Mitigating Urban-Rural Disparities in Contrastive Representation Learning with Satellite Imagery

Miao Zhang[1] **Rumi Chunara**[2]

[1]Tandon School of Engineering, New York University
[2] Tandon School of Engineering; School of Global Public Health, New York University
New York, USA
{miaozhng, rumi.chunara}@nyu.edu

Abstract

Satellite imagery is being leveraged for many societally critical tasks across climate, economics, and public health. Yet, because of heterogeneity in landscapes (e.g. how a road looks in different places), models can show disparate performance across geographic areas. Given the important potential of disparities in algorithmic systems used in societal contexts, here we consider the risk of urban-rural disparities in identification of land-cover features. This is via semantic segmentation (a common computer vision task in which image regions are labelled according to what is being shown) which uses pre-trained image representations generated via contrastive self-supervised learning. We propose fair dense representation with contrastive learning (FairDCL) as a method for de-biasing the multi-level latent space of convolution neural network models. The method improves feature identification by removing spurious model representations which are disparately distributed across urban and rural areas, and is achieved in an unsupervised way by contrastive pre-training. The obtained image representation mitigates downstream urban-rural prediction disparities and outperforms state-of-the-art baselines on real-world satellite images. Embedding space evaluation and ablation studies further demonstrate FairDCL's robustness. As generalizability and robustness in geographic imagery is a nascent topic, our work motivates researchers to consider metrics beyond average accuracy in such applications.

Introduction

Dense pixel-level image recognition via deep learning for tasks such as segmentation have a variety of applications in landscape feature analysis from satellite images. For example, regional water quality analysis (Griffith 2002) or dust emission estimation (Von Holdt et al. 2019). Success of the methods rely on powerful visual representations that include both local and global information. However, since pixel-level annotations are usually costly, fully supervised learning is challenging when the amount and variety of labeled data is scarce. Therefore, self-supervised learning is a promising alternative via pre-training a image encoder and transferring learnt representations to downstream problems. As a mainstream, contrastive self-supervised techniques have shown state-of-the-art performance in learning image representations for land cover semantic segmentation across locations (Ayush et al. 2021; Scheibenreif et al. 2022). In particular, as labeled images are hard to obtain for satellite images, and contrastive approaches do not require labels, they have demonstrated benefits in many real-world tasks including monitoring dynamic land surface (Saha et al. 2020), irrigation detection from uncurated and unlabeled satellite images (Agastya et al. 2021), and volcanic unrest detection with scarce image label (Bountos et al. 2021).

Importantly, recent attention in machine learning systems has highlighted performance inequities including those by geographic area (de Vries et al. 2019; Xie et al. 2022; Setianto and Gamal 2021; Majumdar, Flynn, and Mitra 2022; Aiken, Rolf, and Blumenstock 2023), and how prediction inequities would compromise policy-making goals (Kondmann and Zhu 2021). Disparities at a geographic level fall into the fairness literature due to implications of unequal distribution of resources, opportunities, and essential services, leading to disparities in quality of life and opportunities for those who live in specific areas (Hay 1995). As recent work has reinforced, disparities of machine learning model prediction at a geographic level often show disparate performance with respect to minoritized groups or already under-resourced areas (Kondmann and Zhu 2021). Therefore, given the increased potential of self-supervised contrastive learning, here we turn attention to disparity risks in recognition outcomes between urban and rural areas. The consequences of such recognition tasks have wide usage for societal decisions including urban planning, climate change and disaster risk assessment (Mehrabi et al. 2021; Soden et al. 2019), so disparity shapes an important concern. Further, while recent work has identified disparities with satellite image representation, specifically across urban and rural lines, and shown the negative consequence on poverty prediction (Aiken, Rolf, and Blumenstock 2023), there is limited work on mitigating urban-rural disparities with state-of-the-art vision recognition schemes for landscape analysis, despite their wide applications.

To bridge this gap, we examine the task of land-cover segmentation and identify disparities across urban and rural areas on satellite images from different locations. As previous work shows, segmentation performance can be disparate across geography types. For example, in areas where land-cover objects have higher density or heterogeneity, per-

formance will be lower even for similar training sample sizes (Zhang et al. 2022c). Moreover, identifying and thus addressing disparities for geographic object segmentation is different from classification tasks in other image types such as facial images (Wang et al. 2019; Ramaswamy, Kim, and Russakovsky 2021; Jung et al. 2021). De-biasing classification outcomes relies on robust image-level global representations, which are not ideal for segmentation in which local features are important, thus may not apply to satellite data and relevant tasks. Instead, to our knowledge, we present the first exploration on learning generalizable and robust local landscape features, while reducing spurious features that are unequally correlated with areas of different urbanization or economic development. (referred to as "bias" or "spurious information"). In this way, our work addresses disparity issues in contrastive self-supervised learning for satellite image segmentation. The specific contributions are:

1. We propose a causal model depicting the relationship between landscape features and urban/rural property of images, to unravel the type of implicit bias that a model might learn from data. This framework enables us to identify and address unique disparity challenges in deep learning application for satellite images.

2. For the described bias scenario, we design a fair representation learning method which regularizes the statistical association between pixel-level image features and sensitive variables, termed FairDCL. The methods includes a novel feature map based local mutual information estimation module which incorporates layer-wise fairness regularization into the contrastive optimization objective. Given characteristics of satellite images, this work serves to mitigate performance disparities in downstream landscape segmentation tasks.

3. On real-world satellite datasets, FairDCL shows advantages for learning robust image representation in contrastive pre-training; it surpasses state-of-the-art methods demonstrating smaller urban-rural performance differences and higher worst-case performance, without sacrificing overall accuracy on the target tasks.

Scope and Limitation. This work specifically focuses on image representation learning without supervision of labels for the objects to be segmented (also referred to as pre-training), motivated by the approach's effectiveness and low annotation cost as described. Therefore, we do not cover other image analysis schemes, such as supervised or semi-supervised learning and focus on comparison to unsupervised robust representation learning baselines, including gradient reversal learning, domain independent learning, and global representation debiasing with mutual-information. The evaluation of learnt representation quality is achieved by applying a lightweight decoder for the semantic segmentation target to obtain the final downstream task performance, on segmenting common landscape objects, following previous work (Wang et al. 2021b; Ziegler and Asano 2022).

Related Work

Self-Supervised Methods for Satellite Images

Semantic segmentation, which quantifies land-cover location and boundary at pixel level, is a fundamental problem in satellite data analysis (Lv et al. 2023). Given that per-pixel segmentation annotation required for supervised training is expensive, a growing body of literature leverages self-supervised methods to extract useful image features from large-scale satellite image datasets (Li, Chen, and Shi 2021; Wang et al. 2022a; Li et al. 2022a). Contrastive learning is used as a self-supervised pre-training approach for various downstream vision tasks including classification, detection and segmentation (Chen et al. 2020a,b; Hendrycks et al. 2019; Misra and Maaten 2020; Reed et al. 2022; Vu et al. 2021; Ayush et al. 2021). Though most work in this area focuses on optimizing global representations for a single prediction for each image, such as presence of an animal species in an image, (Wu et al. 2018; Chen et al. 2020a,b), recent work has turned to learning representations suitable for dense predictions (i.e., a prediction for each pixel); such approaches train the model to compare local regions within images, thus preserving pixel-level information (Wang et al. 2021b; O Pinheiro et al. 2020; Chaitanya et al. 2020; Xie et al. 2021). Other work (Xiong, Ren, and Urtasun 2020) uses overlapped local blocks to increase depth and capacity for decoders that improves local learning. Such methods show the importance of local image representations on dense visual problems like satellite image segmentation, which we leverage for the first time to mitigate disparities on such problems.

When satellite data is collected in multiple temporal resolutions, studies have included contrastive learning methods to learn the representations invariant to subtle landscape variations across the short-term (Mall, Hariharan, and Bala 2023; Ayush et al. 2021). However, this type of work requires multi-temporal satellite data and does not consider the same question regarding generalizability with respect to urbanization.

Disparities in Image Recognition

Fairness-promoting approaches are being designed in multiple visual recognition domains, generally with human objects and demographic characteristics as the sensitive attributes. For example, in face recognition applications, methods are proposed for mitigating bias across groups like age, gender or race/ethnicity. Such methods include constraining models from learning sensitive information by adversarially training sensitive attribute classifiers (Raff and Sylvester 2018; Morales et al. 2020), using penalty losses (Xu et al. 2021; Serna et al. 2022), sensitive information disentanglement (Creager et al. 2019; Park et al. 2021), and augmenting biased data using generative networks (Ramaswamy, Kim, and Russakovsky 2021). Related to healthcare data and practice, methods have shown reduction in bias by altering sensitive features such as skin color but preserve relevant features to the clinical tasks (Yuan et al. 2022; Deng et al. 2023), by augmentation (Burlina et al. 2021), and by adversarial training (Abbasi-Sureshjani et al. 2020; Puyol-

Antón et al. 2021). In comparison, investigation on satellite imagery is limited; Xie *et al.* (Xie et al. 2022) formulate disparity among sub-units with linked spatial information, using spatial partitionings instead of sensitive attributes. Aiken *et al.* (Aiken, Rolf, and Blumenstock 2023) illustrate that urban-rural disparities exist in wealth prediction with satellite images. However, no work has explored disparity in image representation learning for satellite images nor proposed a method to mitigate the same.

A few recent studies have examined robustness and fairness in contrastive learning generally, including an adjusting sampling strategy to restrict models from leveraging sensitive information (Tsai et al. 2021). However, this approach could lose task-specific information by only letting the model differentiate samples from the same group to avoid learning group boundaries. Two stage training with balanced augmentation (Zhang et al. 2022a), fairness-aware losses to penalize sensitive information used in positive and negative pair differentiation (Park et al. 2022), and using hard negative samples for contrast to improve representation generalization (Robinson et al. 2020) are other proposed approaches, yet such methods both only apply for classification based on image-level representations opposed to object-level segmentation which is the focus of this work.

Summary of Gaps in the Literature

Existing work in robustness and disparity mitigation for image recognition tasks is limited in multiple ways, with important gaps specifically for satellite imagery. First, some robustness methods assume that spurious feature properties are known, such as skin colors, hair colors, presence of glasses (Wang et al. 2020; Ramaswamy, Kim, and Russakovsky 2021; Yuan et al. 2022), and remove their influences on model performance. The analog of such a property is not available in satellite images, nor is it homogeneous (e.g. each country has unique landscapes relating to urban/rural). Therefore, we do not explicitly define spurious features in our model but automatically extract them with urban/rural discriminators during training. Second, since the existing methods are mostly designed for classification problems, they use image-level representation approaches. However, fairness at an image level would not necessarily extend to pixel-level dense predictions. Third, there is very little work on robust and fair satellite image analysis, for which biased features are harder to discover, interpret and remove, compared to human-object images. Existing work on generalizable satellite representations across temporal changes train models with acquisition date, which is not always available for satellite datasets.

Problem Statement

Selection of Sensitive Attributes

In algorithmic fairness studies, sensitive attributes are those that are historically linked to discrimination or bias, and should not be used as the basis for decisions (Dwork et al. 2012). Commonly, for example in studies focused on face detection, demographic factors such as race and gender are used as sensitive attributes due to their potential, but unwanted influence on decision-making processes (Lee 2018; Pessach and Shmueli 2022). For satellite imagery, while individual-level attributes such as race and gender are not of concern, there are geographical properties such as urban-rural disparities which have precedence both for historical disparities and legal precedence for the need for protection from such disparities (Ananian and Dellaferrera 2024). Indeed, rural areas in the United States and globally have lower resources such as health care services (Peek-Asa et al. 2011; Lin et al. 2014), higher education disadvantage (Roscigno, Tomaskovic-Devey, and Crowley 2006; Li et al. 2022b) and lower investment in other areas such as communication technology (Nazem et al. 1996). These factors can all significantly impact outcomes for populations in these areas, and it is critical that future decisions impacting rural and urban places do not promulgate such disparities. In terms of operationalizing these attributes, while features of specific urban or rural areas can vary globally, there is consensus that urban areas demarcate cities and their surroundings. Urban areas are very developed, meaning there is a density of human structures, such as houses, commercial buildings, roads, bridges, and railways (NationalGeographic 2020). In sum, examining urban and rural designations as sensitive attributes can unveil systemic inequalities and aid in creating more equitable algorithms and policies globally.

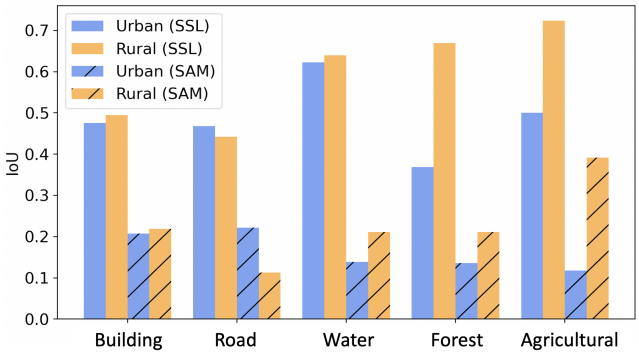

Figure 1: Model segmentation performance on urban and rural images of LoveDA (Wang et al. 2021a), measured by intersection-over-union (IoU). Two types of upstream feature encoders are used: (1) CNN encoder trained on unlabeled satellite images with contrastive self-supervised learning (SSL), and (2) pre-trained foundation model Segment Anything (SAM) (Kirillov et al. 2023). Urban-rural disparities are observed for land-cover classes with both encoders, and the disadvantaged groups are consistent across learning models.

Urban-Rural Disparities with Feature Encoders

We perform satellite image feature extraction with the studied contrastive self-supervised learning (SSL) method, MOCO-v2 (Chen et al. 2020b), and report the semantic segmentation fine-tuning results in Figure 1. There are several major disparities visible, especially for the class of "Forest" and "Agricultural". To further expose the issue, we evaluate

with a general-purpose feature encoder, Segmenting Anything Model (SAM) (Kirillov et al. 2023). It is a vision foundation model trained on a large image dataset (11 millions) of wide geographic coverage for learning comprehensive features. Therefore, the model can transfer zero-shot to image segmentation for our dataset. The results show similar disparities to SSL for each land-cover class (Figure 1). Motivated by the problem, we propose a causal model to unravel feature relationships in satellite images and the design to utilize robust features to mitigate disparity.

Causal Model for Feature Relationships

Land-cover objects in satellite images, such as residential building, roads, vegetation, etc, often have heterogeneous shapes and distributions in urban and rural areas even within the same geographic region. These distributions are affected by varying levels of development (infrastructure, greening, etc). Considering an attribute $S = \{s_0, s_1\}$ denoting urban/rural area, we define visual representation (high-dimensional embeddings output by model intermediate layers) as $X = \{X_{spurious}, X_{robust}\}$, where $X_{spurious}$ includes information that varies across urban/rural groups in S, for example, the contour, color, or texture of "road" or "building" class. X_{robust}, on the other hand, includes generalizable information, for example, "road" segments are narrow and long, while "building" segments are clustered. When model output Y is drawn from both X_{robust} and $X_{spurious}$, it can lead to biased performance. For example, roads in grey/blue color (Figure 3 A), with vehicles on them (Figure 3 B), and with lane markings (Figure 3 C), are segmented better (blue circle) than the others (red circle, Figure 3 D). Examples of more classes' spurious and robust components are in the supplementary material A [1].

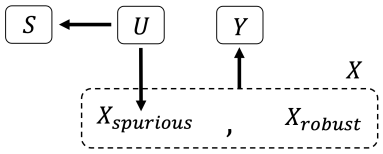

Figure 2: Diagram of defined causal relationships between representation X learnt with contrastive pre-training, target task prediction outputs Y, and urban/rural attribute S. X contains two parts, $X_{spurious}$ generated from features spuriously correlated to S and X_{robust} generated from independent and unchangeable features. U is unmeasured confounders which cause both S and $X_{spurious}$ thus result in correlations between S and $X_{spurious}$.

As a result, urban and rural representations containing disproportionate spurious information levels will cause group-level model performance disparities in semantic segmentation. Note that there are other factors not uniformly distributed across urban and rural areas, such as the number of pixels by land object class. Since we focus on representation learning, we denote such factors as unmeasured confounders U. The problem is illustrated, in terms of causal relationships, in Figure 2. Different from methods which directly alter spurious features which are defined a priori, *the goal here is to reduce the part of model representations that are correlated to the urban and rural split, an important delineation which has strong disparities globally*. That is, to obtain $\hat{X}_{robust}$ which promotes $\hat{Y} \perp S|\hat{X}_{robust}$, where the model prediction $\hat{Y}$ is independent to group discrepancy.

Accordingly, there is a need for model to (1) focus on robust and generalizable landscape features, and (2) capture local features of the image in the pre-training stage. With contrastive self-supervised pre-training as the framework, we propose an intervention algorithm to achieve the goals and promote urban-rural downstream segmentation equity.

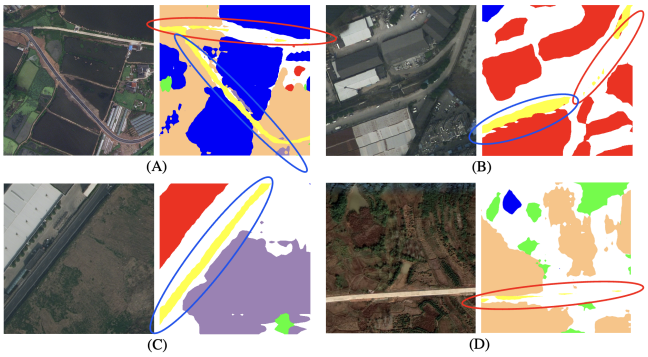

Figure 3: Examples of segmentation bias for "road" class due to spurious landscape features; the model segments certain patterns well, like straight and paved road (blue circles), but segments the variations poorly, like curvy and sand road (red circles).

Methodology

Datasets

While several standard image datasets used in fairness studies exist, datasets with linked group-level properties, specifically, urbanization, for real-world satellite imagery are very limited. We identified two datasets which had or could be linked with urban/rural annotations for disparity analyses, collected from Asia and Europe respectively, and with different spatial resolutions:

LoveDA (Wang et al. 2021a) is composed of 0.3m spatial resolution RGB satellite images collected from three cities in China. Images are annotated at pixel-level into 7 land-cover object classes, also with a label based on whether they are from an urban or rural district. Notably, images from the two groups have different class distributions. For example, urban areas contain more buildings and roads, while rural areas contain larger amounts of agriculture (Wang et al. 2021a). Moreover, it has been shown that model segmentation performances differ across urban and rural satellite images (Zhang et al. 2022c). We split the original images into 512×512 pixel tiles, take 18% of the data for testing,

[1]Code and supplementary material can be accessed at: https://github.com/ChunaraLab/FairDCL-mitigating-urban-rural-disparity

and for the rest, 90% are for contrastive pre-training (5845 urban tiles and 5572 rural tiles) and 10% for fine-tuning the pre-trained representation to generate predictions.

EOLearn Slovenia (Sinergise 2022) is composed of 10m spatial resolution Sentinel-2 images collected from whole region of Slovenia for the year 2017, with pixel-wise land cover annotations for 10 classes. We only use the RGB bands for the consistency with other datasets, remove images that have more than 10% of clouds, and split images into 256×256 pixel tiles to enlarge the training set. Labels are assigned by assessing if the center of each tile is located in urban boundaries or not (using urban municipality information[2] and administrative boundaries from OpenStreetMap[3]). This process generates 1760 urban tiles and 1996 rural tiles in total. Similar to the LoveDA process, 18% of the data are used for testing, and 90% of the rest of the data are used for pre-training and 10% for fine-tuning.

Metrics

The quality of representations learnt from self-supervised pre-training is usually evaluated by its transfer-ability to downstream tasks (Jing and Tian 2020; Wang et al. 2021b). On the downstream semantic segmentation, we use Intersection-over-Union (IoU) as the accuracy metric, calculated using pixel-wise true positives (TP), false positives (FP), and false negatives (FN),

$$\text{IoU} := \frac{TP}{TP + FP + FN}.$$

Group accuracy for group g^i is computed via the mean of class-wise IoUs (referred to as μ_{g^i}). Model overall accuracy is the averaged group results (mIoU).

We use two fairness metrics: First, the group difference with regard to accuracy (Raff and Sylvester 2018; Gong, Liu, and Jain 2021; Szabó, Jamali-Rad, and Mannava 2021; Zietlow et al. 2022) (Diff). Diff for a 2-element sensitive attribute group $\{g^1, g^2\}$ is defined as:

$$\text{Diff}\,\{g^1, g^2\} := \frac{|\mu_{g^1} - \mu_{g^2}|}{\min\{\mu_{g^1}.\mu_{g^2}\}}.$$

Second, the worst group results (Wst), which is the lower group accuracy between urban and rural. This is motivated by the problem of worsening overall performance for zero disparity (Zhang et al. 2022b).

Multi-Level Representation De-biasing

The idea of constraining mutual information between representation and sensitive attribute, also referred to as bias, to achieve attribute-invariant predictions has multiple applications (Zhu et al. 2021; Ragonesi et al. 2021; Kim et al. 2019), which all operate on a global representation $\mathbf{z} = F(\mathbf{d})$, output from image encoder F. However, invariance constraints only on the global output layer do not guarantee that sensitive information is omitted from representation hierarchies of intermediate layers or blocks in a model (herein

[2]https://www.gov.si/en/topics/towns-and-protected-areas-in-slovenia/

[3]https://www.openstreetmap.org/#map=12/40.7154/-74.1289

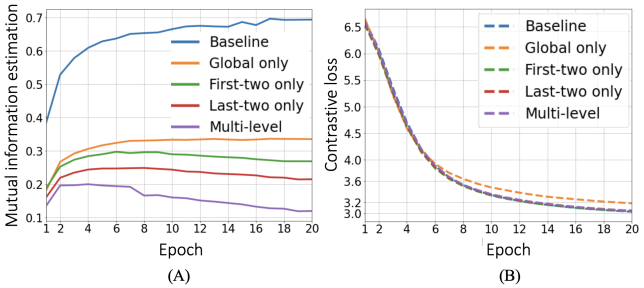

Figure 4: Bias accumulation during contrastive pre-training. (A) Sum of mutual information estimation, and (B) the contrastive loss of ResNet50 model with MoCo-V2 pre-training. The baseline method with no intervention (Baseline), regularizing only on the global feature vector (Global only), first two layers of feature maps (First-two only), last two layers of feature maps (Last-two only) all show bias residuals compared to the multi-level method proposed as part of FairDCL.

we use the term "multi-level representation" for simplicity). As has been shown, the distribution of bias in terms of its category, number and strength is not constant across layers in contrastive self-supervised models (Sirotkin, Carballeira, and Escudero-Viñolo 2022). Besides, layer-wise regularization is necessary to constrain the underlying representation space (Jin et al. 2016; Jiang et al. 2017; Li et al. 2019). Pixel-level image features in representation hierarchies are important (O Pinheiro et al. 2020; Wang et al. 2021b), especially when transferring to dense downstream tasks such as semantic segmentation, where representations are aggregated at different resolution scales in order to identify objects in pixel space. Given the evidences in sum, we design a feature map based local mutual information estimation module and incorporate layer-wise regularization into the contrastive optimization objective.

To measure mutual information $MI(X, S)$ between local feature X and the urban/rural attribute $S = \{s_0, s_1\}$, we adapt the concat-and-convolve architecture in (Hjelm et al. 2018). Notating the i^{th} layer as li, we first build a one-hot encoding map $\mathbf{c}^{li}$ for attributes S whose size is same as the feature map $\mathbf{x}^{li}$ output by li, and channel is the size of S. For each $\mathbf{x}^{li}$, a $\mathbf{c}^{li}$ is built from the joint distribution of representation space X and attribute space S, and the marginal distribution of S separately, then the $\mathbf{c}^{li}$ built in the two ways are concatenated with $\mathbf{x}^{li}$ to form an "aligned" feature map pair, denoted as $P_{XS}(\mathbf{x}^{li} \| \mathbf{c}^{li})$, and a "shuffled" feature map pair, denoted as $P_X P_S(\mathbf{x}^{li} \| \mathbf{c}^{li})$. The mutual information between the aligned and shuffled feature map pairs will be estimated by a three-layer 1×1 convolutional discriminator D_i, using the JSD-derived formation (Hjelm et al. 2018):

$$MI_{JSD}(X^{li}; S) := E_{P_{XS}}[-\text{sp}(-D_i(\mathbf{x}^{li} \| \mathbf{c}^{li}))] - E_{P_X P_S}[\text{sp}(D_i(\mathbf{x}^{li} \| \mathbf{c}^{li}))],$$

where $\text{sp}(a) = log(1 + e^a)$, and D_i uses separate optimization to converge to the lower bound of MI_{JSD}.

We empirically validate the necessity to apply multi-

level constraints to reduce bias accumulation across layers. We run self-supervised contrastive learning on LoveDA data using MoCo-v2 (Chen et al. 2020b) with ResNet50 (He et al. 2016) as the base model. Simultaneous to model contrastive training, four independent discriminators are optimized to measure the mutual information $MI_{JSD}(X^{l1}; S), ..., MI_{JSD}(X^{l4}; S)$ between representation output from the four residual layers and one output layer of ResNet50 and sensitive attributes: urban/rural. MI_{JSD} are summed to measure the total amount of model bias for the data batch. As plotted in Figure 4 (A), the baseline training without MI_{JSD} intervention shows continually increasing and significantly higher bias than other methods as the number of epochs increase. Adding a penalty loss which encourages minimizing MI_{JSD} only on the global representation or on subsets of layers both control bias accumulation, but their measurements are still high compared to multi-level, showing that global level regularization might remove partial bias but leave significant residual from earlier layers. The running loss during training indicates all methods' convergence (Figure 4 (B)); mutual information constraints in latent space do not affect the contrastive learning objective.

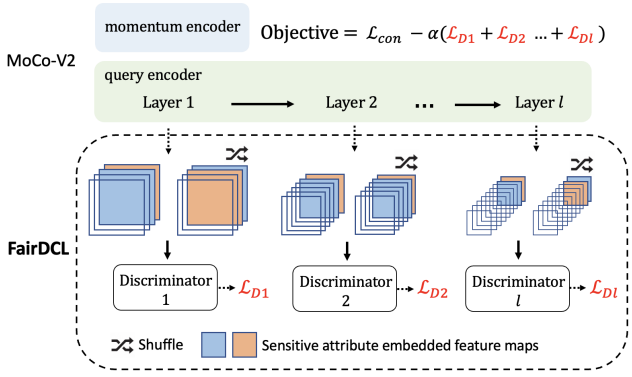

Figure 5: Overview of FairDCL. It captures spurious information $X_{spurious}$ learnt by urban/rural discriminators, and applies regularization on image representations at multiple levels. We build one-hot feature maps to encode urban/rural attribute and estimate mutual information by neural discriminators. Penalty loss $\mathcal{L}_{D_i}$ are computed accordingly and added into the final contrastive pre-training objective.

FairDCL Pipeline

Figure 5 provides an overview of the proposed fair dense representations with contrastive learning (FairDCL) method and the training process (Detailed algorithm steps are in Algorithm 1). For each iteration of contrastive pre-training, latent space representation $\mathbf{x}^{li}$ is yielded at layer li of the encoder F. Layer discriminators D_i are optimized by simultaneously estimating and maximizing MI_{JSD} with the loss:

$$\mathcal{L}_{D_i}(\mathbf{x}^{li}, S; D_i) = -MI_{JSD}(\mathbf{x}^{li}; S). \quad (1)$$

Following (Ragonesi et al. 2021), each MI discriminator is optimized for multiple inner rounds before encoder weights get updated. More rounds are desirable for discriminators

Algorithm 1: FairDCL. $F = \{l1, l2, ..li.., lN\}$ is the contrastive learning encoder. E is iterations per epoch; B is discriminators updating rounds; η is learning rate. α is regularization strength.

for each iteration a from 1 to E **do**:
  Image encoder forward propagation:
  $\mathbf{x}^{lN}, \mathbf{x}^{lN-1}, ..., \mathbf{x}^{l1} \leftarrow F(x)$ ▷ $\mathbf{x}^{li}$ is the query representation output of the layer li
  Discriminators updating:
  for each round b from 1 to B **do**:
    for each discriminator D_i **do**:
      $\mathcal{L}_{Di} \leftarrow D_i(\mathbf{x}^{li} \parallel P_{XS}(\mathbf{c}^{li}), \mathbf{x}^{li} \parallel P_S(\mathbf{c}^{li}))$
      ▷ Forward aligned and shuffled feature pairs
      $W_{D_i} \leftarrow W_{D_i} - \eta \bigtriangledown \mathcal{L}_{D_i}$ ▷ Optimize D_i
  Image encoder updating:
  $\mathbf{x}^{lN}, \mathbf{x}^{lN-1}, ..., \mathbf{x}^{l1}, q, k \leftarrow F(x)$ ▷ q, k is the query and key global representation
  $\mathcal{L}_{con} \leftarrow q, k$ ▷ Compute contrastive loss
  $\mathcal{L}_D \leftarrow \sum_{i=1}^{N} \mathcal{L}_{D_i}$ ▷ Compute MI loss
  $W_F \leftarrow W_F - \eta(\mathcal{L}_{con} - \alpha\mathcal{L}_D)$ ▷ Update encoders

to estimate mutual information with increased accuracy, and based on resource availability, we set a uniform round number $B = 20$. After discriminator optimization completes, one iteration of image encoder training is conducted wherein discriminators infer the multi-stage mutual information by loss in (1), and the losses are combined with the contrastive learning loss with a hyper-parameter α adjusting the fairness constraint strength. The final training objective is:

$$\mathcal{L}_F(X, S; D, F) = \mathcal{L}_{con} - \alpha(\sum_{li} \mathcal{L}_{D_i}(X^{li}, S; D_i)), \quad (2)$$

With the training objective, the image encoder is encouraged to generate representation X with high $\mathcal{L}_D$, thus low MI_{JSD} (low spurious information). We apply FairDCL on the state-of-the-art contrastive learning framework MoCo-v2 (Chen et al. 2020b). The loss used for learning visual representation is InfoNCE (Oord, Li, and Vinyals 2018):

$$\mathcal{L}_{con}(F) = -log\frac{\exp(qk/\tau)}{\exp(qk/\tau) + \sum_j (q\hat{k_j}/\tau)}. \quad (3)$$

Here F consists of a query encoder and a key encoder, which outputs representations q and k from two augmented views of the same image. $\hat{k_j}$ is a queue of representations encoded from different images in the dataset. $\mathcal{L}_{con}$ encourages the image encoder to distinguish positive and negative keys so it can extract useful visual representations.

Generalizability to contrastive frameworks. We note that the proposed locality-sensitive de-biasing scheme applying intervention on embedding space can be integrated with any state-of-the-art convolution feature extractors, thus has the potential to be further promoted with different contastive learning frameworks. Empirically, we experiment with DenseCL (Wang et al. 2021b), which designs pixel-level positive and negative keys to better learn local feature correspondences. Since the method fills the gap between pre-training and downstream dense prediction, it is suitable as an alternative contrastive learning framework for our proposed method. The results are attached in the supplementary material.

	Method	LoveDA			Slovenia		
		Diff(↓)	Wst(↑)	mIoU(↑)	Diff(↓)	Wst(↑)	mIoU(↑)
Moco-v2	Vanilla	0.165 (± 0.010)	0.498 (± 0.004)	0.539 (± 0.002)	0.150 (± 0.028)	0.205 (± 0.003)	0.220 (± 7e-4)
	GR	0.155 (± 0.013)	0.501 (± 0.005)	0.540 (± 0.002)	0.128 (± 0.023)	0.208 (± 6e-4)	0.222 (± 0.002)
	DI	0.144 (± 0.009)	0.499 (± 0.004)	0.535 (± 0.004)	0.136 (± 0.010)	0.208 (± 0.005)	0.222 (± 0.005)
	UnbiasedR	0.150 (± 0.008)	0.502 (± 0.003)	0.540 (± 0.003)	0.130 (± 0.017)	0.205 (± 0.004)	0.219 (± 0.004)
	FairDCL	**0.127** (± 0.005)	**0.508** (± 0.002)	0.540 (± 0.002)	**0.076** (± 0.011)	**0.217** (± 0.002)	0.225 (± 0.003)
DenseCL	Vanilla	0.154 (± 0.013)	0.503 (± 0.004)	0.542 (± 0.002)	0.122 (± 0.017)	0.207 (± 0.006)	0.219 (± 0.005)
	GR	0.148 (± 0.012)	0.498 (± 0.006)	0.534 (± 0.004)	0.120 (± 0.020)	0.201 (± 0.003)	0.216 (±0.002)
	DI	0.140 (± 0.007)	0.501 (± 0.003)	0.536 (± 0.002)	0.120 (± 0.008)	0.206 (± 0.001)	0.218 (± 5e-4)
	UnbiasedR	0.157 (± 0.008)	0.495 (± 0.005)	0.534 (± 0.003)	0.128 (± 0.014)	0.206 (± 0.004)	0.219 (± 0.003)
	FairDCL	**0.108** (± 0.008)	**0.518** (± 0.003)	0.546 (± 0.004)	**0.079** (± 0.016)	**0.215** (± 0.001)	0.223 (± 0.003)

Table 1: Downstream semantic segmentation results on LoveDA and Slovenia datasets, using an FCN-8s model with the backbone learnt with comparison pre-training methods. Our FairDCL shows consistent improvements on fairness metrics (Diff and Wst) (bold) over other de-biasing methods , also we do not see a decreased accuracy (mIoU) (underlined) than the vanilla baseline, on all datasets. Results and standard deviations are reported over 5 independent runs.

Experiments

Implementation Details

The first stage of contrastive pre-training. The base model for the image encoders is ResNet50 (He et al. 2016). The mutual information discriminators D_i are built with 1×1 convolution layers (architecture details in supplementary material D). The contrastive pre-training runs for 10k iterations for each dataset with a batch size of 64. Data augmentations used to generate positive and negative image view pairs are random greyscale conversion and random color jittering (no cropping, flips or rotations in order to retain local feature information). Hyper-parameter α, which scales the amount of mutual information loss $\mathcal{L}_D$ in the total loss, is set to 0.5. Adam optimizer is used with a learning rate of 10^{-3} and weight decay of 10^{-4} for both encoders and discriminators.

Comparison methods include state-of-the-art fair representation learning approaches: (1) gradient reversal training (GR) (Raff and Sylvester 2018), which follows the broad approach of removing bias or sensitive information from learnt representations by inverse gradients of attribute classifiers. This approach has been adapted to multiple image recognition tasks (Zhang, Lemoine, and Mitchell 2018; Wang et al. 2019, 2022b). (2) Domain independent training (DI) which samples data with a consistent group attribute in each training iteration to avoid leveraging spurious group boundaries (Tsai et al. 2021; Wang et al. 2020). (3) Unbiased representation learning (UnbiasedR) (Ragonesi et al. 2021) which uses single-level de-biasing only for the global image representation. All comparison methods use the same learning architectures, and are trained with the same settings.

The second stage of semantic segmentation fine-tuning. Following the protocal of previous work (Wang et al. 2021b; Zhang et al. 2021; Ziegler and Asano 2022), we train a FCN-8s (Long, Shelhamer, and Darrell 2015) model on top of the fixed ResNet50 backbone learnt from the pre-training stage for 60 epochs with a batch size of 16, and evaluate on the testing split for each dataset. We use cross-entropy (CE) loss as the training objective, and stochastic gradient descent (SGD) as the optimizer with a learning rate of 10^{-3} and a momentum of 0.9. The learning rate is decayed using a polynomial learning rate scheduler implemented in PyTorch. Image data augmentations used in the fine-tuning include random horizontal/vertical flips and random rotations. For both stages, the dimension of the input image to the model is $512 \times 512 \times 3$ where 3 indicates the RGB bands. NVIDIA RTX8000 GPU is used for training.

Results

Downstream Performances Table 1 summarizes model fine-tuning results pre-trained with baseline: vanilla MoCo-v2, and de-biasing methods: GR, DI, UnbiasedR, and FairDCL, on the two satellite image datasets.

We first note that FairDCL consistently outperforms other approaches in terms of fairness, that it obtains the smallest cross-group difference (Diff) and highest worst group result (Wst). To reveal model decision process, we draw class activation maps for "road" and "forest" in Figure 6 using Grad-CAM (Selvaraju et al. 2017). Compared to the vanilla baseline, the model pre-trained with FairDCL better activates and recognizes tricky land-cover segments: curved part of roads (more seen in rural area), and sparse river-side forests (more seen in urban area). The method can learn robust representations that are generalizable to object shape and context variations, using the multi-scale representation regularization. Therefore, it reduces segmentation disparity caused by landscape discrepancy between groups.

Another advantage of learning better features to ensure generalizability is no subsequent target task degradation, which is crucial for real-world applications. Importantly, FairDCL does not show degraded accuracy (the improved "Diff" metric or "Wst" metric does *not* cause a worse

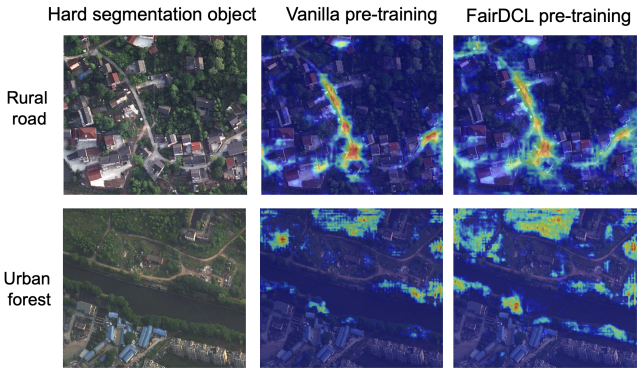

Figure 6: Class activation mapping. Detailed image locations that most impact model's prediction for "road" (first row) and "forest" (second row). FairDCL better recognizes land-cover segments particular to sensitive attributes.

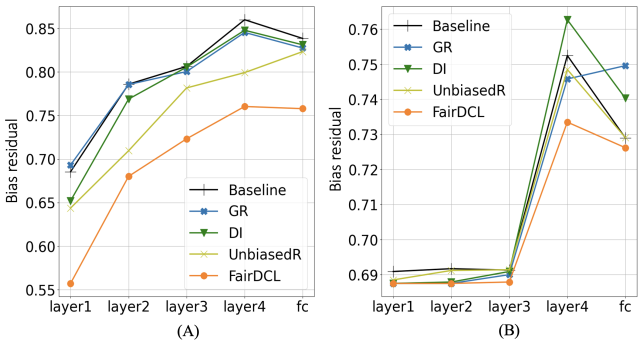

Figure 7: Linear separation evaluation: We train a linear neural layer on top of each level of representations output from different model layers. They include four residual modules ("layer1" - "layer4") that encode intermediate representations and a global output layer ("fc") that encodes the global representation. The linear layer is to classify sensitive attributes: urban/rural on (A) LoveDA dataset, and women/-men on (B) MS-COCO dataset. *Lower accuracy is good*: it indicates harder to predict sensitive attributes using the pre-trained representations, thus lower bias residual.

"mIoU" metric) on all cases. Obtaining comparable or better overall accuracy to Baseline demonstrates robustness in addition to disparity reduction. In contrast, DI allows image contrastive pairs only from a fraction of data which can discount model learning (Wang et al. 2020). The adversarial approach used in GR can be counter-productive if the adversary is not trained enough to achieve the infimum (Moyer et al. 2018), which could all potentially degrade model quality for group equalization. Our adapted mutual information constraints use information-theoretic objectives, proved to be able to optimize without competing with the encoder so can match or exceed state-of-the-art adversarial de-biasing methods (Moyer et al. 2018; Ragonesi et al. 2021). FairDCL further shows that applying the mutual information constraints on multi-level latent representations can better extend fairness to pixel-level applications, which outperforms the image-level only constraints used in UnbiasedR.

Embedding Spaces To further trace how the image representations learnt with proposed method improves fairness, we analyze a linear separation property (Reed et al. 2022) on model embedding spaces. Specifically, we assess how well a linear model can differentiate urban/rural sensitive attributes using learnt representations. High separation degree indicates that the encoder model's embedding space and attribute are differentiable (Oord, Li, and Vinyals 2018; Chen et al. 2020a; Reed et al. 2022), which could be used as a short-cut in prediction and cause bias, thus is not desirable here. We freeze the trained ResNet50 encoder and use a fully connected layer on top of representation output from different layers for urban/rural attribute classification. Figure 7 (A) presents the classification score on urban/rural attribute on LoveDA: FairDCL obtains the lowest attribute differentiation results for all embedding stages and global stage of representation, indicating that the encoder trained with FairDCL has favorably learnt the least sensitive information at pixel-level features during the contrastive pre-training. Though we focus on satellite images, we check the method's generalizability to a different image domain by conducting contrastive pre-training on MS-COCO (Lin et al. 2014), a dataset commonly used in fairness studies (Wang et al. 2019; Tang et al. 2021; Wang, Liu, and Wang 2021). Sensitive attribute gender, categoriezed as "women" or "men", is obtained from (Zhao, Wang, and Russakovsky 2021). There are 2901 images with "women" and 6567 images with "men" labels. Linear analysis results show that FairDCL again produces the desired lowest classification accuracies (Figure 7 (B)), but unlike in Figure 7 (A), it does not surpass the other comparison methods much. This is likely because while geographic attributes are represented at a pixel-level, human face/object, as a foreground, may not be represented through local features throughout the image, thus gender attributes are less pronounced as pixel-level biases in dense representation learning, which is what our proposed approach focuses on.

$\alpha = 0.1$			$\alpha = 0.5$		
Diff	Wst	mIoU	Diff	Wst	mIoU
0.138	0.506	**0.541**	0.127	**0.508**	0.540
$\alpha = 1$			$\alpha = 10$		
Diff	Wst	mIoU	Diff	Wst	mIoU
0.127	0.506	0.538	**0.126**	0.506	0.538

Table 2: Ablation study for discriminator weights. The fine-tuning result is shown for the representation pre-trained with $\alpha = 0.1, 0.5, 1, 10$, of the proposed fairness objective.

Ablation Studies We perform an ablation study for hyper-parameter α which scales discriminator loss $\mathcal{L}_D$, thus the fairness regularization strength. The method is overall robust to the parameter (Table 2); a large α like 10 will not corrupt the downstream accuracy, and a small α like 0.1 has lower fairness gain but still shows advantage over comparison methods in Table 1. We select $\alpha = 0.5$ for a bal-

	Urban:68% Rural:32%			Urban:35% Rural:65%		
Method	Diff(↓)	Wst (↑)	mIoU(↑)	Diff(↓)	Wst (↑)	mIoU(↑)
Baseline	0.147	0.500	0.537	0.170	0.497	0.539
GR	0.154	0.497	0.535	0.144	0.511	0.547
DI	0.145	0.498	0.534	0.148	0.499	0.535
UnbiasedR	0.146	0.500	0.537	0.145	0.503	0.539
FairDCL	**0.128**	**0.511**	0.543	**0.122**	**0.518**	0.549

Table 3: Ablation study for unbalanced group data. The proportion of urban/rural samples in the pre-training data is adjusted such that one group has much less samples. FairDCL performs consistently with the data distirbution shifts.

ance. Furthermore, urban and rural groups have comparable training samples in earlier experiments (LoveDA is 5.8k and 5.5k, Slovenia is 1.7k and 1.9k for urban/rural). We intentionally reduce pre-training samples for certain groups to generate more unbalanced subsets. Shown in Table 3, the proposed method shows robustness under the two less even group distributions.

Discussion and Conclusion

Among the broader fairness literature in visual recognition, work focusing on satellite imagery that depicts physical environments has been limited. This limitation is largely due to the difficulty in identifying population level biased landscape features. Also, disparity problems in satellite image recognition may get categorized as domain adaptation or transfer learning problems, other popular computer vision fields; though they share similar technical methods in bias mitigation and invariant representation learning, the specific objective of fair urban/rural satellite image recognition is to remove spatially disproportionate features that favor one subgroup over the others, beyond addressing covariate shift.

Here we define the scenario with a causal graph, showing that contrastive self-supervised pre-training can utilize spurious land-cover object features, thus accumulate urban/rural attribute-correlated bias. The biased image representations will result in disparate downstream segmentation accuracy between subgroups within a specific geographic area. Then, we address the problem via a mutual information training objective to learn robust local features with minimal spurious representations. Experimental results show fairer segmentation results pre-trained with the proposed method on real-world satellite datasets. In addition to disparity reduction, the method consistently avoids a trade-off between model fairness and accuracy.

As future directions, a wider set of satellite datasets can be explored. The fairness analysis can be scaled to a greater number of attributes relevant to geography, in addition to urbanization. Methods to encode sensitive attributes in the model embedding space in addition to one-hot feature maps can also be explored. We encourage experimenting with different encoding mechanisms and mutual information estimators to improve fairness regularization performance across different real-world settings.

Acknowledgments

We acknowledge funding from NSF award number 1845487. We also thank Harvineet Singh and Vishwali Mhasawade for helpful discussions.

References

Abbasi-Sureshjani, S.; Raumanns, R.; Michels, B. E.; Schouten, G.; and Cheplygina, V. 2020. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, 183–192. Springer.

Agastya, C.; Ghebremusse, S.; Anderson, I.; Vahabi, H.; and Todeschini, A. 2021. Self-supervised contrastive learning for irrigation detection in satellite imagery. *arXiv preprint arXiv:2108.05484*.

Aiken, E.; Rolf, E.; and Blumenstock, J. 2023. Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23) Special Track on AI for Good*.

Ananian, S.; and Dellaferrera, G. 2024. *Employment and wage disparities between rural and urban areas*. 107. ILO Working Paper.

Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10181–10190.

Bountos, N. I.; Papoutsis, I.; Michail, D.; and Anantrasirichai, N. 2021. Self-supervised contrastive learning for volcanic unrest detection. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Burlina, P.; Joshi, N.; Paul, W.; Pacheco, K. D.; and Bressler, N. M. 2021. Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, 10(2): 13–13.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33: 12546–12558.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

de Vries, T.; Misra, I.; Wang, C.; and van der Maaten, L. 2019. Does Object Recognition Work for Everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Deng, W.; Zhong, Y.; Dou, Q.; and Li, X. 2023. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, 158–169. Springer.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gong, S.; Liu, X.; and Jain, A. K. 2021. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3414–3424.

Griffith, J. A. 2002. Geographic techniques and recent applications of remote sensing to landscape-water quality studies. *Water, Air, and Soil Pollution*, 138: 181–197.

Hay, A. M. 1995. Concepts of equity, fairness and justice in geographical studies. *Transactions of the Institute of British Geographers*, 500–508.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Jiang, Z.; Wang, Y.; Davis, L.; Andrews, W.; and Rozgic, V. 2017. Learning discriminative features via label consistent neural network. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 207–216. IEEE.

Jin, X.; Chen, Y.; Dong, J.; Feng, J.; and Yan, S. 2016. Collaborative layer-wise discriminative learning in deep neural networks. In *European Conference on Computer Vision*, 733–749. Springer.

Jing, L.; and Tian, Y. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4037–4058.

Jung, S.; Lee, D.; Park, T.; and Moon, T. 2021. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12115–12124.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9012–9020.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kondmann, L.; and Zhu, X. X. 2021. Under the Radar–Auditing Fairness in ML for Humanitarian Mapping. *arXiv preprint arXiv:2108.02137*.

Lee, N. T. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3): 252–260.

Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; and Tao, C. 2022a. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Li, H.; Zeng, Y.; Gan, L.; Tuersun, Y.; Yang, J.; Liu, J.; and Chen, J. 2022b. Urban-rural disparities in the healthy ageing trajectory in China: a population-based study. *BMC Public Health*, 22(1): 1406.

Li, W.; Chen, H.; and Shi, Z. 2021. Semantic segmentation of remote sensing images with self-supervised multitask representation learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 6438–6450.

Li, Z.; Brendel, W.; Walker, E.; Cobos, E.; Muhammad, T.; Reimer, J.; Bethge, M.; Sinz, F.; Pitkow, Z.; and Tolias, A. 2019. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lv, J.; Shen, Q.; Lv, M.; Li, Y.; Shi, L.; and Zhang, P. 2023. Deep learning-based semantic segmentation of remote sensing images: a review. *Frontiers in Ecology and Evolution*, 11: 1201125.

Majumdar, S.; Flynn, C.; and Mitra, R. 2022. Detecting Bias in the Presence of Spatial Autocorrelation. In *Algorithmic Fairness through the Lens of Causality and Robustness workshop*, 6–18. PMLR.

Mall, U.; Hariharan, B.; and Bala, K. 2023. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5261–5270.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6707–6717.

Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Tolosana, R. 2020. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2158–2164.

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31.

NationalGeographic. 2020. Urban Area. https://education.nationalgeographic.org/resource/urban-area/. Accessed: 2024-08-03.

Nazem, S. M.; Liu, Y.-H.; Lee, H.; and Shi, Y. 1996. Implementing telecommunications infrastructure: a rural America case. *Telematics and Informatics*, 13(1): 23–31.

O Pinheiro, P. O.; Almahairi, A.; Benmalek, R.; Golemo, F.; and Courville, A. C. 2020. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33: 4489–4500.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2403–2411.

Park, S.; Lee, J.; Lee, P.; Hwang, S.; Kim, D.; and Byun, H. 2022. Fair Contrastive Learning for Facial Attribute Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10389–10398.

Peek-Asa, C.; Wallis, A.; Harland, K.; Beyer, K.; Dickey, P.; and Saftlas, A. 2011. Rural disparity in domestic violence prevalence and access to resources. *Journal of women's health*, 20(11): 1743–1749.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Puyol-Antón, E.; Ruijsink, B.; Piechnik, S. K.; Neubauer, S.; Petersen, S. E.; Razavi, R.; and King, A. P. 2021. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 413–423. Springer.

Raff, E.; and Sylvester, J. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 189–198. IEEE.

Ragonesi, R.; Volpi, R.; Cavazza, J.; and Murino, V. 2021. Learning unbiased representations via mutual information backpropagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2729–2738.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9301–9310.

Reed, C. J.; Yue, X.; Nrusimha, A.; Ebrahimi, S.; Vijaykumar, V.; Mao, R.; Li, B.; Zhang, S.; Guillory, D.; Metzger, S.; et al. 2022. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2584–2594.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Roscigno, V. J.; Tomaskovic-Devey, D.; and Crowley, M. 2006. Education and the inequalities of place. *Social forces*, 84(4): 2121–2145.

Saha, S.; Mou, L.; Qiu, C.; Zhu, X. X.; Bovolo, F.; and Bruzzone, L. 2020. Unsupervised deep joint segmentation of multitemporal high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8780–8792.

Scheibenreif, L.; Hanna, J.; Mommert, M.; and Borth, D. 2022. Self-Supervised Vision Transformers for Land-Cover Segmentation and Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1422–1431.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Serna, I.; Morales, A.; Fierrez, J.; and Obradovich, N. 2022. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305: 103682.

Setianto, M.; and Gamal, A. 2021. Spatial justice in the distribution of public services. In *IOP Conference Series: Earth and Environmental Science*, volume 673, 012024. IOP Publishing.

Sinergise. 2022. Modified Copernicus Sentinel data 2017/Sentinel Hub. https://sentinel-hub.com/. Accessed: 2024-08-03.

Sirotkin, K.; Carballeira, P.; and Escudero-Viñolo, M. 2022. A study on the distribution of social biases in self-supervised learning visual models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10442–10451.

Soden, R.; Wagenaar, D.; Luo, D.; and Tijssen, A. 2019. Taking ethics, fairness, and bias seriously in machine learning for disaster risk management. *arXiv preprint arXiv:1912.05538*.

Szabó, A.; Jamali-Rad, H.; and Mannava, S.-D. 2021. Tilted cross-entropy (TCE): Promoting fairness in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2305–2310.

Tang, R.; Du, M.; Li, Y.; Liu, Z.; Zou, N.; and Hu, X. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, 633–645.

Tsai, Y.-H. H.; Ma, M. Q.; Zhao, H.; Zhang, K.; Morency, L.-P.; and Salakhutdinov, R. 2021. Conditional contrastive learning: Removing undesirable infor-

mation in self-supervised representations. *arXiv preprint arXiv:2106.02866*.

Von Holdt, J.; Eckardt, F.; Baddock, M.; and Wiggs, G. F. 2019. Assessing landscape dust emission potential using combined ground-based measurements and remote sensing data. *Journal of Geophysical Research: Earth Surface*, 124(5): 1080–1098.

Vu, Y. N. T.; Wang, R.; Balachandar, N.; Liu, C.; Ng, A. Y.; and Rajpurkar, P. 2021. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, 755–769. PMLR.

Wang, J.; Liu, Y.; and Wang, X. E. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021a. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5310–5319.

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021b. Dense contrastive learning for self-supervised visual pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3024–3033.

Wang, Y.; Albrecht, C. M.; Braham, N. A. A.; Mou, L.; and Zhu, X. X. 2022a. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*.

Wang, Z.; Dong, X.; Xue, H.; Zhang, Z.; Chiu, W.; Wei, T.; and Ren, K. 2022b. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10379–10388.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xie, Y.; He, E.; Jia, X.; Chen, W.; Skakun, S.; Bao, H.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-level Learning Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12208–12216.

Xie, Z.; Lin, Y.; Zhang, Z.; Cao, Y.; Lin, S.; and Hu, H. 2021. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16684–16693.

Xiong, Y.; Ren, M.; and Urtasun, R. 2020. Loco: Local contrastive representation learning. *Advances in neural information processing systems*, 33: 11142–11153.

Xu, X.; Huang, Y.; Shen, P.; Li, S.; Li, J.; Huang, F.; Li, Y.; and Cui, Z. 2021. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 578–586.

Yuan, H.; Hadzic, A.; Paul, W.; de Flores, D. V.; Mathew, P.; Aucott, J.; Cao, Y.; and Burlina, P. 2022. EdgeMixup: Improving Fairness for Skin Disease Classification and Segmentation. *arXiv preprint arXiv:2202.13883*.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, F.; Kuang, K.; Chen, L.; Liu, Y.; Wu, C.; and Xiao, J. 2022a. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*.

Zhang, H.; Dullerud, N.; Roth, K.; Oakden-Rayner, L.; Pfohl, S.; and Ghassemi, M. 2022b. Improving the Fairness of Chest X-ray Classifiers. In *Conference on Health, Inference, and Learning*, 204–233. PMLR.

Zhang, M.; Singh, H.; Chok, L.; and Chunara, R. 2022c. Segmenting across places: The need for fair transfer learning with satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2916–2925.

Zhang, Y.; Hooi, B.; Hu, D.; Liang, J.; and Feng, J. 2021. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Advances in Neural Information Processing Systems*, 34: 29848–29860.

Zhao, D.; Wang, A.; and Russakovsky, O. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14830–14840.

Zhu, W.; Zheng, H.; Liao, H.; Li, W.; and Luo, J. 2021. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15002–15012.

Ziegler, A.; and Asano, Y. M. 2022. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14502–14511.

Zietlow, D.; Lohaus, M.; Balakrishnan, G.; Kleindessner, M.; Locatello, F.; Schölkopf, B.; and Russell, C. 2022. Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10410–10421.