A WEAKLY-SUPERVISED, MULTITASK DEEP LEARNING FRAMEWORK FOR SHADOW MITIGATION IN REMOTE SENSING IMAGERY

Scott D. Couwenhoven[a], Emmett J. Ientilucci[b], Byung H. Park[c], and David Hughes[c]

[a]Rochester Institute of Technology, School of Mathematical Sciences, Rochester, NY, USA
[b]Rochester Institute of Technology, Center for Imaging Science, Rochester, NY, USA
[c]Oak Ridge National Laboratory, Oak Ridge, TN, USA

ABSTRACT

We propose a weakly-supervised, multitask framework for training a convolutional neural network to solve the problem of cloud shadow mitigation given only cloud and shadow masks as labels. The network minimizes the Wasserstein distance between shadows and their proximal sunlit neighborhoods, generating a supervisory signal directly from within the input image. We extract further utility from the shadow mask through multitask learning by introducing an auxiliary task of shadow segmentation. Our approach is advantageous since it performs mitigation in an end-to-end framework which requires only a shadowed image for inference. We apply this process to the Landsat 8 OLI SPARCS validation data set and demonstrate plausible results.

Index Terms— cloud shadow mitigation, deep learning, weak supervision, multitask learning, satellite imagery

1. INTRODUCTION AND OVERVIEW

Cloud shadow mitigation is the task of reversing the radiometric impact of a cloud shadow such that the shadowed and sunlit portions of an image appear to be lit by a consistent source. The proposed framework aims to train a convolutional neural network (CNN) to perform this task in an end-to-end manner so that inference can be performed given a partially shadowed image and no additional inputs.

Ideally, this would be conducted under the paradigm of supervised learning. However, obtaining the required labeled data is infeasible in the context of cloud shadow mitigation. We address this issue through *weak supervision*, using sunlit portions from within the input image to serve as pseudo-labels. We train the network to minimize the Wasserstein distance between the radiometric distributions of a shadowed region and its proximal sunlit neighborhood in an image. This assumes that the shadowed region and its proximal sunlit neighborhood have the same scene content. As this is an approximation, we classify this approach as weakly supervised.

For a given shadow, we define its proximal sunlit neighborhood as the set of sunlit pixels within a user-specified radius of the shadowed region. Computing this requires *a priori* knowledge of the cloud and shadow locations, which are provided in the form of segmentation masks. These are only used during training and, other than partitioning imagery, are not used directly in supervision. We incorporate the availability of masks into the learning framework through multitask learning, training the CNN not only to estimate a shadow mitigated image but also an estimate of the shadow mask. This ensures a more robust mapping of a shadow pixels to sunlit pixels by imposing an additional regularizing constraint.

2. RELATED WORK

Cloud shadow mitigation can be viewed as a special case of cloud and cloud shadow removal, which is the general task of replacing cloud-contaminated regions of imagery. There are several deep learning based methods for cloud and cloud shadow removal in the literature, but most require the use of external information as opposed to recovering the information present in the original input image. For example, [1] proposes an iterative, spatio-temporal CNN framework which requires the use of multi-temporal data. [2] leverages a deep residual network to perform EO-SAR data fusion to recover ground scene content. The literature related to deep learning approaches to cloud shadow mitigation is relatively unexplored, especially when assuming only the original image as input.

3. METHODS

Fig. 1 represents an overview of our weakly-supervised, multitask framework employed during training. The neural network is comprised of three components: a shared encoder, G, and two task-specific decoders, D_1, D_2, tasked

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

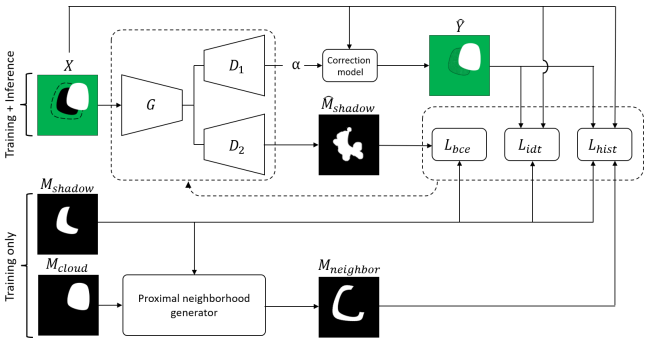

Fig. 1. Schematic of our framework used to train the CNN. The three major loss terms computed are binary cross-entropy L_{bce}, identity L_{idt}, and histogram L_{hist}.

with mitigation and segmentation, respectively. We define X as the partially shadowed image with N_c channels. M_{shadow}, M_{cloud}, and $M_{neighbor}$ are binary masks denoting the shadowed, clouded, and proximal sunlit neighborhood pixels, respectively. $\hat{M}_{shadow} = D_2(G(X))$ is the output of a sigmoid activation and represents a semantic prediction of M_{shadow}. $\alpha = D_1(G(X))$ is the output of a ReLU activation and represents a per-pixel, per-channel adjustment which is applied through the correction model in Eq. (1), inspired by [3], to generate a prediction of the mitigated image, $\hat{Y}$

$$\hat{Y} = (\alpha + 1)X \tag{1}$$

3.1. Loss Formulation

For notation, the region of an image indexed by a mask will be represented with []. For example, $X[M_{shadow}]$ represents the region of X corresponding to M_{shadow}.

The network is tasked with optimizing three separate loss terms. The primary loss, denoted L_{hist}, is generated by computing the Wasserstein $\mathcal{L}_1$ distance, per channel, between the code value distributions of the shadowed region in the mitigation prediction, $\hat{Y}[M_{shadow}]$, and the proximal sunlit neighborhood in the input, $X[M_{neighbor}]$. To compute this distance in a differentiable manner, the cumulative distribution functions (CDF) are determined via the methods of [4], using their suggested hyper-parameters. We compute the loss across $k = 256$ bins and across each image channel as

$$L_{hist} = \sum_{i=1}^{N_c} \sum_{j=1}^{k} |CDF_j(\hat{Y}_i[M_{shadow}]) - CDF_j(X_i[M_{neighbor}])| \tag{2}$$

We also introduce an identity loss, L_{idt}, which penalizes adjustments outside of the shadowed region. The sunlit region is defined as $M_{sunlit} = 1 - M_{shadow}$. L_{idt} is then computed as the mean squared error between $\hat{Y}[M_{sunlit}]$ and $X[M_{sunlit}]$.

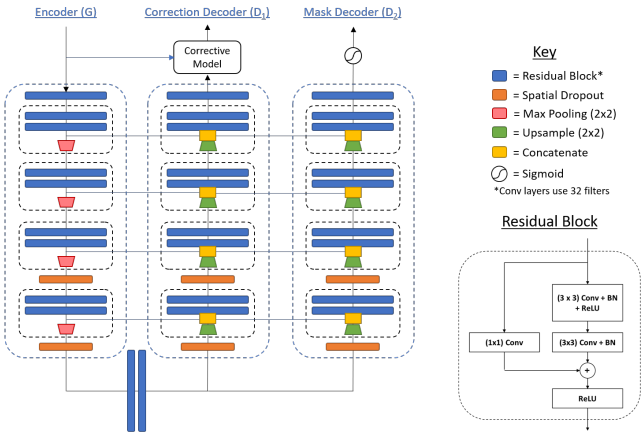

Fig. 2. The CNN architecture employed. A standard residual U-net design inspired by [5, 6] with hard parameter sharing in the encoder, G.

We leverage multitask learning by having the auxiliary decoder, D_2, predict the shadow mask. We enforce hard-parameter sharing in the network encoder which helps produce more robust features and reduce the risk of over-fitting the training data. The loss for this task, L_{bce}, is computed as the binary cross-entropy between M_{shadow} and $\hat{M}_{shadow}$.

The total loss, L_{total}, is a weighted summation of the histogram, identity, and binary cross-entropy losses with an additional regularization constraint on the $\mathcal{L}_1$ norm of the network weights, W, as

$$L_{total} = L_{hist} + \lambda_{idt} L_{idt} + \lambda_{bce} L_{bce} + \lambda_{reg} ||W||_1 \tag{3}$$

The loss coefficients are empirically set as $\lambda_{idt} = 10^3$, $\lambda_{bce} = 4$, and $\lambda_{reg} = 10^{-4}$ so that all loss terms have approximately the same order of magnitude.

3.2. Proximal Sunlit Neighborhood

The proximal sunlit neighborhood is computed by a binary dilation of M_{shadow} for r iterations. We then use M_{shadow} and a cloud mask, M_{cloud}, to exclude shadowed and clouded pixels, respectively, resulting in a ring of sunlit pixels around the perimeter of the shadow. The optimal value of r is dependent upon the ground sample distance of the imagery. In practice, it is set empirically to provide enough samples to form a representative distribution.

If a sample contains multiple shadows, the proximal sunlit neighborhood, as defined, will be the union of the individual sunlit neighborhoods. This effectively groups all shadowed regions together, and means that they may not be compared directly with their corresponding neighborhoods. Correcting this was intractable from a computational standpoint, but can be mitigated by reducing tile size.

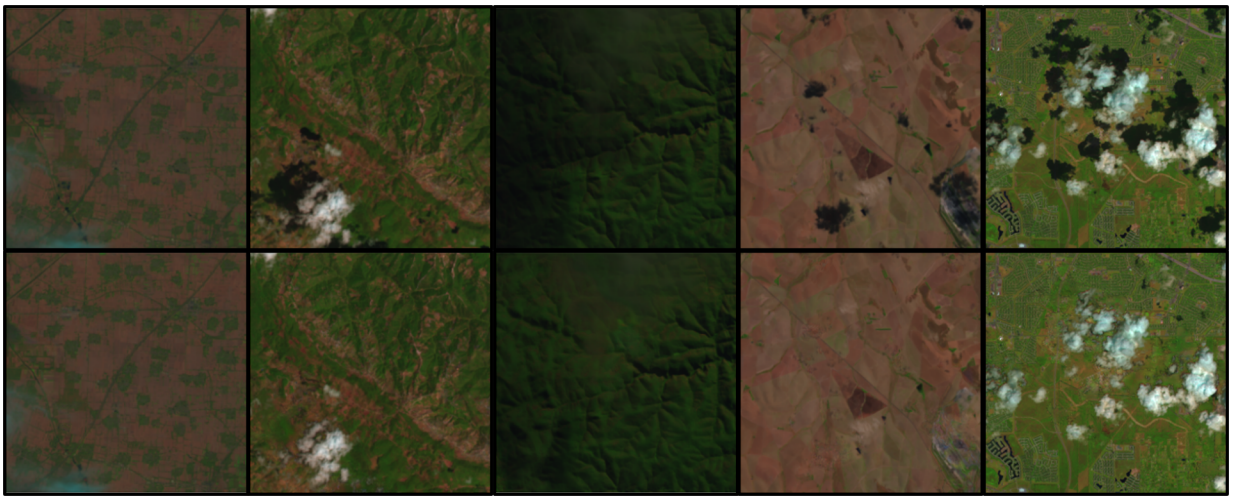

Fig. 3. Results from the proposed procedure applied to the validation partition of the SPARCS data set [7]. The top row represents the input (X) while the bottom row represents the resulting mitigation ($\hat{Y}$).

3.3. Architecture

The network architecture, shown in Fig 2, is based upon a U-net architecture with depth 4 with residual blocks [6, 5]. In an effort to limit the trainable parameters while maintaining the receptive field provided by a depth 4 U-net, all convolutional layers (excluding output layers) have 32 filters. Additionally, bilinear interpolation is selected as the up-sampling operation. Spatial dropout is used to discourage over-fitting and is applied with a rate of 0.5. The total architecture contains 640,000 trainable parameters.

4. RESULTS

4.1. Experimentation

For experimentation, the proposed method is applied to the Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) Validation data set [7], which contains 80 Landsat-8 OLI images of size 1000×1000 pixels. For the results presented, we utilize false-color renderings with bands 6, 5, and 4 mapped to the red, green, and blue channels, respectively. The proposed method is capable of running over an arbitrary number of spectral bands. The decision to use these three is made for ease of visibility, as in [8], with a secondary motive of computational efficiency.

For training, we partition the data into 256×256 chips, 879 of which contain cloud shadows. We randomly select 80% of the data for training, applying rotation and flip data augmentation, and reserve the remaining 20% for validation. The network is trained for 500 epochs using an Adam optimizer [9] with an initial learning rate of 10^{-4}. Proximal neighborhoods are computed using a radius of $r = 10$ pixels.

4.2. Evaluation

4.2.1. Mitigation

Due to the lack of ground truth, the most effective way to evaluate results at this time is visually. A selection of validation results can be seen in Fig 3. In general, the results demonstrate plausible mitigation across a wide variety of cloud and shadow characteristics and scene content. In particular, the result in the middle column demonstrates the ability to distinguish between shadows caused by clouds and those caused by ground terrain (hills, in this case).

Despite encouraging results, there are still a few areas for improvement. In particular, partially clouded regions tend to be over-corrected. Additionally, despite heavy constraints from the multiplicative correction model, the mitigated shadow regions tend to exhibit a fair amount of blurring compared to the sunlit portions. Currently, there is no term in the loss function which accounts for this and we plan on addressing this in our work moving forward.

4.2.2. Segmentation

The primary function of the shadow segmentation decoder, defined as D_2 in Fig 1, is to provide multitask context to enhance the feature space produced by the encoder. As a result, we do not discuss its performance in great detail. On validation data, we report a peak Intersection over Union of $IoU = 0.686$.

4.3. Corrective Model and Constraints

The objective of minimizing the Wasserstein distance between shadowed regions and their proximal sunlit neighborhoods is under-constrained. In particular, any spatial

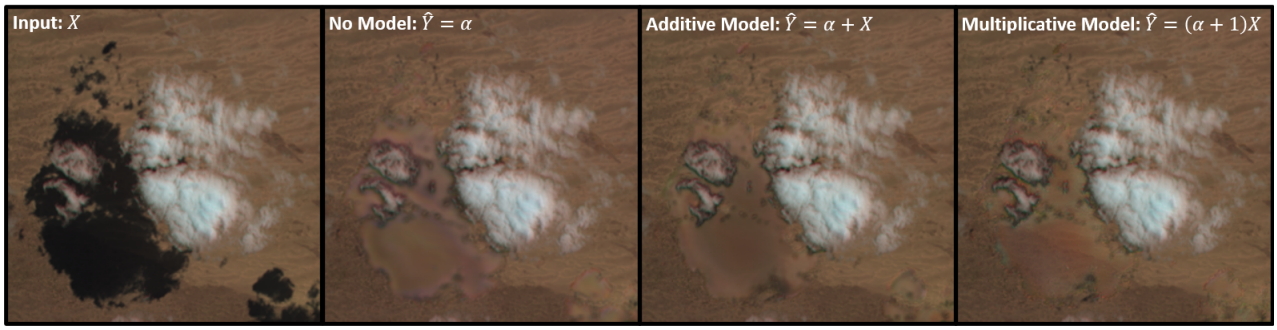

Fig. 4. A comparison of results from identical experiments using different correction models. All solutions yield similar histogram matches despite massive differences in mitigation quality.

permutation of the shadow region will produce an equivalent response as measured by this objective. To discourage spatial permutations, we apply heavy constraints through network weight regularization and a corrective model.

Three forms of corrective model are considered: no model, additive, and multiplicative. Fig. 4 demonstrates subjectively that the multiplicative model better preserves spatial features. We conjecture that this is because the multiplicative model provides a more convenient way of restoring contrast in the shadow regions compared to the others.

Theoretically, each correction model is capable of performing radiometric restoration, at least to a first order. This may seem counter-intuitive, but keep in mind that the corrective model ultimately imposes constraints on the network output, α, which is the result of a complex, context dependent non-linear mapping. Changing the correction model alters the interpretation of α accordingly.

5. CONCLUSIONS

We propose a novel framework for training a convolutional neural network to perform cloud shadow mitigation in satellite imagery *in spite* of the lack of directly labeled data. Our approach is advantageous in multiple ways: 1) It is an end-to-end framework which only requires a shadowed image as input for inference. 2) For training, the only labels required are cloud and cloud shadow segmentation masks, which are feasibly obtained. 3) It can be generalized for arbitrary multispectral data. The algorithm is still considered a work in progress and there are issues to be addressed in our future efforts. Of chief concern is constraining the network output. We still require a solution which preserve structure in the mitigated shadow regions while providing sufficient constraints to ensure the histograms are matched in a desirable fashion.

6. REFERENCES

[1] Qiang Zhang, Qiangqiang Yuan, Jie Li, Zhiwei Li, Huanfeng Shen, and Liangpei Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 148–160, 2020.

[2] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[3] Hui Fan, Meng Han, and Jinjiang Li, "Image shadow removal using end-to-end deep convolutional neural networks," *Appl. Sci. (Basel)*, vol. 9, no. 5, pp. 1009, Mar. 2019.

[4] Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv, "Deephist: Differentiable joint and color histogram layers for image-to-image translation," 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[7] M Joseph Hughes, "L8 SPARCS cloud validation masks," 2016.

[8] M. Joseph Hughes and Robert Kennedy, "High-quality cloud masking of landsat 8 imagery using convolutional neural networks," *Remote Sensing*, vol. 11, no. 21, 2019.

[9] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.