Cloud Masking in Satellite Images: A Comparison Between U-Net and K-Means Clustering

Alessio De Luca
ID: 919790
University of Milan-Bicocca
a.deluca63@campus.unimib.it

***Abstract*—This paper explores cloud segmentation in satellite imagery by comparing a fine-tuned ResNet-based U-Net and K-Means clustering. U-Net, leveraging a ResNet-101 encoder learns spatial dependencies for accurate cloud masking, while K-Means provides a simpler, unsupervised alternative. To improve clustering, a Guided K-Means approach incorporates NDVI and NDWI, enhancing cloud differentiation from land and snow. Results show that U-Net significantly outperforms standard K-Means but the guided version improves clustering by leveraging spectral features. While deep learning remains the best-performing approach, K-Means could be used to generate training masks, enabling semi-supervised learning for cloud segmentation.**

I. Introduction

A. Paper overview

Cloud segmentation is a fundamental task in satellite image processing, as clouds often obstruct important ground features, making data analysis and interpretation more challenging. Effective cloud masking techniques are essential in various applications, including weather forecasting, climate change studies, and remote sensing.

In this project, I explore two different approaches for cloud segmentation in satellite images:

- U-Net, a deep learning-based semantic segmentation model.
- K-Means clustering, an unsupervised learning method that partitions pixels into distinct groups.

U-Net is a powerful supervised model that uses convolutional neural networks (CNNs) to learn patterns from labeled data, while K-Means relies solely on pixel similarity and clustering, making it a computationally simpler alternative.

One of the main challenges of cloud segmentation lies in distinguishing clouds from other bright surfaces, such as snow or water reflections. To address this, I introduce a feature-guided version of K-Means, incorporating spectral indices like the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) to enhance cloud classification.

The objective of this study is to compare these two methods in terms of accuracy, computational efficiency, and interpretability. Through quantitative evaluation using Intersection over Union (IoU), Dice Score, and Accuracy, I analyze the advantages and limitations of each approach, providing insights into when deep learning outperforms traditional clustering and when a simpler method might be sufficient.

B. Dataset

For this study, I use the 38-Cloud Dataset, a publicly available dataset designed for cloud segmentation tasks in satellite imagery. The dataset consists of multispectral images captured by the Landsat 8 satellite, which provides valuable spectral information across different wavelength bands.

Each image in the dataset contains four spectral bands:

- Red: Captures visible red light, useful for vegetation and land classification.
- Green: Captures visible green light, commonly used in vegetation indices.
- Blue: Captures visible blue light, contributing to standard RGB compositions.
- Near-Infrared (NIR): Captures infrared light, which is particularly useful for distinguishing clouds, water bodies, and vegetation.

Along with these spectral bands, the dataset includes binary ground truth masks, where white pixels represent clouds and black pixels represent non-cloud areas. These masks serve as reference labels for supervised learning models such as U-Net.

To better understand the dataset, Fig. 1 shows some sample images along with their corresponding ground truth masks, highlighting the spectral variations and cloud structures present in different scenes.

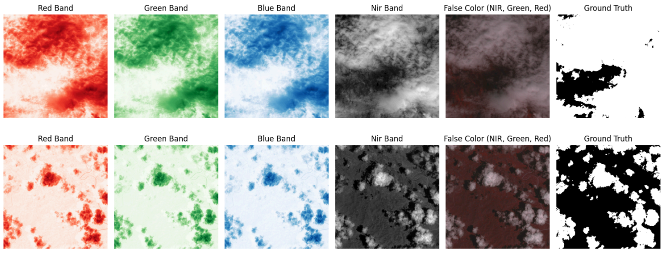

Fig. 1: Visualization of the dataset.

C. *Landsat 8 Satellite*

Landsat 8 is a satellite operated by NASA and the USGS, designed for Earth observation and environmental monitoring. It carries the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), which together capture multispectral and thermal data from the Earth's surface.

Landsat 8 provides high-quality spectral imagery used in applications such as land cover classification, vegetation monitoring, water quality assessment, and atmospheric studies. In this project, the spectral information from its Red, Green, Blue, and Near-Infrared bands is used for cloud detection.

II. Preprocessing

Due to the well-structured nature of the 38-Cloud Dataset, minimal preprocessing was required before training the models. The dataset was already labeled, and the spectral bands were properly aligned, eliminating the need for extensive data cleaning. Additionally, data augmentation was not necessary, as the dataset provided a sufficient number of diverse cloud formations. Morover, he Dataset presented some wrong masks for some images, having a completely black ground truth even if the image had clouds, however, I decided to keep all the images.

For computational reasons, I focused only on the training set, which contained 8400 images. The dataset was split as follows:

- Train set: 5,880 images
- Validation set: 1,680 images
- Test set: 840 images

This split allowed the models to learn effectively while keeping an independent validation and test set for unbiased evaluation. Each image was standardized by normalizing pixel values across all spectral bands, ensuring consistency in input data.

III. Methods

A. *Fine-Tuned ResNet-Based U-Net*

U-Net is a CNN widely used for semantic segmentation tasks. Originally designed for biomedical image segmentation, it has proven highly effective in various domains, including satellite imagery. Its encoder-decoder structure enables it to capture both global context and fine-grained details, making it particularly well-suited for pixel-level classification tasks such as cloud segmentation.

In this project, I implement a fine-tuned U-Net model with a ResNet-101 backbone, importing the pre-trained weights from ImageNet. Fine-tuning allows the model to leverage the powerful feature extraction capabilities of ResNet-101 while adapting to the specific task of cloud segmentation.

The U-Net model consists of two main components:

- Encoder: The pre-trained ResNet-101 backbone extracts hierarchical features from the input image. The encoder progressively reduces spatial resolution while capturing complex patterns useful for segmentation. Fine-tuning is applied to adjust the pretrained layers to the cloud segmentation task while retaining general feature extraction capabilities.
- Decoder: A series of upsampling and convolutional layers reconstructs the spatial information and refines the segmentation mask. Skip connections between the encoder and decoder help retain high-resolution details that would otherwise be lost during the downsampling process.

The model processes four spectral bands—Red, Green, Blue, and Near-Infrared—to enhance cloud detection. The inclusion of the NIR band is particularly useful since clouds reflect strongly in the near-infrared spectrum, making them more distinguishable from bright surfaces such as snow or water.

The fine-tuned ResNet-based U-Net is trained using Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), a standard loss function for binary classification tasks.

To optimize the model, I use the Adam optimizer with L2 weight decay, which helps improve generalization and prevent overfitting by penalizing large weight values.

The training process is conducted for 10 epochs, using a batch size of 8.

Since the model produces continuous probability outputs, a thresholding approach is applied to generate the final binary segmentation masks. Pixels with a probability greater than 0.8 (80%) are classified as clouds, while those below the threshold are classified as non-cloud regions.

Fig. 2 shows both the training and validation losses. The training curve reflects the one of a model that did not over fit, meaning that the parameters were chosen correctly.

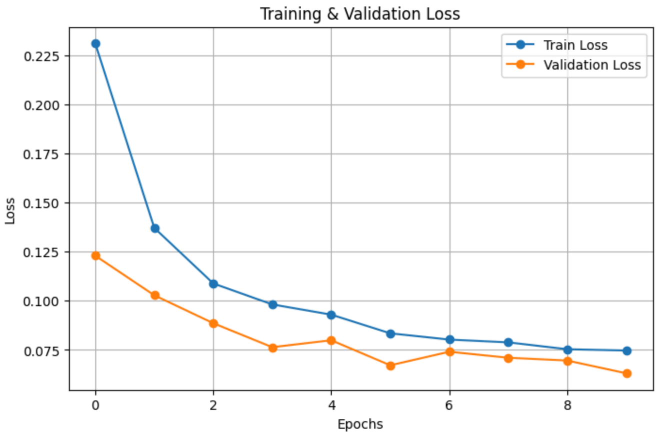

Fig. 2: Evolution of the Training and Validation losses.

B. *K-Means Clustering*

K-Means is an unsupervised machine learning algorithm used for clustering data points into a predefined number of groups. It is widely applied in image segmentation tasks, as it assigns each pixel to one of the clusters based on its spectral characteristics. Unlike deep learning approaches such as U-Net, K-Means does not require labeled data, making it a computationally efficient alternative.

In this project, I explore two different implementations of K-Means for cloud segmentation:

- Basic K-Means clustering, which directly groups pixels based on their spectral values.
- Guided K-Means clustering, which incorporates domain-specific spectral indices to improve cloud classification.

a) *Basic K-Means Approach:*
The first approach applies K-Means clustering directly to the raw spectral data, the same used for training the U-Net. Each image is reshaped into a matrix where each pixel is treated as a feature vector consisting of four spectral bands (Red, Green, Blue, and Near-Infrared). The algorithm then clusters the pixels into two groups, corresponding to cloud and non-cloud regions.

The process follows these steps:

1) Convert the image into a matrix of pixel values, where each row represents a pixel and each column corresponds to a spectral band.
2) Normalize the pixel values to a 0-1 range to prevent any single band from dominating the clustering process.
3) Apply K-Means clustering with K=2 to separate the pixels into two groups.
4) Reshape the clustered output into an image representing the predicted segmentation mask.

However, this method is prone to incorrect classifications due to the reliance on raw spectral values without considering cloud-specific characteristics. This leads to misclassifications, especially in areas with high surface reflectance, such as snow-covered regions or bright water bodies.

b) *Guided K-Means with Spectral Indices:*
To improve the performance of K-Means, I introduce a feature-guided approach that incorporates additional spectral indices specifically designed for distinguishing clouds from other bright surfaces. Instead of relying solely on the four raw spectral bands, the improved model extracts key features that enhance cloud discrimination:

- Normalized Difference Vegetation Index (NDVI)
- Normalized Difference Water Index (NDWI)
- NIR/Red Ratio

These indices provide additional information about the physical properties of clouds, making the clustering process more effective.
NDVI is a widely used index for identifying vegetation by measuring the difference between Near-Infrared (NIR) and Red reflectance. Since vegetation strongly reflects NIR light but absorbs Red light, NDVI can distinguish vegetated areas from other surfaces.

$$\mathrm{NDVI} = \frac{\mathrm{B8} - \mathrm{B4}}{\mathrm{B8} + \mathrm{B4}} \tag{1}$$

Where: B8 is the Near-Infrared (NIR) band. B4 is the Red band.

Clouds typically have low NDVI values (close to or below zero), while vegetation exhibits high NDVI values. This helps differentiate clouds from vegetated regions.

NDWI is an index primarily used to detect water bodies by measuring the contrast between the Green and NIR bands. Water strongly reflects in the Green band while absorbing in NIR, making NDWI a useful feature for distinguishing water from land.

$$\mathrm{NDWI} = \frac{\mathrm{B3} - \mathrm{B8}}{\mathrm{B3} + \mathrm{B8}} \tag{2}$$

Where: B3 is the Green band.

While NDWI is primarily used for water detection, incorporating it into the clustering model can help improve robustness in areas with water bodies, where cloud reflectance patterns might differ. Additionally, NDWI may provide indirect benefits for distinguishing snow from clouds, as snow-covered regions can exhibit unique spectral properties when interacting with nearby water surfaces.
NIR/Red Ratio captures the relative reflectance between NIR and Red bands, providing an additional measure for cloud detection.

$$\mathrm{NIR\_\{ratio\}} = \frac{\mathrm{B8}}{\mathrm{B4}} \tag{3}$$

Since clouds typically reflect more in the NIR spectrum, this ratio is higher for clouds compared to other land surfaces.

Additionally, predefined centroid values are used to initialize the K-Means clustering process. These centroid values, shown in Table I, are selected based on the typical spectral properties of clouds and non-cloud regions.

TABLE I: Predefined centroids values for Guided K-Means clustering.

Category	Red (B4)	Green (B3)	Blue (B2)	Near-Infrared (B8)	NDVI	NDWI	NIR/ Red Ratio
Cloud Reference	0.3	0.3	0.3	0.8	−0.1	−0.5	3.0
Non-Cloud Reference	0.6	0.6	0.6	0.3	0.5	0.3	0.5

These centroids ensure that cloud pixels are assigned to the correct cluster from the start, significantly improving segmentation accuracy compared to the basic K-Means approach.

Compared to the naive K-Means clustering method, the guided approach achieves higher segmentation accuracy because it incorporates physical properties of clouds, rather than relying only on raw spectral values. The use of predefined centroids speeds up convergence and ensures a more stable clustering process. By leveraging spectral indices specifically designed for atmospheric analysis, the guided K-Means method provides a more interpretable and efficient alternative to deep learning-based segmentation.

IV. Results and Discussion

Evaluating the performance of the different cloud segmentation approaches is essential to understand their strengths and limitations. In this section, I compare the fine-tuned ResNet-based U-Net, the basic K-Means clustering, and the guided K-Means approach using standard segmentation metrics. Additionally, I analyze why the guided K-Means performs better than the basic version and discuss scenarios where each method is preferable.

A. Segmentation Metrics

To quantitatively compare the models, I use the following evaluation metrics:

- Intersection over Union (IoU) – Measures how well the predicted segmentation overlaps with the ground truth:

$$\mathrm{IoU} = \frac{|\mathrm{P} \cap \mathrm{T}|}{|\mathrm{P} \cup \mathrm{T}|} \tag{4}$$

Where: P is the predicted mask, and T is the ground truth mask.

- Dice Coefficient – A similarity measure between the prediction and the ground truth:

$$\mathrm{DICE} = \frac{2 \times |\mathrm{P} \cap \mathrm{T}|}{|\mathrm{P}| + |\mathrm{T}|} \tag{5}$$

- Precision – The proportion of correctly predicted cloud pixels relative to all predicted cloud pixels:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

Recall – The proportion of correctly predicted cloud pixels relative to the actual cloud pixels:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

Accuracy – The overall correctness of the model:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

Where: TP (True Positives) are correctly predicted cloud pixels, FP (False Positives) are incorrectly predicted cloud pixels, TN (True Negatives) are correctly predicted non-cloud pixels and FN (False Negatives) are incorrectly predicted non-cloud pixels.

B. Performance Comparison

To assess and compare the effectiveness of each approach, I compute the above metrics for all three methods, and reported the results in Table II.

TABLE II: Performance comparison of U-Net, basic K-Means, and guided K-Means.

Model	IoU	DICE	Precision	Recall	Accuracy
Fine-Tuned U-Net	**0.8416**	**0.8737**	**0.9357**	**0.8488**	**0.9670**
Basic K-Means	0.3228	0.3900	0.4575	0.4268	0.4409
Guided K-Means	0.6567	0.7222	0.7789	0.7708	0.7895

The basic K-Means model struggles because it clusters pixels based solely on their spectral values without incorporating knowledge about cloud characteristics. This leads to misclassification of bright surfaces, such as snow and water, which may appear similar to clouds in visible bands.

Instead, guided K-Means approach significantly improves performance by integrating spectral indices such as NDVI and NDWI, which better differentiate clouds from other bright surfaces.

However, while guided K-Means significantly outperforms the basic version, it still is less effective than the U-Net, which learns more complex spatial dependencies and fine-grained segmentation patterns.

It is also important to note that, sometimes, the ground truth labels were partially or totally incorrect, with full cloud images having a completely black ground truth. The negative effect this implies on the results didn't stop the U-Net from reaching very good accuracies.

The following images show some predictions made by the U-Net, Fig. 3, and by the guided K-Means, Fig. 4.

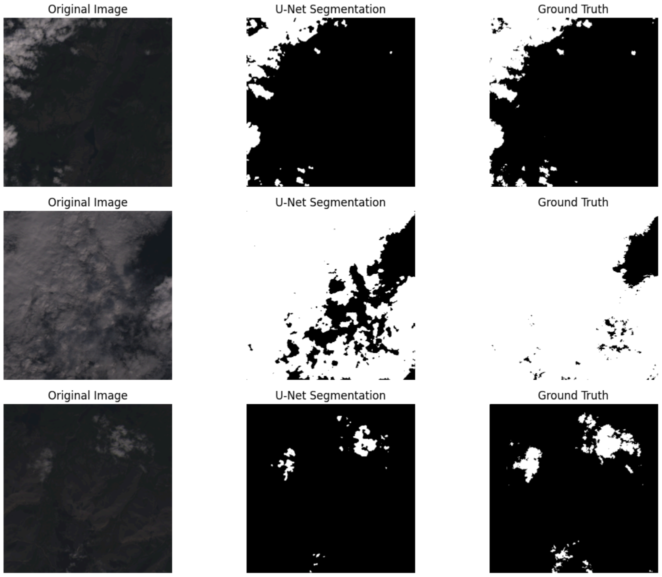

Fig. 3: Predictions made by U-Net model.

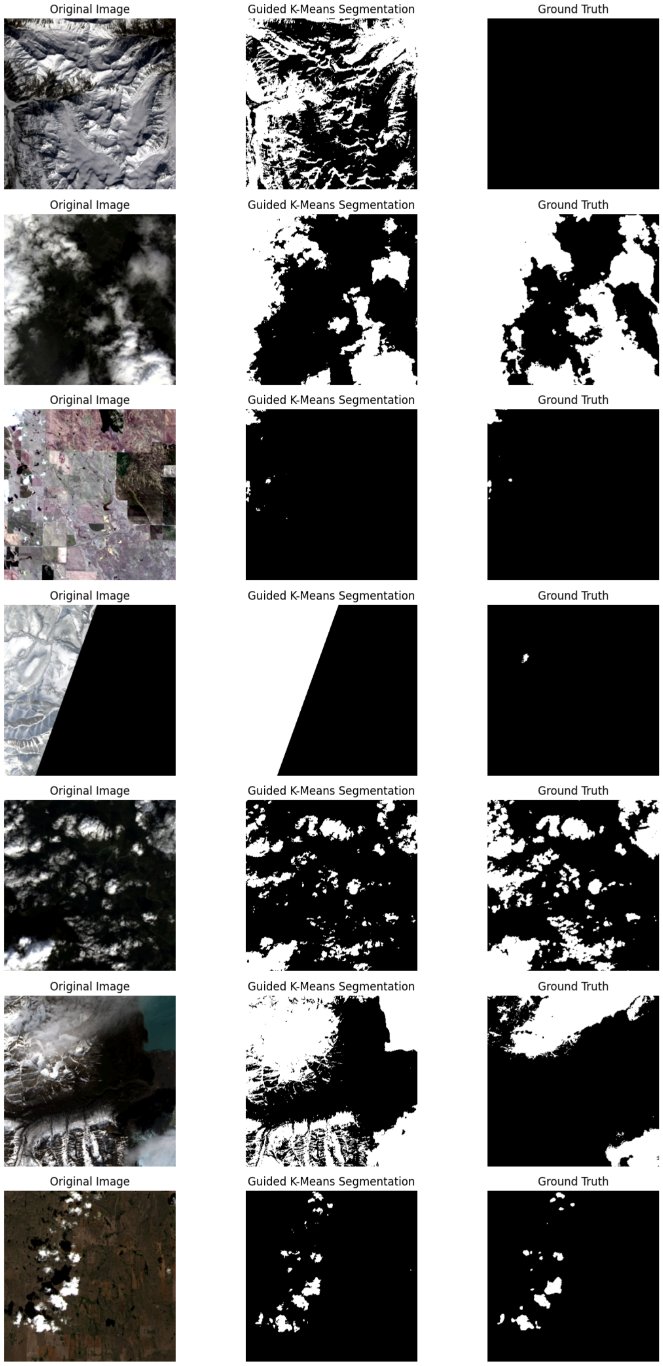

Fig. 4: Predictions made by guided K-Means model.

In Fig. 4 I decided to display the colored images, in order to understand what the background is and to draw some conclusions on the guided K-means algorithm: the third and bottom images show that clustering works well whenever the background consists in vegetation/soil, this means that the difference in NDVI is useful for clustering. However the algorithm still has trouble identifying clouds and snow, as shown in the first, fourth and sixth images, where the results are unsatisfactory, meaning that NDWI features did not help us much as expected. This considerations are further confirmed by the images in Fig. 5: clouds are clearly distinguishable in red/orange color, while vegetation appears green/yellow. Instead, NDWI doesn't really help the algorithm in terms of distinguishing clouds and snow (visible in the left image, a mountainous scenery), which both appear in dark blue values.

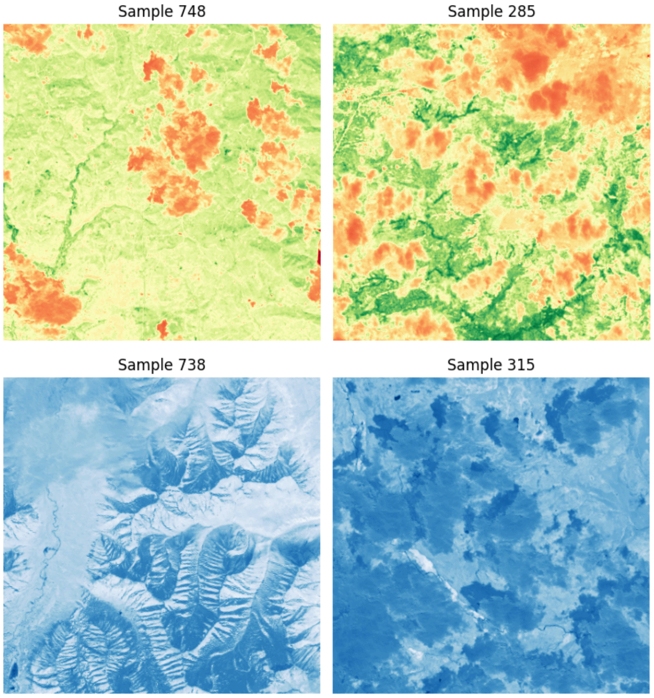

Fig. 5: Visualization of NDVI (top images) and NDWI (bottom images).

V. Conclusion

In this study, I explored and compared two different approaches for cloud segmentation: Fine-tuned U-Net with a ResNet-101 backbone and K-Means clustering (basic and guided versions).

The results demonstrate that U-Net achieves the highest segmentation accuracy, with an IoU of 0.8416 and an overall accuracy of 96.7%. However, guided K-Means, which incorporates spectral indices such as NDVI, significantly improves over the basic K-Means approach, reaching an IoU of 0.6567 and an accuracy of 78.95%.

While U-Net remains the best choice for precise segmentation, guided K-Means offers a practical, unsupervised alternative that can be used when labeled data is unavailable. Additionally, K-Means clustering could serve as an automatic mask generation tool to train deep learning models in a semi-supervised setting.

A. *Future Improvements*

Although the fine-tuned U-Net already provides high segmentation accuracy, improvements can be made in efficiency and data utilization rather than model architecture. Meanwhile, K-Means clustering can be further refined to make it a more viable alternative in real-world applications.

- For U-Net: Instead of modifying the architecture, efforts could be directed toward reducing computational costs. Techniques such as knowledge distillation or quantization could make the model lighter and more suitable for deployment on edge devices. Another practical improvement could be using K-Means-generated masks as pseudo-labels, enabling a semi-supervised training approach that reduces dependency on manually labeled data.
- For K-Means: While guided K-Means already improves over the basic version, it could be further refined by incorporating additional spectral indices relevant to cloud detection, such as Cloud Optical Thickness (COT), however it was not implementable with this dataset because I did not have access to all the bands from Landsat 8. Another potential improvement is experimenting with alternative unsupervised clustering methods, such as Gaussian Mixture Models (GMMs) or Self-Organizing Maps (SOMs), which might provide a better representation of cloud structures.

B. Real-World Applications

The methods explored in this study have several real-world applications, including:

- Weather Forecasting: Cloud masks can improve the accuracy of weather prediction models.
- Satellite Data Preprocessing: Removing cloud interference in satellite images can enhance land cover analysis.
- Climate Monitoring: Tracking cloud patterns over time helps in studying climate change and atmospheric dynamics.
- Automated Training Data Generation: Guided K-Means can generate approximate cloud masks, reducing manual labeling efforts for deep learning models.

This study highlights the trade-offs between deep learning and traditional clustering approaches in satellite image segmentation. While U-Net remains the state-of-the-art solution, guided K-Means offers a lightweight, interpretable, and label-free alternative, demonstrating its potential in various practical scenarios.