

Modèles d'attention pour la classification et reconnaissance de cibles

Abstract—Abstract abstract abstract

Index Terms—keywords, keywords, keywords

I. INTRODUCTION

Synthetic Aperture Radar (SAR) is a high-resolution remote sensing technology that generates detailed images of the Earth's surface by using microwave radar signals. Unlike optical sensors, SAR can operate in all weather conditions, day and night, making it particularly valuable for a wide range of applications, including environmental monitoring, military surveillance, disaster management, and oceanographic studies. The technology's ability to penetrate clouds and work under various environmental conditions provides an advantage over traditional imaging methods, but SAR data comes with its own set of challenges, especially when applied to automatic target recognition (ATR) and classification tasks.

One of the key challenges in SAR image processing is speckle noise, a multiplicative noise that degrades image quality and makes it difficult to extract meaningful features for classification. Speckle is an inherent characteristic of coherent imaging systems like SAR and complicates tasks such as object detection and target recognition by masking fine details and introducing random variations in pixel intensity. As a result, robust noise-reduction techniques are often necessary to improve image clarity before applying classification algorithms.

Another significant challenge is the high-dimensional nature of SAR data. SAR images are often captured at very high resolutions, leading to large amounts of data that must be processed efficiently. In addition, SAR images represent surface features differently from optical images, emphasizing structural information rather than texture or color. This calls for specialized feature extraction techniques to capture the unique characteristics of SAR images for classification tasks.

Moreover, SAR imaging systems often need to function accurately in diverse terrains and environmental conditions. The variability in terrain types—such as urban areas, forests, deserts, and oceans—affects the radar backscatter and further complicates target classification. Additionally, different sensor configurations (such as varying incidence angles or polarizations) lead to further variability in the captured data. These factors make traditional classification approaches, which often rely on handcrafted features, less effective for SAR images.

Given these challenges, machine learning techniques, especially deep learning, have become increasingly popular for SAR image classification. Convolutional neural networks

(CNNs) have been widely applied to extract hierarchical features from SAR images, improving classification accuracy by learning spatial representations automatically. However, CNNs still face difficulties in capturing long-range dependencies and global context in SAR data. This has motivated the adoption of advanced architectures like transformers and attention mechanisms, which can more effectively model global relationships between image patches, addressing some of the limitations of traditional deep learning methods in SAR imaging.

II. STATE OF THE ART

A. Deep Learning for SAR Image Classification

Deep learning has significantly advanced the field of SAR image classification, offering improved performance over traditional machine learning methods by automatically learning hierarchical features from raw data. Among the most prominent deep learning approaches used for SAR classification are Convolutional Neural Networks (CNNs). This architecture has been widely adopted due to its success in image recognition tasks in computer vision.

CNNs have been the backbone of SAR image classification, particularly due to their ability to capture spatial hierarchies in images. A CNN consists of multiple layers of convolutional filters that automatically detect low- to high-level features such as edges, shapes, and patterns, making them ideal for visual tasks. In the context of SAR imagery, CNNs have been effective in extracting relevant features such as texture and geometric shapes of targets, which are crucial for classification. For instance, CNN-based methods have shown success in automatic target recognition (ATR) by learning representations that distinguish between military and civilian targets in SAR data.

Despite their widespread use, CNNs face several limitations when applied to SAR imagery. CNNs rely on local receptive fields and hierarchical feature extraction, which means they are effective at capturing local spatial patterns but struggle to model long-range dependencies or global context across the image. This can be problematic for SAR images, where objects of interest may be spatially distributed or influenced by large-scale structural variations in the scene.

Additionally, CNNs often require large amounts of labeled data for training to generalize effectively, but obtaining annotated SAR datasets can be challenging due to the specialized nature of SAR data collection. Transfer learning and data augmentation techniques have been employed to mitigate this issue, but these approaches have their own limitations, espe-

cially when the distribution of SAR data differs significantly from the pre-trained models used in transfer learning.

In light of these limitations, there has been growing interest in leveraging transformer architectures, which can overcome the drawbacks of CNNs by modeling global relationships between image patches through their attention mechanism. As highlighted in the work Exploring Deep Learning Methods for Classification of SAR Images: Towards NextGen Convolutions via Transformers, transformers are able to capture long-range dependencies in SAR images, providing a more complete representation of the scene. Unlike CNNs, transformers do not rely on fixed local receptive fields, allowing them to account for complex spatial relationships over the entire image, which is critical for accurately classifying SAR data where target features may be subtle or spread out.

B. Transformers in SAR Imaging

The transformer model was first introduced by Vaswani et al. in the landmark paper "Attention Is All You Need" (2017), revolutionizing the field of natural language processing (NLP) by replacing traditional sequence models like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. The transformer's key innovation lies in its self-attention mechanism, which allows it to model long-range dependencies in data more efficiently than previous architectures. By calculating attention weights, the transformer can focus on different parts of the input sequence simultaneously, capturing both local and global relationships without the sequential limitations of RNNs.

The self-attention mechanism works by computing weighted averages of input embeddings, assigning higher importance to specific elements in the sequence based on their relevance to the current task. This ability to selectively focus on important elements makes the transformer highly effective for tasks where context across the entire input sequence is crucial. For instance, in language translation, a word at the beginning of a sentence might depend on a word at the end. The transformer's attention mechanism excels at capturing these dependencies regardless of the distance between them.

One of the key advantages of the transformer is its parallelizability. Unlike RNNs, which process input sequentially, the transformer processes all elements of the input simultaneously, which greatly speeds up training and inference. This property made transformers the dominant architecture in NLP, leading to state-of-the-art performance on tasks such as machine translation, text classification, and language generation.

Following its success in NLP, researchers began exploring the application of transformers to computer vision tasks, culminating in Dosovitskiy et al.'s "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE" (2020). This work introduced the Vision Transformer (ViT), which adapts the transformer architecture for image classification by treating images as sequences of patches, analogous to words in a sentence. Instead of processing images pixel by pixel, ViT divides an image into fixed-size

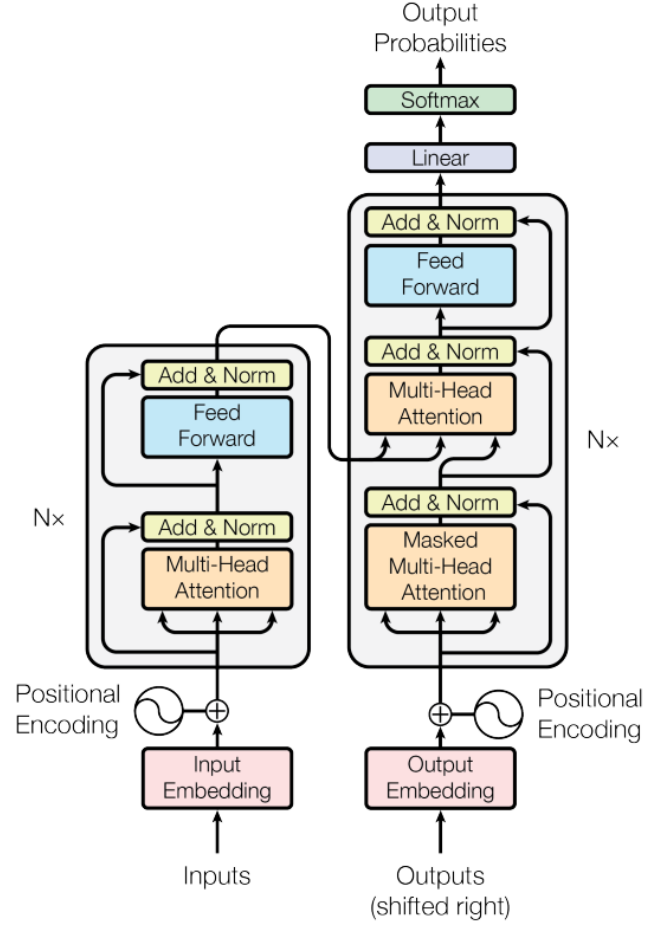


Fig. 1. Self-attention head mechanism.

patches (e.g., 16x16 pixels), each of which is linearly embedded into a vector representation. These patch embeddings are then fed into the transformer model, where self-attention mechanisms capture relationships between patches, allowing the model to learn global image features effectively.

The key benefit of applying transformers to vision tasks is their ability to capture long-range dependencies between image regions, which traditional convolutional neural networks (CNNs) struggle to do efficiently. CNNs use local receptive fields and pooling layers to progressively build feature hierarchies, focusing primarily on nearby pixels. While effective for many tasks, this approach limits their capacity to model global relationships, which can be crucial for tasks like object detection and image recognition in highly structured or cluttered scenes. By contrast, transformers excel at capturing both local and global patterns simultaneously, making them especially suited for tasks that require holistic image understanding.

Moreover, transformers are inherently more flexible than CNNs in terms of handling multi-scale features. Since the self-attention mechanism operates over the entire image at once, the transformer can naturally attend to objects of different sizes and scales without needing specialized multi-scale architec-

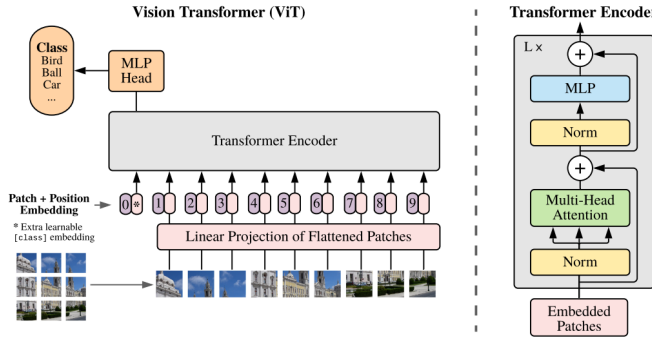


Fig. 2. Vision transformer mechanism.

tures, as is often the case with CNN-based approaches.

C. Vision Transformers for SAR Classification

The success of transformers in natural language processing (NLP) has paved the way for their application in computer vision, leading to the development of vision transformers (ViTs). One of the key motivations for adapting transformers to vision tasks is their ability to model long-range dependencies and capture global context—two aspects where traditional convolutional neural networks (CNNs) often fall short. CNNs excel at extracting local features through convolutional layers with small receptive fields, but as images grow in size and complexity, CNNs struggle to integrate information from distant parts of the image efficiently. This limitation is especially pronounced in tasks that require a global understanding of the scene, such as object detection and target recognition in SAR images.

In the paper "Vision Transformers: A Survey," the evolution of transformers in vision tasks is comprehensively explored, highlighting the growing trend of using attention mechanisms to overcome the limitations of convolution-based approaches. Vision transformers (ViTs) process images by dividing them into patches, which are treated as tokens, akin to words in NLP. These patches are then embedded and fed into a transformer architecture that learns the relationships between different patches using self-attention mechanisms. This approach allows the model to capture both local details and global patterns simultaneously, which is crucial for vision tasks that require understanding of both small objects and larger spatial structures.

One of the key benefits of using transformers in vision tasks is their ability to handle global context more effectively than CNNs. While CNNs rely on progressively larger receptive fields through stacked layers to capture broader context, transformers can attend to all regions of an image at once, regardless of their spatial distance. This makes transformers particularly useful in applications like object detection and semantic segmentation, where understanding the relationship between different parts of the image is critical. For example, in SAR target recognition, transformers can help identify targets that may be small and dispersed across a noisy background,

as they can weigh the importance of distant features without the constraints imposed by local convolutions.

One notable advancement in vision transformers is the development of Pyramid Vision Transformers (PVT), introduced to address some of the challenges associated with applying transformers to dense prediction tasks like object detection and segmentation. The original Vision Transformer (ViT) model struggles with high-resolution images due to its quadratic complexity in relation to the input size, which makes it computationally expensive for large-scale vision tasks. PVT addresses this issue by introducing a pyramid structure similar to that found in CNNs, where feature maps are progressively downsampled to reduce computational cost while maintaining the ability to capture multi-scale information. This design allows PVT to efficiently handle high-resolution images and perform dense predictions, which is particularly beneficial in SAR image analysis, where the spatial resolution and scale of the targets can vary significantly.

In addition to their computational efficiency, transformers offer a more flexible and scalable architecture compared to CNNs. CNNs require specific designs (such as hierarchical layers and pooling mechanisms) to capture multi-scale features, whereas transformers can naturally handle varying scales by adjusting the attention mechanism. This makes transformers better suited for tasks where objects or features appear at different scales, as is often the case in SAR imagery. Furthermore, transformers are more robust to noise and occlusions, two major challenges in SAR data, because the attention mechanism can focus on the most relevant regions of the image and ignore less informative or noisy parts.

D. SAR Image Classification with Transformers

As the adoption of transformers in computer vision progresses, researchers have increasingly explored their application to Synthetic Aperture Radar (SAR) imaging, addressing many of the challenges unique to this modality. SAR imaging, with its high-dimensional data, speckle noise, and complex spatial relationships, benefits from the transformer's ability to capture long-range dependencies and global context through attention mechanisms. While traditional models like convolutional neural networks (CNNs) have made strides in SAR automatic target recognition (ATR), transformers are showing promise in advancing the state of the art by overcoming the limitations of CNNs.

One of the earlier applications of transformers to SAR imagery is explored in "Towards SAR Automatic Target Recognition: MultiCategory SAR Image Classification Based on Light Weight Vision Transformer." This work illustrates the specific use cases of transformers in SAR image classification, focusing on how transformers can handle the unique challenges of SAR data more effectively than CNNs. The proposed architecture in this paper adapts the transformer for lightweight SAR image classification, demonstrating its ability to manage SAR's high-resolution data and varying target scales. By applying attention mechanisms, the model can focus on key parts of the image, reducing the impact

of noise and irrelevant features, which are common in SAR datasets. This improves classification accuracy, particularly in challenging scenarios where targets may be small or obscured by background clutter.

Another early adaptation of transformers in SAR is presented in "SAR Image Classification Based on Transformer Network," where the authors highlight the performance improvements gained by using transformers in place of traditional CNNs. This work demonstrates how the transformer architecture can be modified to suit SAR-specific tasks, such as automatic target recognition and scene classification, by focusing on long-range dependencies within the radar backscatter patterns. Unlike CNNs, which primarily rely on local filters and pooling operations, transformers in this context use attention to weigh the importance of different parts of the image, allowing for more robust feature extraction. The study highlights the significant boost in performance transformers can provide, particularly in terms of precision and recall, compared to conventional deep learning models.

More recently, the "ViT-SAR: Transformer-Based SAR Image Classification" paper has shown further advancements in applying transformers to SAR imaging. This work leverages the Vision Transformer (ViT) architecture, initially developed for optical image classification, and adapts it specifically for SAR data. The authors explore how ViT can be modified to better capture the distinctive features of SAR images, including the introduction of positional embeddings and additional layers to handle SAR's high-resolution and high-dimensional characteristics. The model outperforms traditional CNN-based methods by offering better generalization across varying terrain types and sensor configurations, which are common in SAR imaging. The ability of transformers to incorporate global context and focus on key image regions allows ViT-SAR to achieve higher accuracy in SAR target classification, even when faced with significant noise and variability in the data.

A significant advantage of using transformers in SAR applications is their ability to capture multi-scale features without the need for complex architectural modifications. CNNs typically require additional mechanisms like multi-scale feature pyramids or specialized pooling layers to handle targets of different sizes or resolutions. In contrast, transformers can naturally attend to objects at different scales through their attention mechanism, which operates across the entire image. This is especially beneficial in SAR imagery, where targets may vary significantly in size and often appear at different spatial resolutions depending on the scene or sensor.

Another key aspect of transformer-based approaches in SAR is their ability to reduce the impact of speckle noise, which is a common challenge in SAR image processing. Through the use of attention mechanisms, transformers can focus on more informative regions of the image and ignore noisy areas, improving overall classification performance. This selective focus is particularly useful in SAR automatic target recognition, where distinguishing small targets from noisy backgrounds is critical for success.

E. Attention Mechanisms for SAR Data

Attention mechanisms, which lie at the core of transformer architectures, have proven highly effective in addressing some of the unique challenges presented by Synthetic Aperture Radar (SAR) data. By dynamically focusing on the most relevant parts of the data, attention mechanisms provide a powerful tool for improving SAR image classification and recognition, allowing models to selectively attend to critical features while mitigating the effects of noise and irrelevant information.

In "A Semi-Supervised SAR Image Recognition Algorithm Based on Attention Mechanism," the authors demonstrate how attention mechanisms can be applied specifically to SAR data to enhance recognition performance, even in semi-supervised settings where labeled data is limited. SAR images are often plagued by speckle noise, a granular interference that complicates the identification of targets. Traditional models like CNNs struggle to filter out this noise effectively. The attention mechanism in this context helps the model focus on the more informative regions of the image, reducing the influence of noisy areas. By learning to prioritize key spatial features and ignore irrelevant ones, the model significantly improves its ability to classify SAR targets, particularly in cases where only a small portion of the image contains useful information.

The spatial attention mechanism in particular has been shown to be highly effective for SAR data. It operates by learning to assign different weights to various parts of the image, giving more importance to regions that are likely to contain targets or relevant features while down-weighting noisy or redundant areas. This selective focus helps mitigate the impact of speckle noise, which can otherwise overwhelm traditional feature extraction methods. By highlighting the most important regions, attention mechanisms provide more robust and accurate target classification, especially in noisy and cluttered environments.

In "Attention-Based Transformer Networks for High-Resolution SAR Image Classification," the authors take this further by exploring both channel attention and spatial attention mechanisms for high-resolution SAR data. In SAR imaging, the high-dimensionality of the data can be challenging, as traditional models may struggle to efficiently process such large and detailed images. Channel attention mechanisms help by weighing the importance of different channels (or features) in the data, allowing the model to focus on the most relevant spectral information. Meanwhile, spatial attention mechanisms prioritize the most significant parts of the image in terms of spatial location. Together, these attention mechanisms allow the model to effectively manage the complexity of high-resolution SAR data, improving the precision of target classification.

This combination of spatial and channel attention is particularly useful in SAR target recognition, where distinguishing small, often subtle targets within a larger scene is critical. For example, in high-resolution SAR imagery, targets may be small and sparsely distributed, making it difficult for

traditional models to detect them amidst background noise. Attention-based transformers, however, can attend to the subtle patterns that indicate the presence of a target, while filtering out irrelevant or misleading information, resulting in a more accurate recognition process.

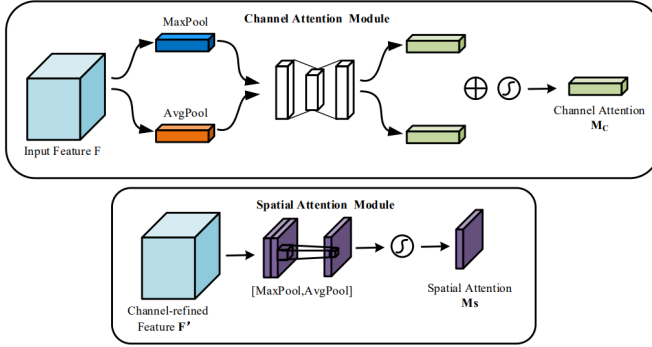


Fig. 3. Spatial and channel attention mechanism

Additionally, attention mechanisms are particularly well-suited for addressing the multi-scale nature of SAR data. In many SAR applications, targets can vary significantly in size depending on the resolution of the sensor and the distance from the radar. Attention-based methods are inherently capable of handling these variations, as they do not rely on fixed-size convolutional kernels like CNNs, but instead dynamically adjust their focus based on the context of the data. This flexibility allows attention mechanisms to effectively capture both small, detailed features and larger, more global patterns, improving the model's ability to classify targets of various scales within SAR images.

F. Hybrid Models Combining Transformers and CNNs

Recent advances in Synthetic Aperture Radar (SAR) image classification have seen the emergence of hybrid models that combine the strengths of transformers with other architectures, such as convolutional neural networks (CNNs). These innovations aim to leverage the complementary strengths of both transformers and CNNs, resulting in models that are better suited to the challenges posed by SAR data, such as handling high-dimensionality, noise, and the need for robust feature extraction across varying scales.

One significant contribution to this field is the integration of self-supervised learning with transformer-based architectures. In "Rethinking SAR ATR with Self-Supervised Learning and Vision Transformers," [16] the authors propose a hybrid approach that combines the unsupervised learning capabilities of transformers with CNN-like architectures for automatic target recognition (ATR) in SAR. Self-supervised learning is a technique that allows models to learn useful feature representations from unlabelled data, a particularly valuable approach for SAR imagery, where labeled data can be scarce or difficult to obtain. By using transformers in a self-supervised framework, the model is able to extract meaningful patterns from SAR images without the need for extensive labeled

training datasets. This hybrid approach enhances SAR ATR by making it more robust to variations in data quality and sensor configurations, and it reduces the reliance on large labeled datasets—a key limitation in SAR applications.

In this model, CNNs are often used to perform initial feature extraction from the SAR images, leveraging their ability to handle fine-grained local patterns. The transformer component is then applied to capture long-range dependencies and global context across the image, compensating for the CNN's limitations in handling large-scale spatial relationships. The result is a system that benefits from both the local feature extraction of CNNs and the global context modeling of transformers, leading to improved performance in SAR target recognition tasks. This hybrid model demonstrates how the combination of CNNs and transformers can overcome the challenges posed by SAR data, such as noise and complex terrain.

Another noteworthy innovation is the application of transformer-based architectures to object detection tasks, which can be directly linked to SAR target recognition. In "End-to-End Object Detection with Transformers," the authors present an architecture that replaces the traditional region proposal network (RPN) found in object detection models with a transformer, allowing for an end-to-end approach. This model, known as DETR (Detection Transformer), is particularly relevant to SAR target recognition, where accurate object detection is crucial for identifying targets in complex environments. DETR eliminates the need for hand-crafted anchors, which are commonly used in CNN-based detection models, and instead relies on the transformer's attention mechanism to directly predict bounding boxes and classify objects. This approach offers several advantages for SAR target recognition, including improved detection of small or partially occluded targets, which are common in SAR imagery.

The hybridization of transformers with other architectures not only enhances object detection but also improves the overall robustness of SAR image classification systems. In many SAR applications, targets are small, sparse, or camouflaged within complex terrain, making them difficult to detect using traditional models. By leveraging the attention mechanisms of transformers, these hybrid models can more effectively differentiate between targets and background clutter, significantly improving detection accuracy.

The adoption of transformer architectures in Synthetic Aperture Radar (SAR) image classification has sparked significant interest, particularly in how these models compare to traditional deep learning approaches like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Each model offers unique advantages, but they differ in their ability to handle SAR-specific challenges such as high-dimensionality, speckle noise, and the requirement for robust feature extraction across varying scales.

When comparing the accuracy of transformers and CNNs, several studies show that transformers consistently outperform CNN-based models in SAR classification tasks. In "SAR Image Classification Based on Transformer Network" and "ViT-SAR: Transformer-Based SAR Image Classification,"

transformers demonstrated superior performance, particularly in terms of classification accuracy on high-resolution SAR datasets. This is largely attributed to the transformer’s ability to model global dependencies within the image through its attention mechanisms, enabling it to capture long-range relationships and subtle spatial features more effectively than CNNs, which rely on local convolutional filters. CNNs tend to struggle with recognizing small or distant objects in SAR imagery, especially when there is significant noise or clutter, whereas transformers excel by focusing attention on the most relevant parts of the image.

However, this improvement in accuracy often comes with increased computational cost. Transformers, particularly in their standard form, are known for their heavy computational requirements, especially when applied to high-resolution SAR data. The self-attention mechanism, which is at the core of transformers, scales quadratically with the input size, making it more computationally expensive than the localized operations in CNNs. This is a key limitation, particularly when processing large SAR datasets in real-time or resource-constrained environments. Efforts to address this include the development of more efficient variants of transformers, such as the lightweight vision transformer introduced in “Towards SAR Automatic Target Recognition.” These models aim to reduce the computation load by using fewer parameters while maintaining high accuracy, though CNNs may still offer an advantage in scenarios where computational efficiency is paramount.

In terms of data efficiency, transformers have historically been more data-hungry than CNNs, requiring large amounts of labeled data to achieve optimal performance. This poses a challenge for SAR applications, where labeled datasets are often limited due to the complexity of manually annotating SAR images. However, recent developments in self-supervised learning and semi-supervised approaches have made transformers more viable for SAR tasks with limited data. For example, “Rethinking SAR ATR with Self-Supervised Learning and Vision Transformers” showed that self-supervised pre-training can significantly reduce the amount of labeled data required for transformers to perform well on SAR classification tasks. These approaches enable the model to learn meaningful representations from unlabeled data, improving its performance on SAR datasets that would otherwise be insufficient for training a transformer from scratch.

In contrast, CNNs have traditionally been more data-efficient, as they can achieve reasonable performance even with smaller training sets. This is partly due to the inductive biases built into CNN architectures, such as translation invariance and local connectivity, which help them generalize better with limited data. However, these same inductive biases limit CNNs in their ability to capture complex global patterns in SAR imagery, particularly when dealing with long-range dependencies or non-localized features, which transformers handle more effectively.

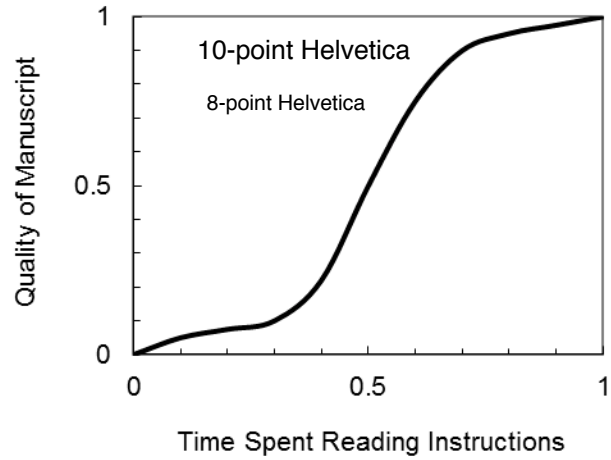


Fig. 4. Estimated relationship between the time an author spends reading these instructions and the quality of the author’s digest article.

III. FIGURES

IV. CITING PREVIOUS WORK

When referencing a journal article [?], a conference digest article [?] or a book [?], place the reference numbers within square brackets. To simultaneously cite these references [?] - [?] use the format just demonstrated. The reference list is the last section and references are listed in the order cited. Use 9 point Times. The paragraph description is set for a line spacing of exactly 10 points with 0 point spacing before and after. A 0.25 inch hanging indentation should be specified. Generally speaking, references should be very detailed. For journal articles, list all authors by initials and last name, the title of the paper in quotations (capitalizing only the first letter of the first word, unless it would be capitalized in a sentence, e.g., a proper noun), the journal name in italics, the volume number, the issue number, the page numbers, and the date. Use the examples provided [?] - [?] as a guide.

For the double-blind submission, citations of the authors’ own work should be worded in a way that avoids identifying any connection to the authors. Simply note your prior work in the same way as work by other authors. For example, **do not write**, “We (or the authors) demonstrated in [x] that. . .” Rather write, “It was demonstrated in [x] that. . .”

V. CREATION OF THE PDF FILE

VI. CONCLUSION

Following these instructions will improve the quality of your paper and the IMS Digest. If you have comments, please contact one of the Steering Committee editors.

ACKNOWLEDGMENT

For the Summary paper submission only, no acknowledgements are allowed.

REFERENCES

- [1] R. K. Raney, "Synthetic aperture radar and its applications," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 2, pp. 117-132, 1998.
- [2] J.-S. Lee and E. Pottier, *Polarimetric Radar Imaging: From Basics to Applications*, CRC Press, 2009.
- [3] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images*, Artech House, 2004.
- [4] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the future," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8-36, 2017.
- [5] H. Zhang et al., "Exploring deep learning methods for classification of SAR images: Towards next-gen convolutions via transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6207-6221, 2020.
- [6] Y. Liu, X. Wang, and J. Gao, "A survey of deep learning-based SAR target recognition," *Remote Sensing*, vol. 11, no. 9, p. 1069, 2019.
- [7] A. Vaswani et al., "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000-6010, 2017.
- [8] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2020.
- [9] S. Khan et al., "Vision Transformers: A Survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10, pp. 1-34, 2021.
- [10] W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568-578, 2021.
- [11] L. Wang et al., "Towards SAR Automatic Target Recognition: Multi-Category SAR Image Classification Based on Light Weight Vision Transformer," *Remote Sensing*, vol. 13, no. 15, p. 3027, 2021.
- [12] Y. Zhang et al., "SAR Image Classification Based on Transformer Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11079-11091, 2021.
- [13] H. Zhao et al., "ViT-SAR: Transformer-Based SAR Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2022.
- [14] S. Li et al., "A Semi-Supervised SAR Image Recognition Algorithm Based on Attention Mechanism," *IEEE Access*, vol. 10, pp. 22285-22296, 2022.
- [15] X. Ma et al., "Attention-Based Transformer Networks for High-Resolution SAR Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-12, 2022.
- [16] H. Wang et al., "Rethinking SAR ATR with Self-Supervised Learning and Vision Transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-10, 2022.
- [17] N. Carion et al., "End-to-End Object Detection with Transformers," in *European Conference on Computer Vision (ECCV)*, pp. 213-229, 2020.

Note: For the Summary paper submission only, references to the authors own work should be cited as if done by others to enable a double-blind review. **Citations must be complete and not redacted, allowing the reviewers to confirm that prior art has been properly identified and acknowledged.**