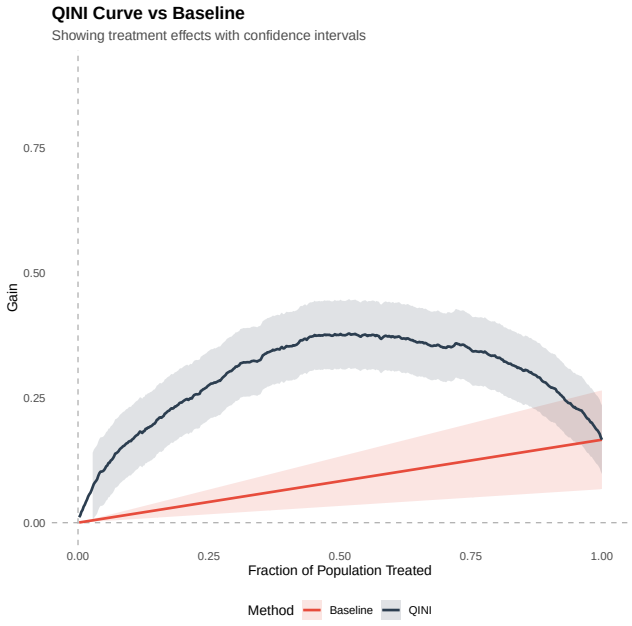

Figure 14: *Peru Qini results with CLIP-RSICD representations ($s_1^* = 64$ and $s_2^* = 64$).*