Transferability and Generalization Investigation of Multiclass Cloud Masking Networks for unseen Biomes and Sensors - A Study on PlanetScope, Platero and Sentinel-2

Michael Greza[1], Tianyi You[2], Boris Jutzi[3]

[1] Technical University of Munich, Photogrammetry & Remote Sensing, Munich - michael.greza@tum.de
[2] Technical University of Munich, Photogrammetry & Remote Sensing, Munich - tianyi.you@tum.de
[3] Karlsruhe Institute of Technology, Institute for Photogrammetry & Remote Sensing, Karlsruhe - boris.jutzi@kit.edu

Keywords: Cloud Masking, Transformer, Satellite Imagery, U-Net, Multispectral.

Abstract

With the rising amount of small satellite Earth observation missions, robust model transferability and generalization is becoming more important in satellite remote sensing image processing pipelines to enable a faster and more efficient processing setup. In this work, the transferability and generalization capabilities of two multiclass cloud masking approaches are tested on the tropical rainforest biome in the Amazon, that is unknown to the models, and on two new satellite systems, Platero and Sentinel-2. This is developed as an examplary test, if the models can be used as a baseline for transfer learning and finetuning for new satellite mission cloud masking processors. The evaluation is on a qualitative level due to the lack of ground truth data and small sample size. The results are promising for the new biome but show that regionally occurring phenomena like snow can mislead the origin networks. Furthermore, the experiments show the importance of finetuning on datasets for new sensor systems, especially when facing high discrepancies in the spectral channels.

1. Introduction

Small satellites, especially CubeSats (Johnstone et al., 2020) drastically lower the costs of satellite missions compared to classic missions like Sentinel-2 (Drusch et al., 2012), enabling a rapidly rising amount of small Earth observation (EO) missions (Kulu, 2025). This is accompanied by a need for new data processing techniques that are well-adapted to their respective mission-, satellite- and sensor-characteristics. To speed up the development of these processing pipelines, neural networks (NNs) could be utilized as an easily adaptable solution. However, their generalization capability is heavily dependent on the diversity of the training dataset. Building a big enough and diverse dataset in turn can be costly. This diversity also encompasses e.g. the biomes, and the sensor platforms that the networks are trained on. Good generalization in these two domains can help reduce the amount of data needed for transfer learning.

In satellite remote sensing imagery processing pipelines, cloud masking is a crucial step. In this work, the generalization capabilities of two neural networks for cloud masking (You et al., 2025) is investigated.

2. Related Work

Past literature contains numerous studies on cloud masking models that utilize traditional machine learning methods, such as Support vector machine (Li et al., 2015) (Yuan et al., 2020) or Random forest (Chen et al., 2020) (Fu et al., 2019). The input for the models are individual pixels or locally segmented regions in an image. One of the drawbacks of such models is the limited ability to mine spatial features despite sufficient training, which can result in lower performance in complex scenarios (Li et al., 2022) with additional spatial feature diversity. The S2cloudless cloud detector is based on gradient boosting. Copernicus provides single-class cloud masks with cloud probability (Skakun et al., 2022) on a per-pixel basis. Furthermore, the Sentinel-2 cloud detector relies on SWIR-channel information that is not available for most EO satellite missions equipped with sensors in the VIS-NIR-range.

In comparison, deep-learning-based (DL-based) cloud-masking models in general prove to succeed in feature representation and the highest accuracy with the same amount of training data. In recent years, they are constantly improving with advanced new techniques and hold potential for further development. In early works, a simple convolutional neural network (CNN) model is designed for cloud detection in multispectral Proba-V images (Mateo-García et al., 2017) and it is then proposed with a CNN for more detailed detection, including haze, cloud, and cloud shadow (Xie et al., 2017). The focus on differentiation of clouds from snowy regions is also addressed (Zhan et al., 2017). The DL-based detection is embedded with an advanced technique of multi-scale convolutional feature fusion (Li et al., 2019). In more recent works, the focus of CNN development shifts towards the inner neural network architecture that best fits the application. Some notable works include Cloud-Net (Mohajerani and Saeedi, 2019), MF-CNN (Shao et al., 2019), and KappaMask (Domnich et al., 2021). The advantages of CNN models mainly lie in the capabilities of pixel-wise segmentation and multi-layer convolutional feature hierarchies.

Due to the increasing use of transformers and their strong performance in vision tasks, transformer models have also been explored in this domain, leading to several studies in recent years. Several of these methods inherit the basic transformer architecture, such as ViTs (Dosovitskiy et al., 2020) or PVT (Wang et al., 2021) and adapt them to cloud masking. Compared to traditional CNN-based approaches, there exist fewer transformer-based studies but contain several state-of-the-art approaches. A novel strategy combining both PVT and a cyclic refinement architecture is developed to emphasize both low- and high-resolution image features for attaining a high standard of cloud detection (Tan et al., 2024). It is also proposed with a new cloud segmentation algorithm (Li and Wang, 2024) based on the SegFormer architecture (Xie et al., 2021). This new transformer-

https://doi.org/10.5194/isprs-annals-X-3-W4-2025-185-2026 | © Author(s) 2026. CC BY 4.0 License.

based architecture has equally good performance in terms of accuracy compared to other traditional CNN-based methods.

The usage of transfer learning in cloud masking has seen much advancement in recent years. For example, transfer learning between different sensors is developed for fully convolutional neural network on a comparably small dataset (Mateo-García et al., 2020). In addition, domain adaptation is also utilized in (Mateo-García et al., 2021) and (Guo et al., 2018) as a solution to limited data availability and to enhance generalization. Further studies include cross-sensor transferability of CNN-based models as in (Mateo-García et al., 2020), (Mateo-García, 2020) or (Pang et al., 2023), who show significant gains in model accuracy with the transfer from Landsat-8 to other multispectral satellite instruments, such as Sentinel-2 and Proba-V. In comparison, the transferability and generalization of the transformer-based models are underexplored and subject to further investigations. Recent works in this direction have laid foundations with attention-based models, including CLiSA (Paul and Gupta, 2023), Cloud-Adapter model (Zou et al., 2024) and CD-CTFM (Ge et al., 2023). A hybrid CNN-transformer model is presented with orthogonal cross-attention to enhance model generalization for multiple satellites (Paul and Gupta, 2023). In addition, vision foundation models are leveraged to enable efficient domain adaptation (Zou et al., 2024). However, the research on transferability and generalization of transformer-based models remain scarce in the cloud masking domain, thus allowing advanced development that holds further potential in model fine-tuning, cross-sensor adaptability, and multi-modal frameworks.

3. Neural Networks

In this work, the transferability and generalization capabilities of two cloud masking approaches (You et al., 2025) are tested. This encompasses a CNN-based and a Transformer-based model, both adapted to multispectral PlanetScope imagery. For the architecture, the CNN-based model consists of the U-Net model (Ronneberger et al., 2015) adapted to eight-channel multispectral inputs, the Transformer-based model is based on Maskformer (Cheng et al., 2021) with a Swin Transformer (Liu et al., 2021) backbone adapted to nine spectral channels. The output classes encompass *background (for blank areas), clear, snow, cloud shadow, light haze, heavy haze* and *cloud*. The defined classes are primarily based on the operational needs and satellite product requirements, which includes physical distinction in radiometric properties, as demonstrated in the past established frameworks such as Fmask v4.0 and the Copernicus CMIX protocol.

The two cloud masking networks are trained on the same multispectral PlanetScope dataset. The dataset consists of eight-channel EO imagery taken by the SuperDove satellite constellation in year 2023 and 2024. The regions covered are worldwide cities, high mountains, the Mediterranean, polar regions, and Bavaria. Most of the data are temperate deciduous biome scenes. It does not contain data from tropical rainforest biome regions except one scene from the city of Belém, Brasil which is an urban environment. The ground sampling distance (GSD) of the dataset lies between 3.7 m and 4.2 m. Subject to wavelength-aligned spectral band allocation, the input order of the spectral channels from the Platero and Sentinel-2 imagery for the models are aligned respectively to the nearest neighbors in the spectrum of PlanetScope imagery.

4. Transferability Tests

To evaluate generalization and transferability of the NNs, different test sets are used with both NNs in inference mode. None of the data was part of training, testing or validation. As the radiometric and geometric properties of the dataset vary and different types of clouds can be fringy, a proper definition of a ground truth mask for clouds is non-trivial. Furthermore, there is high ambivalence in the definition of similar classes like clouds, heavy haze and light haze. PlanetScope deprecated the differentiation of light and heavy haze (Labs, 2025) because of this. Following this ambivalence between the datasets, small sample sizes, and non-overlapping segmentation classes in other datasets, the authors mainly evaluate the results on a qualitative basis. The Planet masks are assumed as ground truth but face the same aforementioned challenges. For this work, the bands in Table 1 are used for inference, as they are the closest to the data the networks are trained on. The bands are ordered such that they are closest to the original PlanetScope spectra.

4.1 PlanetScope Amazon

First, transferability to a new biome is tested on the PlanetScope Amazon dataset. Original training scenes mostly consisted of the temperate deciduous biome, while this experiment's dataset is taken from tropical rainforests. The satellite system is the same as in the original training dataset.

4.2 Platero Bavaria

Second, transferability to a new, but comparable satellite system with a different sensor in a known biome is tested on the Platero dataset. Platero's and PlanetScope's GSD is similar but there are differences in the radiometric characteristics of their sensors. The Platero dataset consists of a region that is part of the original training dataset.

4.3 Sentinel-2 Amazon

Third, transferability to a new satellite system and a new biome is tested on the Sentinel-2 dataset. This is considered as especially challenging, as the GSD difference between the training data of the NNs and the Sentinel-2 dataset is considerably higher compared to the second experiment and the tropical rainforest biome is not part of the NN training dataset. Additionally, the sensors of PlanetScope and Sentinel-2 are more diverse than in the previous experiment. As the SWIR channels of Sentinel-2 are useful for cloud detection, a separate experiment is conducted where a SWIR-channel is included although the NNs are not trained on this.

5. Datasets

This Section gives an overview of the datasets that are used for training and testing. First, the PlanetScope dataset used for training the neural networks is described, followed by the datasets for testing transferability. Examples from the datasets are shown in Section 6. The datasets are comprised of L0-L1B data.

5.1 PlanetScope Amazon

This PlanetScope dataset contains scenes from the Amazon rainforest region that were acquired between June and August 2024. This dataset and its acquistion region were never shown to the network before the experiments.

Spectral Channels and Central Wavelength [nm]		
PlanetScope	Sentinel-2	Platero
Coastal Blue: 442 Blue: 490 Green I: 531 Green: 565 Yellow: 610 Red: 665 Red Edge: 705 NIR: 865	Blue: 490 Green: 560 Red: 665 VNIR I: 705 VNIR II: 865 For one test SWIR: 1610	Blue: 480 Green: 560 Red: 662 Red Edge: 703 NIR: 841

Table 1. Overview of PlanetScope spectral channels and their wavelengths.

5.2 Platero Bavaria

Platero is a CubeSat mission by Open Cosmos and the Junta de Andalucia. Its payload consists of a seven-channel multispectral sensor with an additional PAN channel. The scenes were collected between July 2024 and April 2025 with a GSD of 4.75 m. The region of interest is Ansbach in Bavaria, which was one of the regions included in the original PlanetScope training dataset.

5.3 Sentinel-2 Amazon

The Sentinel-2 dataset contains scenes from the Amazon rainforest region that were acquired between June and August 2020. The GSD of Sentinel-2 is channel-dependent and either 10 m, 20 m, or 60 m. Only channels with a GSD of 10 m and 20 m were used. The 20 m channels were resampled to a 10 m resolution, utilizing a nearest neighbor approach. The region of interest is the Amazon rainforest in Brazil. It is to note that none of the SWIR channels of the Sentinel-2 dataset were used during inference as the neural networks are not trained on SWIR data.

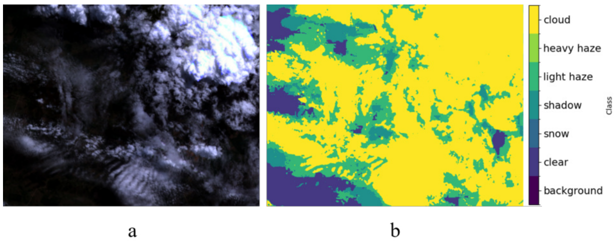

Figure 1. Examplary results for the Maskformer on PlanetScope data in the Amazon. a) Input image, b) Maskformer cloud mask.

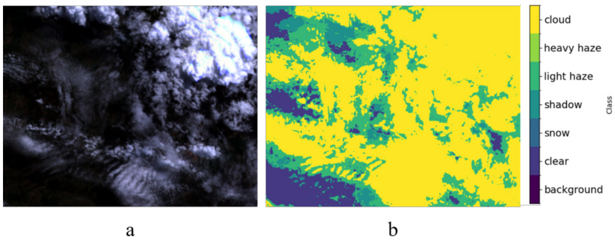

Figure 2. Examplary results for the U-Net on PlanetScope data in the Amazon. a) Input image, b) U-Net cloud mask.

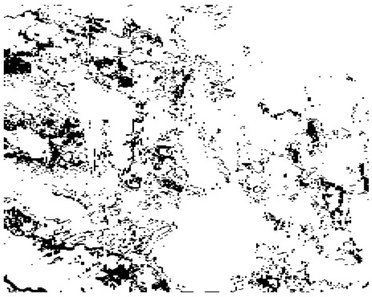

Figure 3. Pixelwise comparison between Maskformer and U-Net results of 1 and 2. Black pixels show differences. 98 % fit.

6. Results

The results of the three different experiments with both NNs are displayed in this Section. Additional results are displayed in the Appendix.

6.1 PlanetScope Amazon

Figure 1 shows the result for an Amazonian PlanetScope result of Maskformer inference. Similarly, Figure 2 depicts the U-Net results. The masks appear almost identical, with visible differences only in the details.

6.2 Platero Bavaria

Figure 4 depicts a result for the Maskformer inference on the Platero dataset. For this case, rectangular artifacts appear. The U-Net result in Figure 5 is free of artifacts but mistakenly segments the clouds as haze.

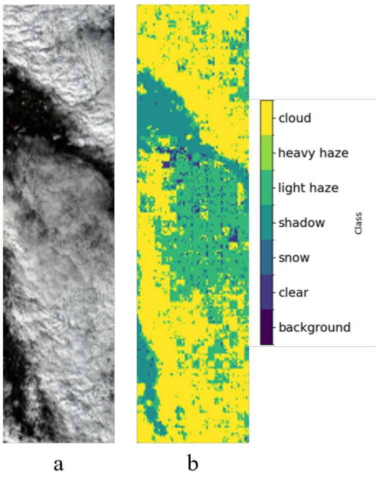

Figure 4. Examplary results for the Maskformer on Platero data in the Ansbach region. a) Input image, b) Maskformer cloud mask.

6.3 Sentinel-2 Amazon

Examplary results for the nine-channel Transformer network are shown in Figure 6, with a zoomed-in view containing a side-by-side comparison with combined classes in Figure 8. Figure 9 shows the inference on an input image including a SWIR channel. It displays larger swaths of hazy regions that are only visible in the input image upon a more detailed inspection. U-Net results are depicted in Figure 7.

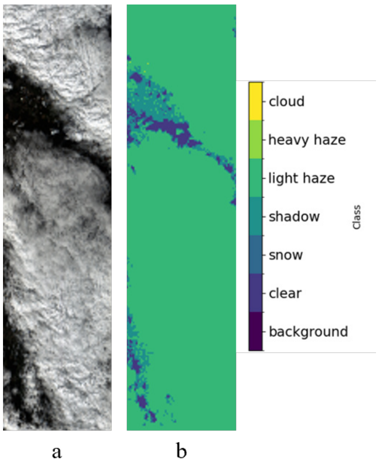

Figure 5. Examplary results for the U-Net on Platero data in the Ansbach region. a) Input image, b) U-Net cloud mask.

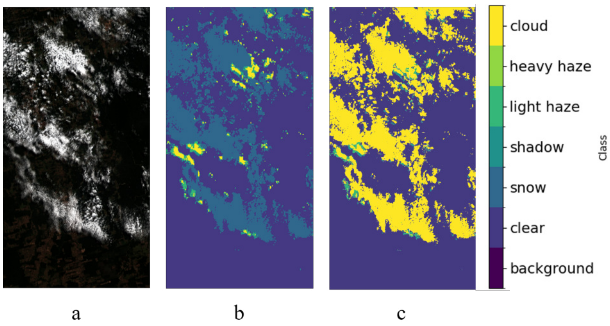

Figure 6. Examplary results for the Maskformer on Sentinel-2 data. a) Input image, b) Model output, c) Combined cloud and snow classes.

7. Discussion

Both networks lean towards a more conservative cloud segmentation, as the ground truth input for training provided by PlanetScope is also similarly leaning towards larger masks.

7.1 PlanetScope Amazon

On the Planet Amazon dataset, the two networks performed comparably well to the performance on the original test dataset (You et al., 2025). The transfer to a new biome did not pose a challenge. None of the artifacts that occured in the other experiments are visible and none of the clouds were mistaken for snow by the Maskformer. It is to note that both networks produced almost identical masks that differ only slightly which is reflected by an average agreement of 98 % between Transformer and U-Net (Figure 3).

7.2 Platero Bavaria

For Platero, the quadratic artifacts appearing in the Maskformer's results shown in Figure 4 and the appendix worsen the quality of the outcome. These artifacts resemble the tokenization pattern of the Transformer model. Still, the reason why these artifacts appear in this case is unknown. The U-Net performs worse on Figure 5. The model mistakes clouds with haze. Transfer learning or finetuning the models for new sensors is advised.

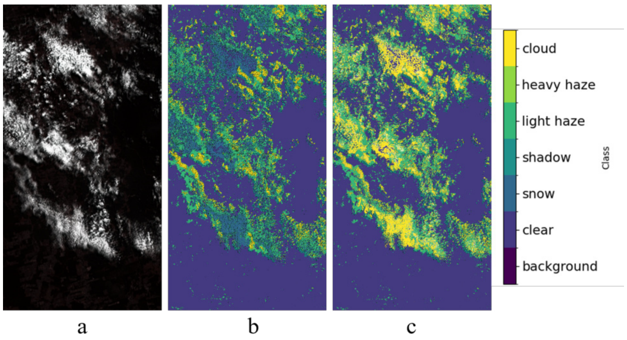

Figure 7. Examplary results for the U-Net on Sentinel-2 data. a) Input image, b) Model output, c) Combined cloud and snow classes.

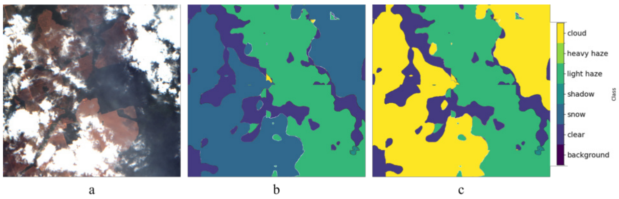

Figure 8. Examplary results for the Maskformer on Sentinel-2 data. a) Input image, b) Model output, c) Combined cloud and snow classes.

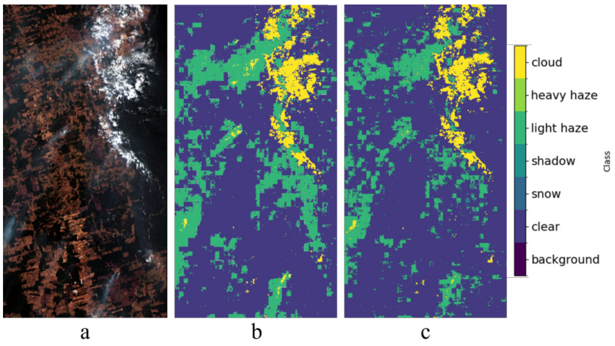

Figure 9. Examplary results for the Maskformer on Sentinel-2 data. a) Input image, b) Combined cloud and snow classes without SWIR, c) Combined cloud and snow classes utilizing SWIR.

7.3 Sentinel-2 Amazon

Results for the Sentinel-2 data are not as straightforward to interpret as for Planet and Platero. The most apparent difference to the former is the detection of large areas of snow instead of clouds by the Maskformer as depicted in Figure 6. The occurence of snow in the Amazon is heavily unlikely therefore the Maskformer misclassified these areas. Knowing that snow is very unlikely and clouds are mistaken for snow, both classes can be fused to a single cloud class, resulting in a good segmentation of the clouds. As these false-positive snow classes occur only for the Sentinel-2 dataset, a possible explanation would be that the high difference in geometric and radiometric characteristics of Sentinel-2 and PlanetScope data results in these misclassifications. Transfer learning for the new Sentinel-2 modality is highly recommended in this case. Including a SWIR channel as in Figure 9 does not result in significant differences but slightly reduces artifacts. As the model was not trained on SWIR, the model may not be able to extract reasonable information from this channel. Fine-tuning the networks to enable the inclusion of the SWIR channels may also increase performance. Inference quality on sensor platforms that are very different from the training dataset in the case of Sentinel-2 declines due to the models relying only on a specific set of spectral channels during training which do not match the ones fed to the model during inference. A possible approach to tackle this issue is to introduce greater sensor diversity in the training dataset. Neural networks like VSISR (Greza et al., 2024) show that training on data from three to five different satellite systems already improve the transferability to new sensors.

8. Conclusion

Transferability of deep learning models is a continuously important feature to enable faster, more data-efficient and less ecologically impactful reutilization of these models. Training data diversity is important, but as the experiments show, satisfying results can also be obtained with limited resources. Overall, the results throughout the experiments are good and show that the models generalize well to the tropical rainforest biome. The transfer to new sensors proves to be more challenging than the biome transfer. Still, pleasing results can be achieved. Recent experiments on foundational models for remote sensing (Le Lain and Lefèvre, 2024) show that pretrained models have an advantage over models trained from scratch, even when a domain transfer is needed. Therefore, the NNs tested in this work could act as a potential baseline for transfer learning. In case of the Maskformer, transfer learning for Sentinel-2 or comparable platforms is advised.

Acknowledgements

The authors want to thank Felipe Ferrari and Raul Feitosa for providing data from their Sentinel-2 dataset.

References

Chen, X., Liu, L., Gao, Y., Zhang, X., Xie, S., 2020. A Novel Classification Extension-Based Cloud Detection Method for Medium-Resolution Optical Images. *Remote Sensing*, 12(15).

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (eds), *Advances in Neural Information Processing Systems*, 34, Curran Associates, Inc., 17864–17875.

Domnich, M., Sünter, I., Trofimov, H., Wold, O., Harun, F., Kostiukhin, A., Järveoja, M., Veske, M., Tamm, T., Voormansik, K., Olesk, A., Boccia, V., Longepe, N., Cadau, E. G., 2021. KappaMask: AI-Based Cloudmask Processor for Sentinel-2. *Remote Sensing*, 13(20).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25–36.

Fu, H., Shen, Y., Liu, J., He, G., Chen, J., Liu, P., Qian, J., Li, J., 2019. Cloud Detection for FY Meteorology Satellite Based on Ensemble Thresholds and Random Forests Approach. *Remote Sensing*, 11(1). https://www.mdpi.com/2072-4292/11/1/44.

Ge, W., Yang, X., Zhang, L., 2023. CD-CTFM: A Lightweight CNN-Transformer Network for Remote Sensing Cloud Detection Fusing Multiscale Features. *arXiv preprint arXiv:2308.14992*.

Greza, M., Bhattacharya, I., Hoegner, L., Jutzi, B., 2024. GAN-Based Dual Image Super Resolution for Satellite Imagery Decreasing Radiometric Uncertainty. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 155–162.

Guo, H., Liu, G., Liang, D., Zhang, L., Xiao, H., 2018. Progress of Earth Observation in China. *Chinese Journal of Space Science*, 38(5), 797–809.

Johnstone, A., Mehrparvar, A., Pignatelli, D., Carnahan, J., Munakata, R., Lan, W., Toorian, A., Hutputanasin, A., Lee, S., 2020. Cubesat design specification.

Kulu, E., 2025. World's largest database of nanosatellites. https://www.nanosats.eu/. Accessed: 2025/05/03.

Labs, P., 2025. Usable data mask. https://docs.planet.com/data/imagery/udm/. Accessed: 2025/05/20.

Le Lain, M., Lefèvre, S., 2024. When Visual Foundation Models Meet Astronomical Data. *EAS2024*, 759.

Li, J., Wang, Q., 2024. CSDFormer: A cloud and shadow detection method for landsat images based on transformer. *International Journal of Applied Earth Observation and Geoinformation*, 129, 103799.

Li, P., Dong, L., Xiao, H., Xu, M., 2015. A cloud image detection method based on SVM vector machine. *Neurocomputing*, 169, 34-42. Learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis.

Li, Z., Shen, H., Cheng, Q., Liu, Y., You, S., He, Z., 2019. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 197-212.

Li, Z., Shen, H., Weng, Q., Zhang, Y., Dou, P., Zhang, L., 2022. Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Mateo-García, G., Gómez-Chova, L., Camps-Valls, G., 2017. Convolutional neural networks for multispectral image cloud masking. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2255–2258.

Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., 2020. Transferring deep learning models for cloud detection between Landsat-8 and Proba-V. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160, 1-17.

Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., 2021. Cross-Sensor Adversarial Domain Adaptation of Landsat-8 and Proba-V Images for Cloud Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 747-761.

Mateo-García, G., 2020. Cloud masking of Proba-V multispectral images using convolutional neural networks. PhD thesis, Universitat de València. Doctoral thesis.

Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., 2020. Cross-Sensor Adversarial Domain Adaptation of Landsat-8 and Proba-V images for Cloud Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Mohajerani, S., Saeedi, P., 2019. Shadow Detection in Single RGB Images Using a Context Preserver Convolutional Neural Network Trained by Multiple Adversarial Examples. *IEEE Transactions on Image Processing*, 28, 4117-4129.

Pang, S., Sun, L., Tian, Y., Ma, Y., Wei, J., 2023. Convolutional Neural Network-Driven Improvements in Global Cloud Detection for Landsat 8 and Transfer Learning on Sentinel-2 Imagery. *Remote Sensing*.

Paul, S., Gupta, A., 2023. CLiSA: A Hierarchical Hybrid Transformer Model using Orthogonal Cross Attention for Satellite Image Cloud Segmentation. *arXiv preprint arXiv:2305.01829*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Shao, Z., Pan, Y., Diao, C., Cai, J., 2019. Cloud Detection in Remote Sensing Images Based on Multiscale Features-Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6), 4062-4076.

Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O. et al., 2022. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274, 112990.

Tan, H., Sun, S., Cheng, T., Shu, X., 2024. Transformer-Based Cloud DetectionMethod for High-Resolution Remote Sensing Imagery. *Computers, Materials and Continua*, 80, 661-678.

Wang, Z., Yuan, L., Chen, Q., Li, X., Wang, M., Zhang, H., Liu, Y., Xie, L., Yang, Y., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *International Conference on Computer Vision (ICCV)*, 206–215.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34, 12077–12090.

Xie, F., Shi, M., Shi, Z., Yin, J., Zhao, D., 2017. Multilevel cloud detection in remote sensing images based on deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10, 3631-3640.

You, T., Greza, M., Jutzi, B., 2025. U-net- and transformer-based cloud masking for multispectral earth observation satellite missions. *IEEE International Geoscience and Remote Sensing Symposium*.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241, 111716.

Zhan, Y., Wang, J., Shi, J., Cheng, G., Yao, L., Sun, W., 2017. Distinguishing Cloud and Snow in Satellite Images via Deep Convolutional Network. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1785-1789.

Zou, X., Zhang, S., Li, K., Ren, Y., Zhu, H., Li, J., Wang, F., 2024. Adapting Vision Foundation Models for Robust Cloud Segmentation in Remote Sensing Images. *arXiv preprint arXiv:2401.08479*.

Appendix

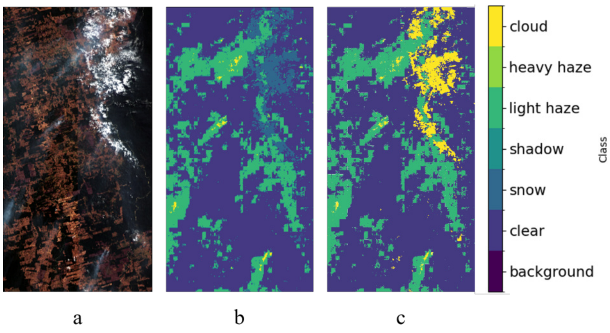

Figure 10. Examplary results for the Transformer on Sentinel-2 data. a) Input image, b) Model output, c) Combined cloud and snow classes.

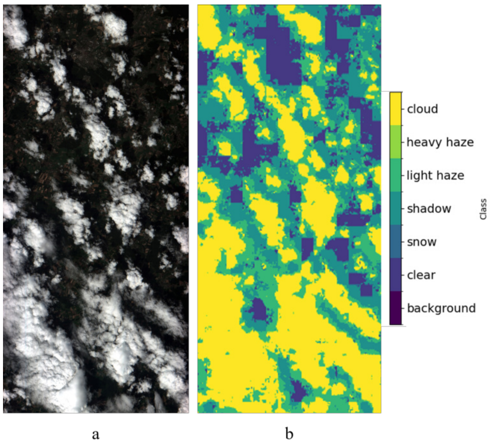

Figure 11. Examplary results for the Transformer on Platero data in the Ansbach region. a) Input image, b) Maskformer cloud mask.