Data Augmentation Approaches for Satellite Imagery

Laurel M. Hopkins[1], **Weng-Keen Wong**[1], **Hannah Kerner**[2], **Fuxin Li**[1], **Rebecca A. Hutchinson**[1]

[1]Oregon State University, Corvallis, OR 97331
[2]Arizona State University, Tempe, AZ 85281
{hopkilau,wongwe}@oregonstate.edu, hkerner@asu.edu, {lif,rah}@oregonstate.edu

Abstract

Deep learning models commonly benefit from data augmentation techniques to diversify the set of training images. When working with satellite imagery, it is common for practitioners to apply a limited set of transformations developed for natural images (e.g., flip and rotate) to expand the training set without overly modifying the satellite images. There are many techniques for natural image data augmentation, but given the differences between the two domains, it is not clear whether data augmentation methods developed for natural images are well suited for satellite imagery. This paper presents an extensive experimental study on three classification and three regression tasks over four satellite image datasets. We compare common computer vision data augmentation techniques and propose three novel satellite-specific data augmentation strategies. Across tasks and datasets, we find that geometric transformations are beneficial for satellite imagery while color transformations generally are not. Additionally, our novel Sat-SlideMix, Sat-CutMix, and Sat-Trivial methods all exhibit strong performance across all tasks and datasets.

Introduction

Machine learning allows us to harness satellite data for a variety of applications with societal impact. For example, convolutional neural networks can analyze greenhouse gas emissions detected in satellite images and detect where large emissions are occurring (Rolnick et al. 2022). Machine learning models trained on satellite images in data-rich areas can be transferred to data-scarce areas to inform socioeconomic indicators, such as poverty (Jean et al. 2016). Neural networks have the capacity to model complex interactions between species and their environments which can be used to monitor biodiversity at large scales (Chen et al. 2016; Davis et al. 2023). Population predictions can be produced more frequently than censuses (Robinson, Hohman, and Dilkina 2017). Predicting water quality from satellite imagery has enabled higher-resolution (both spatial and temporal) mapping and monitoring compared to traditional methods (Peterson, Sagan, and Sloan 2020). Crop type maps produced from satellite images in areas where labeled data is scarce can aid in humanitarian efforts (You et al. 2017; Kerner et al. 2020).

One limitation to more widespread modeling of satellite imagery is the lack of annotated labels. While satellite data archives greatly outweigh the amount of data in "large" machine learning datasets, satellite data paired with ground-truth labels is scarce (Rolf et al. 2024) and collecting more data is costly in time and/or financially (Zhu et al. 2018). Methods designed for small datasets and methods that do not require labeled data have been applied to satellite imagery, such as transfer learning (Jean et al. 2016), representation learning (Jean et al. 2019; Neumann et al. 2019; Rolf et al. 2021; Klemmer et al. 2023), meta-learning (Tseng et al. 2021), and more recently, foundation models (Tseng et al. 2023; Lacoste et al. 2024; Stewart et al. 2024). An alternative method for handling the lack of annotated data is data augmentation.

To date, data augmentation methods for images have largely been developed for natural imagery (i.e., images that compose common computer vision datasets like ImageNet (Deng et al. 2009) and CIFAR (Krizhevsky, Hinton et al. 2009)). Many of these data augmentation strategies involve transforming the colors and/or geometry of the images. However, the differences between natural and satellite images may complicate direct transfer of such methods to satellite imagery. For example, some operations common to natural image data augmentation distort topology (e.g., shear), which may be undesirable for satellite imagery (Yu et al. 2017). Modifying spectral intensities for satellite imagery has been discouraged as certain tasks may rely on the absolute pixel values (i.e., spectral information may be more sensitive to changes in satellite images than natural images) (Neumann et al. 2019; Hao et al. 2023). Moreover, many color-distorting operations are only defined for 3-band images, which renders them useless on datasets with more than three spectral channels (i.e., multi- and hyperspectral images) (Stewart et al. 2024). There is currently no consensus on which image data augmentation strategies to use for satellite imagery (Hao et al. 2023; Lacoste et al. 2024); many studies use simple geometric transformations such as flipping, rotating, and occasionally translating and random cropping (Yu et al. 2017; Neumann et al. 2019; Abdelhack 2020; Lacoste et al. 2024).

In this paper, we experimentally compare data augmentation techniques on a variety of satellite imagery analysis tasks. Our main contributions are:

[1] We compare several existing methods from natural image analysis on three classification and three regression tasks. We find that data augmentation methods common in natural image analysis should not be directly applied to satellite imagery. Our results indicate that geometric augmentations consistently outperform color augmentations.
[2] We present three new satellite-specific data augmentation strategies: Sat-CutMix, Sat-Trivial, and Sat-SlideMix. We show that all three methods are the top performing augmentation methods across the tasks.
Our results can guide practitioners in selecting data augmentation methods for satellite imagery analysis and may inspire further development of appropriate data augmentation approaches for this domain.

Data Augmentation for Natural Images

Given an (x, y) data instance where x is an image and y is the associated label, data augmentation can be broken into three main categories: 1) *label-preserving* augmentations which modify only x while maintaining y, 2) *non-label-preserving* augmentations which modify both x and y in tandem, and 3) *generative methods* which create entirely new (x, y) data instances.

Label-preserving augmentations consist of basic image manipulations. Geometric and color operations are the most common, but other methods include sharpness transformations (e.g., sharpen, blur), noise operations, and erasing procedures (Hao et al. 2023). Color augmentations manipulate the color space of an image (e.g., modifying brightness or contrast). Geometric operations maintain the color properties but modify the physical location of pixels (e.g., translating the image some number of pixels or horizontally/vertically flipping an image). Geometric operations are useful for combating positional biases but can become detrimental if the image is modified so much that the label is no longer present in the image (Shorten and Khoshgoftaar 2019).

Rather than applying these label-preserving augmentations individually, it is common to apply collections of them or to use more complex methods which automatically determine which transformations to apply. Three common automated augmentation methods are AutoAugment (Cubuk et al. 2019), RandAugment (Cubuk et al. 2020), and TrivialAugment (Müller and Hutter 2021). These automated methods rely on the same set of image transformations, but have different approaches for selecting which transformations to apply. AutoAugment is a reinforcement learning method which learns a policy dictating the set of transformations to apply, the probability of applying each transformation, and the transformation's magnitude (Cubuk et al. 2019). Two disadvantages of AutoAugment are that the search phase is computationally expensive and the learned policy is dataset-specific. RandAugment was developed as a more computationally efficient method which replaces AutoAugment's learned policy with two tunable parameters: the number of sequential transformations to apply to a given image and the magnitudes of those transformations (Cubuk et al. 2020). TrivialAugment is even more efficient as it is parameter-free and randomly selects a single transformation and a corresponding magnitude for a given image (Müller and Hutter 2021).

Non-label-preserving techniques generally consist of mixing techniques and feature space interpolation. Mixing techniques usually combine two images and produce a new label that is scaled based on how the images are mixed. Mixing images commonly results in unnatural looking images (Shorten and Khoshgoftaar 2019). CutMix (Yun et al. 2019) was proposed to overcome this issue. Rather than linearly or non-linearly combining two images, CutMix samples a random patch from one image and pastes it onto another. A new label is then established based on the proportions of the two classes in the resulting image (Yun et al. 2019).

Lastly, generative methods create new (x, y) pairs through methods such as Generative Adversarial Networks (GANs), flow models, and stable diffusion (Shorten and Khoshgoftaar 2019; Trabucco et al. 2023). These methods have shown success in the satellite imagery domain as well. GANs have commonly been used to increase the training set size by generating synthetic aerial (Ma, Tang, and Zhao 2019), RGB (Adedeji et al. 2022), multi- and hyperspectral (Zhu et al. 2018; Mohandoss et al. 2020), and synthetic aperture radar (SAR) images (Guo et al. 2017; Hughes, Schmitt, and Zhu 2018). Abady et al. (2020) and Marín and Escalera (2021) have performed image generation and style transfer for multispectral images and Khanna et al. (2023) used a generative foundation model for satellite image creation. W

Data Augmentation for Satellite Images

While it is common to apply some level of data augmentation to satellite imagery, it is not clear which augmentations are best suited for this domain. The geometric operations of flipping, rotating, and clipping are most commonly used on satellite imagery (Ding et al. 2016; Yang et al. 2016; Ghaffar et al. 2019) and some studies discourage using color-based augmentations (Neumann et al. 2019; Hao et al. 2023; Lacoste et al. 2024). However, a direct comparison of augmentation methods is lacking. Color augmentations incentivize deep models to learn object shapes and outlines and to deemphasize color. Modifying the color properties of natural images is generally acceptable since images are captured in many different lighting conditions. With satellite images however, the spectral information is more sensitive to changes and altering it can lead to misinterpretations (Hao et al. 2023). Additionally, many color augmentations (e.g., saturation and grayscale) are simply not defined for images with more than three bands (Stewart et al. 2024), and therefore, cannot be applied to multi- or hyperspectral imagery. It is also thought that correlations across channels in satellite images are more important than in most computer vision tasks (Illarionova et al. 2021).

Of the data augmentation strategies that have been proposed for satellite imagery, many of the techniques have been developed for specific applications. Oubara et al. (2022) and Hao et al. (2023) present literature reviews on such work. Some custom solutions include methods for super resolution (Ghaffar et al. 2019) and SAR target recognition (Ding et al. 2016). Illarionova et al. (2021) developed MixChannel which leverages the fact that satellites capture

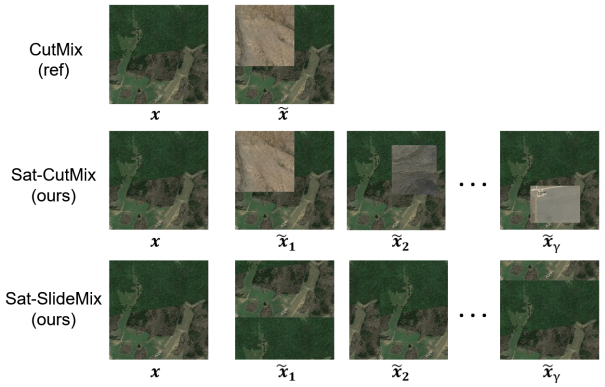

Figure 1: **CutMix versus our proposed Sat-CutMix and Sat-SlideMix.** CutMix creates one mixed image for every original image in the batch while Sat-CutMix and Sat-SlideMix create γ mixed images per original image. Sat-CutMix combines two images by pasting a patch from another image in the batch onto the original image. Sat-SlideMix translates the original image in the H or W dimension and pastes the remaining portion onto the other side of the image.

multiple images over the same location by randomly swapping channels from a secondary image taken over the same location with channels from the original image. A limitation to MixChannel is that some regions simply do not have enough cloud-free images over a given time period to provide a secondary image. Additionally, some tasks vary too much over time (e.g., crops during the growing season) that swapping channels from images taken at different dates may in fact be detrimental. In contrast to these task or dataset-specific methods, we propose three general-purpose data augmentation techniques for satellite imagery.

Proposed Data Augmentation Methods

We present three general-purpose data augmentation strategies tailored for satellite image analysis. We take inspiration from CutMix, as it provides strong data augmentation while maintaining the color properties of images. *Sat-CutMix* modifies the CutMix method to be more suitable to satellite imagery and is non-label-preserving. *Sat-SlideMix* draws inspiration from CutMix, but is label-preserving. Lastly, *Sat-Trivial* extends the label-preserving method TrivialAugment with satellite-specific augmentations. All three methods are designed for classification and regression tasks, but could be extended for image segmentation. The proposed methods are straightforward to implement and do not alter the original loss functions. Implementation details are in the appendix.

Sat-CutMix

Since Sat-CutMix is an extension to CutMix, we first describe the CutMix approach.

CutMix generates new images for classification tasks by patching images together and scaling the class label accordingly (Yun et al. 2019). Given a training image $x \in R^{W \times H \times C}$ and training label y, CutMix produces a new training image, label pair ($\tilde{x}$, $\tilde{y}$) by combining two training samples within a batch: (x_A, y_A) and (x_B, y_B). Let $\mathbf{M} \in \{0,1\}^{W \times H}$ represent a binary mask which indicates which pixels of an image to keep. $\mathbf{1}$ is a binary mask of ones, $\odot$ denotes element-wise multiplication, and λ is the combination ratio. CutMix combines the two training samples as:

$$\begin{aligned} \tilde{x} &= \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \\ \tilde{y} &= \lambda y_A + (1 - \lambda) y_B \end{aligned}$$

where λ is sampled from from a beta distribution Beta(α, α). In essence, a patch from another image in the batch is pasted onto the base image and the class label is scaled by the proportions of the two image classes (Fig. 1). Sat-CutMix modifies the original CutMix algorithm in the following three ways.

First, we extend CutMix to the regression setting. Mixup-type data augmentation techniques are far less common for regression than classification, in large part due to challenges with establishing "meaningful" labels (Hwang and Whang 2021; Yao et al. 2022). Depending on the regression task, it can be challenging to create new, mixed images that are semantically meaningful. For example, in rotation prediction tasks, mixing two images of objects having different degrees of rotation produces an unnatural image. While the issues in creating meaningful samples likely holds for many computer vision tasks, satellite images can be mixed with minimal concern for similarity. To extend CutMix to the regression setting, we calculate $\tilde{y}$ in the same manner as the classification setting but without converting the labels to one-hot encodings; we scale the image labels by λ and 1 - λ, respectively, and add them. For classification tasks, we leave the original CutMix label generation procedure unchanged.

Second, we introduce a parameter controlling how many new samples are created from each instance pair. In the traditional CutMix algorithm, each image in a batch is mixed with one other batch image, producing a single ($\tilde{x}$, $\tilde{y}$) sample. For sufficiently large datasets, adding variety but not volume via data augmentation is a sound approach, but given the prevalence of smaller satellite imagery datasets, it may be valuable to increase the number of training samples. Therefore, we introduce γ which defines the number of new samples to produce for every image in a batch (Fig. 1). With γ = 3, every image in a batch will be mixed with three other images to produce ($\tilde{x_1}$, $\tilde{y_1}$), ($\tilde{x_2}$, $\tilde{y_2}$) & ($\tilde{x_3}$, $\tilde{y_3}$) within the batch.

Third, we make a small modification to the sampling procedure for the combination ratio λ. In CutMix, λ is sampled from a beta distribution Beta(α, α). In their experiments, Yun et al. (2019) samples λ from Beta(1,1), meaning that λ is sampled from a uniform distribution from (0,1). We swap the Beta(α, α) distribution for a uniform(α, 1) distribution to make it easier to control the amount of mixing that is performed. With our formulation, α sets the lower bound for how much of the base image to keep.

Sat-SlideMix

Taking inspiration from Sat-CutMix, we propose Sat-SlideMix, which produces slightly more realistic looking

Algorithm 1: **Sat-SlideMix**

Input: imgs: [batch, C, H, W] tensor
Parameters: γ, β
Output: [$\gamma\times$ batch,C,H,W] tensor

```
magnitudes = sample γ× batch magnitudes from [0.0, ..., β]
imgs = imgs.repeat(γ)
for (m, img) in (magnitudes, imgs) do
    dim: randomly select to shift img in H or W dimension
    randomly multiply m by -1 {randomly shift in opposite direction}
    img = torch.roll(img, m, dim)
    rolled_imgs = torch.cat(rolled_imgs, img)
end for
return rolled_imgs
```

images than Sat-CutMix (Fig. 1) and is label-preserving. Rather than mixing two images, we translate an image by some number of pixels and paste the portion of the image pushed outside of the image bounds onto the other side of the image (Fig. 1). In other words, we roll the image along its height or width axis. Sat-SlideMix has two parameters: γ (the number of new samples to create within the batch for every image in the batch) and β (the maximum percentage of the image to shift). See Algorithm 1 for Sat-SlideMix pseudocode.

Sat-Trivial

We introduce Sat-Trivial as an automated augmentation technique with satellite-specific augmentations. TrivialAugment was developed as a computationally efficient alternative to the more costly automated methods of AutoAugment and RandAugment. TrivialAugment has the simplest scheme for selecting augmentations and yet, somewhat surprisingly, outperforms the other two methods (Müller and Hutter 2021). Given a set of augmentations, TrivialAugment samples a single transformation uniformly at random along with a corresponding magnitude (i.e., distortion strength) for every image in the batch.

Our method modifies the set of augmentations in TrivialAugment to be satellite-specific. We kept the geometric augmentations (flip, rotate, translate, and shear) and added three transformations to model types of variation common to satellite imagery. Specifically, we added *random erasing* to mimic missing data, *random saturation* to represent oversaturation, and *Gaussian noise* to model sensor noise.

Comparative Framework

We evaluated the three proposed augmentation methods along with several common data augmentation techniques across three classification and three regression tasks. We selected tasks with datasets derived from a range of different sensors to evaluate the methods on a variety of satellite data sources. See Tab. 1 for a summary of the tasks and the appendix for a detailed description of the tasks.

Data Augmentation Methods

We evaluated four common image augmentation methods; three automated strategies (AutoAugment, RandAugment, and TrivialAugment) and a standard natural image augmentation set (MoCo v2's aug-plus). MoCo v2's aug-plus (Chen et al. 2020b) randomly applies the following transformations to every image: *ResizedCrop, color jitter, grayscale, Gaussian blur, and horizontal and vertical flips*. Learning policies for AutoAugment requires a significant amount of computational resources (Müller and Hutter 2021), therefore, we used the three prelearned policies: ImageNet, CIFAR10, and SVHN. RandAugment is a more efficient extension to AutoAugment with two tunable parameters and TrivialAugment is a further improvement which is parameter free. The transformations used in the automated methods are: *identity, rotate, shear-x, shear-y, translate-x, translate-y, autoContrast, brightness, color, contrast, equalize, posterize, sharpness, solarize.*

We additionally investigated groupings of the individual transformations used in the automated methods. We split the set of transformations into a geometric set and a color set of transformations. We reduced the set of geometric transformations down to the two most basic operations of flip and rotate. To determine if distorting the topology (e.g., shear) degraded performance, we evaluated a geometric set without shear and a geometric set with a maximum amount of shear. Following TrivialAugment, we randomly sample one transformation from the set of possible transformations for each image in the batch. Similar to the automated methods, we used a linear scale for determining the magnitude of each transformation. Again following TrivialAugment, we scale the magnitudes from $\{0, ..., 30\}$ and uniformly at random sample one of the strengths for each (image, transformation) pair. Below is a summary of the augmentation methods we compared. We did not include MixChannel in our comparison as many satellite imagery datasets (to include those analyzed in our study) do not have multiple images taken over the same location, a requirement for implementing MixChannel. Our three proposed methods are underlined. See the appendix for

Flip & Rotate: Random horizontal and vertical flips and random rotations in increments of 90°.

Geometric: Flip, rotate, translate, and shear transformations. Rotations were again confined to increments of 90°. We sampled translation magnitudes from a range of 0-10% of the image width and height and shear magnitudes between 0-0.3 (the smallest range of shear in the automated methods).

- **No shear**: Geometric set without shear.
- **Large shear**: Geometric set with shear magnitude between 0-0.99 (the largest range of shear in the automated methods).

Color: AutoContrast, brightness, color, contrast, equalize, posterize, and solarize transformations.

Auto_ImageNet: AutoAugment with ImageNet policy. Automated method which learns sub-policies that define which two transformations to apply in sequential order, the probability of applying the transformations, and the

Dataset	Task	Classes	Dataset size	Spatial res.	Num. bands
UC Merced Land Use (Yang and Newsam 2010)	Classification	21	2100	0.3 m	3
Brazilian Coffee Scenes (Penatti, Nogueira, and Dos Santos 2015)	Classification	2	2876	10 m	3
EuroSAT (Helber et al. 2019)	Classification	10	27000	10-60 m	13
Forest cover (Rolf et al. 2021)	Regression	n/a	1000	$\sim$ 4 m	3
Nighttime light intensity (Rolf et al. 2021)	Regression	n/a	1000	$\sim$ 4 m	3
Elevation (Rolf et al. 2021)	Regression	n/a	1000	$\sim$ 4 m	3

Table 1: **Description of tasks.** While the original datasets for the three regression tasks consist of 100k images, we significantly downsampled the datasets to evaluate performance in the small data regime.

magnitudes of the given operations.

Auto_CIFAR: AutoAugment with CIFAR10 policy.

Auto_SVHN: AutoAugment with SVHN policy.

Rand: RandAugment. A policy-free automated method with two tunable parameters: N (number of randomly selected image transformations to apply to each image) and M (magnitude of the transformations). We used the default values of PyTorch's implementation (N=2 and M=9).

Trivial: TrivialAugment. A parameter-free automated method. Randomly selects a transformation and a magnitude for each image.

MoCo v2's aug-plus: ResizedCrop, color jitter, grayscale, Gaussian blur, and horizontal and vertical flips. All transformations are applied to each image with randomly sampled magnitudes.

CutMix: Parameter-free mixing method that combines two images and scales the label based on the proportions of the classes present in the mixed image. We included flip and rotate transformations.

Sat-CutMix: Proposed extension of CutMix. We found $\gamma = 3$ and $\alpha = 0.9$ to perform best (Fig. A1, A2). Similar to CutMix, we also included flip and rotate transformations[1].

Sat-SlideMix: Proposed label-preserving extension to Sat-CutMix. We set $\gamma = 3$ and set $\beta = 1$ (images can be shifted along their full H or W dimension). We included flip and rotate transformations for fair comparisons to CutMix and Sat-CutMix.

Sat-Trivial: Proposed modification to TrivialAugment with satellite-specific augmentations. The set of augmentations includes flip, rotate, translate, shear, noise, random erase, and random saturate.

Model, Tuning, and Evaluation

For each task and augmentation pair, we fine-tuned ResNet-18 models with pre-trained ImageNet weights. For the EuroSAT task which contained 13 input channels, we repeated the ImageNet weights in the first convolutional layer (RGBRGB. . .) and multiplied them by 3/C (C = 13 for the input channels) as was done by Stewart et al. (2024) to handle multispectral images. For each task, we performed hyperparameter tuning on the learning rate, batch size, and weight decay. We also performed early stopping based on a validation set. Details about final parameter settings can be found in the appendix. To quantify variation across training runs and stochasticity inherent to some of the data augmentation techniques, we trained each model five times. We characterized model performance by evaluating the models on a held-out test set (i.e., we did not report the highest performance achieved during training). We measured percent change in accuracy and mean absolute error (MAE) from a baseline of no data augmentation for the classification and regression tasks respectively. Source code and supplementary material are available at https://github.com/Hutchinson-Lab/Data-Augmentation-Approaches-for-Satellite-Imagery.

[1]We did not include additional transformations because augmentations are applied in the dataloader which occurs before the Sat-CutMix procedure. Mixing two images with different augmentations (e.g., an image with shear and an image with noise) would lead to obscure combinations.

Results

Sat-SlideMix. Overall, Sat-SlideMix was the best performing augmentation method (Tab. 2) and outperformed all other methods in three out of the six tasks (Fig. 2). Sat-SlideMix significantly outperformed all other methods in the elevation task.

Sat-CutMix. Sat-CutMix is the second best performing augmentation method (Tab. 2) and exhibited the highest performance for forest cover (Fig. 2). For the classification tasks, Sat-CutMix had performance similar to CutMix for Brazilian Coffee Scenes and EuroSAT and significantly outperformed CutMix for UC Merced Land Use (Fig. 2). CutMix is not defined in the regression setting and, therefore, a comparison is not possible.

Sat-Trivial. Sat-Trivial was the third best performing augmentation method (Tab. 2) and had the highest performance for UC Merced Land Use (Fig. 2). Across all tasks, Sat-Trivial significantly outperformed TrivialAugment, its natural imagery equivalent.

Automated methods. The automated methods showed mixed results across the tasks. In general, the automated techniques degraded model performance for the classification tasks, but the results were mixed for the regression

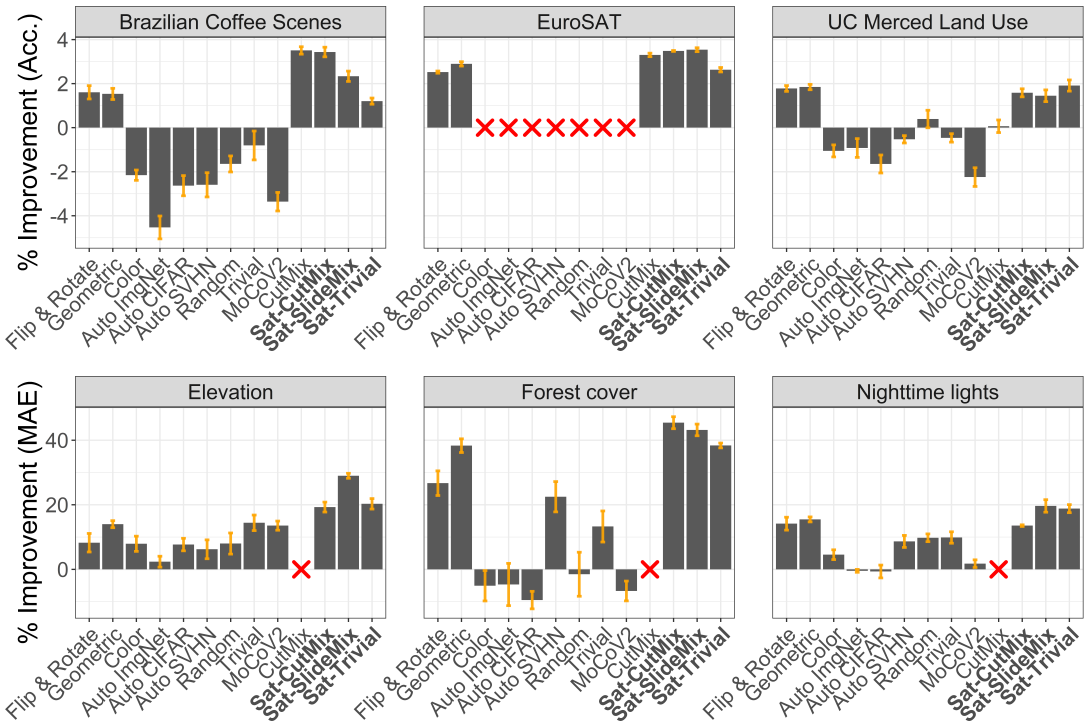

Figure 2: **Percent improvement over no augmentation.** We compared model performance across different augmentation methods where the results are shown as percent improvement over no augmentation. Our three proposed methods are in bold. We characterized stochasticity in model training and in the augmentation methods by training five models per task and augmentation method and evaluating on a held-out test set. The bar plots represent the mean model performance over the five trained models with error bars showing the standard error. Positive values indicate improved model performance over models with no augmentation while negative values indicate worsened performance. Statistical significances are reported in Tab. A3. The red x marks indicate situations in which augmentation methods are undefined: traditional CutMix is not defined for regression tasks and most augmentation methods are only defined for 3-band images and, therefore, cannot be applied to multi- or hyperspectral images (e.g., EuroSAT). We evaluated the classification tasks (top row) with accuracy, precision, and recall and the regression tasks (bottom row) with mean absolute error (MAE), mean squared error (MSE), and R^2. We only present accuracy and MAE as the other metrics showed similar trends (Fig. A5, A6).

tasks. Elevation was the only task in which all forms of augmentation were beneficial.

Geometric. In all cases, except for Brazilian Coffee Scenes, Geometric outperformed Flip & Rotate (Fig. 2). In all tasks except for forest cover, removing shear from the geometric set had either no impact (zero in the error bars) or a very small negative impact (Fig. 3). Removing shear from forest cover had a significant negative impact. Increasing the amount of shear had mixed results. Larger amounts of shear were detrimental to Brazilian Coffee Scenes, EuroSat, and nighttime lights, beneficial for elevation, and had no impact for the other two tasks (Fig. 3).

Color. The color augmentation set was only beneficial for elevation and had degraded model performance for all other tasks (Fig. 2). Over the classification tasks, the five other methods which involve color transformations (all three AutoAugment policies, Random, and Trivial) displayed degraded performance or had error bars overlapping zero. However, the results were mixed for the regression tasks. Any type of augmentation was beneficial for elevation, augmentation was either equivalent to no augmentation or beneficial for nighttime lights, and forest cover had mixed results (Fig. 2). Since the color augmentation set did not consistently perform well, we did not pursue characterizations of their individual contributions.

Discussion

Our comparative study supports the idea that augmentation techniques designed for natural images should not be applied to satellite imagery without careful consideration. For a given task or dataset, some natural image augmentation methods are beneficial, but those same methods do not necessarily perform well on a different task or dataset (Fig. 2). All three regression tasks are based on the same imagery dataset, and yet, the automated augmentation methods exhibit varying performance across the tasks, whereas our three proposed methods consistently perform well on the three regression tasks (Fig. 2). Across all tasks, Sat-SlideMix, Sat-CutMix, and Sat-Trivial's performances are

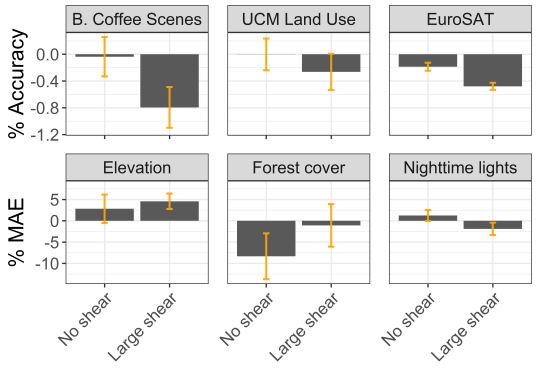

Figure 3: **Percent improvement over Geometric set.** The bar plots represent mean model performance over five trained models with the associated standard errors. Positive values indicate improved model performance over the geometric set while negative values indicate worsened performance.

Augmentation method	Average rank	Standard deviation
Sat-SlideMix	**2.17**	**1.60**
Sat-CutMix	**2.83**	**1.47**
Sat-Trivial	**3.17**	**1.94**
CutMix	3.67	3.06
Geometric	3.83	1.17
Flip & Rotate	4.83	1.47
Trivial	6.40	1.52
Random	7.40	0.89
Auto SVHN	8.80	1.92
Color	9.60	0.89
MoCov2's aug-plus	10.40	2.70
Auto ImgNet	11.00	1.58
Auto CIFAR	11.40	0.89

Table 2: **Mean and standard deviation rankings of the augmentation methods across tasks.** Our proposed methods are in bold.

notable since they are the top three performing methods (Tab. 2). **In fact, in all but one task, the best performing method is either Sat-SlideMix, Sat-CutMix, or Sat-Trivial (Fig. 2).** Furthermore, Sat-Trivial, which has the same augmentation scheme as TrivialAugment but with satellite-specific augmentations, significantly outperforms Trivial-Augment across all tasks. This highlights the benefit of domain-specific augmentation methods.

Some tasks may benefit from the increased diversity in the label space created by the non-label-preserving methods. Across our three proposed methods, our non-label-preserving method (Sat-CutMix) has the best performance for Brazilian Coffee Scenes and forest cover, while our label-preserving methods (Sat-SlideMix and Sat-Trivial) have the best performance among the remaining four tasks. The performance differences between Sat-CutMix and Sat-SlideMix could be due to the non-label-preserving and label-preserving characteristics of the two methods.

While Illarionova et al. (2021) suggest geometric augmentations, such as rotation and translation, do not provide enough variability for medium resolution data (10m/pixel) and, therefore, have somewhat limited influence, our results show that flip and rotate are better than no augmentation whatsoever (Fig. 2). Our results indicate that augmentations beyond geometric transformations can further improve performance (Sat-CutMix, Sat-SlideMix, and Sat-Trivial compared to Geometric, Fig. 2). Additionally, when randomly sampling from a set of transformations (i.e., images sometimes have translate or shear applied, for example), adding small amounts of translation and shear generally improves performance over only flipping and rotating (Geometric versus Flip & Rotate, Fig. 2). However, care should be taken in how much shear is applied. Increasing the amount of shear was detrimental to Brazilian Coffee Scenes, EuroSat, and nighttime lights, but beneficial for elevation (Fig. 3).

As several studies have suggested (Neumann et al. 2019; Hao et al. 2023; Lacoste et al. 2024), we found that color augmentations were generally not beneficial for satellite imagery (Fig. 2). While we cannot determine the root cause for the mixed (and often negative) impacts of the automated methods, we know that all automated methods rely on color transformations. Conversely, the methods which do not involve color augmentations (Flip & rotate, Geometric, CutMix, Sat-CutMix, Sat-SlideMix, and Sat-TrivialAugment) almost always exhibited improved performance over no augmentation. However, it should be noted that in cases like contrastive learning where two different data views are necessary, a significant amount of data augmentation, to include varying color properties, may be desirable (Chen et al. 2020a; Greenstreet et al. 2023).

Conclusion

While larger datasets are preferred for training deep models, establishing large, labeled datasets for satellite imagery analyses is not always feasible (Adedeji et al. 2022). When transferring machine learning methods to satellite imagery, the main approach is to transfer techniques developed for other domains with little work done to tailor the methods to satellite imagery (Rolf et al. 2024). Others have highlighted the need to develop satellite-specific methods, especially in regard to data augmentation (Abdelhack 2020; Lacoste et al. 2024).

We present an analysis of how data augmentation techniques developed for natural imagery transfer to satellite imagery and propose three satellite-specific methods. We demonstrate the benefit of our methods by evaluating them on a variety of tasks. Our results indicate that color-based augmentations and natural image augmentation techniques are not directly transferable to satellite imagery, but that geometric operations are beneficial. Additionally, we show that our proposed Sat-SlideMix, Sat-CutMix, and Sat-Trivial methods are the top performing augmentation methods.

Acknowledgments

Laurel Hopkins received funding from the National Aeronautics and Space Administration under Future Investigators in NASA Earth and Space Science and Technology (FINESST) (Grant No: 80NSSC20K1664).

References

Abady, L.; Barni, M.; Garzelli, A.; and Tondi, B. 2020. GAN generation of synthetic multispectral satellite images. In *Image and signal processing for remote sensing XXVI*, volume 11533, 122–133. SPIE.

Abdelhack, M. 2020. A comparison of data augmentation techniques in training deep neural networks for satellite image classification. In *Proc. Comput. Vis. Pattern Recognit.*, 1–4.

Adedeji, O.; Owoade, P.; Ajayi, O.; and Arowolo, O. 2022. Image augmentation for satellite images. *arXiv preprint arXiv:2207.14580*.

Chen, D.; Xue, Y.; Chen, S.; Fink, D.; and Gomes, C. 2016. Deep multi-species embedding. *arXiv preprint arXiv:1609.09353*.

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

Chen, X.; Fan, H.; Girshick, R. B.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. *CoRR*, abs/2003.04297.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 113–123.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Davis, C. L.; Bai, Y.; Chen, D.; Robinson, O.; Ruiz-Gutierrez, V.; Gomes, C. P.; and Fink, D. 2023. Deep learning with citizen science data enables estimation of species diversity and composition at continental extents. *Ecology*, 104(12): e4175.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Ding, J.; Chen, B.; Liu, H.; and Huang, M. 2016. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and remote sensing letters*, 13(3): 364–368.

Ghaffar, M.; McKinstry, A.; Maul, T.; and Vu, T. 2019. Data augmentation approaches for satellite image super-resolution. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4: 47–54.

Greenstreet, L.; Fan, J.; Pacheco, F. S.; Bai, Y.; Ummus, M. E.; Doria, C.; Barros, N. O.; Forsberg, B. R.; Xu, X.; Flecker, A.; et al. 2023. Detecting Aquaculture with Deep Learning in a Low-Data Setting.

Guo, J.; Lei, B.; Ding, C.; and Zhang, Y. 2017. Synthetic aperture radar image synthesis by using generative adversarial nets. *IEEE Geoscience and Remote Sensing Letters*, 14(7): 1111–1115.

Hao, X.; Liu, L.; Yang, R.; Yin, L.; Zhang, L.; and Li, X. 2023. A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition. *Remote Sensing*, 15(3): 827.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Hughes, L. H.; Schmitt, M.; and Zhu, X. X. 2018. Generative Adversarial Networks for Hard Negative Mining in CNN-Based SAR-Optical Image Matching. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 4391–4394. IEEE.

Hwang, S.-H.; and Whang, S. E. 2021. RegMix: Data Mixing Augmentation for Regression. *arXiv preprint arXiv:2106.03374*.

Illarionova, S.; Nesteruk, S.; Shadrin, D.; Ignatiev, V.; Pukalchik, M.; and Oseledets, I. 2021. MixChannel: Advanced augmentation for multispectral satellite images. *Remote Sensing*, 13(11): 2181.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3967–3974.

Kerner, H.; Tseng, G.; Becker-Reshef, I.; Nakalembe, C.; Barker, B.; Munshell, B.; Paliyam, M.; and Hosseini, M. 2020. Rapid response crop maps in data sparse regions. *arXiv preprint arXiv:2006.16866*.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D. B.; and Ermon, S. 2023. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*.

Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; and Rußwurm, M. 2023. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lacoste, A.; Lehmann, N.; Rodriguez, P.; Sherwin, E.; Kerner, H.; Lütjens, B.; Irvin, J.; Dao, D.; Alemohammad, H.; Drouin, A.; et al. 2024. Geo-bench: Toward foundation

models for earth monitoring. *Advances in Neural Information Processing Systems*, 36.

Ma, D.; Tang, P.; and Zhao, L. 2019. SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro. *IEEE Geoscience and Remote Sensing Letters*, 16(7): 1046–1050.

Marín, J.; and Escalera, S. 2021. SSSGAN: Satellite style and structure generative adversarial networks. *Remote Sensing*, 13(19): 3984.

Mohandoss, T.; Kulkarni, A.; Northrup, D.; Mwebaze, E.; and Alemohammad, H. 2020. Generating synthetic multispectral satellite imagery from sentinel-2. *arXiv preprint arXiv:2012.03108*.

Müller, S. G.; and Hutter, F. 2021. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 774–782.

Neumann, M.; Pinto, A. S.; Zhai, X.; and Houlsby, N. 2019. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*.

Oubara, A.; Wu, F.; Amamra, A.; and Yang, G. 2022. Survey on remote sensing data augmentation: Advances, challenges, and future perspectives. In *International Conference on Computing Systems and Applications*, 95–104. Springer.

Penatti, O. A.; Nogueira, K.; and Dos Santos, J. A. 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 44–51.

Peterson, K. T.; Sagan, V.; and Sloan, J. J. 2020. Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing. *GIScience & Remote Sensing*, 57(4): 510–525.

Robinson, C.; Hohman, F.; and Dilkina, B. 2017. A deep learning approach for population estimation from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, 47–54.

Rolf, E.; Klemmer, K.; Robinson, C.; and Kerner, H. 2024. Mission Critical–Satellite Data is a Distinct Modality in Machine Learning. *arXiv preprint arXiv:2402.01444*.

Rolf, E.; Proctor, J.; Carleton, T.; Bolliger, I.; Shankar, V.; Ishihara, M.; Recht, B.; and Hsiang, S. 2021. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Commun*, 12(4392).

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A. S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2): 1–96.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Stewart, A.; Lehmann, N.; Corley, I.; Wang, Y.; Chang, Y.-C.; Ait Ali Braham, N. A.; Sehgal, S.; Robinson, C.; and Banerjee, A. 2024. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Tseng, G.; Cartuyvels, R.; Zvonkov, I.; Purohit, M.; Rolnick, D.; and Kerner, H. 2023. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*.

Tseng, G.; Kerner, H.; Nakalembe, C.; and Becker-Reshef, I. 2021. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1111–1120.

Yang, N.; Tang, H.; Sun, H.; and Yang, X. 2016. Dropband: A convolutional neural network with data augmentation for scene classification of VHR satellite images.

Yang, Y.; and Newsam, S. 2010. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, 270–279. New York, NY, USA: Association for Computing Machinery. ISBN 9781450304283.

Yao, H.; Wang, Y.; Zhang, L.; Zou, J. Y.; and Finn, C. 2022. C-mixup: Improving generalization in regression. *Advances in Neural Information Processing Systems*, 35: 3361–3376.

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Yu, X.; Wu, X.; Luo, C.; and Ren, P. 2017. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, 54(5): 741–758.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhu, L.; Chen, Y.; Ghamisi, P.; and Benediktsson, J. A. 2018. Generative adversarial networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9): 5046–5063.