Multispectral Contrastive Learning with Viewmaker Networks

Jasmine Bayrooti, Noah Goodman, Alex Tamkin
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
{jbayrooti, ngoodman, atamkin}@stanford.edu

Abstract

Contrastive learning methods have been applied to a range of domains and modalities by training models to identify similar "views" of data points. However, specialized scientific modalities pose a challenge for this paradigm, as identifying good views for each scientific instrument is complex and time-intensive. In this paper, we focus on applying contrastive learning approaches to a variety of remote sensing datasets. We show that Viewmaker networks, a recently proposed method for generating views without extensive domain knowledge, can produce useful views in this setting. We also present a Viewmaker variant called Divmaker, which achieves similar performance and does not require adversarial optimization. Applying both methods to four multispectral imaging problems, each with a different format, we find that Viewmaker and Divmaker can outperform cropping- and reflection-based methods for contrastive learning in every case when evaluated on downstream classification tasks. This provides additional evidence that domain-agnostic methods can empower contrastive learning to scale to real-world scientific domains. Open source code can be found at https://github.com/jbayrooti/divmaker.

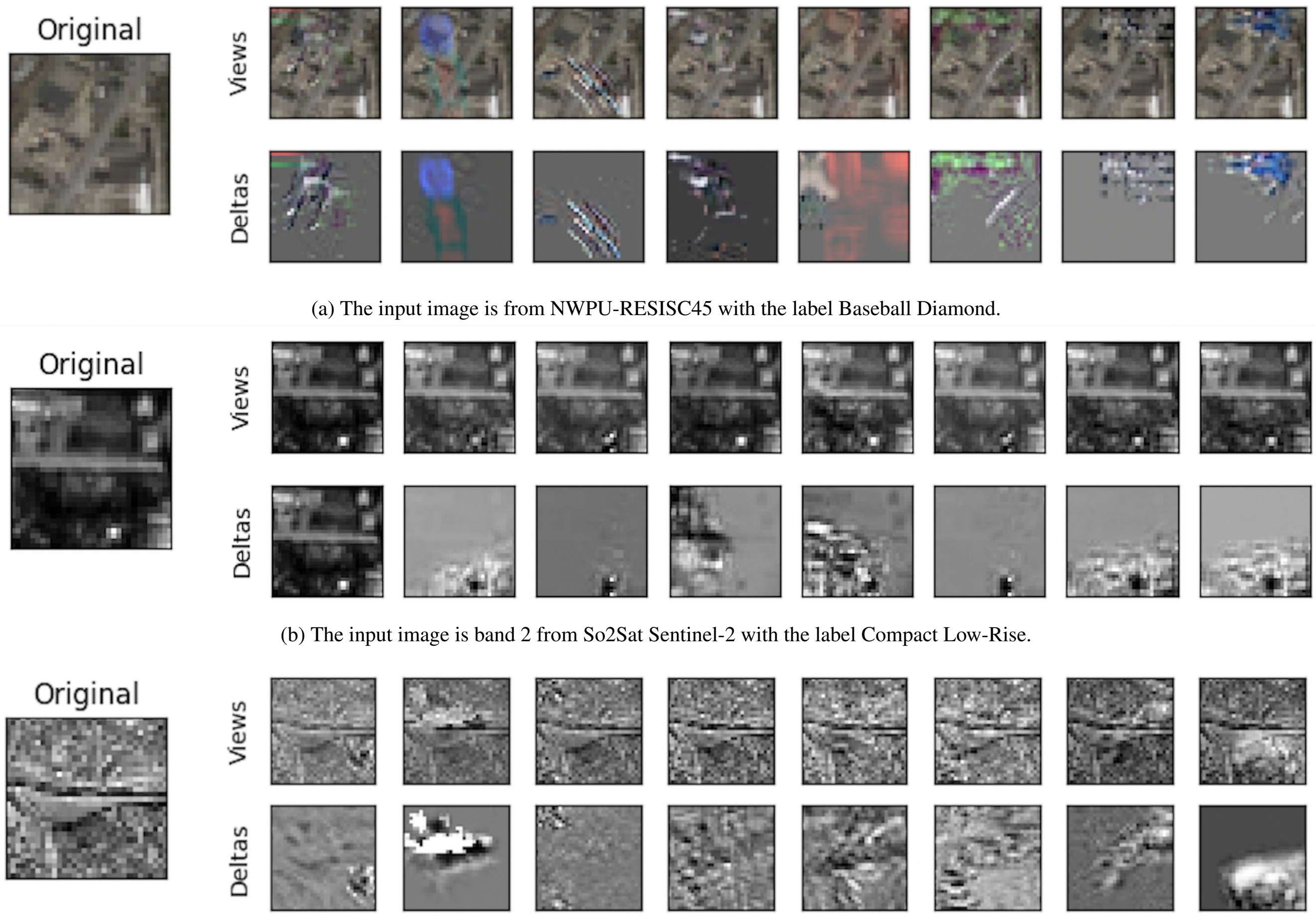
1. Introduction

Contrastive learning methods have demonstrated remarkable ability to learn high-quality representations without relying on labels, often achieving equivalent or higher classification accuracy than supervised approaches after pretraining on large unlabeled image datasets [7, 33, 45]. These advances suggest the utility of contrastive learning in settings beyond natural images, including many impactful applications in the sciences, engineering, medicine, and beyond. However, a key barrier to mainstreaming contrastive learning applications is choosing the appropriate "views": the data corruptions that determine the contrastive learning process [7, 37]. It is challenging to develop effective views for new applications, as this process requires domain knowledge and trial and error for each setting.

Aerial satellites measure terabytes worth of multispectral image data in various forms every day. They use remote sensors to measure distinct wavelengths of light and stack outputs into n-channel images, which are used to analyze light beyond the human-visible spectrum. Such images can offer insights into natural phenomena like temperature variation that RGB data cannot. Robust and accurate self-supervised learning techniques for satellite images could aid advances in agricultural growth efficiency, understanding of climate change, tracking urban development, and environmental monitoring [15, 18, 24–26]. In this paper, we investigate contrastive learning with multispectral satellite images.

The dominant views for contrastive learning on natural images were identified via extensive trial and error, and involve applying augmentations such as color jitter and horizontal flipping to RGB images [38, 44]. However, these RGB views do not transfer well to multispectral images since each channel has different numeral ranges and semantics, making it impossible to directly apply such transformations. Domain-agnostic generative Viewmaker networks [35] propose to *learn* such data transformations with a generative model trained adversarially with an encoder. However, Viewmaker networks have not yet been applied to a broader range of scientific data.

To address this gap, we evaluate Viewmaker networks on multispectral satellite image datasets. Furthermore, we introduce a generative network called Divmaker, which produces views optimized for diversity and does not require adversarial optimization. While Divmaker performs slightly worse than Viewmaker, our results confirm that both domain-agnostic view generation methods enable higher quality contrastive learning than with reasonable, hand-designed augmentations. We demonstrate this on three different large-scale multispectral satellite datasets and compare with an RGB satellite dataset for additional context.



(a) The input image is from NWPU-RESISC45 with the label Baseball Diamond.

(b) The input image is band 2 from So2Sat Sentinel-2 with the label Compact Low-Rise.

(c) The input image is band 1 from BigEarthNet with the multi-label: Mixed Forest, Transitional Woodland/Shrub, Non-Irrigated Arable Land, Broad-Leaved Forest.

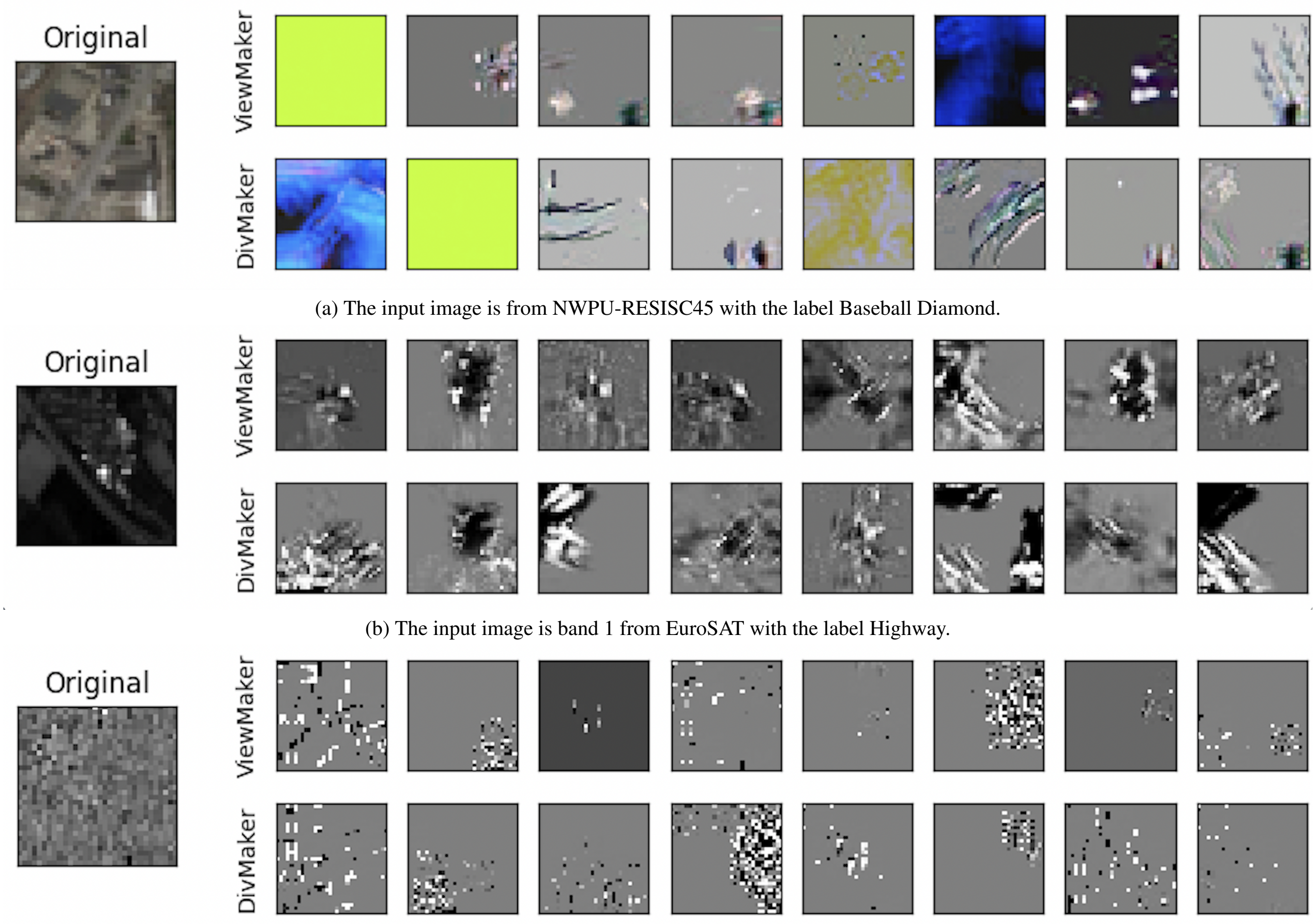
Figure 1. **Learned perturbations (bottom right) appear varied in shape, intensity, and placement to augment semantic features in the original input in targeted ways.** The input image is on the left and resulting views are given on the top right. Perturbations (deltas) were generated using Viewmaker and then linearly scaled to the full image range for clear visualization with corresponding views.

2. Related Work

Contrastive learning. Self-supervised learning enables learning representations from large, unlabeled datasets which can be used for downstream tasks like classification, object detection, semantic segmentation, or visual navigation. Contrastive learning methods have been shown to produce good representations by identifying transformed positive "views" of the same inputs [6–8, 14, 42, 45]. Such approaches rely on good choices of data augmentations [38, 44] that yield representations discriminative with respect to the downstream tasks and yet general enough to be applied to new tasks. Such augmentations may not always be known a priori. This problem is especially pertinent in less common domains and modalities like multi-spectral satellite images, 3D images, tabular data, and voice recordings.

Learning from satellite images. Deep learning research on satellite imagery contends with a variety of factors including vast unlabeled datasets, spatial-temporal heterogeneity within classes, cloud interference, and texture and color discontinuities between image tiles [36]. Learning methods have been productively applied to tasks such as poverty mapping [13], local climate zone classification [47], water temperature prediction [39], food safety analysis [27], enhancing agricultural yield [4], everyday scene classification [9, 17, 29], and sustainable development monitoring [46]. In this paper, we investigate classification of land use, local climate zones, and everyday scenes.

Contrastive learning for satellite images. Multispectral satellite images often contain heterogeneous backgrounds with significant structural information and varying resolutions depending on the remote sensor's settings



(a) The input image is from NWPU-RESISC45 with the label Baseball Diamond.

(b) The input image is band 1 from EuroSAT with the label Highway.

(c) The input image is band 2 from So2Sat Sentinel-1 with the label Dense Trees.

Figure 2. **Divmaker manages to generate diverse and effective views without adversarial optimization.** We see this as perturbations from the Divmaker (bottom right) appear to act in similar ways to perturbations from the Viewmaker (upper right). The original input image is given on the left and perturbations (deltas) have been linearly scaled to the full image range for clear visualization.

[3, 10]. Since these characteristics differ from those of standard RGB images (i.e. in ImageNet or CIFAR-10), traditional contrastive learning augmentations do not transfer well to multispectral images. Features like geographical distance between patches have been shown to produce useful views for datasets [19], although this requires having access to the coordinate location of each image or very-high resolution allowing the images to be divided into smaller patches during training [3]. Another work splits input images into two views based on their channels, passes them as inputs to two different encoder networks, and uses the embeddings themselves as positive views for contrastive learning [1]. Since each remote sensing application has different bands, the splitting process would need to be customized. There have been additional methods proposed like sharpness transformation and random erasure [16, 30], however many of these depend on dataset-specific properties (i.e. high resolution and number of bands) or have varying effectiveness across datasets. Due to the diversity of multispectral image formats, it is less feasible to exhaustively search for the most effective augmentation strategy as in [7], which isolated the best augmentation pairs on RGB images. Thus, we focus on general out-of-the-box view-generating methods that need little customization.

Domain-agnostic machine learning. Domain-specific self-supervised learning algorithms have enabled significant gains in fields such as natural language processing [11, 12], computer vision [7, 8], and speech processing [2]. However, many other domains with rich, unlabeled datasets could also benefit from self-supervised approaches. Recent work responds to this interest in advancing domain-agnostic self-supervised learning with new benchmarks [32, 34] and learning algorithms [23, 41]. In this work, we apply and build on Viewmaker networks and demonstrate that such domain-agnostic learning approaches can be fruitfully applied in a range of different remote sensing applications.

3. Methods

In this section, we discuss Viewmaker and introduce a new variant called Divmaker, which optimizes for diverse view generation.

3.1. Viewmaker

The Viewmaker [35] is a generative network trained to produce augmented images, or views, that are useful for contrastive learning. The Viewmaker V takes an input image X and gives a perturbation $V(X)$, which is added to the input to obtain the view $X+V(X)$. Adversarial training encourages perturbations to be complex and strong enough to necessitate encoding useful representations. Perturbations are constrained to an l_1 sphere (with size controlled by a distortion budget hyperparameter) around the input to maintain faithfulness to the original features. Lastly, the network injects random noise so perturbations differ from each other. Viewmaker-learned views have seen recent success, outperforming baseline augmentations on speech recordings and wearable sensor data and attaining comparable downstream accuracy on natural images. Note that the Viewmaker's adversarial training setup with the encoder requires a fully differentiable objective.

3.2. Diversity Viewmaker (Divmaker)

We introduce the Divmaker, a domain-agnostic view generation approach that does not require adversarial training and optimizes for diverse rather than challenging views. Like the Viewmaker, Divmaker offers a way to generate new views and uses the same contrastive loss.

Before formalizing the Divmaker view-generation loss, we introduce the following notation. Consider a temperature parameter τ, an anchor input $x \in \mathcal{D}$ with embedding z, and associated views x_i with embeddings z_i for $1 \leq i \leq K$. We use the cosine similarity measure on embeddings:

$$\text{sim}(z, z') = \frac{z^T z'}{||z||||z'||} \quad (1)$$

And define:

$$h(z, z') = \exp\left(\frac{\text{sim}(z, z')}{\tau}\right) \quad (2)$$

Then the Divmaker loss is:

$$\mathcal{L} = \mathbf{E}_{x \sim \mathcal{D}}\left(-\sum_{k=1}^{K} \log \frac{h(z_k, z)}{h(z_k, z) + \sum_{l \neq k} h(z_k, z_l)}\right) \quad (3)$$

The intuition is to optimize diversity by maximizing cosine similarity between the original input and its generated views while minimizing similarity between two views ($K = 2$) for the same input. This is in contrast to the Viewmaker, which tries to create challenging views via an adversarial loss without explicitly optimizing for the diversity of the views. This diversity objective was previously used for self-supervised anomaly detection [28], where the views learned were a finite set of learned masks. Instead, we generate dynamic and input-conditioned views with a stochastic neural network, as Viewmaker does. The Divmaker loss is also closely related to the Triplet loss [43], which has been widely used in other applications and shares parallels with standard contrastive learning losses [20].

Like the Viewmaker, the Divmaker network outputs a bounded perturbation, which is added to the input to produce a view that can be used for contrastive learning. The strength of Divmaker-generated views is controlled by the distortion budget, which specifies the magnitude of Divmaker's perturbation output as done in the Viewmaker. Training with diverse views could help capture a wider range of augmentations encountered in practice and enable the encoder to learn useful representations earlier in training. Furthermore, by separating the Divmaker and encoder objectives, we eliminate the differentiable restriction on the encoder, allowing Divmaker to work with state-of-the-art non-differentiable contrastive learning methods [5, 6]. Finally, while we did not experience training instabilities with Viewmaker, Divmaker's avoidance of adversarial training may enable more stable training for larger models [21].

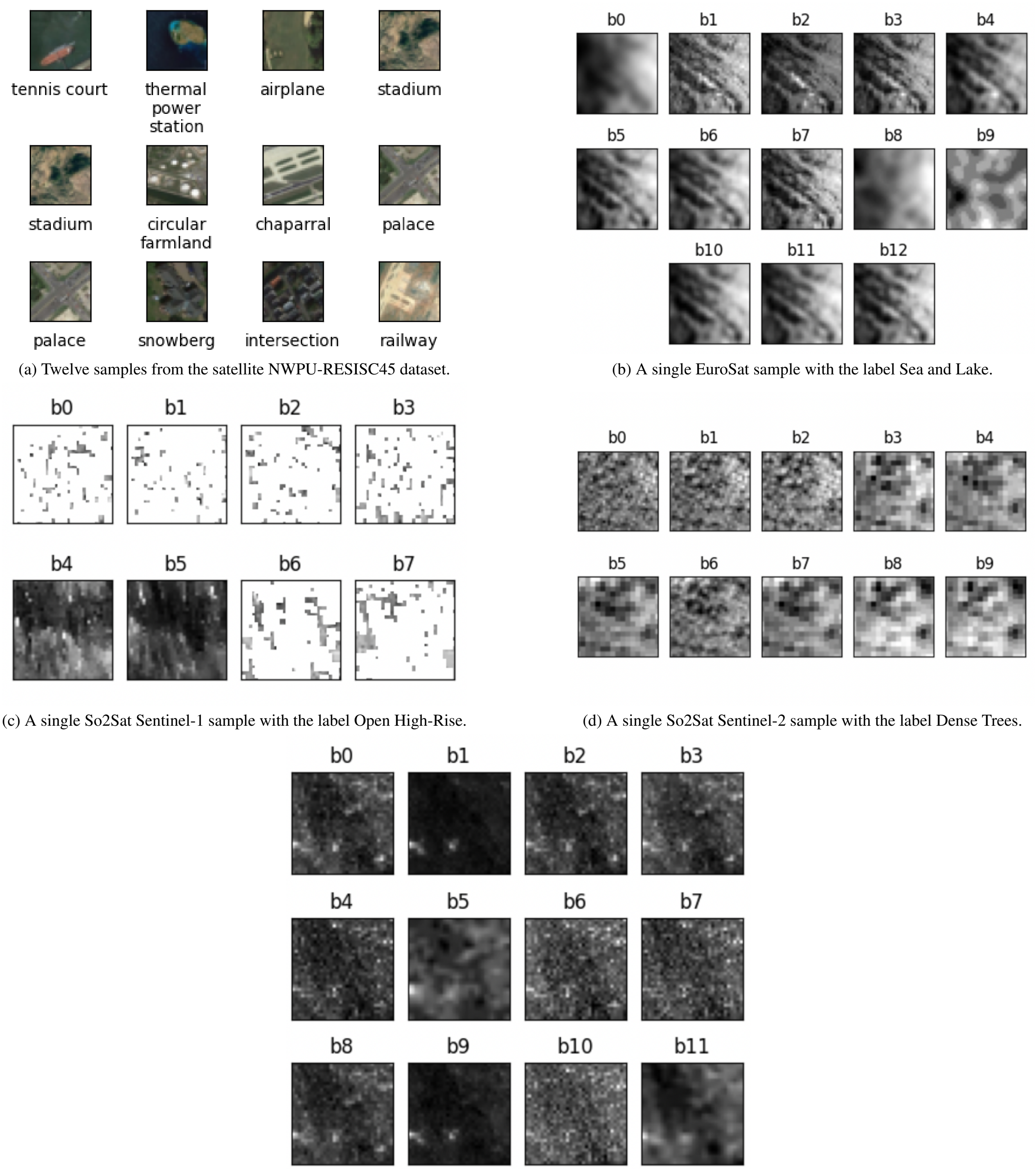
3.3. Datasets

We apply Viewmaker and Divmaker to four large-scale satellite datasets. Examples from each dataset are given in Figure 3. Note that each band is shown separately for multispectral images.

EuroSAT is a dataset of 27,000 images consisting of low cloud-cover satellite Sentinel-2 images from 34 European countries. Images are labeled with one of 10 classes describing land use such as Sea and Lake, Industrial, and Pasture with 2,000 to 3,000 images per class. Each image includes 13 spectral bands in the visible, near infrared, and short wave infrared part of the light spectrum [17].

So2Sat LCZ42 Sentinel-1 and Sentinel-2 is a benchmark dataset composed of 400,673 low cloud-cover images collected by Sentinel-1 and Sentinel-2 satellites over 42 large cities and 10 smaller areas spanning six continents. Images are labeled with one of 17 classes in the Local Climate Zones (LCZ) classification scheme, which are based on climate-relevant surface properties such as structure, surface cover, and anthropogenic parameters. Some examples include Open High-Rise, Dense Trees, Heavy Industry, and Water [47]. Sentinel-1 data contains 8 real-valued bands Sentinel-2 data contains 10 real-valued bands.

BigEarthNet is a multi-label dataset made up of 590,326 Sentinel-2 image patches collected from 10 European coun-



(a) Twelve samples from the satellite NWPU-RESISC45 dataset.

(b) A single EuroSat sample with the label Sea and Lake.

(c) A single So2Sat Sentinel-1 sample with the label Open High-Rise.

(d) A single So2Sat Sentinel-2 sample with the label Dense Trees.

(e) A single BigEarthNet sample with the multi-label: Sea and Ocean, Water Bodies.

Figure 3. **Multispectral satellite images have very different characteristics from RGB images, and thus require different strategies for creating views.** We display randomly selected images from each dataset considered. All channels are shown for multispectral images with band number prefaced with ”b”.

Dataset	Metric	Basic Expert	Expert	Viewmaker	Divmaker
NWPU-RESISC45	Accuracy	N/A	**76.07**	67.61	71.72
EuroSAT	Accuracy	90.93	90.68	**96.4**	95.67
So2Sat Sentinel-1	Accuracy	31.12	29.33	**38.25**	36.39
So2Sat Sentinel-2	Accuracy	51.6	51.54	**60.08**	59.67
BigEarthNet	F1 Score	4.03	4.21	**12.99**	10.83

Table 1. **Domain-agnostic methods outperform domain-specific approaches on satellite datasets.** We measure performance as linear classification accuracy over pre-trained representations except on BigEarthNet, for which we use F1 score of multi-classification accuracy. For multispectral datasets, random cropping makes up the basic expert augmentations with horizontal flipping added for expert views. For NWPU-RESISC45, we use standard expert RGB augmentations. Results are averaged over four seeds with tuned distortion budgets.

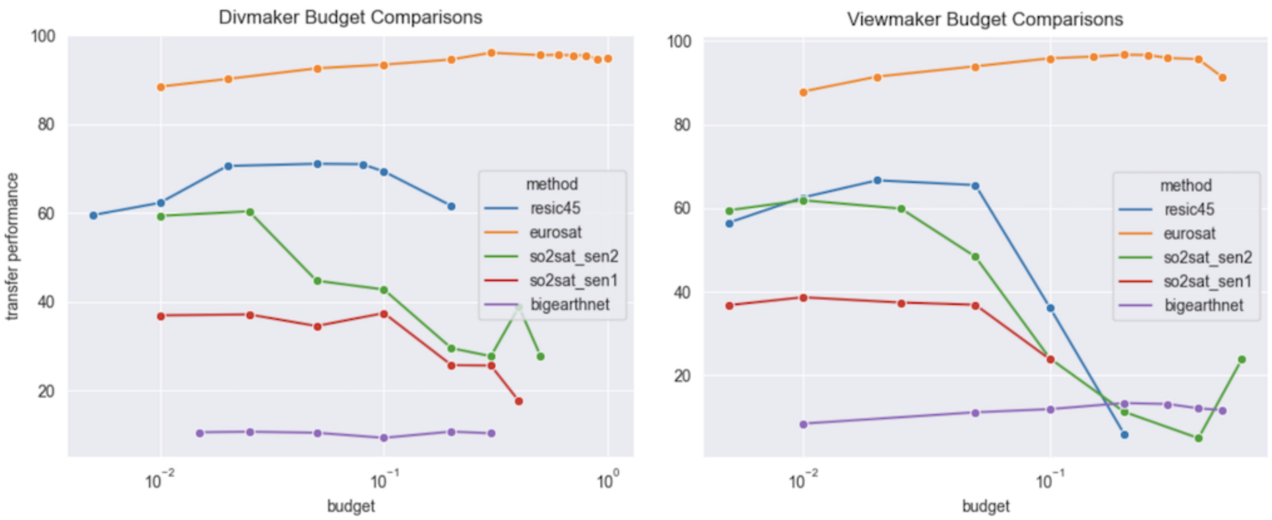


Figure 4. Downstream task performance can change drastically with the distortion budget, but a wide range of settings enable good performance.

tries. Each image is labeled with a subset of 43 land-cover classes such as Pastures, Water Bodies, Agro-Forestry Areas, and Green Urban Areas [31]. Note that, in the BigEarthNet dataset, two bands measure 20 × 20 images, six bands measure 60 × 60 images, and four bands measure 120 × 120 images. To standardize the full size, we resize all bands to 120 × 120 resolution and then stack the bands.

NWPU-RESISC45 is a dataset of 31,500 RGB satellite images sourced from Google Earth. Images are labeled with one of 45 classes such as Tennis Court, Thermal Power Station, Airplane, Stadium, Circular Farmland, Chaparral, Palace, Snowberg, Intersection, and Railway. The dataset includes 700 samples for each scene class with large variations in translation, spatial resolution, viewpoint, object pose, illumination, background, and occlusion. This dataset is challenging due to large variance within-class and high inter-class similarities [9]. Since the NWPU-RESISC45 dataset consists of RGB satellite images, we use it as a control to compare Viewmaker performance when we have more domain knowledge and better expert views.

4. Experiments

In this section, we explore whether Viewmaker and Divmaker can outperform domain-specific methods by learning to generate appropriate views on four well-known satellite datasets.

4.1. Experimental Details

We first train the encoder and view-generating networks simultaneously by pretraining on a dataset. Then we evaluate the quality of the learned representations using the widely used linear transfer protocol [7, 22, 40]. We evaluate on the RGB NWPU-RESISC45 dataset as a baseline to confirm that well-researched expert views from the RGB image domain can give greater performance gains [7]. We use random cropping as the basic expert transformation and combine this with horizontal flipping for expert views for the other datasets. While there exist more complex augmentation methods [16, 19, 30], we choose these because they are generalizable to any large-scale satellite dataset and hence useful benchmarks in the domain-agnostic setting.

For pretraining and downstream task training, we use a learning rate of 0.005 and the same encoder and viewmaker architectures, temperature parameter, batch size, and other parameters as [35]. We find normalizing multispectral images before passing through the Viewmaker and Divmaker networks to be useful for all datasets except NWPU-RESISC45, which has low standard deviation across channels before normalization. We also clamp all pixels in generated views symmetrically between -1 and 1. For the Divmaker loss, we experimented with $K = 2$ and $K = 3$, finding that $K = 3$ gave minimal performance gains for greater computational cost. Hence, we used $K = 2$ in all reported experiments throughout the paper. All implementation details are available in our open source code at https://github.com/jbayrooti/divmaker.

4.2. Results and Interpretation

In Table 1, we report the best performance for every method on each dataset. For the NWPU-RESISC45 dataset, learning from expert RGB transformations results in the highest linear classification accuracy over learned Viewmakers. This is not surprising, considering the abundance of research into these RGB transformations [7, 38]. For the multispectral datasets, we find much stronger gains from Viewmaker and Divmaker methods, compared to the handcrafted augmentations: horizontal flipping and cropping. This demonstrates the utility of domain-agnostic methods like Viewmaker and Divmaker when working with less popular data forms.

Divmaker performs better than Viewmaker on the RGB dataset and nearly as well on the multispectral datasets, indicating that a broader range of objectives for generative views can enable success on multispectral data. We provide illustrations of views and perturbations on NWPU-RESISC45 in Figure 1 and a comparison of perturbations produced by Viewmaker and Divmaker in Figure 2. Related work demonstrates that, in some cases, the Viewmaker network can identify and alter semantic features in an input to aid learning [33]. Although it is hard to pinpoint exact interpretations of the perturbations, our work corroborates this as perturbations appear correlated across channels and sometimes with simple image features.

We also compare downstream classification performance for Viewmaker and Divmaker with different distortion budgets in Figure 4. This demonstrates that, while the budget should be tuned for optimal performance, a wide range of budgets allow for good performance. These results also confirm that Divmaker can achieve comparable performance with Viewmaker for slightly higher budgets. This is expected since Divmaker does not use adversarial training.

5. Conclusion

In this paper, we examine whether domain-agnostic approaches to contrastive learning can scale to an important scientific domain: multispectral satellite images. Our experiments show that domain-agnostic methods can outperform existing domain-specific contrastive learning methods using out-of-the-box baseline views. This insight is important considering the utility of multispectral satellite images, and suggests that domain-agnostic self-supervised methods may enjoy success across a wider array of scientific applications. Additionally, we demonstrate another successful strategy for domain-agnostic view learning with the Divmaker, which avoids adversarial optimization. Future directions of research include comparing against more sophisticated domain-specific view-generation methods, analyzing further downstream differences between the Viewmaker and Divmaker, and considering additional multispectral applications like water temperature prediction [39] and sustainable development monitoring [46].

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10161–10170, 2021. 3

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020. 3

[3] Paul Berg, Minh-Tan Pham, and Nicolas Courty. Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives. *Remote Sensing*, 14(16), 2022. 3

[4] Sebastian Candiago, Fabio Remondino, Michaela De Giglio, Marco Dubbini, and Mario Gattelli. Evaluating multispectral images and vegetation indices for precision farming applications from uav images. *Remote Sensing*, 7(4):4026–4047, 2015. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020. 4

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2, 4

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings*

of Machine Learning Research, pages 1597–1607. PMLR, 13–18 Jul 2020. 1, 2, 3, 6, 7

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[9] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2, 6

[10] Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020. 3

[11] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[13] Bradley J Gram-Hansen, Patrick Helber, Indhu Varatharajan, Faiza Azam, Alejandro Coca-Castro, Veronika Kopackova, and Piotr Bilinski. Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 361–368, 2019. 2

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020. 2

[15] M Hanif, B G Putra, K Nizam, H Rahman, and A Y Nofrizal. Multi spectral satellite data to investigate land expansion and related micro climate change as threats to the environment. *IOP Conference Series: Earth and Environmental Science*, 303(1):012030, jul 2019. 1

[16] Xuejie Hao, Lu Liu, Rongjin Yang, Lizeyan Yin, Le Zhang, and Xiuhong Li. A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, 15(3), 2023. 3, 6

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2, 4

[18] J. Holmgren, Å. Persson, and U. Söderman. Species identification of individual trees by combining high resolution lidar data with multi-spectral images. *International Journal of Remote Sensing*, 29(5):1537–1552, 2008. 1

[19] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David B. Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 3, 6

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 4

[21] Naveen Kodali, Jacob D. Abernethy, James Hays, and Zsolt Kira. How to train your DRAGAN. *arXiv preprint arXiv:1705.07215*, 2017. 4

[22] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 6

[23] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. *arXiv preprint arXiv:2010.08887*, 2020. 3

[24] H. Noh and Q. Zhang. Shadow effect on multi-spectral image for detection of nitrogen deficiency in corn. *Computers and Electronics in Agriculture*, 83:52–57, 2012. 1

[25] Stuart Phinn, Chris Roelfsema, Arnold Dekker, Vittoro Brando, and Janet Anstee. Mapping seagrass species, cover and biomass in shallow waters: An assessment of satellite multi-spectral and airborne hyper-spectral imaging systems in moreton bay (australia). *Remote Sensing of Environment*, 112(8):3413–3425, 2008. Earth Observations for Marine and Coastal Biodiversity and Ecosystems Special Issue. 1

[26] Moacir Ponti, Arthur A. Chaves, Fábio R. Jorge, Gabriel B.P. Costa, Adimara Colturato, and Kalinka R.L.J.C. Branco. Precision agriculture: Using low-cost systems to acquire low-altitude images. *IEEE Computer Graphics and Applications*, 36(4):14–20, 2016. 1

[27] Jianwei Qin, Kuanglin Chao, Moon S. Kim, Renfu Lu, and Thomas F. Burks. Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, 118(2):157–171, 2013. 2

[28] Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In *International Conference on Machine Learning*, pages 8703–8714. PMLR, 2021. 4

[29] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-2/W7:153–160, 2019. 2

[30] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021. 3, 6

[31] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 6

[32] Alex Tamkin, Gaurab Banerjee, Mohamed Owda, Vincent Liu, Shashank Rammoorthy, and Noah Goodman. DABS 2.0: Improved datasets and algorithms for universal self-supervision. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[33] Alex Tamkin, Margalit Glasgow, Xiluo He, and Noah Goodman. Feature dropout: Revisiting the role of augmentations in contrastive learning. *arXiv preprint arXiv:2212.08378*, 2022. 1, 7

[34] Alex Tamkin, Vincent Liu, Rongfei Lu, Daniel Fein, Colin Schultz, and Noah Goodman. DABS: a domain-agnostic benchmark for self-supervised learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 3

[35] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning. In *ICLR*, 2021. 1, 4, 7

[36] Chao Tao, Ji Qi, Mingning Guo, Qing Zhu, and Haifeng Li. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. *arXiv preprint arXiv:2211.08129*, 2022. 2

[37] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *CoRR*, abs/2005.10243, 2020. 1

[38] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. 1, 2, 7

[39] C. Tornow, C.C. Borel, and B.J. Powers. Robust water temperature retrieval using multi-spectral and multi-angular ir measurements. In *Proceedings of IGARSS '94 - 1994 IEEE International Geoscience and Remote Sensing Symposium*, volume 1, pages 441–443 vol.1, 1994. 2, 7

[40] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[41] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541. PMLR, 2021. 3

[42] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[43] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. 4

[44] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah D. Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020. 1, 2

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2

[46] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track (Round 2)*, 12 2021. 2, 7

[47] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, Lloyd H. Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat LCZ42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019. 2, 4