R2U++ & DIFFCR: A Novel Deep Learning Framework for Identifying Deforestation Drivers

Sander Braastad[1], Kjell Anders Sæter Grytting[2]
[1]Fakultet for naturvitenskap og teknologi, Universitet i Bergen

***Abstract*—This paper presents R2U++ & DIFFCR, an approach for identifying deforestation drivers using satellite imagery. Unlike traditional methods that rely on raw, cloud-obstructed inputs, R2U++ & DIFFCRincorporates a conditional diffusion model for cloud removal, followed by semantic segmentation using a multiscale recurrent residual U-Net with dense skip connections [1]. The framework is evaluated on high-resolution Sentinel-2 imagery provided by the Solafune Deforestation Drivers competition. Although R2U++ & DIFFCRintroduces an enhanced preprocessing step and a more expressive segmentation backbone, experimental results show that it does not surpass the baseline in terms of F1-score. Nonetheless, the findings offer insights into the trade-offs between model complexity, data quality, and segmentation performance in remote sensing tasks.**

***Index Terms*—Deep learning, Remote sensing, Diffusion models, U-Net, Semantic segmentation, Deforestation, Satellite imagery**

I. INTRODUCTION

This paper addresses the issue of identifying the drivers of deforestation. Deforestation refers to the large-scale clearing of forests, often replaced by other land uses such as agriculture or urban development. The loss of forests is environmentally damaging and contributes to climate change. Deforestation results in biodiversity loss, disruption of water cycles, and soil erosion [2]. Forests play a critical role in photosynthesis, producing oxygen and acting as carbon sinks. In contrast, deforestation releases stored carbon dioxide, accelerating global warming. It is therefore essential to identify both the areas affected and the human activities responsible for this phenomenon. Common drivers of deforestation include mining, logging, and plantation development.

This study focuses on identifying deforestation drivers using high-resolution remote sensing data. Specifically, we utilize Sentinel-2 satellite imagery, which offers a spatial resolution of 1024×1024 pixels and captures 12 spectral bands spanning the visible (RGB), near-infrared, and shortwave infrared regions, as well as additional bands sensitive to atmospheric conditions such as aerosols and water vapor. These spectral bands offer valuable land cover context and are well-suited for detecting human activities in forested areas

We explore the use of deep learning for this task, formulating it as a multi-label semantic segmentation problem. As a baseline, we implement a U-Net architecture with a ResNet-34 encoder pretrained on ImageNet, adapted to process 12-channel Sentinel-2 imagery and output segmentation masks for four deforestation driver classes.

To benchmark model design choices, we compare this baseline to the R2U++ model [1], which enhances U-Net through recurrent residual convolutional blocks and dense skip pathways, potentially enabling better context modeling and gradient flow.

Furthermore, we investigate the impact of preprocessing the imagery with a diffusion-based cloud removal model—DiffCR [3]. Cloud cover often obstructs critical visual features in satellite images, and DiffCR, a conditional guided diffusion model, has demonstrated state-of-the-art performance in generating high-fidelity cloud-free images. We assess whether this preprocessing step improves segmentation accuracy by providing clearer input data to the downstream models.

Deforestation—the large-scale clearing of forests, often to make way for agriculture, mining, or urban development—is a major environmental challenge with wide-reaching consequences. It contributes significantly to climate change by releasing stored carbon dioxide into the atmosphere, and it undermines natural carbon sinks that are essential for balancing the global carbon cycle. Deforestation also results in biodiversity loss, soil degradation, and disruption of local and global water cycles [2]. Forest ecosystems play a vital role in regulating the climate through carbon sequestration and oxygen production via photosynthesis. As such, understanding where deforestation is occurring and what human activities are driving it is essential for effective environmental monitoring and policy intervention.

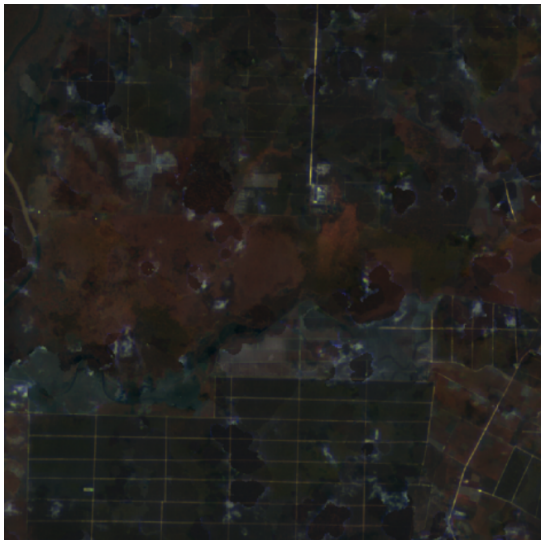


Fig. 1: Example of a Sentinel-2 multispectral satellite image used in the study.

II. Related Work

Recent advances in medical semantic segmentation for remote sensing often build on U-Net and its variants. R2U++ [1] enhances the standard U-Net by introducing recurrent residual blocks and dense skip pathways, improving feature reuse and long-range context modeling. In parallel, cloud occlusion remains a persistent challenge in satellite imagery; DiffCR [3] addresses this with a fast conditional diffusion model for reconstructing cloud-free observations. DiffCR, compared to previous cloud removal approaches, offers faster generation and produces higher-fidelity cloud-free images. Earlier methods often required thousands of iterations or relied on GAN-based models, which suffered from mode collapse and instability [3]. While both R2U++ and DiffCR address key bottlenecks in their respective domains, their combined use for cloud-robust deforestation driver identification has, to our knowledge, not been previously explored.

III. Method

A. Overview

We formulate the identification of deforestation drivers from satellite imagery as a multi-label semantic segmentation problem. Given a 12-band Sentinel-2 image, the goal is to predict a binary segmentation mask for each of four driver classes, indicating the spatial extent of their presence.

Our method consists of three main components:

- An optional cloud removal preprocessing step using the DiffCR framework, applied before segmentation, to assess its impact on performance.
- A deep segmentation model that maps multispectral input to multi-label driver masks.
- A comparative study of two model architectures: a U-Net with ResNet-34 encoder (baseline), and an R2U++ model with recurrent residual and dense skip connections.

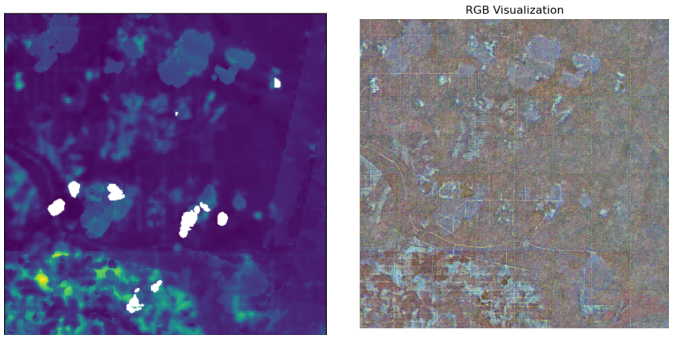
B. Network Architecture

a) Cloud Removal with DiffCR: Satellite imagery provided by Solafune contains missing or occluded data due to cloud cover, which can obscure critical visual information required for accurate segmentation. To mitigate this issue, we investigated whether training a model to reconstruct cloud-covered regions could outperform the naïve approach of setting these pixels to zero.

We adopted the framework presented in "DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal from Optical Satellite Images" [3], which employs a conditional diffusion model to iteratively refine noisy observations into cloud-free outputs. While the original DiffCR model was designed for 3-channel RGB images, we modified it to support 12-channel Sentinel-2 inputs by adapting the UNet backbone and normalization layers. This modification allows the model to exploit information across the full spectral range, including non-visible bands such as near-infrared and shortwave infrared.

Due to computational constraints, we made several simplifications to ensure feasible training:

- **Patch Size: 128 × 128** — Full-resolution images (1024×1024) were divided into non-overlapping tiles of 128×128 to reduce memory requirements.
- **Batch Size: 1** — Reduced from the original batch size of 8 to fit within GPU limits.
- **inner_channel: 32** — Lowered the base channel width to minimize parameter count and computational load.
- **channel_mults: [1, 2, 4]** — Introduced three resolution stages in the UNet to balance model depth with receptive field size.
- **res_blocks: 2** — Maintained two residual blocks per stage to capture fine-grained features.
- **dropout: 0.05** — Applied mild regularization to prevent overfitting given the limited training data.
- **attn_res: []** — Omitted attention layers to save memory.
- **norm_groups: 16** — Used GroupNorm with 16 groups, ensuring stable optimization even with small batch sizes.



(a) Original image (b) Cloud-removed image

Fig. 2: Comparison between the original Sentinel-2 satellite image and the corresponding cloud-removed output generated by DiffCR.

b) Baseline: U-Net with ResNet-34: Our baseline model was inspired by the public baseline shared by *Motokimura* for the Solafune "Identifying Deforestation Drivers" competition [4]. The original model utilized a U-Net backbone combined with a `tf_efficientnetv2_s` encoder sourced from the `timm` library, offering a powerful but computationally intensive feature extractor. In contrast, to better accommodate our hardware limitations and training pipeline, we replaced the encoder with a ResNet-34 pretrained on ImageNet. For training, we employed a composite loss function combining Dice Loss and Binary Cross-Entropy (BCE) Loss, similar to the public baseline. Unlike the original implementation which relied on the `segmentation_models_pytorch` loss wrappers, we implemented the loss computation manually, ensuring identical behavior while allowing greater customization flexibility. Some of the callbacks were also removed to benefit computational cost. Another factor in selecting Motokimura's baseline was the use of PyTorch Lightning, which provided a modular and reproducible training framework that simplified the integration of callbacks, checkpointing, and logging. Additionally, we adapted Motokimura's `generate_masks.ipynb` notebook for generating segmentation masks, integrating it into our own preprocessing with minor modifications.

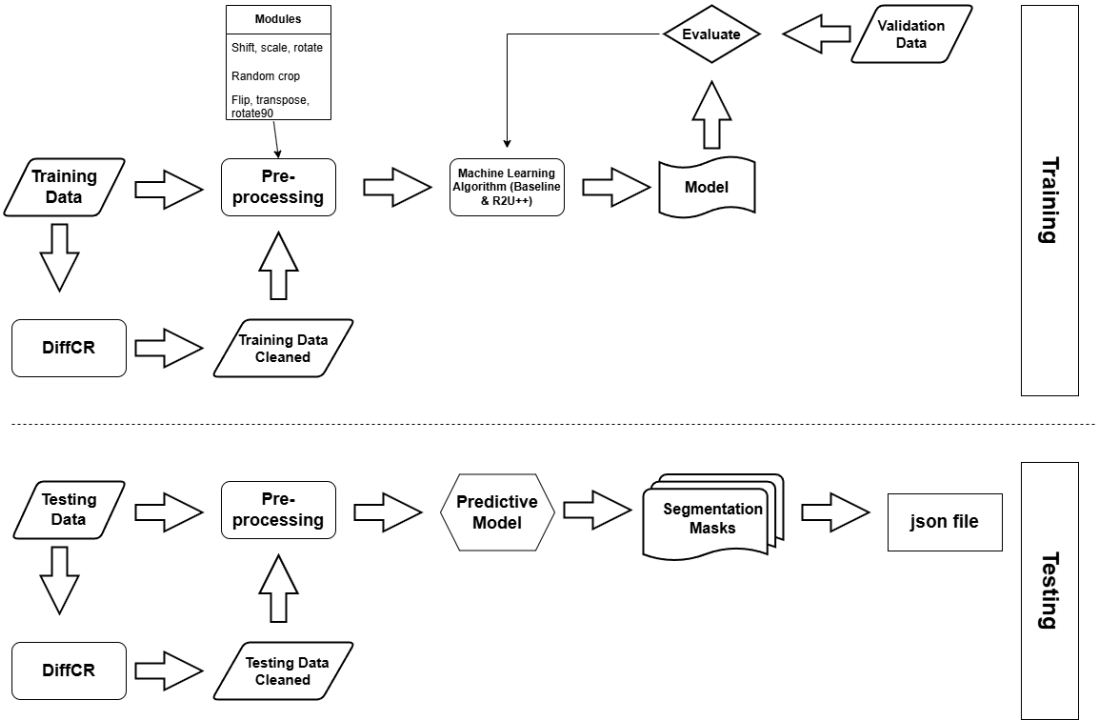


Fig. 3: Overall pipeline for identifying deforestation drivers from satellite imagery. First, Sentinel-2 images are optionally preprocessed using the DiffCR cloud removal framework to produce clearer multispectral inputs. These images are then passed through a deep semantic segmentation model—either the baseline U-Net or an R2U++ model—to generate pixel-wise multi-label predictions for different deforestation driver classes. Finally, the predicted segmentation masks are post-processed and converted into a JSON file format for competition submission.

c) R2U++: Recurrent Residual U-Net with Dense Skip Connections: R2U++ was originally developed for medical image segmentation tasks, such as electron microscopy and computed tomography imaging [1]. The model enhances the standard U-Net by introducing two architectural modifications: *recurrent residual blocks* and *dense skip connections*. The recurrent residual blocks expand the field of view and improve feature extraction by iteratively refining feature maps, while the dense skip pathways help bridge the semantic gap between encoder and decoder features.

Given the similarities between medical imaging (e.g., fine-grained structures, occlusions) and remote sensing imagery, we investigate the applicability of R2U++ for segmenting multispectral satellite data. Our goal was to assess whether the architectural advantages observed in medical tasks could also benefit land cover segmentation in deforestation driver detection.

Recurrent Residual Blocks: Instead of applying a convolution once, R2U++ repeats it several times. This helps the model refine its feature maps. Each block:

- Adjusts input channels using a 1×1 convolution,
- Applies a 3×3 convolution with batch normalization and ReLU activation,
- Repeats this process t times to improve the result.

Dense Skip Connections: In regular U-Net, skip connections link encoder and decoder layers at the same depth. R2U++ adds a denser structure. It:

- Collects features from earlier encoder layers,
- Resizes them to match the current decoder layer,
- Applies a convolution to each feature map,
- Sums them together before passing to the decoder.

This approach reduces the semantic gap between encoder and decoder outputs.

Decoder and Output Fusion: Each decoder stage:

- Upsamples the input using a transposed convolution,
- Concatenates it with the corresponding skip connections,
- Processes it with a recurrent residual block.

The model also produces intermediate outputs at different decoder stages. The intermediate outputs are resized to the original input resolution and averaged to produce the final prediction.

d) Input and Output: The network takes a 12-channel Sentinel-2 satellite image as input. It predicts four binary segmentation maps, each representing one deforestation driver.

e) Loss Function: For training the segmentation networks, we employ a composite loss function that combines Dice loss and soft binary cross-entropy (BCE) loss. The Dice component addresses class imbalance and encourages spatial overlap between predicted and ground truth masks, while BCE provides stable gradient signals for multi-label classification. The total loss is computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}$$

Each loss is computed per class and averaged across all four target classes.

IV. Experiments

A. Dataset

The dataset used in this study consists of high-resolution Sentinel-2 satellite imagery, provided by Solafune as part of the Deforestation Drivers competition [2]. Each image has a spatial resolution of 1024×1024 pixels and contains 12 spectral bands. These bands cover the visible spectrum (RGB), near-infrared (NIR), shortwave infrared (SWIR), and additional bands sensitive to atmospheric conditions such as aerosols and water vapor.

The multispectral nature of this data provides rich contextual information about vegetation health and land cover types, which is particularly useful for identifying human-induced disturbances in forested regions.

In addition to the imagery, the dataset includes pixel-level semantic segmentation annotations corresponding to four deforestation driver classes: *mining*, *logging*, *grassland/shrubland*, and *plantation*. The task is framed as a multi-label semantic segmentation problem, where each pixel may be assigned to one or more deforestation drivers.

We were provided with a total of 176 annotated satellite images for training and 118 additional images for evaluation. Each training image is accompanied by a corresponding multi-label segmentation mask identifying deforestation driver classes at the pixel level. The evaluation set, by contrast, is unlabeled and used exclusively for leaderboard submissions, requiring participants to generate predictions without ground truth feedback.

B. Training Settings

All models in this study were trained using the same data augmentation pipeline to ensure a fair comparison across architectures. Augmentations included standard transformations such as random horizontal flips, random vertical flips, random rotations, and brightness/contrast adjustments. Due to computational resource limitations and time constraints, all models were trained for 50 epochs. Ideally, training for over 200 epochs would have allowed the models to converge further and potentially achieve higher segmentation performance. However, because of practical restrictions, we limited the training duration to ensure feasibility across all experiments.

C. Evaluation Metrics

The evaluation metric used in the Solafune "Identifying Deforestation Drivers" competition is based on the pixel-wise F1 score. This metric is computed separately for each deforestation driver class and reflects the balance between precision and recall.

For each class, the F1 score is calculated using the number of true positives (TP), false positives (FP), and false negatives (FN) between the predicted and ground truth binary segmentation masks. The F1 score is defined as:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

where

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

[2]

The final score is obtained by computing the mean F1 score across the four deforestation driver classes: *mining*, *logging*, *grassland/shrubland*, and *plantation*.

This class-wise F1 evaluation encourages models that generalize across both prevalent and rare deforestation patterns, ensuring robustness and fairness in performance assessment. [2]

D. Results

Table I presents the performance comparison between different model architectures on the evaluation data. Results are reported in terms of Validation F1-score, as well as the Public and Private leaderboard scores provided by Solafune. We observe that R2U++ without DiffCR outperformed the other models, even when training was limited to 50 epochs. We hypothesize that the recurrent residual connections in R2U++ allowed the model to better capture fine-grained spatial dependencies, leading to more accurate segmentation of deforestation drivers despite the constrained training duration.

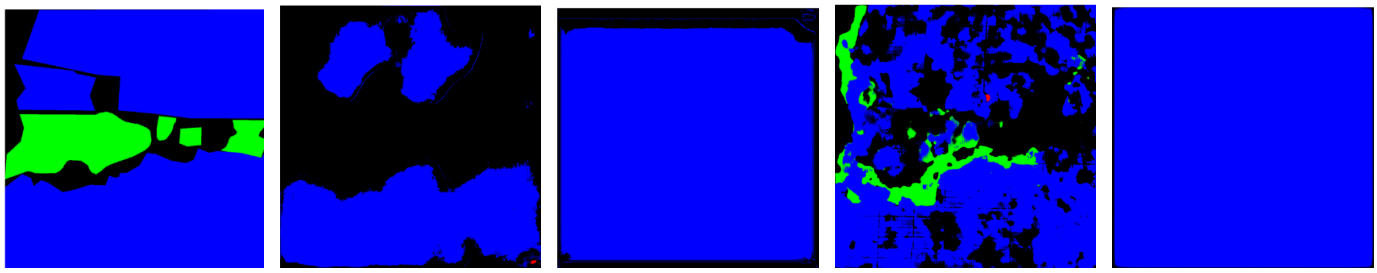
TABLE I: Performance comparison of different model configurations on the evaluation dataset. The table reports the Validation F1-score as well as the Public and Private leaderboard scores from the Solafune competition. Models include a baseline U-Net with ResNet-34 encoder, the R2U++ architecture, and versions of both models incorporating DiffCR cloud removal preprocessing.

Method	Validation F1-score	Public	Private
Baseline (resnet + imagenet)	0.5737	0.2654	0.3167
R2U++	**0.5756**	**0.3635**	**0.3763**
Baseline & DiffCR	0.5611	0.1107	0.1501
R2U++ & DiffCR	0.5664	0.3436	0.3389

V. Limitations and Challenges

While the proposed framework demonstrates potential in combining cloud removal and semantic segmentation, several limitations and challenges must be acknowledged:

- **Cloud Removal as Preprocessing:** The pipeline assumes that cloud removal via generative models improves segmentation quality. However, artifacts introduced by imperfect cloud removal can propagate and negatively affect downstream segmentation performance.
- **Computational Constraints:** All models in this study were implemented under strict hardware limitations, which constrained both memory usage and training throughput. Full-resolution Sentinel-2 images (1024×1024) were cropped into 128×128 tiles, and training was conducted with a batch size of 1 due to GPU memory constraints for the diffusion model. These design decisions affected all architectures, including the U-Net baseline, the recurrent R2U++ model, and the modified DiffCR variant. Notably, the diffusion model's architecture had to be further simplified by reducing internal



(a) Original (b) Baseline (c) Baseline & DiffCR (d) R2U++ (e) R2U++ & DiffCR

Fig. 4: Qualitative comparison of the original satellite image and predicted segmentation masks generated by different models, with and without DiffCR cloud removal. All predictions correspond to the 5th sample from the validation set.

channel widths, omitting attention modules, and minimizing residual block depth. These compromises may have reduced each model's ability to capture global or long-range dependencies, thereby limiting segmentation accuracy.

Moreover, the small batch size made it difficult for the diffusion model to generalize, as noisy gradient estimates hindered stable learning. Inference using DiffCR also required approximately 100 denoising steps to produce acceptable cloud-free images, significantly higher than the 5-step inference target proposed in the original paper [3]. This increased inference time further constrained the practicality of deploying the model in time-sensitive applications.

- **Lack of Multi-temporal Conditioning:** Unlike the original DiffCR framework [3], which utilizes multi-temporal image sequences to infer occluded content, our implementation relies on single-frame conditioning due to dataset and resource constraints. This significantly limits the model's ability to resolve dense cloud coverage where temporal context is crucial.
- **Lack of Source Code:** The original R2U++ paper did not provide any publicly available source code, which introduced an additional challenge. To address this, we re-implemented the R2U++ architecture based on the descriptions in the paper. We used OpenAI's ChatGPT to help interpret the technical descriptions and verify the correctness of the implementation. Through this process, we successfully developed a working version of R2U++ that aligns with the architecture described in the paper.

VI. Conclusion

In this study, we presented R2U++ & DiffCR, a deep learning framework designed to identify deforestation drivers from multispectral satellite imagery. By integrating the DiffCR conditional diffusion model for cloud removal with the baseline UNet and advanced segmentation architectures R2U++, we aimed to enhance the clarity of input data and the expressiveness of the segmentation backbone. The goal was to assess whether enhancing input clarity through cloud removal together with a multiscale recurrent residual U-Net with dense skip connections leads to improved multi-label segmentation performance.

Experimental results showed that the R2U++ architecture achieved the best overall performance, with a validation F1-score of 0.5756 and private leaderboard score of 0.3763. The baseline U-Net model achieved slightly lower scores, with a validation F1-score of 0.5737 and a private score of 0.3167. Preprocessing with DiffCR cloud removal, however, reduced segmentation performance across both models, likely due to limitations in model capacity and training conditions. Under the current implementation and resource constraints, the diffusion model did not consistently generate cloud-free images that benefited segmentation

Overall, our results underscore the trade-offs between computational complexity, cloud handling, and segmentation fidelity in remote sensing workflows. While R2U++ demonstrated consistent improvements over the baseline, further refinements to the cloud removal step—such as using multi-temporal inputs or more stable training schedules—could help fully unlock the benefits of preprocessing. Future work could also explore more lightweight segmentation architectures optimized specifically for multispectral data and low-data regimes.

Acknowledgments

Our work was inspired by the methods described in the DiffCR and R2U++ papers, which offered valuable guidance and implementation insights. We also made extensive use of the source code provided by the authors of DiffCR [5], modifying it to meet the specific requirements of the Solafune competition data. And lastly, throughout this project, we used OpenAI's ChatGPT to assist with writing, coding, and debugging.

References

[1] M. Z. Alom, M. T. Hasan, M. Hasan, and T. M. Taha, "R2u++: an efficient and lightweight network for retinal vessel segmentation with residual recurrent convolutions," *Soft Computing*, vol. 27, pp. 10 701–10 718, 2023. [Online]. Available: https://doi.org/10.1007/s00521-022-07419-7

[2] Solafune, "Identifying Deforestation Drivers Challenge," https://solafune.com/competitions/68ad4759-4686-4bb3-94b8-7063f755b43d, 2024, accessed: 2025-04-23.

[3] X. Zou, K. Li, J. Xing, Y. Zhang, S. Wang, L. Jin, and P. Tao, "Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images," *arXiv preprint arXiv:2308.04417*, 2023. [Online]. Available: https://arxiv.org/abs/2308.04417

[4] Motokimura, "Solafune deforestation drivers baseline," https://github.com/motokimura/solafune_deforestation_baseline/, 2024, accessed: 2025-04-27.

[5] X. Zou, K. Li, J. Xing, Y. Zhang, S. Wang, L. Jin, and P. Tao, "Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images," https://github.com/XavierJiezou/DiffCR, 2024, accessed: 2025-04-27.

[2] [3] [1] [5]