Auto-CM: Unsupervised Deep Learning for Satellite Imagery Composition and Cloud Masking Using Spatio-Temporal Dynamics

Yiqun Xie[1*], Zhili Li[1*], Han Bao[2], Xiaowei Jia[3], Dongkuan Xu[4], Xun Zhou[2], Sergii Skakun[1]

[1]University of Maryland
[2]University of Iowa
[3]University of Pittsburgh
[4]North Carolina State University
{xie, lizhili, skakun}@umd.edu, {han-bao, xun-zhou}@uiowa.edu, xiaowei@pitt.edu, dxu27@ncsu.edu

Abstract

Cloud masking is both a fundamental and a critical task in the vast majority of Earth observation problems across social sectors, including agriculture, energy, water, etc. The sheer volume of satellite imagery to be processed has fast-climbed to a scale (e.g., >10 PBs/year) that is prohibitive for manual processing. Meanwhile, generating reliable cloud masks and image composites is increasingly challenging due to the continued distribution-shifts in the imagery collected by existing sensors and the ever-growing variety of sensors and platforms. Moreover, labeled samples are scarce and geographically limited compared to the needs in real large-scale applications. In related work, traditional remote sensing methods are often physics-based and rely on special spectral signatures from multi- or hyper-spectral bands, which are often not available in data collected by many – and especially more recent – high-resolution platforms. Machine learning and deep learning based methods, on the other hand, often require large volumes of up-to-date training data to be reliable and generalizable over space. We propose an autonomous image composition and masking (Auto-CM) framework to learn to solve the fundamental tasks in a label-free manner, by leveraging different dynamics of events in both geographic domains and time-series. Our experiments show that Auto-CM outperforms existing methods on a wide-range of data with different satellite platforms, geographic regions and bands.

Introduction

Clouds are highly frequent and pervasive atmospheric phenomena lying between satellite sensors and the surface of the Earth. As a result, cloud masking and image composition are among the most fundamental tasks in satellite-based Earth observation, and have a direct impact on the vast majority of important downstream Earth observation applications across social sectors, such as crop monitoring, solar energy budgeting, water resource surveillance, disaster response, carbon emission monitoring, etc. Given the societal importance of these use cases, satellite-based platforms have undergone many revolutions and the imagery is being collected at an ever-growing resolution, scale, frequency, and variety. For example, NASA's Earth Observing System Data and Information System (EOSDIS) collects data at 12 PBs/year by 2020, and the total volume is projected to grow from 42 PBs to 250 PBs by 2025 with new sensors. Similarly, commercial platforms such as Planet constellations can scan the entire Earth on a daily basis, and its SkySat program can capture every location on Earth seven times a day at 0.5m resolution. This sheer volume of data has fast-climbed to a scale that is prohibitive for manual processing, making it critical to develop robust and efficient techniques for cloud masking and cloud-free image composition.

Despite the importance and broad impact, the problem is challenging from several aspects. First, ground-truth samples of clouds have limited availability, especially considering the volume and variety of both existing and incoming satellite data that are needed to meet the demand of large-scale applications. Moreover, unlike other geospatial objects such as buildings and roads, clouds are constantly-moving targets, which means labels are limited to a single snapshot and not usable for future data. Second, the training samples are geographically constrained to specific locations and only cover a very tiny portion of the Earth, making learned models hard to generalize in the heterogeneous Earth surface (Xie et al. 2021; Goodchild and Li 2021; Karpatne et al. 2018). Third, the data distribution is non-stationary due to changes in the Earth's surface environments, resolution (e.g., very-high-resolution imagery), sensors (e.g., new platforms), and more. In addition, the availability and choices of spectral bands often vary across sensors. For example, lower-resolution imaging platforms such as Landsat-8 and Sentinel-2 often have broader (but different) spectrum coverage, whereas higher-resolution imagery from platforms such as SkySat may only have visible RGB channels and near-infrared. There are several directions in related work:

Physics-based methods. Traditional approaches from remote sensing – including Fmask (Qiu, Zhu, and He 2019; Zhu, Wang, and Woodcock 2015; Zhu and Woodcock 2012), LaSRC (Skakun et al. 2019), Sen2Cor (Main-Knorn et al. 2017), MAJA (Hagolle et al. 2010) and others (Tarrio et al. 2020; Baetens, Desjardins, and Hagolle 2019) – often utilize physical modeling of the interactions between clouds and spectral signals to derive their signatures from multi- or hyper-spectral imagery. Physics-based methods enjoy good performance when the satellite product contains the required

*These authors contributed equally.

signals (e.g., thermal) and the physical assumptions are met (e.g., existence of measurable parallax) (Skakun et al. 2019; Frantz et al. 2018); for example, the Fmask algorithm works well for Landsat imagery (Foga et al. 2017). However, the conditions are only met for a few products as many sensors are not designed to capture such spectral details. For example, Sentinel-2, another major multispectral imagery collection platform, does not have thermal bands and this is known to reduce cloud masking performance (Tarrio et al. 2020). This is more common for more recent higher-resolution constellations (e.g., from Planet or Maxar), where many only have four bands: visible (i.e., RGB) and near-infrared.

Machine learning based methods. Various learning techniques have been studied for cloud masking. Earlier attempts mainly focused on traditional machine learning methods such as decision trees, Bayesian methods, SVM, random forests, and more (Hollstein et al. 2016; Li et al. 2015; Wei et al. 2020). Recent developments have switched towards deep learning models, including both CNN-based scene-level classification (Shendryk et al. 2019) frameworks and semantic segmentation with residual learning (Wieland, Li, and Martinis 2019). Generative adversarial networks (GANs) were also used to generate cloud-free scenes (Singh and Komodakis 2018). However, the methods rely on labeled ground truth to deliver reliable performance, which has limited availability and is extremely time-consuming to collect for large-scale applications. As identified by many evaluations, the constraints caused the learning methods to be highly-sensitive to the imagery conditions and can hardly generalize beyond the training samples for broad scenarios (Tarrio et al. 2020). There have also been efforts to reduce the demand of pixel-level labels, such as unpaired learning for image dehazing (Yang, Xu, and Luo 2018; Yang et al. 2022), and GAN variants, such as CycleGAN for single-scene based synthetic cloud removal (Zi et al. 2021). These methods, however, still require scene-level labels, which are challenging to obtain and update for the scale (e.g., global scale) of existing and new Earth observation data.

Other related directions. In addition to the above directly related studies, another related direction is the unsupervised species of intrinsic image decomposition (Li and Snavely 2018; Shen et al. 2011; Yi, Tan, and Lin 2020), which splits an image into a reflectance and a shadow layer. However, the cloud masking problem does not follow the assumptions of the model, including constant reflectance and smoothness. For example, the Earth surface is a dynamic environment with changes in temperature, humidity, sun angle, vegetation, reflectance, etc. Thus, images captured at different timestamps and dates at the same location often exhibit lots of differences in conditions. Moreover, intrinsic image decomposition focuses on changes in lights and does not consider events with blockage of views. Finally, unsupervised domain adaptation methods can learn domain-invariant features. However, as Earth data are highly heterogeneous across space, time and sensors, such features are hard to adapt to very different distributions (Kothandaraman et al. 2022), as we will show in the experiments.

We propose an Autonomous image Composition and Masking (Auto-CM) framework to learn to perform the fundamental task in a completely label-free manner. Specifically, our contributions are:

- We present a DISTANCE prior and corresponding spatio-temporal data representation of satellite imagery.
- We propose a deep learning framework that captures the clouds based on the differences in the spatio-temporal dynamics of the atmospheric events and land surface, without using any labels.
- The framework is very simple to implement but highly robust and self-adaptive for different regions and sensors.
- We perform extensive experiments that cover different geographic regions, sensing platforms and spectral bands.

The experiment results show that the proposed Auto-CM framework offers promising performance and improvements for cloud masking across diverse scenarios. The results are also similar to those from supervised models for test data that are very similar to training and better for test data with different distributions.

Problem Formulation

Definition 1 *A **satellite image tile** I is a full scene captured by a satellite at a location s and timestamp t in the orbit.*

Definition 2 ***Image composites*** I^{com} *(e.g., weekly or monthly composites) are cloud-free images generated from a time-series of image tiles.*

The input to the cloud-masking problem is a time-series of satellite image tiles $S_{img} = \{I_i\}$ covering a common spatial region. The output is a cloud mask M_i for each image tile in S_{img}. We also output one complimentary cloud-free image composite I_T^{com} for a set of image tiles in a local time window of the time-series. While the training uses a time-series of images, a learned model can make classifications using a single snapshot.

Method

We propose an Auto-CM framework to learn to generate cloud masks in a completely label-free manner. In the following, we first introduce a DISTANCE prior for the spatio-temporal (ST) dynamics. Then, we present the corresponding ST data representation, and DISTANCE-informed model designs. Finally, we discuss test-time generalization.

Spatio-Temporal Dynamics by A DISTANCE Prior

As our Auto-CM framework does not rely on any labeled samples to learn cloud masking, it is important to design a mechanism to guide the learning process. Ideally, the optimal solution trained from such a mechanism should correspond to all cloud pixels being masked and non-cloud pixels passing through. To "approximate" such a mechanism, we introduce a Difference In the Spatio-TemporAl dyNamiCs of Events (DISTANCE) prior as follows:

Definition 3 ***Events*** E *are different types of phenomena or processes happening on Earth. Here we consider two major types of dynamic events: (1) atmospheric events E_A,*

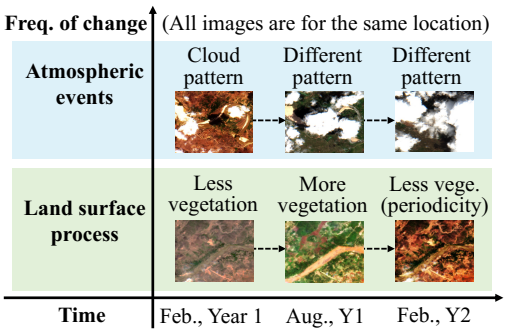


Figure 1: Spatio-temporal dynamics.

which are mainly represented by clouds in satellite imagery, and (2) land surface processes E_S, which include changes in land covers (e.g., vegetation growth), environment conditions (e.g., temperature), etc.

Definition 4 *The **DISTANCE prior** focuses on the differences in the dynamics of the events (i.e., E_A and E_S) in space and time:*

- ***Differences in temporal dynamics:** Comparing E_A and E_S, clouds in the atmosphere are constant-changing events, where the pattern tends to be highly different in different image tiles covering the same location. In contrast, surface processes E_S change at a much lower frequency (e.g., it may take weeks for vegetation to change). Finally, changes in E_S exhibit temporal periodicity whereas those in E_A do not (Fig. 1).*
- ***Differences in spatial dynamics:** As the changes in clouds E_A are constant and have high-degrees of local randomness, the patterns of clouds are different both across different locations at the same timestamp, and across different timestamps at the same location. In contrast, E_S results in different patterns (e.g., different land cover types or layouts) at different locations, but the expressions are similar in adjacent timestamps for the same location.*

The DISTANCE prior represents the key differences in ST-dynamics of E_A and E_S. In the next sections, we will leverage the prior to design the corresponding data representation and network structures, which together establish the desired mechanism to guide the label-free training.

Spatio-Temporal Data Representation

Fig. 2 shows the overall ST-representation of the input data to facilitate the needs of the DISTANCE prior. We create local ST-packs of data based on the following definitions:

Definition 5 *A **temporal pack** P_T is a local subset of L images in a time-series, where $T = \{t_j, t_{j+1}, ..., t_{j+L-1}\}$. L needs to be sufficiently large so that each pixel in the image is not blocked by clouds in at least one of the images in P_T; it is not a hard constraint and it is okay to have pixels covered by clouds in a pack P_T. This will introduce noises in the training but is not expected to have a major impact as long as it only accounts for a small proportion.*

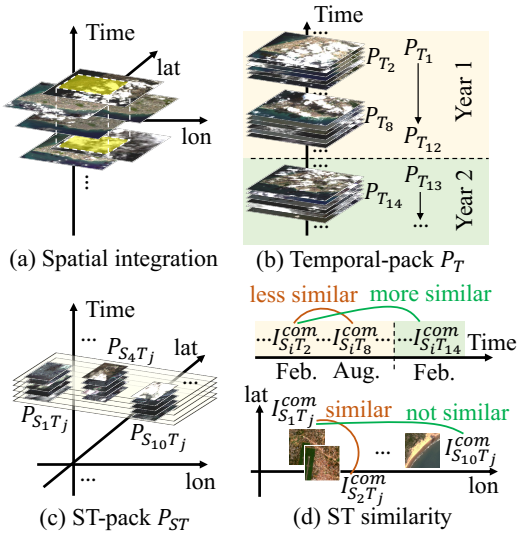


Figure 2: ST data representation.

Definition 6 *A **ST-pack** P_{ST} is a temporal-pack P_T in a local spatial window, where S has a size of $W \times W$ (measured by pixels). ST-packs P_{ST} have overlaps in both space and time to be able to represent the DISTANCE prior and evaluate similarity between them:*

- ***Temporal similarity** sim_T: This is evaluated between the composites I^{com} (Def. 2) generated from two ST-packs $P_{S_iT_j}$ and $P_{S_iT_{j'}}$ for the same spatial window. The similarity $sim_T(I^{com}_{S_iT_j}, I^{com}_{S_iT_{j'}})$ is expected to gradually increase as $dist_T(T_j, T_{j'})$ increases, where the distance function $dist_T(\cdot,\cdot)$ needs to include corrections for temporal periodicity (e.g., Fig. 1 and 2), which will be shown in Eq. (3).*
- ***Spatial similarity** sim_S: This is evaluated between composites I^{com} for the same temporal range T_j but different spatial windows S_i and $S_{i'}$. The value of $sim_S(I^{com}_{S_iT_j}, I^{com}_{S_{i'}T_j})$ is expected to increase as $dist_S(S_i, S_{i'})$ increases, where $dist_S(\cdot,\cdot)$ evaluates the spatial overlaps between two windows. sim_S reaches the maximum value for completely overlapping pairs.*

DISTANCE-Informed Design of Training

We use the DISTANCE prior to guide the training process of Auto-CM without ground-truth labels. Fig. 3 shows the general design of the network architecture, where the CNN component can be a user-selected backbone. As we can see, the DISTANCE-informed model is easy to implement.

Overall network flow. The inputs of the network include four ST-packs P_{ST} split into two pairs: $(P_{S_1T_1}, P_{S_2T_2})$ and $(P_{S_3T_3}, P_{S_4T_4})$. There are two requirements on the pairs: (1) Within each pair, we keep either the spatial window S or temporal range T the same (not both) between the two packs, and vary the other. For example, we can have $S_1 = S_2$ and $T_1 \neq T_2$. (2) Between the two pairs, the choice of varying S or T must be consistent. For example, we cannot have $S_1 = S_2$ but $S_3 \neq S_4$. Satisfying the requirements

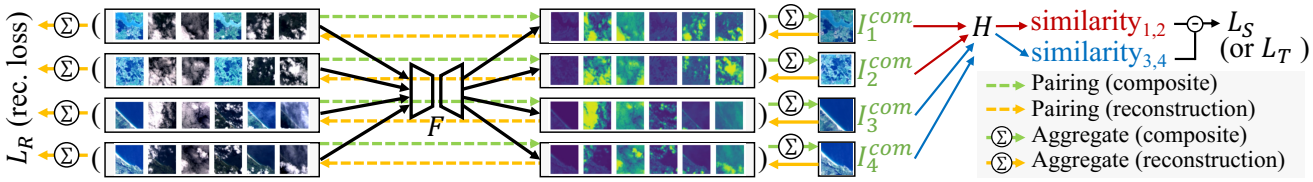


Figure 3: Overall network flow of Auto-CM (using $\mathcal{L}_S$ as an example).

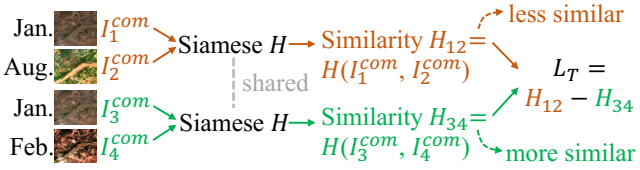


Figure 4: Example of the temporal similarity loss $\mathcal{L}_T$.

allows us to evaluate the spatial or temporal similarity for each of the two pairs, and then compare their relationships according to the expectations from Def. 6.

All the four packs share the same CNN backbone to generate the four image composites (I_1^{com}, I_2^{com}) and (I_3^{com}, I_4^{com}), which will be used to evaluate the similarity-based losses in the following sections. Meanwhile, the network also generates a cloud mask for each image I in a pack, which is both a final output of Auto-CM and an important intermediate result for generating an image composite for each pack. As the training is in a label-free unsupervised setting, the cloud masks will not be directly used in loss calculation. Instead, the similarity between the image composites from the packs will be used to construct the loss functions and guide the training, as shown in Fig. 3.

Using one ST-pack $P_{S_1T_1}$ as an example, each image composite is generated by:

$$I_{S_1T_1}^{com} = G\Big(\big(\sum_{i \in T_1} I_i' \otimes (1 - M_i) \big) \oslash \big(\sum_{i \in T_1} (1 - M_i) \big), \mathbf{\Theta}_G \Big) \tag{1}$$

where T_1 is the local time-series of the ST-pack $P_{S_1T_1}$; $I_i' \in \mathbb{R}^{W \times W \times d}$ (d is the image depth) is the i^{th} image in the pack, which is a network-processed version of the original image I_i and the network-processing can be considered as learned-standardization; $M_i \in \mathbb{R}^{W \times W \times 1}$ is the cloud mask predicted for image I_i, and each value in M_i is in range $[0, 1]$ where 0 means no cloud and 1 means complete blockage by clouds (some thin clouds may not fully block the view); G and $\mathbf{\Theta}_G$ are the network layers generating the image composite and the network parameters, respectively; and $\otimes$ and $\oslash$ are Hadamard (element-wise) product and division, respectively (the same M_i is used for all d channels of I_i').

Temporal similarity loss. The inputs to the temporal similarity loss function are the similarity predictions from the deep network for the image composites generated using each pair of the ST-packs (Eq. (1)). It implements the expected relationships between similarities sim_T in Def. 6, where a pair of temporally closer ST-packs are expected to have a higher similarity than a more distant pair:

$$\mathcal{L}_T = H(I_{S_1T_1}^{com}, I_{S_1T_2}^{com}, \mathbf{\Theta}_H) - H(I_{S_1T_3}^{com}, I_{S_1T_4}^{com}, \mathbf{\Theta}_H) \tag{2}$$

where H and $\mathbf{\Theta}_H$ represent the network layers and parameters used to evaluate the similarity between the two image composites generated from one pair of ST-packs (e.g., $P_{S_1T_1}$ and $P_{S_1T_2}$), respectively; H is a Siamese structure shared by the two composites; all image composites share the same location S_1; and the distance between the temporal ranges follows $dist_T(T_1, T_2) > dist_T(T_3, T_4)$, so minimizing $\mathcal{L}_T$ will enforce that the similarity between $I_{S_1T_1}^{com}$ and $I_{S_1T_2}^{com}$ is smaller than that between $I_{S_1T_3}^{com}$ and $I_{S_1T_4}^{com}$. Fig. 4 illustrates $\mathcal{L}_T$ calculation ($I_{S_1T_j}^{com}$ is simplified as I_j^{com}). To consider the temporal periodicity, the distance $dist_T$ between two time ranges T_1 and T_2 (T_2 is later than T_1) is defined as:

$$dist_T = (T_2[0] - T_1[0]) \,\%\, \beta \tag{3}$$

where $T[0]$ is the start timestamp of a time range $T = \{t_j, t_{j+1}, ..., t_{j+L-1}\}$; $\%$ is the modulo; and β is the length of a period. The value of β depends on the format of the timestamps. For example, the periodicity in land surface patterns mostly happens at the year level (e.g., the repetition of the four seasons), and we can set $\beta = 12$ if the timestamp of an image is indexed by month.

To avoid noises caused by large changes over years (e.g., major changes to the landscapes), in our training data generation, we only generate pairs of T_1 and T_2 where the difference in the years they belong to is smaller than two (e.g., 06/2020 to 06/2022 will not be generated); for convenience, all timestamps in each T are from the same year.

Spatial similarity loss. Similarly, the spatial similarity loss function also takes the similarity predictions for the two pairs of image composites as inputs. The difference is that here we vary spatial window S for the same T. Moreover, according to Def. 4, patterns of the land surface are different across locations, which is true even if two locations are nearby (e.g., two neighboring districts of a city have different layouts). This is different from temporal patterns of the land surface, which remain similar for adjacent time periods.

Utilizing this new characteristic, we consider two forms of the spatial similarity loss $\mathcal{L}_S$. The first form has the same format as $\mathcal{L}_T$ from Eq. (2), except that all image composites share the same time window T_1, and $dist_S(S_1, S_2) > dist_S(S_3, S_4)$, where $dist_S(\cdot, \cdot)$ evaluates the overlap ratio between two spatial windows of size $W \times W$ (e.g., 1 means full overlap and 0 means disjoint). The Siamese network layers H are shared between the calculation of $\mathcal{L}_T$ and $\mathcal{L}_S$ in this case. The second and better form (used in Auto-CM)

takes advantage of the characteristic and evaluates the loss in a more controlled manner. Specifically, the network directly predicts the overlap ratio between a pair of cloud-free image composites predicted from the ST-packs, as we can control the level of overlaps during ST-pack generation. The loss is then:

$$\begin{aligned}\mathcal{L}_S = &\big(H'(I^{com}_{S_1T_1}, I^{com}_{S_2T_1}, \mathbf{\Theta}_{H'}) - O_{12}\big)^2 + \\ &\big(H'(I^{com}_{S_3T_1}, I^{com}_{S_4T_1}, \mathbf{\Theta}_{H'}) - O_{34}\big)^2\end{aligned} \tag{4}$$

where O_{12} and O_{34} are scalars denoting the overlap ratio between spatial windows S_1 and S_2, and S_3 and S_4, respectively; H' is a separate set of network layers used to estimate the overlap ratios based on the composites. Note that this form of $\mathcal{L}_S$ can be evaluated using only one pair of composites. We keep the two-pair format just to be consistent with the input structure in Fig. 3.

Auxiliary reconstruction loss. Finally, we include an additional reconstruction loss, which is commonly used to regulate the training. In our case, we reconstruct each image from a ST-pack using the pack's image composite and the corresponding cloud mask:

$$I^{recon}_i = (1 - M_i) \otimes I^{com} + M_i \tag{5}$$

The loss function is then:

$$\mathcal{L}_R = \frac{1}{|T| \cdot H \cdot W} \cdot \sum_{i=1}^{|T|} \|I^{recon}_i - I_i\|^2_F \tag{6}$$

where $|T|$ is the number of images in the ST-pack, and F denotes the Frobenius norm. The overall loss function is then:

$$\mathcal{L} = \lambda_T \cdot \mathcal{L}_T + \lambda_S \cdot \mathcal{L}_S + \lambda_R \cdot \mathcal{L}_R \tag{7}$$

where λ_T, λ_S and λ_R are scaling factors.

On-the-Go Test-Time Generalization

As Auto-CM is a label-free method, it is robust and can automatically adapt to different data distributions. In addition, the deep network trained by the DISTANCE prior does not need to be re-trained when we switch to a new dataset or new region. Instead, it can be easily fine-tuned using new observations as needed. As there is no need for ground-truth label data, any amount of new imagery can be used for training as needed, increasing the generalizability of the framework.

On the other hand, from the computational perspective, we prefer to have a better understanding of whether fine-tuning is needed or how much new data we should use for fine-tuning. Thus, we introduce a statistical testing based on-the-go generalization approach to help determine this, and only perform fine-tuning as necessary.

Specifically, we consider the fine-tuning process as a sequence of phases with new data subsets: D_1, D_2, D_3, ..., and the subsets follow the same distribution, which can be the same or different from the data distribution used for training. Starting from D_1, the fine-tuning process continues to include new subsets one by one. The goal of the statistical testing is to determine when to terminate the tuning.

Denote D_{test} as a left-out subset, which is from the same test data but not included as part of the sequence D_1, D_2, ...

Further, denote $M_i = F(I_i, \mathbf{\Theta}_F)$ as the cloud mask generated by the sub-network F of Auto-CM (Fig. 3) for an image $I_i \in D_{test}$, where $\mathbf{\Theta}_F$ are the learned parameters from training data. Similarly, denote $M^J_i = F(I_i, \mathbf{\Theta}^J_F)$ as the cloud mask generated with $\mathbf{\Theta}^J_F$, which are fine-tuned by $\cup^J_{j=1} D_j$. We use the paired T-test to evaluate if the masks generated are significantly different after each phase of fine-tuning, with the following test statistic τ:

$$\tau = \mu_M / \big(\sigma_M / \sqrt{W^2 \cdot |D_{test}|}\big) \tag{8}$$

$$\mu_M = \sum_{i=1}^{|D_{test}|} \big(\mathbf{e}^T (M^J_i - M^{J-1}_i)\mathbf{e}\big) / (W^2 \cdot |D_{test}|) \tag{9}$$

$$\sigma_M = \sum_{i=1}^{|D_{test}|} \big(\|M^J_i - M^{J-1}_i - \mu_M \mathbf{e}\mathbf{e}^T\|^2_F\big) / (W^2 \cdot |D_{test}|) \tag{10}$$

where M^J_i equals M_i when $J = 0$, and $M^J_i \in \mathbb{R}^{W \times W}$. The T-test value is compared with the corresponding critical values from the look-up table to determine the significance under level α (defaulted to the standard choice of 0.01).

Experiments

Satellite Datasets

We consider three satellite sensing platforms:

- **Landsat-8** is a multispectral sensing platform with 11 bands covering wavelengths of 0.43 to 12.51µm, where visible bands are from 0.45 to 0.67µm. The spatial resolution is 15m for the panchromatic band, and 30m for all other bands except the thermal bands, which are at 100m.
- **Sentinel-2** also delivers multispectral imagery but covers a different set of 13 bands. For example, its instrument captures additional red edge bands but does not include the thermal bands, which are important for physics-model-based cloud detection.
- **PlanetScope** is a more recent platform with 3-4m high resolution. As a trade-off, it covers a smaller set of bands: the visible bands (red, green and blue) and near-infrared, making it more challenging for cloud masking.

Locations and bands: The datasets cover a diverse range of geographic areas over the world with different landscapes. We use Landsat-8 data in New Zealand (islands; **L1**) and central Australia (desert; **L2**). and Sentinel-2 data in Kenya (urban and agriculture; **L3**) and eastern United States (urban; **L4**); and finally PlanetScope data in Ethiopia (urban area and rivers; **L5**) and Brazil (urban peninsula; **L6**). We also consider different sets of spectral bands using Landsat-8. Specifically, we use three subsets to evaluate the proposed approach: single band (panchromatic), visible bands (RGB), and RGB + near-infrared. This will help show the method's capacity in detecting clouds with limited band information. More details and temporal information are available in the Appendix. **Labels:** (1) As Landsat-8 has been deployed for a longer time, efforts have been made to develop labeled cloud masks for evaluation (e.g., by USGS (USGS 2021)), which we use for quantitative comparisons. (2) For Sentinel-2 and PlanetScope, there is a lack of benchmark data that

are similar to that of the Landsat-8, so we manually labeled two tiles at the pixel-level for evaluation purposes. As the amount of training data is smaller, for supervised baselines, we initialize them using weak labels that come together with the imagery (also evaluated in experiments). **Cloud-free image composites:** As there is no gold-standard for the cloud-free image composites, we use visualizations to qualitatively evaluate the results. As defined in the scope, this is a complimentary output and a necessary intermediate result for Auto-CM (e.g., ST-similarity evaluation during training).

Results

We consider the following methods in the comparison:

- **Physics-based:** (1) **FML8**: The Fmask algorithm for Landsat-8, one of the most adopted physics-based models for cloud masking (Zhu, Wang, and Woodcock 2015). It relies on specific spectral bands (e.g., thermal) in Landsat-8, and is not applicable for high-resolution imagery with fewer bands including PlanetScope (4 bands). (2) **FMS2:** A Fmask variant for Sentinel-2 that uses other spectra due to the missing thermal bands.
- **Supervised learning:** (1) **UNet:** An encoder-decoder semantic segmentation network (Ronneberger, Fischer, and Brox 2015; Zhang et al. 2021). (2) **UNet-DA**: UNet with domain adaptation. An adversarial setup is used to learn domain-invariant features for better generalizibility (Tzeng et al. 2017; Fan et al. 2020); (3) **D3**: DeepLabV3+, a multi-scale segmentation network (Chen et al. 2018). (4) **D3-DA**: DeepLabV3+ with domain adaptation.
- **Unsupervised learning:** (1) **K-means**: K-means++ with k set to 2 (best from experiments). (2) **HD**: Hierarchical density-based clustering that can handle arbitrary shapes, densities and number of clusters (Campello, Moulavi, and Sander 2013). (2) **DEC**: Deep embedding clustering with initial k set to two (best from experiments) (Xie, Girshick, and Farhadi 2016; Obeid, Elfadel, and Werghi 2021).
- **Default masks:** These are approximate cloud masks included as part of the imagery products due to their necessity in most downstream applications. They are typically generated to the best of a provider's ability with the existing methods and engineering (e.g., customized Fmask). **L8M**, **S2M** and **PSM** are masks from Landsat-8, Sentinel-2 and PlanetScope, respectively. For example, **PSM** is created by supervised convolutional networks.
- **Auto-CM:** Our proposed approach (unsupervised).

Results for different geographic regions. Tables 1 to 3 show the results of the methods on Landsat-8, Sentinel-2 and PlanetScope data, respectively. Each table includes results from different geographic regions. For supervised methods, when evaluating a method in a testing region (e.g., New Zealand in Table 1), we use data from the other regions for training. Additionally, to show the effect of distribution shifts, we conduct another test where use 50% samples in the test region as training and the other 50% for testing, where the results are shown in parentheses (**for UNet and DeepLabV3+**). Note that these numbers in parentheses are not used for method comparison (as it is using true labels from the test data) and are only used to help understand supervised models' robustness to region changes (Tarrio et al. 2020). For physics and unsupervised methods, the results are directly obtained from the test region. Note that the physics methods require information from certain spectral bands and cannot be used on the PlanetScope data (Table 3). We can see that physics-based Fmask algorithms show relatively stable performances for different regions in Landsat-8 and Sentinel-2, though the performance is reduced for Sentinel-2. UNet-DA has the best performance among the supervised methods, which require labeled samples. In general, they can reach similar performances as the physics-based approach on left-out test samples from the same training regions (in the parentheses in the tables), but the scores decrease quickly when applied to a different area or landscape. Their results on L3 are poorer than other regions potentially due to the weak labels that come together with the imagery product have low quality around that region (e.g., 0.22 for S2M); reliance on high-quality labels is a limitation for supervised models. For both UNet and DeepLabV3+, unsupervised domain adaption shows improvements but are not sufficient to bridge the distribution gap without labels from the target domain. The reason may be that domain-invariant features can reduce the variance of performance but may be sub-optimal for individual areas. For the unsupervised methods, the performance is not very stable on different types of landscapes. They tend to work better for large and thick clouds in areas with homogeneous landscape, and worse with small or thin clouds on complex landscapes (e.g., urban). Finally, Auto-CM is consistently among the top results in most regions.

Results for different sensing platforms. Comparing Tables 1 to 3, we can see that the performance of the physics-based Fmask algorithms decreases from Landsat-8 to Sentinel-2 due to the missing thermal bands, which the physical rules rely on the most. They are no longer applicable when it comes to the PlanetScope imagery which has four high-resolution bands: RGB + near-infrared (NIR). Supervised approaches have more stable performances for different sensors as they do not have the physical assumptions. However, as we analyzed before, they do not generalize well to different geographic regions for all three sensing platforms. The performance of unsupervised methods decreases for higher-resolution sensors, which is potentially caused by the greater variation under clouds with the increased local details. Finally, Auto-CM shows more consistent performance for different types of sensors.

Results for different spectral bands. Here we perform a controlled experiment for three subsets of spectral bands using Landsat-8, which has 11 bands. As we can see in Table 4, Auto-CM does not rely on specific bands that are needed by physics-based methods and can generate similar quality or better masks even with a single band. The results with fewer bands are also better than the current default masks from the imagery product which are generated using full-band information in the test regions. This may potentially open new opportunities in band prioritization for future satellite sensor design and deployment.

Test area	FML8	UNet	UNet-DA	D3	D3-DA	Kmeans	HD	DEC	L8M	Auto-CM
L1	**0.959***	**0.964*** (0.985)	0.961	0.775 (0.904)	0.920	0.829	0.723	0.668	0.849	**0.966**
L2	0.891	0.908 (0.961)	0.919	0.857 (0.902)	0.810	0.849	0.694	0.923	0.873	**0.940**
Mean	0.925	0.936 (0.973)	0.940	0.816 (0.903)	0.865	0.839	0.709	0.795	0.861	**0.953**

Table 1: F1-scores of cloud masks on Landsat-8 multispectral imagery (results within 1% of the best are denoted by *)

Test area	FMS2	UNet	UNet-DA	D3	D3-DA	Kmeans	HD	DEC	S2M	Auto-CM
L3	**0.741**	0.552 (0.686)	0.655	0.081 (0.457)	0.460	0.643	0.482	0.678	0.220	**0.737***
L4	0.675	0.754 (0.92)	0.883	0.669 (0.794)	0.565	0.808	0.814	0.885	0.790	**0.913**
Mean	0.708	0.653 (0.803)	0.769	0.375 (0.626)	0.512	0.726	0.648	0.782	0.505	**0.825**

Table 2: F1-scores of cloud masks on Sentinel-2 multispectral imagery (results within 1% of the best are denoted by *)

Test area	Fmask	UNet	UNet-DA	D3	D3-DA	Kmeans	HD	DEC	PSM	Auto-CM
L5	-	0.818 (0.923)	0.899	0.55 (0.772)	0.855	0.823	0.297	0.893	0.486	**0.930**
L6	-	0.563 (0.907)	**0.904**	0.8 (0.811)	0.800	0.753	0.772	0.825	0.739	**0.898***
Mean	-	0.69 (0.915)	0.902	0.675 (0.791)	0.827	0.788	0.534	0.859	0.613	**0.914**

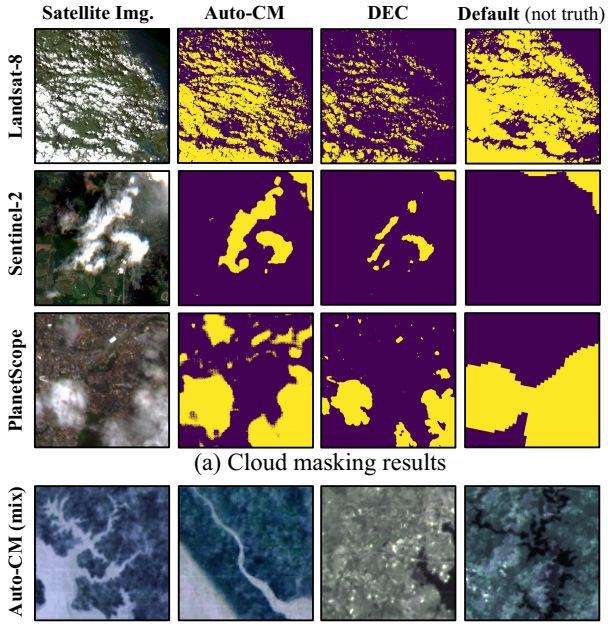
Table 3: F1-scores of cloud masks on PlanetScope high-resolution imagery (results within 1% of the best are denoted by *)

Bands	F1 (L1)	Sig. level α	F1 (L1→L2)
Pan. (single)	0.959	0.001	0.900
RGB	0.965	0.005	0.900
RGB+NIR	0.966	0.01	0.905

Table 4: Sensitivity analysis (Landsat-8)

Significance-based test-time generalization. Table 4 also evaluates the effectiveness of the significance-based test-time generalization. This is for users who prefer to generalize existing Auto-CM models from other regions using only a proportion of unlabelled ST-packs – to the degree that is necessary – from the test region. Here we train Auto-CM on L1 and then finetune to L2 with the phased significance testing. As we can see, the module is not sensitive to the choices of significance levels α (p-value thresholds). In the experiments, it self-decided to use only 5%, 5% and 10% of samples for the fine-tuning, respectively. We can see the results are consistent with the version that is directly trained (unsupervised) on all data (Table 1).

Qualitative visual comparisons. **(1) Cloud masks:** Fig. 5(a) shows the cloud masks generated by different types of methods in the test areas for the three sensing platforms. Interestingly, we can see the results of Auto-CM (here using only four-bands) are often better than existing default masks (e.g., S2M) that require physical information in other bands as well; it also improves over PSM from supervised deep learning. **(2) Cloud-free composites:** Fig. 5(b) shows examples of the composites by Auto-CM. Here the composites are considered complimentary outputs from Auto-CM, which can be generated using the cloud masks from a ST-pack. Additionally, the composites are also necessary intermediate results in Auto-CM, which are used to model the spatio-temporal dynamics based on the DISTANCE prior and calculate the ST-similarity (e.g., used in $\mathcal{L}_S$ and $\mathcal{L}_T$).



(a) Cloud masking results

(b) Cloud-free composites

Figure 5: Example results.

Conclusions

We proposed an unsupervised Auto-CM approach to generate cloud masks and complimentary image composites for general satellite datasets. Our deep learning model uses a new DISTANCE prior, showing promising ability to detect clouds with limited bands and without any labeled samples. A test-time generalization is also proposed to facilitate adaptation to new areas. Our future work will explore scenarios where the surface process is highly dynamic such as polar regions (Yu et al. 2021). We will also consider the fairness and robustness of the masking results for different landscapes and geographic regions (Xie et al. 2022).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2105133, 2126474 and 2147195; NASA under Grant No. 80NSSC22K1164 and 80NSSC21K0314; USGS under Grant No. G21AC10207; US-DOT under Grant No. 69A3551747131 (through SAFER-SIM); Google's AI for Social Good Impact Scholars program; the DRI award at the University of Maryland; Pitt Momentum Funds award and CRC at the University of Pittsburgh; and the ISSSF grant from the University of Iowa.

References

Baetens, L.; Desjardins, C.; and Hagolle, O. 2019. Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. *Remote Sensing*, 11(4): 433.

Campello, R. J.; Moulavi, D.; and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, 160–172. Springer.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Fan, R.; Wang, H.; Bocus, M. J.; and Liu, M. 2020. We learn better road pothole detection: from attention aggregation to adversarial domain adaptation. In *European Conference on Computer Vision*, 285–300. Springer.

Foga, S.; Scaramuzza, P. L.; Guo, S.; Zhu, Z.; Dilley Jr, R. D.; Beckmann, T.; Schmidt, G. L.; Dwyer, J. L.; Hughes, M. J.; and Laue, B. 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote sensing of environment*, 194: 379–390.

Frantz, D.; Haß, E.; Uhl, A.; Stoffels, J.; and Hill, J. 2018. Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. *Remote sensing of environment*, 215: 471–481.

Goodchild, M. F.; and Li, W. 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35).

Hagolle, O.; Huc, M.; Pascual, D. V.; and Dedieu, G. 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sensing of Environment*, 114(8): 1747–1755.

Hollstein, A.; Segl, K.; Guanter, L.; Brell, M.; and Enesco, M. 2016. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sensing*, 8(8): 666.

Karpatne, A.; Ebert-Uphoff, I.; Ravela, S.; Babaie, H. A.; and Kumar, V. 2018. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8): 1544–1554.

Kothandaraman, D.; Guan, T.; Wang, X.; Hu, S.; Lin, M.; and Manocha, D. 2022. Fourier Disentangled Space-Time Attention for Aerial Video Recognition. In *European Conference on Computer Vision*.

Li, P.; Dong, L.; Xiao, H.; and Xu, M. 2015. A cloud image detection method based on SVM vector machine. *Neurocomputing*, 169: 34–42.

Li, Z.; and Snavely, N. 2018. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9039–9048.

Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; and Gascon, F. 2017. Sen2Cor for sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII*, volume 10427, 37–48. SPIE.

Obeid, A.; Elfadel, I. M.; and Werghi, N. 2021. Unsupervised Land-Cover Segmentation Using Accelerated Balanced Deep Embedded Clustering. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Qiu, S.; Zhu, Z.; and He, B. 2019. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*, 231: 111205.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Shen, J.; Yang, X.; Jia, Y.; and Li, X. 2011. Intrinsic images using optimization. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 3481–3487.

Shendryk, Y.; Rist, Y.; Ticehurst, C.; and Thorburn, P. 2019. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157: 124–136.

Singh, P.; and Komodakis, N. 2018. Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 1772–1775. IEEE.

Skakun, S.; Vermote, E. F.; Roger, J.-C.; Justice, C. O.; and Masek, J. G. 2019. Validation of the LaSRC cloud detection algorithm for Landsat 8 images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2439–2446.

Tarrio, K.; Tang, X.; Masek, J. G.; Claverie, M.; Ju, J.; Qiu, S.; Zhu, Z.; and Woodcock, C. E. 2020. Comparison of cloud detection algorithms for Sentinel-2 imagery. *Science of Remote Sensing*, 2: 100010.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, 4.

USGS. 2021. Cloud Cover Assessment Validation Datasets. https://www.usgs.gov/landsat-missions/cloud-cover-assessment-validation-datasets. Accessed: 2022-06-30.

Wei, J.; Huang, W.; Li, Z.; Sun, L.; Zhu, X.; Yuan, Q.; Liu, L.; and Cribb, M. 2020. Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches. *Remote Sensing of Environment*, 248: 112005.

Wieland, M.; Li, Y.; and Martinis, S. 2019. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sensing of Environment*, 230: 111203.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xie, Y.; He, E.; Jia, X.; Bao, H.; Zhou, X.; Ghosh, R.; and Ravirathinam, P. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*, 767–776. IEEE.

Xie, Y.; He, E.; Jia, X.; Chen, W.; Skakun, S.; Bao, H.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-level Learning Framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12208–12216.

Yang, X.; Xu, Z.; and Luo, J. 2018. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yang, Y.; Wang, C.; Liu, R.; Zhang, L.; Guo, X.; and Tao, D. 2022. Self-Augmented Unpaired Image Dehazing via Density and Depth Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2037–2046.

Yi, R.; Tan, P.; and Lin, S. 2020. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12685–12692.

Yu, J.; Xie, Y.; Duncan, K.; and Farrell, S. 2021. Apache Sedona in Action: Analyzing Large-scale Arctic Observations Using an Open-source Big Data Platform. In *AGU Fall Meeting Abstracts*, volume 2021, C55B–0580.

Zhang, L.; Sun, J.; Yang, X.; Jiang, R.; and Ye, Q. 2021. Improving deep learning-based cloud detection for satellite images with attention mechanism. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Zhu, Z.; Wang, S.; and Woodcock, C. E. 2015. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote sensing of Environment*, 159: 269–277.

Zhu, Z.; and Woodcock, C. E. 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote sensing of environment*, 118: 83–94.

Zi, Y.; Xie, F.; Song, X.; Jiang, Z.; and Zhang, H. 2021. Thin Cloud Removal for Remote Sensing Images Using a Physical-Model-Based CycleGAN With Unpaired Data. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.