

# Fair contrastive pre-training for geographic image segmentation

Anonymous submission

## ABSTRACT

Contrastive representation learning, a method for learning features that distinguish dissimilar data samples, is widely employed in visual recognition for geographic image data (remote-sensing such as satellite imagery or proximal sensing such as street-view imagery). However because of heterogeneity in landscapes (e.g. how a road looks in different places), models can show disparate performance across spatial units. In this work, we consider fairness risks in identification of land-cover features (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 assess model prediction disparities across selected sensitive groups: urban and rural scenes for satellite image datasets and city GDP level for a street view image dataset. We propose fair dense representation with contrastive learning (FairDCL)<sup>1</sup> as a method for de-biasing the multi-level latent space of a convolution neural network. The method improves feature identification by removing spurious latent representations which are disparately distributed across groups, and is achieved in an unsupervised way by contrastive pre-training. The pre-trained image representation improves downstream task fairness and outperforms state-of-the-art methods for the absence of a fairness-accuracy trade-off. Image embedding evaluation and ablation studies further demonstrate FairDCL's robustness. As fairness in geographic imagery is a nascent topic, our work motivates researchers to consider fairness metrics in such applications, especially reinforced by our results showing no accuracy degradation.

## CCS CONCEPTS

- Computing methodologies → Machine learning

## KEYWORDS

Fairness, Urbanization, Geographic images, Semantic segmentation, Computer vision, Contrastive self-supervised learning

## ACM Reference Format:

Anonymous submission. 2018. Fair contrastive pre-training for geographic image segmentation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (EAAMO '23)*. ACM, New York, NY, USA, 14 pages. <https://doi.org/XXXXXX.XXXXXX>

<sup>1</sup>The code for our method is published at <https://anonymous.4open.science/r/FairDCL-1283>

Permission to make digital or hard copies of all or part of this work for personal or  
**Unpublished working draft. Not for distribution.** attributed  
for profit or commercial advantage and that copies bear this notice and the full citation  
on the first page. Copyrights for components of this work owned by others than ACM  
must be honored. Abstracting with credit is permitted. To copy otherwise, or republish,  
to post on servers or to redistribute to lists, requires prior specific permission and/or a  
fee. Request permissions from permissions@acm.org.

EAAMO '23, October 30–November 1, 2023, Boston, MA

© 2018 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM/...\$15.00

<https://doi.org/XXXXXX.XXXXXX>

2023-05-10 22:03. Page 1 of 1-14.

## 1 INTRODUCTION

Dense pixel-level image recognition via deep learning for tasks such as segmentation have a variety of applications in landscape feature analysis from geographic images. For example, regional water quality analysis [16] or dust emission estimation [60]. Success of the methods rely on powerful visual representations that include both local and global information. However, since pixel-level annotations are 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 feature encoder and transferring learnt representation to downstream problems. Indeed, the performance of such self-supervised approaches has been shown to match supervised learning on a large range of image tasks [12, 59]. Moreover, as a mainstream, contrastive self-supervised techniques have shown state-of-the-art performance in learning image representation for land cover semantic segmentation across locations [2, 48]. In particular, since labeled images are hard to obtain for geographic images, and contrastive approaches do not require labeled images, they have demonstrated benefits in many real-world tasks including analyzing multitemporal images to monitor dynamic land surface [47], irrigation detection from uncurated and unlabeled satellite images [1], and volcanic unrest detection with scarce label and imbalanced classes [4].

Recent attention in machine learning systems has highlighted performance inequities including those by geographic area [10, 31, 33, 50]. Therefore, given the increased potential of self-supervised contrastive learning, here we turn attention to fairness risks in recognition outcomes from geographic images. Algorithmic fairness is an increasingly important concern in computer vision, considering its usage for societal decisions across many areas such as health, urban planning, climate change and disaster risk assessment [32, 33, 53]. Further, while recent work calls for more focus on fairness studies for imagery with human as objects [57], there is very limited fairness work on model performance disparities for geographic images despite their wide applications.

To bridge this gap, we examine model semantic segmentation prediction and identify disparities across geographic subgroups on satellite image and city street view datasets from three 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 [74], performance will be lower even for similar data representations. Moreover, identifying and thus addressing disparities for geographic object segmentation is different from fairness tasks in other image types such as facial images. First, facial recognition (mostly classification) relies on image-level global representations, which are not ideal for segmentation in which local features are important. Second, specifying sensitive attributes for geographic images is challenging because sensitive land-cover features are generally task-related, unlike many previous fairness studies in which bias is easier to

117 identify and separate out (e.g. skin color, gender) because the sensitive  
 118 attributes are less relevant with respect to the specific task (e.g.  
 119 skin color is not relevant to facial recognition) [24, 44, 64].

120 In sum, existing fairness methods do not always apply to geo-  
 121 graphic data and relevant tasks. Instead, we develop unbiased  
 122 representation learning by relying on generalizable and robust  
 123 landscape features, while reducing spurious features that are un-  
 124 equally correlated with sensitive groups (referred to as “bias” or  
 125 “sensitive information”). By constructing regularization terms on  
 126 the statistical association between pixel-level image features and  
 127 sensitive variables, the approach directly mitigates performance  
 128 disparities in the downstream landscape semantic segmentation  
 129 tasks. The specific contributions of the work are:

- 131 (1) We propose a causal model depicting the relationship be-  
 132 tween image features and sensitive attributes to unravel  
 133 the type of implicit bias caused by spurious correlations in  
 134 geographic images. This framework enables us to identify  
 135 and address unique fairness challenges in geographic image  
 136 recognition.
- 137 (2) For the described bias scenario, we design a fair representa-  
 138 tion learning method which focuses on local image features  
 139 of multiple resolution levels, termed FairDCL. This novel  
 140 method leverages mutual information guided de-biasing with  
 141 regularization at multiple levels for pixel-level visual recog-  
 142 nition such as segmentation, which is specifically relevant  
 143 for geographic data and tasks.
- 144 (3) On three diverse geographic imagery datasets, FairDCL shows  
 145 the advantages for learning fair visual representation in con-  
 146 trastive pre-training; it surpasses state-of-the-art fairness  
 147 methods with smaller group difference, higher worst-case  
 148 group performance, and without sacrificing overall accuracy  
 149 on downstream semantic segmentation tasks.

## 150 2 LITERATURE REVIEW

153 *Contrastive learning for dense representation.* Contrastive learning is  
 154 used as a self-supervised pre-training approach for various down-  
 155 stream vision tasks including classification, detection and segmen-  
 156 tation [2, 6, 7, 18, 34, 45, 61]. Though most work in this area focuses  
 157 on optimizing a global representation for image-level tasks [6, 7, 67],  
 158 recent work has turned to learning representation suitable for pixel-  
 159 level predictions; Wang *et al.* [65] design a dense projection through  
 160 local feature vectors for contrastive objectives, thus preserving  
 161 spatial information. Xiong *et al.* [69] use overlapped local blocks  
 162 to increase depth and capacity for decoders that improves local  
 163 learning. Others [5, 37, 68] apply local contrastive loss to enlarge  
 164 representation dissimilarities across regions and similarities across  
 165 augmented views. Such methods show the importance of local  
 166 features on dense visual problems like object detection and segmen-  
 167 tation, which motivates method design in this study.

168 *Fairness in image recognition.* Fairness-promoting approaches  
 169 are being designed in multiple visual recognition domains. In face  
 170 recognition applications, methods are proposed towards mitigating  
 171 bias across groups such as by age, gender or race/ethnicity. Meth-  
 172 ods include constraining models from learning sensitive informa-  
 173 tion by adversarially training sensitive attribute classifiers [35, 42],

175 using penalty losses [49, 70], sensitive information disentangle-  
 176 ment [9, 39], and augmenting biased data using generative net-  
 177 works [44]. These methods achieve fairer recognition, but only  
 178 focus on global classification tasks. Related to healthcare data and  
 179 practice, Puyol-Antón *et al.* [41] improve cardiac MR segmentation  
 180 by jointly training a racial meta-attribute classifier. Yuan *et al.* [71]  
 181 reduce skin tone bias in skin disease classification and segmentation  
 182 by altering colors in images but preserve lesion structure edges.  
 183 Efforts on geographic imagery include achieving even class-level  
 184 segmentation results on street scenes using tilted cross entropy  
 185 loss [55] and fair underwater object detection by simulating better  
 186 quality images for scarce species [11]. A few recent studies look at  
 187 fairness in contrastive learning; Tsai *et al.* [58] propose sampling  
 188 positive and negative pairs from the same group to restrict models  
 189 from leveraging sensitive information, but could potentially lose  
 190 task-specific information of the data contrasts with different groups.  
 191 Park *et al.* [40] propose fairness-aware losses to penalize sensitive  
 192 information used in positive and negative pair differentiation, but  
 193 in a supervised setting with target labels fully available.

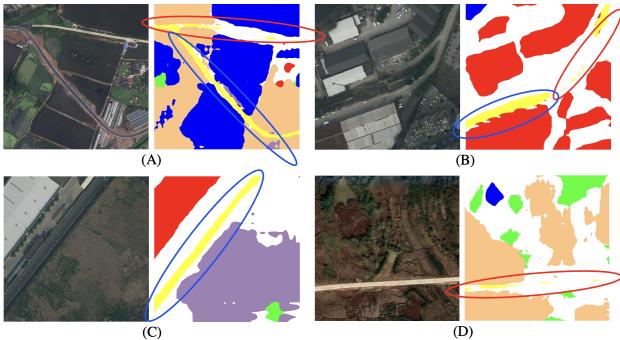
194 *Mutual information for de-biasing.* For de-biasing purposes, mu-  
 195 tual information as a statistical association measure between vari-  
 196 ables [26] has recently been used to instruct training objectives  
 197 to minimize model dependence on irrelevant variables. In deep  
 198 networks, mutual information lower bound estimators are adapted,  
 199 such as MINE [3] and DeepInfoMax [19]. Ragonesi *et al.* [43] use  
 200 MINE loss to model optimization which shows utility in remov-  
 201 ing irrelevant information in digit and face classifications. Zhu *et*  
 202 *al.* [77] compute cross-sample mutual information for reducing  
 203 different bias variations. Previous work shows leveraging mutual  
 204 information can ease the requirement of pre-defined bias properties  
 205 between groups, which informs the de-biasing approach in this  
 206 study on geographic imagery.

### 207 2.1 Gaps in the Literature

209 Existing fairness work in image recognition is limited in multiple  
 210 ways, leaving gaps for adaptation to geographic imagery. First,  
 211 some methods require prior knowledge of sensitive attribute prop-  
 212 erties, such as skin colors for ethnicity groups, hair colors, presence  
 213 of glasses [44, 66, 71]. The analog of such a property is not avail-  
 214 able in satellite images, nor are such properties homogeneous (e.g.  
 215 each country has unique urban-rural landscape pattern changes).  
 216 Second, since the existing methods are mostly designed for classifi-  
 217 cation problems, they use image-level representation approaches.  
 218 However, fairness at an image level would not necessarily extend  
 219 to pixel-level dense predictions. Third, there is very little work on  
 220 fairness for geographic images, for which biased features are harder  
 221 to discover, interpret and remove, compared to facial images.

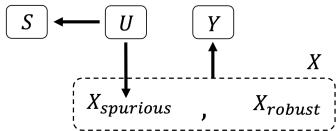
223 Beyond the contributions to fairness in geographic images, an-  
 224 other gap our work addresses is in fair contrastive learning. Fair-  
 225 ness has been less studied in contrastive approaches. Such work  
 226 is needed, as shown by Sirotnik *et al.* [52], self-supervised models  
 227 can incorporate implicit bias from data and potentially transfer  
 228 to downstream tasks. Despite the limitations of existing fair con-  
 229 trastive approaches, such as losing task-specific information which  
 230 might harm model accuracy, and requiring target task labels, all  
 231 work shows increased downstream fairness scores. Considering

limited research in this area, we are motivated to further increase representation fairness via contrastive learning.



**Figure 1: Examples of Segmentation bias due to spurious visual representation; the model segments certain road patterns well (blue circles), but segments the variations poorly (red circles).**

### 3 PROBLEM STATEMENT

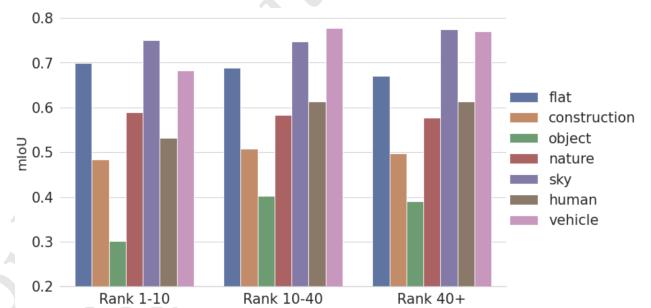


**Figure 2: Diagram of defined causal relationships between representation  $X$  learnt with contrastive pre-training, target task prediction outputs  $Y$ , and sensitive attribute  $S$ .  $X$  contains two parts,  $X_{spurious}$  generated from features spuriously correlated to sensitive attributes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

#### 3.1 Unfairness in satellite image recognition

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 learnt by model as  $X = \{X_{spurious}, X_{robust}\}$ , where  $X_{spurious}$  includes information that varies across sensitive 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 1 A), with vehicles on them (Figure 1 B), and with lane markings (Figure 1 C), are segmented better (blue circle) than the others (red circle, Figure 1 D). Examples of

more classes’ spurious and robust components are in the Appendix A. As a result, urban and rural embedding containing disproportionate spurious information levels will cause group-level model performance disparities in semantic segmentation. This problem is sketched out, in terms of causal relationships, in Figure 2. Different from previous fair methods which directly remove sensitive information, which is not available in the problem we study, we define the fairness goal to be to remove spurious information from representation correlated to sensitive features. That is, to obtain  $\hat{X}_{robust}$  which promotes  $\hat{Y} \perp S | \hat{X}_{robust}$ . The instinctive fairness benefit of learning  $X_{robust}$  over  $X_{spurious}$  is that it can relieve the need to expose a model to enough data to extract relevant target representation from non-generalizable feature variations, especially when the underrepresented group’s data is scarce.



**Figure 3: CityScapes semantic segmentation model unfairness: Class-level Intersection-over-Union (IoU) on the testing set differs significantly across the 3 city groups with different GDP attributes.**

### 4 METHODOLOGY

#### 4.1 Datasets and sensitive groups

While several standard image datasets used in fairness studies exist, datasets with linked sensitive attribute information for real-world geographic imagery are very limited. We identified three datasets which had or could be linked with attribute annotations for bias analyses. Next we describe sensitive group definitions for each dataset, based on location-based disparity context and attributes which could encode disproportionate spurious information, below.

*LoveDA* [63], *urban or rural designation as sensitive attribute*, 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 [63]. Moreover, it has been shown that model segmentation performances differ across urban and rural satellite images [74]. We split the original images into 512×512 pixel tiles, take 18% of the data for testing, 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 [51], urban or rural designation as sensitive attribute*, 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  $256 \times 256$  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<sup>2</sup> and administrative boundaries from OpenStreetMap<sup>3</sup>). 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.

*Cityscapes [8], GDP levels as sensitive attribute*, is an urban street view dataset with pixel-level annotations available for 30 classes. The train and validation split from the original dataset are merged, then 18% of the data (900 images) are randomly selected for testing, and for the rest, 90% (3690 images) are used for pre-training and 10% (410 images) for fine-tuning. The group unfairness categories are identified by performing supervised training and evaluate on the testing images, where segmentation accuracy differences between cities are observed. In particular, we find that metropolitan cities tend to have worse accuracy than small cities. Therefore, we split the 21 cities into 3 groups by GDP level<sup>4</sup>, denoted as Rank 1-10, Rank 10-40 and Rand 40+, since GDP is a comprehensive factor related population density, infrastructure, and others that can affect street views. Group-level results are shown in Figure 3.

There are several major disparities visible. For instance, Rank 1-10 cities have lower accuracy on object, human, and vehicle, which could be because these classes exist in higher proportion in the images (statistics in Figure 10, Appendix) and their positions commonly overlap, increasing segmentation difficulty. Slightly lower accuracy on the sky class for Rank 1-10 and Rank 10-40 cities may occur due to a higher prevalence of trees (nature class), traffic light and signs (object class) obstructing the sky and disturbing detection. In terms of overall accuracy, Rank 1-10 group has worst performance (IoU: 0.577) than the other two groups (0.613 for Rank 10-40 and 0.617 for Rank 40+).

## 4.2 Metrics

The quality of representations learnt from self-supervised pre-training is usually evaluated based on downstream task performance [23]. The principle behind this approach is using limited supervision and fine-tuning in assessment [15, 54, 72]. In line with this precedent, for comprehensively assessing representation quality, we report target task results every 1k iterations of fine-tuning for the pre-trained representation. On the downstream semantic segmentation task, we use Intersection-over-Union (IoU) as the accuracy metric, calculated using pixel-wise true positives ( $TP$ ), false positives ( $FP$ ), and false negatives ( $FN$ ),

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

<sup>2</sup><https://www.gov.si/en/topics/towns-and-protected-areas-in-slovenia/>

<sup>3</sup><https://www.openstreetmap.org/#map=12/40.7154/-74.1289>

<sup>4</sup>GDP level is measured by gross domestic product: [https://en.wikipedia.org/wiki/List\\_of\\_German\\_cities\\_by\\_GDP](https://en.wikipedia.org/wiki/List_of_German_cities_by_GDP)

Group accuracy for group  $g^i$  is computed via the mean of class-wise IoUs (referred to as  $\mu_{g^i}$ ).

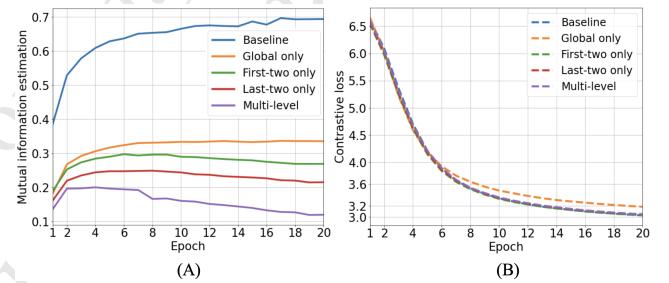
We use two fairness metrics, both motivated by previous work in algorithmic fairness. First, the group difference with regard to accuracy [14, 42, 55, 78] (Diff). Diff for a 2-element sensitive attribute group  $\{g^1, g^2\}$  is defined as:

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

And for  $K \geq 3$  element group  $\{g^1, \dots, g^K\}$ , we measure group distances from parity [13]:

$$\text{Diff } \{g^1, \dots, g^K\} := \frac{K}{K-1} \sum_{i=1}^K \left| \frac{\mu_{g^i}}{\mu_{g^1} + \dots + \mu_{g^K}} - \frac{1}{K} \right|.$$

The second metric is worst group results (Wst), motivated by the problem of worsening overall performance for zero disparity [73]. Additionally, we define a fairness-accuracy trade-off criteria at optimal fine-tuning. We indicate that there is no trade-off (No TO) when fairness increases with no decrease in accuracy as compared to the vanilla baseline.



**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.

## 4.3 Multi-level representation de-biasing

The idea of constraining mutual information between representation and sensitive attribute to achieve fair learning has multiple applications [25, 43, 77], which all operate on a global representation  $\mathbf{z} = F(\mathbf{d})$ , output from image encoder  $F$ . However, fairness 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 network (herein 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 [52]. Besides, layer-wise regularization is necessary to constrain the underlying representation space of CNN models [21, 22, 29]. Local features in representation hierarchies are important [37, 65], especially when transferring to dense downstream tasks such as

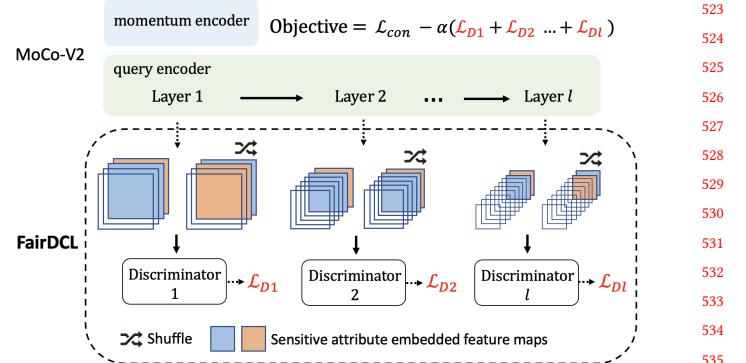
semantic segmentation, where representations are aggregated at different scales in order to let the decoder project predictions onto pixel space. Given the evidences in sum, we design a feature map based local mutual information estimation module and incorporate layer-wise fairness regularization into the contrastive optimization objective.

To measure mutual information  $MI(X, S)$  between local feature  $X$  and the sensitive attribute  $S = \{s_0, s_1, \dots\}$ , we adapt the concat-and-convolve architecture in [19]. Notating the  $i^{th}$  layer as  $l_i$ , we first build a one-hot encoding map  $c^{li}$  for sensitive attributes  $S$  whose size is same as the feature map  $x^{li}$  output by  $l_i$ , and channel is the size of  $S$ . For each  $x^{li}$ , a  $c^{li}$  is built from the joint distribution of representation space  $X$  and attribute space  $S$ , and the marginal distribution of  $S$  separately, then the  $c^{li}$  built in the two ways are concatenated with  $x^{li}$  to form an “aligned” feature map pair, denoted as  $P_{XS}(x^{li} \parallel c^{li})$ , and a “shuffled” feature map pair, denoted as  $P_X P_S(x^{li} \parallel c^{li})$ . The mutual information between the aligned and shuffled feature map pairs will be estimated by a three-layer  $1 \times 1$  convolutional discriminator  $D_i$ , using the JSD-derived formation [19]:

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

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 [7] with ResNet50 [17] 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 of ResNet50 and sensitive attributes: urban/rural.  $MI_{JSD}$  are summed to measure the total amount of model bias for the data batch. Contrastive training is conducted for 7k iterations (around 20 epochs) and the mean bias of iterations for each epoch is 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 effectively control bias accumulation. However, their measurements are still high compared to intervening on all four layers (Multi-level), showing that global level regularization might remove partial bias but leave significant residual from earlier layers. The influence of this problem on model downstream performance will be tested with the method UnbiasedR (Sec. 5.2.1 and Table 1). Additionally, the running contrastive loss during training is plotted in Figure 4 (B). From this plot it can be seen that all methods converge well; mutual information constraints in latent space do not affect the contrastive learning objective. The Global only method converges slower in later epochs than the other four methods, which may be because that the penalty factor is directly added on the target optimization function which affects the convergence. Both the de-biasing and convergence benefits motivate us to learn fair multi-level representations for different image encoder stages.



**Figure 5: Overview of FairDCL. It captures sensitive information and applies fairness regularization on image representation at multiple scales. We build one-hot feature maps to encode sensitive attribute and estimate mutual information by neural discriminators. Penalty loss  $\mathcal{L}_{Di}$  are computed accordingly and added into the final contrastive pre-training objective.**

---

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

---

for each iteration  $a$  from 1 to  $E$  do:

Image encoder forward propagation:

$x^{lN}, x^{lN-1}, \dots, x^{l1} \leftarrow F(x)$   $\triangleright x^{li}$  is the query representation output of the layer  $l_i$

Discriminators updating:

for each round  $b$  from 1 to  $B$  do:

for each discriminator  $D_i$  do:

$\mathcal{L}_{Di} \leftarrow D_i(x^{li} \parallel P_{XS}(c^{li}), x^{li} \parallel P_S(c^{li}))$

$\triangleright$  Forward aligned and shuffled feature pairs

$W_{Di} \leftarrow W_{Di} - \eta \nabla \mathcal{L}_{Di}$   $\triangleright$  Optimize  $D_i$

Image encoder updating:

$x^{lN}, x^{lN-1}, \dots, x^{l1}, q, k \leftarrow F(x)$   $\triangleright q, k$  is the query and key global representation

$\mathcal{L}_{con} \leftarrow q, k$   $\triangleright$  Compute contrastive loss

$\mathcal{L}_D \leftarrow \sum_{i=1}^N \mathcal{L}_{Di}$   $\triangleright$  Compute MI loss

$W_F \leftarrow W_F - \eta(\mathcal{L}_{con} - \alpha \mathcal{L}_D)$   $\triangleright$  Update encoders

---

#### 4.4 FairDCL pipeline

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

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

Following [43], each  $MI$  discriminator is optimized for multiple inner rounds before encoder weights get updated. More rounds are desirable for discriminators to estimate mutual information

523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580

with increased accuracy [43], and based on resource availability, we set a uniform round number  $B = 20$ . The architecture of  $D_i$  for each layer differs based on channel size of the concatenated feature map pair. 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  $\alpha$  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 sensitive information). We apply FairDCL on the state-of-the-art contrastive learning framework MoCo-v2 [7]. The contrastive loss used for learning visual representation is InfoNCE [38], defined as:

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

Here  $F$  consists of a query encoder and a key (or momentum) encoder, which outputs representations  $q$  and  $k$  from two augmented views of the same image, therefore  $k$  is viewed as “positive key”.  $\hat{k}_j$  is a queue of representations encoded from different images in the dataset, viewed as “negative keys” [7]. Further technical details of the framework can be found in [7].  $\mathcal{L}_{con}$  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 contrastive learning frameworks. Empirically, we experiment with the recently proposed DenseCL [65], 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.

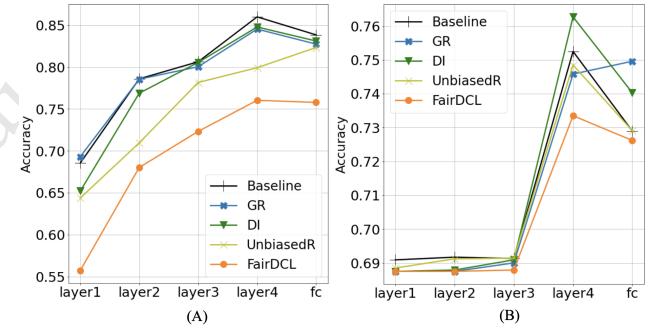
## 5 EXPERIMENTS

### 5.1 Implementation details

*The first stage of contrastive pre-training.* The base model for the image encoders is ResNet50 [17]. The mutual information discriminators  $D_i$  are built with  $1 \times 1$  convolution layers (architecture details in Appendix D). The contrastive pre-training runs for 10k iterations for each dataset with a batch size of 32. 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  $\alpha$ , 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: gradient reversal training (GR) by forcing encoders to generate representations that confuse

a sensitive attribute classifier [42], domain independent training (DI) which samples data from same sensitive group in each training iteration to avoid leveraging sensitive domain boundaries [58, 66], and unbiased representation learning (UnbiasedR) [43] which uses mutual information to de-bias but only in a global image space. All comparison methods use the same learning architectures, and are trained with the same settings.

*The second stage of semantic segmentation fine-tuning.* We use UNet [46] to perform landscape segmentation. The encoder weights of UNet are directly transferred from the pre-training stage. The model is fine-tuned for 5k iterations with a batch size of 16 (around 76 epochs on LoveDA dataset, 284 epochs on Slovenia dataset, and 195 epochs on Cityscapes dataset). Since there lacks fine-tuning benchmark on geographic datasets used in this work, we set the epoch numbers close to the fine-tuning settings in previous contrastive self-supervised learning studies [65, 75]. We evaluate metrics based on iterations rather than epochs to allow for direct computational comparison across datasets, following [27, 45]. Segmentation accuracy and fairness results are reported every 1k iterations. 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 encoder part of UNet is frozen during fine-tuning and no weight decay strategy is applied to avoid feature distortion [28]. Image data augmentations used in the fine-tuning include random horizontal/vertical flips and random rotations.



**Figure 6: Linear separation evaluation:** We train a linear neural layer on top of each representation level, which are feature maps output from different model layers. They include four residual module (“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.

## 5.2 Results

**5.2.1 Downstream performances.** Table 1 summarizes model fine-tuning results on semantic segmentation with the representations pre-trained with Baseline: vanilla MoCo-v2, and fairness-promoting methods: GR, DI, UnbiasedR, and FairDCL, on the three geographic

		Iteration=1k			Iteration=2k			Iteration=3k			Iteration=4k			Iteration=5k			
	Method	Diff(↓)	Wst (↑)	Acc(↑)	Diff(↓)	Wst(↑)	Acc(↑)	No TO									
697 698 699 700 701 702 703 704 705 706 707 708 709 710 711	LoveDA	Baseline	0.206	0.329	0.363	0.136	0.465	0.497	0.103	0.502	0.528	0.087	0.515	<b>0.537</b>	0.106	0.514	0.542
		GR	0.176	0.358	0.390	0.113	0.450	0.474	0.070	0.502	0.519	0.088	0.496	0.517	0.089	0.517	0.540
		DI	0.187	0.323	0.353	0.105	0.461	0.485	0.111	0.477	0.503	0.093	0.507	0.530	0.094	0.512	0.536
		UnbiasedR	0.234	0.351	0.391	0.125	0.449	0.476	0.096	0.506	0.531	0.114	0.495	0.523	0.103	0.520	<b>0.548</b>
		FairDCL	<b>0.161</b>	<b>0.362</b>	<b>0.392</b>	<b>0.098</b>	<b>0.479</b>	<b>0.502</b>	<b>0.069</b>	<b>0.513</b>	<b>0.532</b>	<b>0.080</b>	<b>0.517</b>	<b>0.537</b>	<b>0.084</b>	<b>0.525</b>	0.547
	Slovenia	Baseline	0.248	0.177	0.199	0.168	0.205	0.222	0.122	0.215	0.228	0.144	0.238	0.254	0.125	0.256	0.271
		GR	0.231	0.180	0.200	0.134	0.202	0.216	0.119	<b>0.231</b>	0.244	0.146	0.231	0.249	0.134	0.253	0.268
		DI	0.251	0.173	0.195	0.114	0.214	0.226	0.133	0.223	0.238	0.141	0.233	0.249	0.123	0.255	0.270
		UnbiasedR	0.230	0.183	<b>0.205</b>	0.215	0.197	0.218	0.141	0.212	0.227	0.136	0.235	<b>0.256</b>	0.122	0.249	0.264
		FairDCL	<b>0.226</b>	<b>0.184</b>	<b>0.205</b>	<b>0.109</b>	<b>0.217</b>	<b>0.228</b>	<b>0.117</b>	<b>0.231</b>	<b>0.245</b>	<b>0.122</b>	<b>0.241</b>	<b>0.256</b>	<b>0.0801</b>	<b>0.262</b>	<b>0.273</b>
	CityScapes	Baseline	0.0215	0.509	0.526	0.0253	0.530	<b>0.551</b>	0.0275	0.520	0.534	<b>0.0248</b>	0.541	0.560	0.0246	0.539	0.559
		GR	0.0313	0.486	0.510	0.0298	0.487	0.509	0.0270	0.498	0.518	0.0255	0.510	0.530	0.0252	0.511	0.531
		DI	0.0230	0.476	0.494	<b>0.0229</b>	0.506	0.524	0.0240	0.519	0.538	0.0251	0.523	0.543	0.0243	0.527	0.547
		UnbiasedR	0.0208	0.493	0.508	0.0250	0.520	0.540	0.0237	0.518	0.535	0.0250	0.516	0.537	0.0245	0.518	0.537
		FairDCL	<b>0.0206</b>	<b>0.525</b>	<b>0.541</b>	0.0238	<b>0.536</b>	0.546	<b>0.0236</b>	<b>0.537</b>	<b>0.557</b>	<b>0.0248</b>	<b>0.545</b>	<b>0.566</b>	<b>0.0241</b>	<b>0.545</b>	<b>0.564</b>

**Table 1: Downstream semantic segmentation results on LoveDA, Slovenia, CityScapes dataset for 5k fine-tuning iterations.** The encoder weights are learnt with 5 comparison pre-training methods. Our FairDCL shows clear improvements on fairness metrics (Diff and Wst) and accuracy metric (Acc) over baseline and prior methods throughout the training, also we do not see a fairness-accuracy trade-off (No TO) on all datasets. Results are the mean over 5 independent runs, and the standard deviations are provided in the Appendix.

image datasets: LoveDA, Slovenia, and CityScapes. Segmentation accuracy and fairness metrics are reported every 1k iterations and we include results from 1k to 5k iterations to compare representation quality in term of its influence on the whole fine-tuning procedure (extended iterations are shown in the ablation study). “No TO” indicates if a method waives the fairness-accuracy trade-off problem, that in its best fine-tuning round, whether the improved “Diff” metric does *not* cause a worse “Acc” metric. Such trade-off robustness with respect to the other fairness metric “Wst” is depicted in Figure 13 in the Appendix E.

We first note that across fine-tuning iterations, FairDCL nearly always outperforms other approaches in terms of fairness. FairDCL obtains the smallest cross-group difference (Diff) and highest worst group result (Wst). The results indicate that the representation trained with multi-scale fairness constraints can lead to higher group parity meanwhile maximizing performance of the worst case groups (urban or rural places in LoveDA and Slovenia, or cities with different GDP levels in CityScapes). Results also show the stability of the learnt representation, that the downstream training shifts, such as fine-tuning the decoder for shorter or longer rounds, do not break model performance advantages.

Importantly, FairDCL is the only method that does not show a fairness-accuracy trade-off on all three datasets. FairDCL in general obtains comparable or better overall accuracy to Baseline, demonstrating robust model quality in addition to fairness. In contrast, GR and DI show lower overall accuracy, especially on the CityScapes dataset. This observation resonates with the intrinsic trade-off problem pointed out by previous fairness studies [20, 73, 73, 78]; DI allows building image contrastive pairs only from a fraction of data which can discount model learning [66], while the adversarial approach used in GR can be counter-productive if the adversary is not trained enough to achieve the infimum [36], which could all potentially degrade model quality for group equalization. Our adapted

mutual information discriminators use information-theoretic objectives, which are shown to be optimized without competing with the encoder so can match or exceed state-of-the-art adversarial debiasing methods [36, 43]. FairDCL further illustrates that applying the mutual information constraints on latent representations of multiple resolutions can better extend fairness to pixel-level applications. In contrast, UnbiasedR, the comparison method which applies constraints at the single image level, shows no trade-off on LoveDA dataset, and has worse results on both fairness metrics than FairDCL (Table 1).

Using the alternative self-supervised contrastive learning framework DenseCL, similar advantages can be observed. FairDCL pre-trained feature encoder for the downstream segmentation prediction obtains fairness improvement with no trade-off of task precision (Table 11 in the Appendix E).

**5.2.2 Embedding spaces.** To further trace how the image representations learnt with proposed method improves fairness, we analyze a linear separation property [45] on embedding spaces. Specifically, we assess how well a linear model can differentiate sensitive attributes using learnt representations. High separation degree indicates that the encoder model’s embedding space and sensitive attribute are differentiable [6, 38, 45], 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 sensitive attribute label classification. Figure 6 (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 geographic images, we check the method's generalizability to a different image domain by conducting contrastive pre-training on MS-COCO [30], a dataset commonly used in fairness studies [56, 62, 64]. Sensitive attribute gender, categorized as "women" or "men", is obtained from [76]. 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 6 (B)), but unlike in Figure 6 (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.

Method	Iteration=6k			Iteration=7k			Iteration=8k			No TO
	Diff(↓)	Wst (↑)	Acc(↑)	Diff	Wst	Acc	Diff	Wst	Acc	
Baseline	0.106	0.533	0.566	0.109	0.529	0.558	0.097	0.536	0.562	
GR	0.092	0.538	0.563	0.093	0.535	0.559	0.094	0.538	0.560	✗
DI	<b>0.087</b>	0.537	0.561	0.090	0.533	0.557	0.084	0.530	0.552	✗
UnbiasedR	0.109	0.538	<b>0.567</b>	0.114	<b>0.542</b>	<b>0.570</b>	0.123	0.534	0.567	✗
FairDCL	0.089	<b>0.543</b>	<b>0.567</b>	0.82	<b>0.542</b>	0.564	0.73	<b>0.549</b>	<b>0.569</b>	✓

**Table 2: Ablation study for downstream training shifts for 5k-8k iterations on LoveDA dataset. When training iterations extend, FairDCL method maintains the advantage in fairness improvements (Diff and Wst) over Baseline without a trade-off of mean accuracy (Acc).**

$\alpha = 0.1$	$\alpha = 0.5$			$\alpha = 1$			$\alpha = 10$				
	Diff(↓)	Wst (↑)	Acc(↑)	Diff	Wst	Acc	Diff	Wst	Acc		
0.096	0.519	0.544	0.084	<b>0.525</b>	<b>0.547</b>	0.075	0.521	0.540	<b>0.069</b>	0.521	0.539

**Table 3: Ablation study for discriminator weights. The best fine-tuning result is shown for the encoder pre-trained with different weight,  $\alpha = 0.1, 0.5, 1, 10$ , of the proposed fairness objective.**

Method	Urban:68% Rural:32%			Urban:35% Rural:65%			No TO	
	Diff(↓)	Wst (↑)	Acc(↑)	No TO	Diff(↓)	Wst (↑)	Acc(↑)	
Baseline	0.214	0.337	0.373		0.253	0.341	0.384	
GR	0.209	0.329	0.364	✗	0.218	0.350	0.387	✓
DI	0.203	0.332	0.364	✗	0.206	0.330	0.364	✗
UnbiasedR	0.166	0.363	0.393	✓	0.229	0.338	0.377	✗
FairDCL	<b>0.123</b>	<b>0.394</b>	<b>0.418</b>	✓	<b>0.198</b>	<b>0.352</b>	<b>0.388</b>	✓

**Table 4: Ablation study for unbalanced data distribution. The proportion of sensitive groups in the pre-training data is adjusted such that one group has much less representation. FairDCL performs consistently with data balance changes.**

**5.2.3 Ablation studies.** We run longer training on the labeled downstream data to validate fine-tuning stability of the pre-trained image encoders. Shown in Table 2, in general, all methods converge and reach performance plateau after 8k iterations, so we only show results until that point, and the converged performance should be bounded by the capacity of the decoder component. The results give further evidence that the benefits of the de-biased representations pre-trained with FairDCL do not vanish in longer training regimes.

We then perform an ablation study for hyper-parameter  $\alpha$  which scales discriminator loss  $\mathcal{L}_D$ , thus the fairness regularization strength. As shown in Table 3, the method is overall robust to the parameter; a large weight like 10 will not completely corrupt the downstream accuracy. Using smaller  $\alpha$  have higher Wst, while larger  $\alpha$  have higher Diff, and we select  $\alpha = 0.5$  for a balance.

The sensitive attribute, urban and rural, have comparable training samples in earlier experiments (LoveDA is 5.8k and 5.5k, EOlearn Slovenia is 1.7k and 1.9k for urban/rural, but Cityscapes is highly unbalanced: 1.8k, 0.78k, 1.1k for the three GDP groups). Therefore, we intentionally remove a part of pre-training samples for certain groups to generate more unbalanced subsets. With performances shown in Table 4, the proposed method shows robustness under the two less even attribute distributions.

## 6 DISCUSSION AND CONCLUSION

Among the broader fairness literature in visual recognition, work focusing on geographic imagery that depicts physical environments has been limited. This limitation is largely due to the difficulty in identifying population level biased landscape features. Also, fairness problems in geographic image recognition may get categorized as domain adaptation or transfer learning problems, other popular computer vision fields. Though fairness enhancement and domain adaptation share similar technical methods bias mitigation and invariant feature learning, the specific objective of fair geographic image recognition is to remove spatially disproportionate features that favor one subgroup over the others. Though fairness enhancement and domain adaptation can share similar technical approaches such as invariant feature learning, the specific objective of fair geographic image recognition is distinct and goes beyond addressing covariate shift. The goal is to remove spatially disproportionate spurious features that favor one subgroup over the others.

Here we identify and address unique fairness challenges in semantic segmentation of satellite images and street view images. We theoretically define the scenario with a causal graph, showing that contrastive self-supervised pre-training can utilize spurious land-cover object features, thus accumulate sensitive attribute-correlated bias. The biased image representation 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 good local features with minimal spurious representation. Experimental results show fairer segmentation results pre-trained with the proposed method on multiple geographic datasets and different sensitive groups. In addition to performance gain, the method consistently avoids a trade-off between model fairness and accuracy. Finally, we would like to note that the fairness problem and fair image representation learning

method studied in this work are not exclusive to contrastive self-supervised learning, but are also applicable to other supervised learning settings, such as supervised semantic segmentation, supervised object detection, and others, via incorporation of the fairness regularization term to the target task training objective.

As future directions, first, more geographic domains and data can be explored. The fairness analysis can be scaled to a greater number of attributes. Second, sensitive attribute encoding for model latent space besides one-hot feature map can be explored. We encourage experimenting with different encoding mechanisms and mutual information estimators to improve performance across different real-world settings.

## REFERENCES

- [1] Chitra Agastya, Sirak Ghebremusse, Ian Anderson, Hossein Vahabi, and Alberto Todeschini. 2021. Self-supervised contrastive learning for irrigation detection in satellite imagery. *arXiv preprint arXiv:2108.05484* (2021).
- [2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10181–10190.
- [3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.
- [4] Nikolaos Ioannis Bountous, Ioannis Papoutsis, Dimitrios Michail, and Nantheera Anantrasirichai. 2021. Self-supervised contrastive learning for volcanic unrest detection. *IEEE Geoscience and Remote Sensing Letters* 19 (2021), 1–5.
- [5] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems* 33 (2020), 12546–12558.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.
- [9] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Tonianne Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR, 1436–1445.
- [10] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- [11] Ranjith Dinakaran, Li Zhang, Chang-Tsun Li, Ahmed Bouridane, and Richard Jiang. 2022. Robust and Fair Undersea Target Detection with Automated Underwater Vehicles for Biodiversity Data Collection. *Remote Sensing* 14, 15 (2022), 3680.
- [12] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. 2021. How well do self-supervised models transfer?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5414–5423.
- [13] Rina Friedberg, Stuart Ambler, and Guillaume Saint-Jacques. 2022. Representation-Aware Experimentation: Group Inequality Analysis for A/B Testing and Alerting. *arXiv preprint arXiv:2204.12011* (2022).
- [14] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2021. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3414–3424.
- [15] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. 2019. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*. 6391–6400.
- [16] Jerry A Griffith. 2002. Geographic techniques and recent applications of remote sensing to landscape-water quality studies. *Water, Air, and Soil Pollution* 138 (2002), 181–197.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [18] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems* 32 (2019).
- [19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).
- [20] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.
- [21] Zhulun Jiang, Yaming Wang, Larry Davis, Walter Andrews, and Viktor Rozicic. 2017. Learning discriminative features via label consistent neural network. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 207–216.
- [22] Xiaojie Jin, Yunpeng Chen, Jian Dong, Jiashi Feng, and Shuicheng Yan. 2016. Collaborative layer-wise discriminative learning in deep neural networks. In *European Conference on Computer Vision*. Springer, 733–749.
- [23] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4037–4058.
- [24] Sangwon Jung, Donggyu Lee, Taeon Park, and Taesup Moon. 2021. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12115–12124.
- [25] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 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.
- [26] Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 111, 9 (2014), 3354–3359.
- [27] Simon Kornblith, Jonathon Shlens, and Quoc V Le. 2019. Do better imagenet models transfer better?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2661–2671.
- [28] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054* (2022).
- [29] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. 2019. Learning from brains how to regularize machines. *Advances in neural information processing systems* 32 (2019).
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [31] Subhabrata Majumdar, Cheryl Flynn, and Ritwik Mitra. 2022. Detecting Bias in the Presence of Spatial Autocorrelation. In *Algorithmic Fairness through the Lens of Causality and Robustness workshop*. PMLR, 6–18.
- [32] Narenih Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [33] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence* 3, 8 (2021), 659–666.
- [34] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6707–6717.
- [35] Aytthami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosa. 2020. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 6 (2020), 2158–2164.
- [36] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems* 31 (2018).
- [37] Pedro O O Pinheiro, Amjad Almahairi, Ryan Bennalek, Florian Golemo, and Aaron C Courville. 2020. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems* 33 (2020), 4489–4500.
- [38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [39] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2403–2411.
- [40] Sungho Park, Jiwook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair Contrastive Learning for Facial Attribute Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10389–10398.
- [41] Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steven E Petersen, Reza Razavi, and Andrew P King. 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*. Springer, 413–423.
- [42] Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Big Data*. 1044

- 1045      *on Data Science and Advanced Analytics (DSAA)*. IEEE, 189–198.
- 1046 [43] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. 2021. Learning unbiased representations via mutual information backpropagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2729–2738.
- 1047 [44] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9301–9310.
- 1048 [45] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, 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.
- 1049 [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- 1050 [47] Sudipan Saha, Lichao Mou, Chuning Qiu, Xiao Xiang Zhu, Francesca Bovolo, and Lorenzo Bruzzone. 2020. Unsupervised deep joint segmentation of multitemporal high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing* 58, 12 (2020), 8780–8792.
- 1051 [48] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. 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.
- 1052 [49] Ignacio Serna, Ahythami Morales, Julian Fierrez, and Nick Obradovich. 2022. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence* 305 (2022), 103682.
- 1053 [50] MAS Setianto and A Gamal. 2021. Spatial justice in the distribution of public services. In *IOP Conference Series: Earth and Environmental Science*, Vol. 673. IOP Publishing, 012024.
- 1054 [51] Sinergise. 2022. Modified Copernicus Sentinel data 2017/Sentinel Hub. <https://sentinel-hub.com/>.
- 1055 [52] Kirill Sirokin, Pablo Carballeira, and Marcos Escudero-Viñolo. 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.
- 1056 [53] Robert Soden, Dennis Wagenaar, Dave Luo, and Annegien Tijssen. 2019. Taking ethics, fairness, and bias seriously in machine learning for disaster risk management. *arXiv preprint arXiv:1912.05538* (2019).
- 1057 [54] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. 2020. When does self-supervision improve few-shot learning?. In *European conference on computer vision*. Springer, 645–666.
- 1058 [55] Attila Szabó, Hadi Jamali-Rad, and Siva-Datta Mannava. 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.
- 1059 [56] Ruixiang Tang, Mengnan Du, Yueming Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*. 633–645.
- 1060 [57] Huan Tian, Tianqiang Zhu, Wei Liu, and Wanlei Zhou. 2022. Image fairness in deep learning: problems, models, and challenges. *Neural Computing and Applications* (2022), 1–19.
- 1061 [58] Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv preprint arXiv:2106.02866* (2021).
- 1062 [59] Wouter Van Gansbeke, Simon Vandenhende, Stamatis Georgoulis, and Luc V Gool. 2021. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems* 34 (2021), 16238–16250.
- 1063 [60] JRC Von Holdt, FD Eckardt, MC Baddock, and Giles FS Wiggs. 2019. Assessing landscape dust emission potential using combined ground-based measurements and remote sensing data. *Journal of Geophysical Research: Earth Surface* 124, 5 (2019), 1080–1098.
- 1064 [61] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. 2021. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*. PMLR, 755–769.
- 1065 [62] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433* (2021).
- 1066 [63] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733* (2021).
- 1067 [64] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 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.
- 1068 [65] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3024–3033.
- 1069 [66] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 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.
- 1070 [67] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.
- 1071 [68] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. 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.
- 1072 [69] Yuwen Xiong, Mengye Ren, and Raquel Urtasun. 2020. Loco: Local contrastive representation learning. *Advances in neural information processing systems* 33 (2020), 11142–11153.
- 1073 [70] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. 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.
- 1074 [71] Haolin Yuan, Armin Hadzic, William Paul, Daniella Villegas de Flores, Philip Matthew, John Autcott, Yinzhi Cao, and Philippe Burlina. 2022. EdgeMixup: Improving Fairness for Skin Disease Classification and Segmentation. *arXiv preprint arXiv:2202.13883* (2022).
- 1075 [72] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867* (2019).
- 1076 [73] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfahl, and Marzyeh Ghassemi. 2022. Improving the Fairness of Chest X-ray Classifiers. In *Conference on Health, Inference, and Learning*. PMLR, 204–233.
- 1077 [74] Miao Zhang, Harvineet Singh, Lazarus Chok, and Rumi Chunara. 2022. 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.
- 1078 [75] Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. 2021. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Advances in Neural Information Processing Systems* 34 (2021), 29848–29860.
- 1079 [76] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14830–14840.
- 1080 [77] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. 2021. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15002–15012.
- 1081 [78] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. 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.

## A EXAMPLES OF UNFAIRNESS IN LAND-COVER OBJECT SEMANTIC SEGMENTATION

Besides segmentation disparities for different “road” variations shown in Figure 1, here we provide examples for “building”, “water”, and “forest” classes, and their corresponding spurious and robust features in Figure 7, Figure 8, and Figure 9. In each caption, based on simple visual assessment, we provide examples of possible aspects of the images that could contribute to spurious or robust features.

## B CITYSCAPES: CITY - SENSITIVE GROUP MAPPING

The group assignment of cities in CityScape dataset is shown in Table 5.

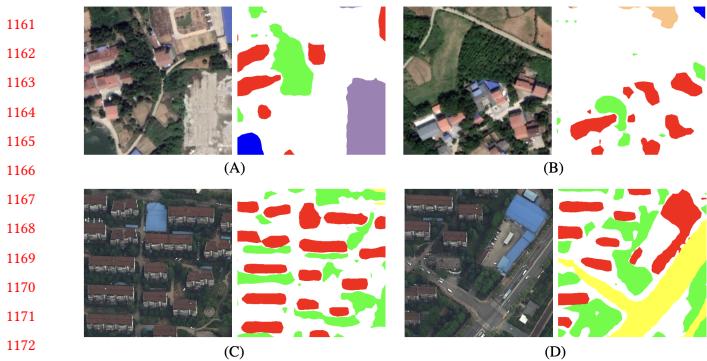


Figure 7: Segmentation bias for “building” class. Compared to the scattered rural buildings in (A) and (B), urban buildings that have uniform shapes and neat arrangements are better segmented (C) and (D), and especially have their edges detected more precisely. Spurious features of building objects could include the shapes, e.g. rectangular or irregular shapes, and/or the colors. Robust aspects of buildings can include the shape of their clustering and/or height in relation to surroundings.

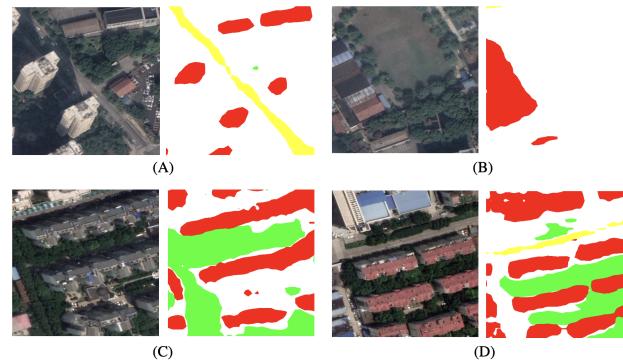


Figure 9: Segmentation bias for “forest” class. Forests/trees that are in open areas are largely neglected by the model in (A) and (B), while in contrast, trees in between building blocks are well detected as in (C) and (D). The development of urban greening promotes growing trees along streets or buildings, which could be a spurious feature for model to use, that the feature does not generalize to trees grown in open fields, a common case rural area. A more robust feature of trees should be texture of tree crowns that is generalizable and distinguishable.

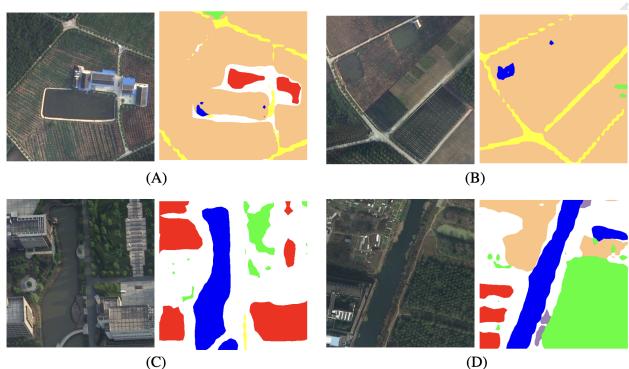


Figure 8: Segmentation bias for “water” class. The model has difficulty distinguish water from agriculture land, especially when the water area is small as in (A) and (B). In comparison, water that is surrounded by more different landscapes are much better segmented as in (C) and (D). Spurious features of water object could include the shape and color: rural areas have more small water ponds, and the water colors can vary potentially based on depth of the water source. Robust features could include the texture of the water surface which is generally more smooth compared to other land-covers.

## C CLASS DISTRIBUTION SHIFTS ACROSS GROUPS

Class distributions also vary across sensitive groups; we plot class distributions in terms of proportion of pixels per class by group for each of the three datasets, in Figure 10, Figure 11, and Figure 12.

Table 5: The list of cities for each sensitive group in CityScapes dataset

Sensitive group	Cities
GDP level Rank 1-10	Hamburg, Cologne, Dusseldorf, Hanover, Stuttgart, Frankfurt, Bremen
GDP level Rank 10-40	Aachen, Ulm, Munster, Tubingen, Bochum, Krefeld, Zurich
GDP level Rank 40+	Darmstadt, Jena, Monchengladbach, Strasbourg, Weimar, Erfurt, Lindau

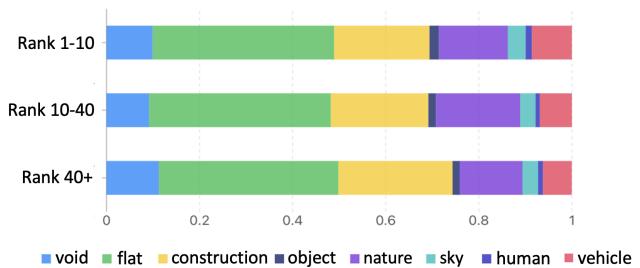
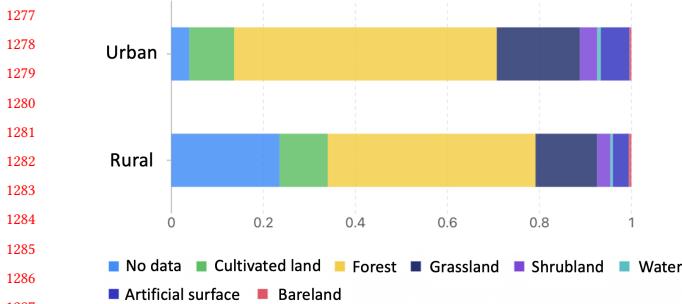
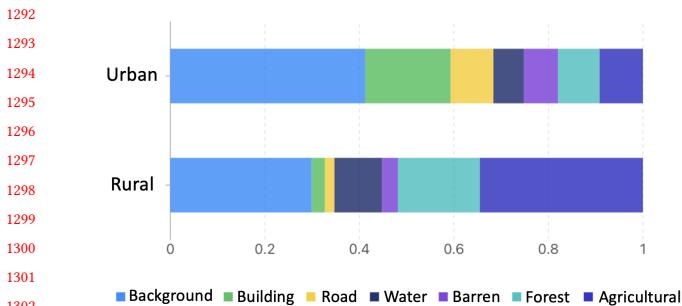


Figure 10: CityScapes dataset class distribution shifts: Class distribution for images from the three groups: Rank 1-10, Rank 10-40, and Rank 40+ of the CityScapes dataset.



**Figure 11: Slovenia dataset class distribution shifts:** Class distribution for images from the two groups: urban, rural of the Slovenia dataset.



**Figure 12: LoveDA dataset class distribution shifts:** Class distribution for images from the two groups: urban, rural of the LoveDA dataset.

## D MUTUAL INFORMATION DISCRIMINATORS

The base model used as the encoder in our experiments is ResNet50. Multi-stage mutual information discriminators are built for representation outputs of four layers (residual blocks). In Table 6 to Table 9 we provide the architecture dimensions for each discriminator and the input and output feature map channel sizes  $C$ , noting that feature map height  $H$  and width  $W$  vary across datasets, depending on the size of images the model is trained on.

Layer	Input	Output
1×1 Conv2D (258, 258)	258	258
1×1 Conv2D (258, 20)	258	20
1×1 Conv2D (20, 1)	20	1

**Table 6: Discriminator for layer1 of ResNet50 encoder.**

Layer	Input	Output
1×1 Conv2D (514, 514)	514	514
1×1 Conv2D (514, 50)	514	50
1×1 Conv2D (50, 1)	50	1

**Table 7: Discriminator for layer2 of ResNet50 encoder.**

Layer	Input	Output
1×1 Conv2D (1026, 1026)	1026	1026
1×1 Conv2D (1026, 100)	1026	100
1×1 Conv2D (100, 1)	100	1

**Table 8: Discriminator for layer3 of ResNet50 encoder.**

Layer	Input	Output
1×1 Conv2D (2050, 2050)	2050	2050
1×1 Conv2D (2050, 200)	2050	200
1×1 Conv2D (200, 1)	200	1

**Table 9: Discriminator for layer4 of ResNet50 encoder.**

The outputs of each discriminator are a  $H \times W$  matrix for the “shuffled” feature map pair and the “aligned” feature map pair respectively. Details of building feature map pairs are described in Section 4.3 and visualized in Figure 5. As the next step, the discriminator outputs are used to compute the mutual information estimation  $M_{JSD}$ <sup>5</sup> and the discriminator loss  $\mathcal{L}_{D_i}$  listed in Section 4.3.

## E SUPPLEMENTARY EXPERIMENT RESULTS

In Table 10, we provide standard deviation results of the 5 independent runs for the results shown in Table 1.

In Table 11, we generalize FairDCL and the comparison fair representation learning methods to a different contrastive pre-training framework DenseCL [65]. The same metrics are reported as with MoCo-v2 framework.

In Figure 13, different to how we have been defining the fairness-accuracy trade-off with the fairness metric: the group difference, which is more commonly used in literature to quantify disparity, we further depict the trade-off with regard to another fairness metric: the worst-case performer.

We use pareto curve following [78], which assumes that computer vision models tend to have a capacity threshold, that a higher maximum group accuracy can lead to a lower minimum group accuracy, and vice versa, with the premise that the fairness intervention does not affect model’s efficiency. Under this assumption, the FairDCL method shows higher minimum group accuracy (better fairness) with model efficiency least affected (Figure 13).

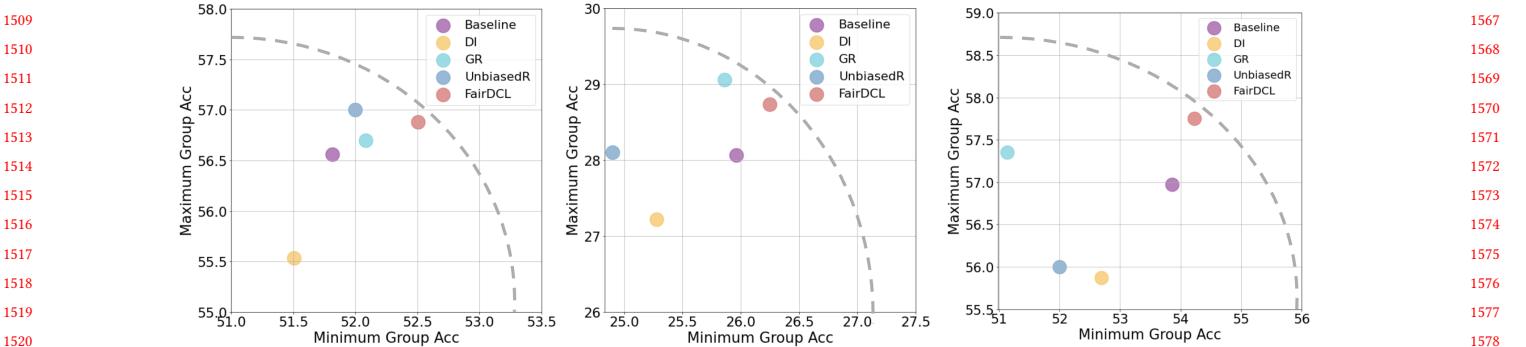
<sup>5</sup>Codes for the loss function are from <https://github.com/rdevon/DIM>

	Method	Iteration=1k			Iteration=2k			Iteration=3k			Iteration=4k			Iteration=5k			
		Diff	Wst	Acc													
1393	Baseline	0.11	0.040	0.056	0.047	0.036	0.044	0.026	0.025	0.033	0.014	0.025	0.035	0.033	0.020	0.031	
1394	GR	0.11	0.051	<b>0.053</b>	0.069	0.040	0.045	0.027	<b>0.021</b>	0.027	0.025	0.040	0.044	0.031	<b>0.011</b>	<b>0.025</b>	
1395	DI	0.057	0.076	0.094	0.029	<b>0.021</b>	0.034	0.064	0.037	0.039	0.023	0.033	0.048	0.025	0.017	0.038	
1396	LoveDA	UnbiasedR	0.078	0.095	0.066	<b>0.022</b>	0.022	0.039	0.052	0.049	0.030	0.019	0.039	0.035	<b>0.024</b>	0.037	0.028
1397		FairDCL	<b>0.054</b>	<b>0.034</b>	0.059	0.031	0.023	<b>0.034</b>	<b>0.012</b>	0.023	<b>0.029</b>	<b>0.006</b>	<b>0.013</b>	<b>0.031</b>	0.029	0.027	0.035
1398		Baseline	0.065	0.015	0.026	0.062	0.012	0.022	0.093	0.020	0.023	0.065	<b>0.010</b>	0.019	<b>0.051</b>	0.007	<b>0.015</b>
1399	GR	0.041	0.012	0.024	0.069	0.022	0.029	0.073	0.021	0.026	0.062	0.023	0.034	0.077	0.006	0.024	
1400	Slovenia	DI	<b>0.011</b>	0.019	0.024	<b>0.059</b>	<b>0.006</b>	0.019	0.015	0.012	<b>0.018</b>	0.079	0.011	0.022	0.074	0.013	0.023
1401		UnbiasedR	0.021	0.033	0.045	0.062	0.106	0.029	0.017	0.051	0.040	0.035	0.018	0.022	0.064	0.033	0.019
1402		FairDCL	0.020	<b>0.008</b>	<b>0.021</b>	<b>0.057</b>	0.009	0.021	<b>0.007</b>	<b>0.010</b>	0.022	<b>0.029</b>	0.012	<b>0.020</b>	0.054	<b>0.004</b>	0.020
1403	Baseline	0.009	0.032	0.032	0.002	<b>0.014</b>	<b>0.020</b>	0.005	0.040	0.047	0.005	0.014	0.024	0.002	<b>0.005</b>	0.016	
1404	GR	0.002	<b>0.009</b>	<b>0.016</b>	0.005	0.027	0.030	0.008	0.045	0.046	0.003	<b>0.011</b>	<b>0.018</b>	0.004	0.007	0.019	
1405	CityScapes	DI	<b>0.002</b>	0.013	0.020	0.003	0.046	0.049	0.005	0.028	0.030	0.010	0.045	0.042	0.004	0.007	<b>0.015</b>
1406		UnbiasedR	0.009	0.019	0.030	0.003	0.035	0.051	<b>0.004</b>	0.037	0.032	0.008	0.025	0.026	<b>0.002</b>	0.019	0.018
1407		FairDCL	0.006	0.020	0.027	<b>0.002</b>	0.027	0.029	0.006	<b>0.010</b>	<b>0.020</b>	<b>0.002</b>	0.016	0.021	0.004	0.011	0.019

**Table 10: Standard deviation of results in Table 1. The smallest deviation results are marked in bold between the comparison methods for each iteration. It shows that FairDCL method produces comparable performance consistency from run to run with the baseline methods.**

	Method	Iteration=1k			Iteration=2k			Iteration=3k			Iteration=4k			Iteration=5k		
		Diff( $\downarrow$ )	Wst ( $\uparrow$ )	Acc( $\uparrow$ )	Diff( $\downarrow$ )	Wst( $\uparrow$ )	Acc( $\uparrow$ )	Diff( $\downarrow$ )	Wst( $\uparrow$ )	Acc( $\uparrow$ )	Diff( $\downarrow$ )	Wst( $\uparrow$ )	Acc( $\uparrow$ )	Diff( $\downarrow$ )	Wst( $\uparrow$ )	Acc( $\uparrow$ )
1414	Baseline	0.204	0.308	0.339	0.089	0.358	0.374	0.118	0.388	0.411	0.116	0.399	0.422	0.103	0.383	0.403
1415	GR	0.170	0.326	0.354	0.104	0.365	0.384	0.114	0.369	0.390	0.110	<b>0.406</b>	0.428	0.092	0.388	0.406
1416	DI	0.187	0.308	0.321	<b>0.048</b>	0.347	0.351	0.089	0.369	0.386	0.097	0.388	0.406	0.088	0.366	0.383
1417	UnbiasedR	0.169	0.299	0.324	0.116	0.380	<b>0.402</b>	<b>0.087</b>	0.386	0.403	0.095	0.396	0.415	0.90	0.386	0.403
1418	FairDCL	<b>0.168</b>	<b>0.336</b>	<b>0.365</b>	0.076	<b>0.385</b>	0.399	0.108	<b>0.394</b>	<b>0.415</b>	<b>0.091</b>	<b>0.406</b>	<b>0.43</b>	<b>0.083</b>	<b>0.391</b>	<b>0.408</b>
1419	Baseline	0.102	0.401	0.421	0.085	0.421	0.329	0.077	0.435	0.452	0.087	0.434	<b>0.453</b>	0.106	0.424	0.446
1420	GR	0.094	0.402	0.421	0.081	0.420	0.437	0.083	0.441	0.459	0.099	0.431	0.452	0.106	0.425	0.448
1421	DI	0.087	0.389	0.405	0.082	0.414	0.424	0.087	0.420	0.436	0.085	0.422	0.435	0.108	0.420	0.441
1422	UnbiasedR	0.086	0.401	0.419	0.092	0.416	0.435	0.098	0.408	0.428	0.092	0.431	0.451	<b>0.104</b>	0.424	0.445
1423	FairDCL	<b>0.084</b>	<b>0.405</b>	<b>0.422</b>	<b>0.080</b>	<b>0.423</b>	<b>0.440</b>	<b>0.074</b>	<b>0.445</b>	<b>0.461</b>	<b>0.083</b>	<b>0.435</b>	<b>0.453</b>	0.104	0.428	0.452
1424																

**Table 11: Testing the generalizability to the different contrastive pre-training framework DenseCL. We run 10k fine-tuning iterations on the downstream semantic segmentation on LoveDA dataset. Similarly, the frozen encoder weights are learnt with different fairness method: Baseline (no intervention), GR, DI, UnbiasedR, and our FairDCL. For every 1k steps, the best results in term of fairness: group difference (Diff) and worst-case group result (Wst), and accuracy: mean IoU (Acc) are marked in bold. The fairness-accuracy trade-off indicator (No TO) shows whether a method, in its best performing round, obtains both better Diff and Acc than the Baseline ( $\checkmark$ ) or fails on either metric ( $\times$ ). FairDCL continuously shows supreme fairness performance on this alternative contrastive learning pipeline, and avoids the undesired trade-off throughout the fine-tuning.**



**Figure 13: Worst case group - Accuracy trade-off depiction:** For each colored dot, a closer distance to the grey dotted curve (maximum model capacity) indicates a higher model efficiency in term of overall accuracy. A larger x-axis value indicates the model is fairer and has a higher minimum group accuracy. The plots are for results on LoveDA (left), Slovenia (middle), and CityScapes (right) dataset. FairDCL shows better fairness performances on the worst case group's aspect meanwhile a higher and stable model efficiency.

1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624