Resolving Super Fine-Resolution SIF via Coarsely-Supervised U-Net Regression

Joshua Fan
Cornell University
jyf6@cornell.edu

Di Chen
Cornell University
di@cs.cornell.edu

Jiaming Wen
Cornell University
jw2495@cornell.edu

Ying Sun
Cornell University
ys776@cornell.edu

Carla Gomes
Cornell University
gomes@cs.cornell.edu

Abstract

Climate change presents challenges to crop productivity, such as increasing the likelihood of heat stress and drought. Solar-Induced Chlorophyll Fluorescence (SIF) is a powerful way to monitor how crop productivity and photosynthesis are affected by changing climatic conditions. However, satellite SIF observations are only available at a coarse spatial resolution (e.g. 3-5km) in most places, making it difficult to determine how individual crop types or farms are doing. This poses a challenging *coarsely-supervised regression* task; at training time, we only have access to SIF labels at a coarse resolution (3 km), yet we want to predict SIF at a very fine spatial resolution (30 meters), a 100x increase. We do have some fine-resolution input features (such as Landsat reflectance) that are correlated with SIF, but the nature of the correlation is unknown. To address this, we propose *Coarsely-Supervised Regression U-Net (CSR-U-Net)*, a novel approach to train a U-Net for this coarse supervision setting. CSR-U-Net takes in a fine-resolution input image, and outputs a SIF prediction for each pixel; the average of the pixel predictions is trained to equal the true coarse-resolution SIF for the entire image. Even though this is a very weak form of supervision, CSR-U-Net can still learn to predict accurately, due to its inherent localization abilities, plus additional enhancements facilitated by scientific prior knowledge. CSR-U-Net can resolve fine-grained variations in SIF more accurately than existing averaging-based approaches, which ignore fine-resolution spatial variation during training. CSR-U-Net could also be useful for a wide range of "downscaling" problems in climate science, such as increasing the resolution of global climate models.

1 Introduction

Crop production is very sensitive to climate change; rising temperatures and increases in drought can negatively impact crop yield [1, 2, 3]. Monitoring crop productivity in the face of climate change is essential for food security purposes [4]. Thus, there have been many attempts to use remote sensing to monitor crop growth from space [5]. A common approach is to use vegetation indices (such as NDVI), which are simple combinations of a few spectral bands measured from satellites. However, it has been shown that satellite-based Solar-Induced Chlorophyll Fluorescence (SIF) can outperform these vegetation indices for crop yield prediction [6, 7], as well as for monitoring the effects of drought [8] and heat stress [9] on vegetation productivity. Unlike vegetation indices, SIF has mechanistic linkages to plant photosynthesis, and can provide more accurate information about plant growth [10, 11].

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

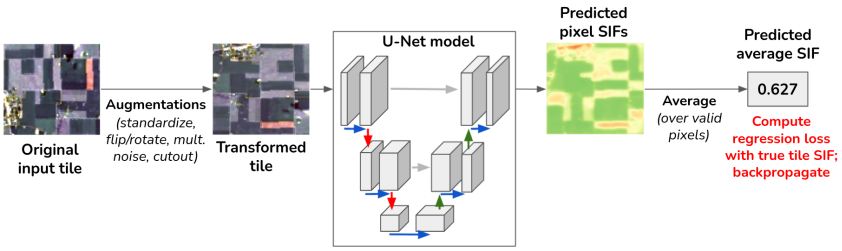

Figure 1: Coarsely-Supervised Regression U-Net (CSR-U-Net). U-Net figure inspired from [21].

.

However, SIF is a very difficult signal to measure from space [12]. Most available SIF measurements are noisy, and are only available at a coarse spatial resolution [13], such as 3×3 km. Within a large region of many square kilometers, SIF can vary dramatically, depending on crop type, management practices, and genetic varieties. This variation cannot be resolved by existing satellite measurements, due to their coarse spatial resolution [11].

There have been several attempts to predict SIF at finer spatial resolutions, given coarse-resolution SIF measurements and additional data [14, 13, 12, 15]. This task is also known as "statistical downscaling." These works train a supervised machine learning model on large coarse-resolution tiles, to map from auxiliary input variables (averaged across the entire large tile) to the SIF of that tile. Then this trained model is applied on fine-resolution input variables to predict SIF at a fine resolution. Since the input variables are averaged across a large region during training, information about fine-grained spatial variation is lost. Thus, these works only attempt to resolve SIF down to a 0.05-degree (≈ 5 km) resolution; we find that these methods are generally inaccurate at super-fine resolutions such as 30 meters.

This problem poses a challenging *coarsely-supervised regression* task. We desire to learn a model that maps from fine-resolution input features to fine-resolution SIF, yet the only supervision available for SIF at training time is at a much coarser resolution. We note that this task (inferring a fine-resolution map of a variable, given only coarse-resolution observations and other auxiliary data) is not exclusive to SIF, but is also applicable to many problems in climate science. For example, global circulation models (GCMs) are frequently used to simulate future climate patterns, but can only predict at a coarse resolution; there is strong interest in downscaling the model outputs to a finer resolution [16, 17]. Other similar problems include downscaling soil moisture [18], evapotranspiration [19], and precipitation [20] data to finer resolutions.

To address this issue, we propose a novel technique, *Coarsely-Supervised Regression U-Net (CSR-U-Net)*, which can produce SIF predictions at a 30-meter resolution, even though it only learns from SIF labels at a 3km resolution (100x coarser). CSR-U-Net takes in a fine-resolution input image, and outputs a SIF prediction for each pixel; the average of the pixel predictions is trained to equal the true coarse-resolution SIF for the entire 3km image. This is an extremely weak form of supervision, and in itself does not provide sufficient constraints for the pixel predictions. However, CSR-U-Net can still learn to predict accurately in this setting through techniques inspired by prior knowledge. The CSR-U-Net architecture inherently provides impressive localization abilities. While the model can overfit (by outputting extreme pixel predictions that happen to average to the right value), we find that two techniques can mitigate this: (1) early stopping based on a small fine-resolution validation set, and (2) multiplicative noise that forces the model to pay attention to the ratio between input bands (which is known to be informative of vegetation growth). To our knowledge, our approach has never been applied to resolving SIF or similar "coarse-to-fine" regression problems before. Experimental results show that CSR-U-Net can predict SIF at high resolutions more accurately than strong existing baselines for statistical downscaling, reducing RMSE by around 8%.

2 Methods

We propose CSR-U-Net, a novel approach for coarsely-supervised regression tasks. CSR-U-Net takes in an input image, transforms it in ways that incorporate prior knowledge, and passes the transformed

image through a U-Net to obtain pixel-level predictions. The average of the pixel predictions is trained to equal the true tile SIF. Figure 1 summarizes the approach.

2.1 Coarsely Supervised Training

CSR-U-Net is trained in a coarsely-supervised fashion. Our training set consists of pairs $(\mathbf{X}_i, y_i)$, where $\mathbf{X}_i \in \mathbb{R}^{H \times W \times C}$ is a large tile "image" with many channels/features, and $y_i \in \mathbb{R}$ is the ground-truth average (coarse) SIF for the entire tile i.

First, we perform data augmentations on the input tile to obtain a transformed tile, $\mathbf{X}'_i$ (see Section 2.2 for details). We train a U-Net model f that takes in the transformed tile $\mathbf{X}'_i$ and predicts SIF for each pixel, $\mathbf{z}_i = f(\mathbf{X}'_i)$. Let $\mathbf{z}_i^{(p)} \in \mathbb{R}$ be the predicted SIF of pixel p. Since we do not know the pixel SIFs while training, the model is trained by encouraging the average pixel SIF prediction to be close to the true total SIF of the entire tile. Specifically, the model is trained to minimize the Mean Squared Error loss function, where P is the set of pixels which are valid (i.e. they contain CFIS SIF observations and were not obscured by clouds), or a random subset of valid pixels:

$$\ell(y_i, \mathbf{z}_i) = \left(y_i - \frac{1}{|P|} \sum_{p \in P} \mathbf{z}_i^{(p)} \right)^2 \tag{1}$$

During training, we pass a batch of large tiles through the model, obtain pixel-level SIF predictions for each tile ($\mathbf{z}_i$), and compute the Mean Squared Error loss between the *average predicted SIF* (over valid pixels) and the true SIF. This error is backpropagated through the U-Net model.

CSR-U-Net does not look at any fine-resolution SIF data during training. However, during validation, CSR-U-Net does peek at the model's performance on the *fine-resolution validation set* to select which epoch's model to use. If CSR-U-Net simply used the last epoch's model, this would often lead to a unique form of overfitting, where the model outputs extreme high and low pixel SIF predictions that are wildly incorrect, but still produce the correct tile average. CSR-U-Net's early stopping based on the fine-resolution validation set helps ensure that the model is well-regularized (so that it outputs similar predictions for similar pixels across different tiles); this is discussed further in Appendix C. Appendix F contains more details on the model architecture and hyperparameters.

2.2 Incorporating Prior Knowledge via Data Augmentations

To artificially increase the size of the dataset, we harness prior knowledge to design augmentations of the input images that should not significantly affect the final tile-level SIF. First, we randomly flip and/or rotate the input tile, randomly permute parts of the tiles, and sometimes randomly erase part of the input [22] to prevent the model from relying too much on a single part of the tile.

We also propose a "multiplicative noise" augmentation, where we multiply all channels of the Landsat image by a random constant $(1 + \epsilon)$, where $\epsilon \sim N(0, \sigma^2)$. The standard deviation σ can be tuned; we used $\sigma = 0.2$ for our results. This is inspired by the fact that many vegetation indices (which aim to track vegetation growth) actually measure the *ratio* between different channels (such as the near-infrared reflectance divided by the red reflectance) [23]. Multiplying all channels by a random constant forces our model to pay attention to the ratio between channels, which is more informative of vegetation growth than the absolute values. In practice, multiplicative noise is crucial for achieving improved prediction accuracy (see Appendix B for ablation studies).

3 Experiments

3.1 Dataset

During training, we want to simulate a setting where only coarse-resolution SIF measurements are available, so we train the model using SIF labels at a coarse 3km resolution from the Chlorophyll Fluorescence Imaging Spectrometer (CFIS) [24] and Orbiting Carbon Observatory-2 (OCO-2) [25]. We split the tiles into 60% train, 20% validation, and 20% test. We use 712 CFIS and 1390 OCO-2 tiles during training, and 261 CFIS tiles for validation. To evaluate the quality of the model's fine-resolution predictions, we compare them against a limited amount of fine-resolution (30m) CFIS data

in the different sets. To reduce noise, we only evaluate on pixels for which CFIS had at least 30 pixels and where the SIF value was at least 0.1. We use Normalized Root Mean Squared Error (NRMSE) and R^2 as evaluation metrics to compare our predictions with the ground-truth fine-resolution SIF. (NRMSE is Root Mean Squared Error, normalized by the average SIF of the training dataset.)

The input features to our model include Landsat surface reflectance bands [26], FLDAS land data such as temperature and rainfall [27], and land cover information [28], which are available at a 30-meter resolution. Thus, each coarse-resolution SIF label corresponds to a tile of dimensions 24 (features) $\times 100 \times 100$ (pixels). More dataset details are Appendix E.

3.2 Baselines

We compare the CSR-U-Net approach with existing "statistical downscaling" baselines that have been used in papers such as [13, 14]: Ridge Regression, Gradient Boosting Regressor, and a fully-connected artificial neural network (ANN). These involve averaging each input feature (channel) over all valid pixels in the tile, and training a model to predict SIF from the feature averages. For each method, a grid-search was performed over hyperparameter values; we selected the hyperparameters that performed best on the fine-resolution validation set. We also include a trivial "predict coarse" baseline where we predict the SIF of every pixel to be the same as the SIF of the entire tile.

3.3 Results

Table 1 presents the results for fine pixels in train tiles (e.g. the coarse-resolution SIF of the entire tile was known during training, but not the fine-resolution SIF of the individual pixels). The model is tasked with interpolating the fine-resolution SIF from the coarse-resolution labels. According to both metrics, CSR-U-Net outperforms standard "statistical downscaling" methods in both settings, reducing NRMSE by 8% and increasing R^2 by 13% (relative) over the best baseline, Ridge Regression. Additional results in Appendix A show that CSR-U-Net outperforms baselines over all 3 major land cover types (corn, soybean, grassland) and over most resolutions, and can also generalize to tiles where the coarse-resolution SIF label was not seen during training.

Method	**NRMSE** (lower better)	R^2 (higher better)
Predict coarse	*0.248*	*0.373*
Ridge Regression	0.213	0.537
Gradient Boosting	0.225	0.486
ANN	$0.244 \pm$ *0.006*	$0.396 \pm$ *0.027*
CSR-U-Net	**0.196** $\pm$ *0.002*	**0.609** $\pm$ *0.007*

Table 1: Results on interpolating CFIS to 30m resolution, on train tiles (the model was trained on SIF labels at 3km resolution for these tiles, but needs to predict at 30m resolution). For methods that involve randomness, we report the average $\pm$ the standard deviation over 3 runs.

4 Conclusion

We presented Coarsely-Supervised Regression U-Net (CSR-U-Net), which is capable of predicting SIF at a super fine resolution (30m), even when only coarse-resolution (3km) SIF measurements are available. Due to its localization properties, CSR-U-Net is able to figure out which farms within a large tile had higher and lower SIF. CSR-U-Net can avoid overfitting, thanks to techniques informed by prior knowledge, such as multiplicative noise and early stopping.

Although CSR-U-Net's predictions are not perfect and are affected by data noise, they clearly outperform existing state-of-the-art methods, and can provide valuable vegetation information. Even noisy fine-resolution SIF estimates can facilitate improvements in crop monitoring, as SIF contains detailed information about plant photosynthesis that is not captured in vegetation indices such as NDVI [7], which are simple combinations of a few spectral bands. Moreover, in addition to SIF, CSR-U-Net could potentially be applied to make fine-resolution predictions of any numerical variable that is only available at coarse resolution, if there exists other auxiliary fine-resolution data that is correlated with the variable of interest. Such applications could include predicting climate model variables (including precipitation, soil moisture, and evapotranspiration) at finer resolutions.

References

[1] Ariel Ortiz-Bobea, Toby R Ault, Carlos M Carrillo, Robert G Chambers, and David B Lobell. Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change*, 11(4):306–312, 2021.

[2] Chuang Zhao, Bing Liu, Shilong Piao, Xuhui Wang, David B Lobell, Yao Huang, Mengtian Huang, Yitong Yao, Simona Bassu, Philippe Ciais, et al. Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, 114(35):9326–9331, 2017.

[3] Wenzhe Jiao, Qing Chang, and Lixin Wang. The sensitivity of satellite solar-induced chlorophyll fluorescence to meteorological drought. *Earth's Future*, 7(5):558–573, 2019.

[4] Liyin He, Troy Magney, Debsunder Dutta, Yi Yin, Philipp Köhler, Katja Grossmann, Jochen Stutz, Christian Dold, Jerry Hatfield, Kaiyu Guan, et al. From the ground to space: Using solar-induced chlorophyll fluorescence to estimate crop productivity. *Geophysical Research Letters*, 47(7):e2020GL087474, 2020.

[5] Kaiyu Guan, Jin Wu, John S. Kimball, Martha C. Anderson, Steve Frolking, Bo Li, Christopher R. Hain, and David B. Lobell. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment*, 199:333–349, 2017.

[6] Kaiyu Guan, Joseph A Berry, Yongguang Zhang, Joanna Joiner, Luis Guanter, Grayson Badgley, and David B Lobell. Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence. *Global change biology*, 22(2):716–726, 2016.

[7] Bin Peng, Kaiyu Guan, Wang Zhou, Chongya Jiang, Christian Frankenberg, Ying Sun, Liyin He, and Philipp Köhler. Assessing the benefit of satellite-based solar-induced chlorophyll fluorescence in crop yield prediction. *International Journal of Applied Earth Observation and Geoinformation*, 90:102126, 2020.

[8] Lifu Zhang, Na Qiao, Changping Huang, and Siheng Wang. Monitoring drought effects on vegetation productivity using satellite solar-induced chlorophyll fluorescence. *Remote Sensing*, 11(4):378, 2019.

[9] Yang Song, Jing Wang, and Lixin Wang. Satellite solar-induced chlorophyll fluorescence reveals heat stress impacts on wheat yield in india. *Remote Sensing*, 12(20):3277, 2020.

[10] J. Joiner, L. Guanter, R. Lindstrot, M. Voigt, A. P. Vasilkov, E. M. Middleton, K. F. Huemmrich, Y. Yoshida, and C. Frankenberg. Global monitoring of terrestrial chlorophyll fluorescence from moderate-spectral-resolution near-infrared satellite measurements: methodology, simulations, and application to gome-2. *Atmospheric Measurement Techniques*, 6(10):2803–2823, 2013.

[11] Oz Kira and Ying Sun. Extraction of sub-pixel c3/c4 emissions of solar-induced chlorophyll fluorescence (sif) using artificial neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161:135 – 146, 2020.

[12] G. Duveiller, F. Filipponi, S. Walther, P. Köhler, C. Frankenberg, L. Guanter, and A. Cescatti. A spatially downscaled sun-induced fluorescence global product for enhanced monitoring of vegetation productivity. *Earth System Science Data*, 12(2):1101–1116, 2020.

[13] L. Yu, J. Wen, C. Y. Chang, C. Frankenberg, and Y. Sun. High-resolution global contiguous sif of oco-2. *Geophysical Research Letters*, 46(3):1449–1458, 2019.

[14] Pierre Gentine and Hamed Alemohammad. Rsif (reconstructed solar induced fluorescence): a machine-learning vegetation product based on modis surface reflectance to reproduce gome-2 solar induced fluorescence. *Geophysical Research Letters*, 45, 03 2018.

[15] J. Wen, P. Köhler, G. Duveiller, N.C. Parazoo, T.S. Magney, G. Hooker, L. Yu, C.Y. Chang, and Y. Sun. A framework for harmonizing multiple satellite instruments to generate a long-term global high spatial-resolution solar-induced chlorophyll fluorescence (sif). *Remote Sensing of Environment*, 239:111644, 2020.

[16] Andrew W Wood, Lai R Leung, Venkataramana Sridhar, and DP Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic change*, 62(1):189–216, 2004.

[17] Kazi Farzan Ahmed, Guiling Wang, John Silander, Adam M Wilson, Jenica M Allen, Radley Horton, and Richard Anyah. Statistical downscaling and bias correction of climate model outputs for climate change impact assessment in the us northeast. *Global and Planetary Change*, 100:320–332, 2013.

[18] Jian Peng, Alexander Loew, Olivier Merlin, and Niko E. C. Verhoest. A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics*, 55(2):341–366, 2017.

[19] Yinghai Ke, Jungho Im, Seonyoung Park, and Huili Gong. Downscaling of modis one kilometer evapotranspiration using landsat-8 data and machine learning approaches. *Remote Sensing*, 8(3):215, Mar 2016.

[20] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew R. Michaelis, Ramakrishna R. Nemani, and Auroop R. Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. *CoRR*, abs/1703.03126, 2017.

[21] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2):207, 2020.

[22] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.

[23] A Bannari, D Morin, F Bonn, and AjRsr Huete. A review of vegetation indices. *Remote sensing reviews*, 13(1-2):95–120, 1995.

[24] Christian Frankenberg, Philipp Köhler, Troy S. Magney, Sven Geier, Peter Lawson, Mark Schwochert, James McDuffie, Darren T. Drewry, Ryan Pavlick, and Andreas Kuhnert. The chlorophyll fluorescence imaging spectrometer (cfis), mapping far red fluorescence from aircraft. *Remote Sensing of Environment*, 217:523 – 536, 2018.

[25] Y. Sun, C. Frankenberg, J. D. Wood, D. S. Schimel, M. Jung, L. Guanter, D. T. Drewry, M. Verma, A. Porcar-Castell, T. J. Griffis, L. Gu, T. S. Magney, P. Köhler, B. Evans, and K. Yuen. Oco-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. *Science*, 358(6360), 2017.

[26] J. G. Masek, E. F. Vermote, N. E. Saleous, R. Wolfe, F. G. Hall, K. F. Huemmrich, Feng Gao, J. Kutler, and Teng-Kui Lim. A landsat surface reflectance dataset for north america, 1990-2000. *IEEE Geoscience and Remote Sensing Letters*, 3(1):68–72, 2006.

[27] Amy McNally, Kristi Arsenault, Sujay Kumar, Shraddhanand Shukla, Pete Peterson, Shugong Wang, Chris Funk, Christa D. Peters-Lidard, and James P. Verdin. A land data assimilation system for sub-saharan africa food and water security applications. *Scientific Data*, 4(1):170012, Feb 2017.

[28] USDA. Usda national agricultural statistics service cropland data layer. published crop-specific data layer [online]. available at http://nassgeodata.gmu.edu/cropscape/ (accessed 13 aug 2020; verified 2 mar 2021). usda-nass, washington, dc., 2016.

[29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

A Additional Results

Results on new tiles. Table 2 presents results for fine pixels where not even coarse-resolution SIF was seen during training; it tests the model's ability to generalize to "gaps" where SIF observations were not directly available.

Table 2: Results on extrapolating CFIS to 30m resolution, on test tiles (model doesn't even know coarse-resolution SIF for these tiles).

Method	**NRMSE**	R^2
Ridge Regression	0.204	0.373
Gradient Boosting	0.215	0.296
ANN	0.226 $\pm$ *0.007*	0.224 $\pm$ *0.048*
CSR-U-Net	**0.187** $\pm$ *0.007*	**0.470** $\pm$ *0.038*

Results by land cover type. Table 3 compares the accuracy of the different methods' predictions for the most common land cover types within the study region (grassland, corn, and soybean). CSR-U-Net outperforms competing baselines for all of these land cover types, and provides an especially large improvement for grassland. CSR-U-Net is effective at understanding the unique SIF patterns for these highly distinct land cover types.

Table 3: NRMSE by land cover type, 30m pixels in train tiles (where 3km coarse-resolution labels were seen during training)

Method	**Grassland**	**Corn**	**Soybean**
Trivial: predict coarse	*0.255*	*0.225*	*0.278*
Ridge Regression	0.246	0.192	0.220
Gradient Boosting	0.238	0.195	0.244
ANN	0.260 $\pm$ *0.015*	0.230 $\pm$ *0.026*	0.239 $\pm$ *0.007*
CSR-U-Net	**0.197** $\pm$ *0.003*	**0.182** $\pm$ *0.009*	**0.214** $\pm$ *0.012*

Effect of resolution. Table 4 examines the performance of the different approaches for different resolutions. CSR-U-Net method outperforms competing baselines for most resolutions (30-300 meters). At coarser resolutions such as 600 meters, CSR-U-Net still performs well, but averaging-based methods such as Ridge Regression and ANN perform similarly well. This demonstrates that the averaging-based methods can provide reasonable results if the resolution is only being increased by a small factor; but for a large increase in resolution, CSR-U-Net is much better.

Table 4: NRMSE by resolution, train tiles (where 3km coarse-resolution labels were seen during training)

Method	**30m**	**90m**	**150m**	**300m**	**600m**
Trivial: predict coarse	*0.248*	*0.245*	*0.229*	*0.203*	*0.165*
Ridge Regression	0.213	0.211	0.191	0.166	**0.127**
Gradient Boosting	0.225	0.217	0.194	0.166	0.130
ANN	0.244 $\pm$ *0.006*	0.217 $\pm$ *0.005*	0.194 $\pm$ *0.004*	0.163 $\pm$ *0.004*	0.128 $\pm$ *0.004*
CSR-U-Net	**0.197** $\pm$ *0.003*	**0.199** $\pm$ *0.004*	**0.184** $\pm$ *0.003*	**0.160** $\pm$ *0.004*	0.128 $\pm$ *0.002*

Scatterplots. Figure 2 presents a scatterplot of the true SIF values versus the model predictions; this shows that the CSR-U-Net predictions align more closely with the ground truth compared to the ridge regression baseline.

Qualitative evaluation. Figure 3 presents an example of the predictions outputted by various measurements. All methods provide useful results; even the most basic linear regression method is able to identify fields with higher and lower SIF. However, the error maps on the top row show that CSR-U-Net's errors are lower in magnitude (such as in the lower right quadrant), and there are fewer dark blue or red pixels (which indicate large errors). In addition, CSR-U-Net is the only method that takes the spatial context of pixels into consideration – the other methods only use features from the pixel itself. Thus, the predictions outputted by CSR-U-Net are much more smooth and consistent across local regions.

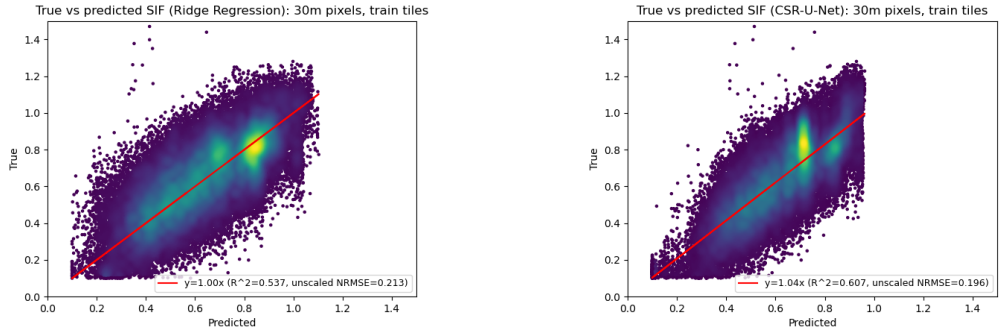

Figure 2: Ground-truth vs. predicted SIF for 30m pixels, train tiles.
Left: Ridge Regression baseline, **Right:** CSR-U-Net

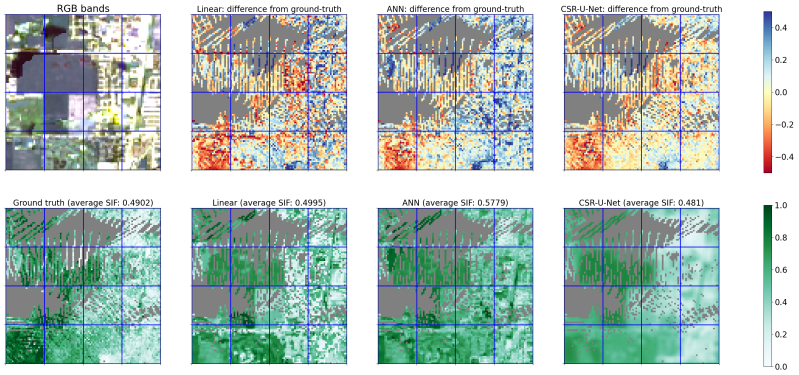

Figure 3: Example result. **Top-left:** RGB bands, used as input features. **Bottom-left:** ground-truth CFIS SIF (gray pixels are missing data). **Other bottom images:** predictions with different methods (CSR-U-Net is rightmost). **Other top images:** error maps (yellow is accurate, blue is over-prediction, red is under-prediction). CSR-U-Net is generally more accurate than other methods – for example, there are fewer dark red or blue pixels (which indicate major errors).

B Effect of Multiplicative Noise

Adding multiplicative noise is important for achieving better prediction accuracy. As described in the "Methods" section, we multiply all channels of the Landsat image by a constant $(1 + \epsilon)$, where $\epsilon \sim N(0, \sigma^2)$. $\sigma = 0$ means no noise is added. Table 5 shows the impact of adding multiplicative noise, holding other hyperparameters constant.

Table 5: Effect of multiplicative noise

Method	**NRMSE** (fine val. set)
ANN without mult. noise	*0.236*
ANN, mult. noise $\sigma = 0.2$	*0.233*
Ridge Regression without mult.noise	*0.214*
Ridge Regression, mult. noise $\sigma = 0.2$	*0.210*
CSR-U-Net without mult. noise	0.199
CSR-U-Net, mult. noise $\sigma = 0.1$	0.189
CSR-U-Net, mult. noise $\sigma = 0.2$	0.184
CSR-U-Net, mult. noise $\sigma = 0.3$	0.184

We tried values of multiplicative noise $\sigma \in \{0, 0.1, 0.2, 0.3\}$ - we found that the fine-resolution validation loss generally stopped improving after $\sigma = 0.2$, so we used that value for our experiments. Note that the difference between CSR-U-Net without multiplicative noise and with multiplicative

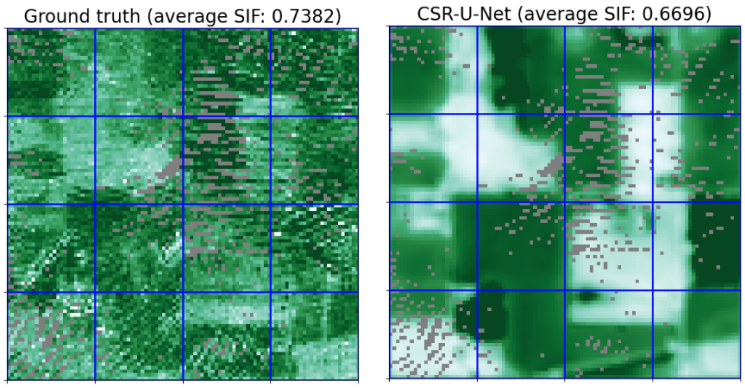

Figure 5: Example of overfitting. **Left:** ground-truth SIF map. **Right:** prediction by a U-Net that is over-trained. Note that the average tile SIFs are not too different, but the model tends to output extremely low and high values that do not reflect reality.

noise is significant (0.015 absolute NRMSE) and accounts for half of the improvement CSR-U-Net has over Ridge Regression (the best-performing baseline). This makes sense, since the ratios between bands are invariant to multiplying the whole image by a constant, and adding this noise forces the model to pay attention to the ratio between bands, which is known to be informative of vegetation growth. On the other hand, multiplicative noise does not seem to help the baselines as much, perhaps because they are already operating on highly averaged data.

C Importance of Early Stopping

As described in the paper, we find that early stopping (based on a small fine-resolution validation set) is critical to producing reasonable results. Without early stopping, the model can overfit in a way that is unique to coarsely-supervised regression tasks. For example, Figure 4 plots losses over time for a model run. Note that the coarse-resolution losses decrease continuously until around epoch 60, but the fine-resolution losses start increasing after epoch 20-30. (This is a relatively extreme example because we turned off weight decay and used a high learning rate, but the same effect occurs in most settings.) In later epochs, the model is making pixel predictions that produce the correct tile average SIF, but are inaccurate for the individual pixels.

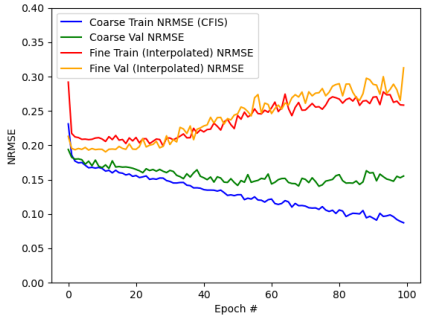

Figure 4: Losses over time. Note that the coarse-resolution losses (blue/green) continue decreasing for while, but the fine-resolution losses (red/orange) go up quickly after epoch 20-30, indicating overfitting to the coarse labels.

Specifically, we observed that over-trained models tend to output extreme maps; the model learns how to keep pushing the predictions for low SIF regions down and high SIF regions up, in a way that maintains the correct average. An example of this is shown in Figure 5.

Given the coarse nature of training labels, the only way for the model to avoid this is to look across multiple tiles, and ensure that pixels with similar features in different tiles have similar SIF predictions. In other words, the model needs to be well-regularized. So far, we found that early stopping (based on a fine-resolution validation set) and multiplicative noise are the most effective forms of regularization for our problem.

D Random Subset Regularization

As another regularization technique, only a random subset of valid pixels were averaged to obtain a predicted coarse SIF. This effectively means that every reasonably-large subset of pixels should have SIF averaging to the total SIF of the area; this is a stronger constraint than simply asking all pixels to average to the total SIF of the area. The effect of this was minor – it slightly improved the robustness of the model.

E Dataset Details

E.1 SIF Labels

The Chlorophyll Fluorescence Imaging Spectrometer (CFIS) [24] provides SIF observations at very high resolution, but only for extremely limited areas. We aggregate the CFIS observations into a 30×30 meter grid, matching the Landsat grid.

Table 6: Summary of SIF datasets used

Dataset	Resolution	# 3km labels in train set
CFIS	30 m	712
OCO-2	3 km	1390

At training time, we want to simulate a setting where only coarse-resolution SIF measurements are available. Thus we further aggregate these CFIS SIF pixels to a 3×3 km resolution; for each 3×3 km tile, we compute the average SIF across all 30-meter pixels in the tile that have at least 1 CFIS measurement. As a geographic coverage requirement, we only include 3×3 km tiles that contain at least 1000 pixels with CFIS data. We create a boolean mask to record which pixels actually had any CFIS observations.

At evaluation time, we compare our algorithms' fine-resolution SIF predictions with the ground-truth CFIS SIF labels at different resolutions, including $\{30, 90, 150, 300, 600\}$ meters. To reduce measurement noise in the fine-resolution SIF labels, we only evaluate on small pixels that have at least 30 measurements and have SIF > 0.1 (because low SIF values are difficult to measure accurately). At resolutions of greater than 30 meters, we also require that 90% of the 30m pixels within the larger pixel have at least 1 CFIS measurement.

OCO-2 [25] provides additional SIF measurements at a medium spatial resolution. The original footprints cover areas of around 1.3×2.25 km at nadir. To reduce noise, we grid them to a 3km resolution, and only include areas with at least 3 measurements.

Table 6 summarizes the SIF datasets used. For all SIF datasets, we again remove tiles with average SIF below 0.1 to reduce noise. In addition, to correct between the differing wavelengths of the different instruments, we multiply OCO-2 SIF values by 1.11 to match CFIS. (This scaling factor was empirically determined by fitting a model on OCO-2, evaluating it on CFIS, and determining what scaling factor produced the optimal fit.)

E.2 Study Region and Time Range

All tiles in our dataset correspond to a coarse-resolution CFIS or OCO-2 SIF label that lies within the Midwest US, from 38 to 48.7 degrees N and 108 to 82 degrees W. The region is plotted in Figure 6. As shown in the figure, even coarse-resolution SIF datasets are sparse geographically, so machine learning is needed to fill in the gaps and predict at a higher resolution. We extracted input feature data (reflectance, crop cover maps, etc.) from the same time periods as the CFIS SIF observations: June 15-29, 2016 and August 1-16, 2016. We also used OCO-2 observations from those time periods.

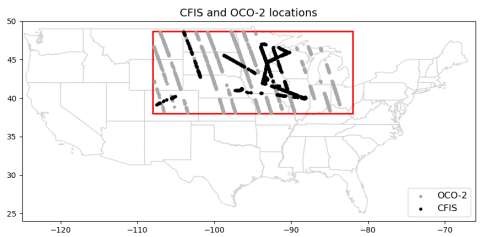

Figure 6: Locations of CFIS (black) and OCO-2 (dark gray) tiles in our study region (red box).

E.3 Input Features

Landsat Surface Reflectance. Satellite imagery is available from the Landsat satellite at a fine resolution of 30×30 meters, every 16 days [26], and can be downloaded from Google Earth Engine. For each pixel, there are 9 values (representing different bands). We zero out pixels with missing data (where the Landsat QA band is False; e.g. if there are clouds). We create a binary mask of "missing reflectance pixels" (which is 1 where reflectance is missing, 0 otherwise) to inform the model of which Landsat data is missing. We only include tiles where less than 50% Landsat pixels are missing.

Table 7: Summary of input features used

Dataset	# vars
Landsat surface reflectance Blue, green, red, near infrared, etc.	10
FLDAS land data Rainfall, temperature, radiation	3
CDL land cover types (binary masks) Corn, soybean, grassland, forest, etc.	11

FLDAS land data. Features such as temperature, rainfall, and surface radiation are available from FLDAS (Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System), for each 10×10 km pixel [27]. We linearly interpolate these variables to a 30-meter resolution to match the other variables.

Land cover data. Land cover data is available from the Cropland Data Layer [28] at a fine resolution of 30×30 meters, and can be downloaded from Google Earth Engine. Each pixel is labelled with a land cover type. For each land cover type that makes up more than 1% of the dataset, we created a binary mask of which pixels were covered by that type. We removed tiles where less than 50% of the tile is covered by one of the common land cover types. See Appendix for a full list of features.

E.4 Data Preprocessing

There are a total of 24 input features, as described in Table 7: 9 Landsat reflectance bands (plus a binary mask indicating missing Landsat data), 3 FLDAS variables (temperature, rainfall, surface radiation), and 11 land cover binary masks. All of these features were resampled to a fine resolution of 30 meters. For each coarse (3km) SIF label, we extracted the corresponding tile of feature data. These tiles are 3-dimensional tensors, of size 24 (features) $\times 100 \times 100$ (pixels). We standardized the continuous variables to have zero mean and unit variance, and clipped them to $[-3, 3]$ standard deviations from the mean. (We do not transform the binary masks.)

After filtering out tiles with missing data, there are a total of 1186 large tiles with enough CFIS SIF labels (both coarse-resolution for the entire tile, and fine-resolution for some 30m pixels), and 2382 large tiles with a coarse-resolution OCO-2 SIF label.

We randomly assign each tile to one of 3 sets: train (60%), validation (20%), and test (20%); we ensure that CFIS and OCO-2 tiles that overlap end up in the same set. The model is trained on coarse-resolution (tile-level) SIF labels from the train tiles; there are a total of 712 CFIS and 1390 OCO-2 tiles in the train set (see Table 6). Then, we select hyperparameters based on its predictive accuracy on the *fine-resolution validation* set. We evaluate the model against the true fine-resolution SIF observations on both the train and test sets, in order to see how accurate the fine-resolution predictions are in settings when the coarse-resolution SIF is known (train), and when the coarse-resolution SIF is unknown (test).

E.5 List of Features

Here is a full list of input features used in our tiles. Each 30m pixel has a value for each of these features.

Landsat surface reflectance [26]:

1. Ultra blue surface reflectance (435-451 nm)
2. Blue surface reflectance (452-512 nm)

3. Green surface reflectance (533-590 nm)
4. Red surface reflectance (636-673 nm)
5. Near infrared surface reflectance (851-879 nm)
6. Shortwave infrared 1 surface reflectance (1566-1651 nm)
7. Shortwave infrared 2 surface reflectance (2107-2294 nm)
8. Band 10 brightness temperature (10.60-11.19 μm)
9. Band 11 brightness temperature (11.50-12.51 μm)
10. Missing reflectance mask (1 if reflectance data is missing, e.g. due to cloud cover)

FLDAS land data features [27]:

1. Rainfall flux (kg m-2 s-1)
2. Surface downward shortwave radiation (W m-2)
3. Surface air temperature (K)

Cropland Data Layer land cover types [28]:

1. Grassland/pasture
2. Corn
3. Soybean
4. Deciduous Forest
5. Evergreen Forest
6. Developed/Open Space
7. Woody Wetlands
8. Open Water
9. Alfalfa
10. Developed/Low Intensity
11. Developed/High Intensity

F Training Details

We train on one NVIDIA Tesla V100 GPU with 16GB memory, on the Linux CentOS 7 operating system. For CSR-U-Net, training a model for 100 epochs takes roughly 30 minutes. We used the following libraries with Python 3.7: Matplotlib 3.3.4, Numpy 1.18.1, Pandas 1.1.3, PyTorch 1.7.0, Scikit-Learn 0.24.1, Scipy 1.4.1. For all methods that involve randomness, we report the average and standard deviation using three random seeds: {0, 1, 2}.

For the baseline methods, we did a grid search over hyperparameters, and chose the configuration that performed best on the fine-resolution validation set. For Ridge Regression, we selected the regularization parameter α from $\{0.01, 0.1, 1, 10, 100, 1000, 10000\}$; we chose $\alpha = 100$. For Gradient Boosting Regressor, we selected the maximum number of iterations from $\{100, 300, 1000\}$, and the maximum depth of the tree from $\{2, 3, None\}$. We chose 100 iterations and max depth of 2. For the fully-connected artificial neural network, we selected hidden layer sizes from $\{(100), (20, 20), (100, 100), (100, 100, 100)\}$, initial learning rate from $\{10^{-2}, 10^{-3}, 10^{-4}\}$, and set the maximum number of iterations to 10,000. We chose hidden layer sizes of $(100, 100, 100)$ and an initial learning rate of 0.001.

For CSR-U-Net, we used the AdamW optimizer [29], and multiplicative noise standard deviation $\sigma = 0.2$ as described earlier. We tried batch sizes of $\{50, 64, 100\}$, and the effect was minimal; we used a batch size of 64. We also used "random subset regularization", where we take the average of a random 20% of valid pixels (instead of all of them) when comparing with the tile SIF; this helped a little. In terms of other augmentations, we used random flip/rotate, jigsaw (cutting the image vertically and swapping the two sides, and same horizontally), and random erase [22] with dimensions of 20×20 and probability of 0.5.

Then we did a grid search, considering learning rates from {1e-4, 2e-4, 5e-4, 1e-5} and weight decay from {1e-4, 3e-4, 1e-3}. We found that a learning rate of 2e-4 and weight decay of 1e-4 produced the best NRMSE on the fine-resolution validation set.

In terms of model architecture, we used a smaller version of the U-Net architecture [30], with 2 downsampling and 2 upsampling blocks, with {64, 128, 256} hidden units. We start with a 1x1 convolution for a pixel encoder, followed by a rectified linear unit (ReLU), and then 2 downsampling blocks. Each downsampling block consists of the following sequence: (2×2 average pooling, convolutional layer with filter size 3, ReLU, convolutional layer with filter size 1, ReLU). Note that reducing the second convolutional layer's filter size to 1 reduces the receptive field of each pixel and ensures better localization. We found that batch normalization caused training to be more unstable, so we removed it.