Published in final edited form as:
Mach Learn Appl. 2024 June ; 16: . doi:10.1016/j.mlwa.2024.100549.

INSTRAS: INfrared Spectroscopic imaging-based TRAnsformers for medical image Segmentation

Hangzheng Lin[a], **Kianoush Falahkheirkhah**[b], **Volodymyr Kindratenko**[a,c], **Rohit Bhargava**[a,b,d,*]

[a]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, United States

[b]Beckman Institute, University of Illinois at Urbana-Champaign, IL, United States

[c]Center for Artificial Intelligence Innovation, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, IL, United States

[d]Departments of Bioengineering, Mechanical Science and Engineering and Cancer Center at Illinois, University of Illinois at Urbana-Champaign, IL, United States

Abstract

Infrared (IR) spectroscopic imaging is of potentially wide use in medical imaging applications due to its ability to capture both chemical and spatial information. This complexity of the data both necessitates using machine intelligence as well as presents an opportunity to harness a high-dimensionality data set that offers far more information than today's manually-interpreted images. While convolutional neural networks (CNNs), including the well-known U-Net model, have demonstrated impressive performance in image segmentation, the inherent locality of convolution limits the effectiveness of these models for encoding IR data, resulting in suboptimal performance. In this work, we propose an INfrared Spectroscopic imaging-based TRAnsformers for medical image Segmentation (INSTRAS). This novel model leverages the strength of the transformer encoders to segment IR breast images effectively. Incorporating skip-connection and transformer encoders, INSTRAS overcomes the issue of pure convolution models, such as the difficulty of capturing long-range dependencies. To evaluate the performance of our model and existing convolutional models, we conducted training on various encoder–decoder models using a breast dataset of IR images. INSTRAS, utilizing 9 spectral bands for segmentation, achieved a remarkable AUC score of 0.9788, underscoring its superior capabilities compared to purely convolutional models. These experimental results attest to INSTRAS's advanced and improved segmentation abilities for IR imaging.

[*]Correspondence to: Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States. rxb@illinois.edu (R. Bhargava).

CRediT authorship contribution statement
Hangzheng Lin: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Kianoush Falahkheirkhah:** Conceptualization, Data curation, Writing – original draft. **Volodymyr Kindratenko:** Conceptualization, Writing – review & editing, Supervision. **Rohit Bhargava:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Keywords

Image segmentation; Transformer; Infrared imaging; Machine learning

1. Introduction

Chemical imaging techniques are very attractive for many scientific disciplines, allowing for understanding both the molecular and morphological structure of specimens. By relying on the intrinsic composition of the sample, these techniques eliminate the need for extrinsic labeling. Such an approach not only maintains sample integrity but also provides a direct and accurate representation of its unique characteristics. Among the available imaging methods, Infrared (IR) imaging distinctly stands out. With its capacity to offer molecular insights through fundamental vibrational modes, and ensuring minimal disruption to samples, IR imaging has been a preferred choice, especially in scenarios where sample integrity is critical (Baker et al., 2014; Bhargava, 2023; Geinguenaud, Militello, & Arluison, 2020; Kedzierski et al., 2019; Zhang, Gao, & Yilmaz, 2020). Traditionally, the analytical approach to IR images, particularly for segmentation tasks, has been heavily pixel-centric (Fernandez, Bhargava, Hewitt, & Levin, 2005; Lasch, Diem, Hänsch, & Naumann, 2006). Each pixel was typically analyzed in isolation, focusing on its spectral properties without considering the broader spatial context. However, the integration of computational advancements, especially in the realm of deep learning, suggests a more holistic approach. Convolutional neural networks (CNNs) and transformers propose a potential for a detailed analysis of IR data, integrating both spectral depth and spatial context. This combined approach can illuminate subtle patterns and richer insights in the IR data.

In the recent past, CNN-based models (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2018; Pang, Du, Orgun, & Yu, 2018; Ronneberger, Fischer, & Brox, 2015) have made significant strides in the domain of biomedical image segmentation. The architecture of such models, with its convolution layers intertwined with skip connections, ensures the retention of crucial feature information even amidst complex data transformations. However, when confronted with IR data's multidimensional nature and its burgeoning band complexity, conventional convolution layers might struggle. The challenge lies in encapsulating the intricate dependencies and relationships inherent in such expansive IR datasets. Transformers offer a potential solution to this challenge. Originally conceptualized for natural language processing (NLP) (Devlin, Chang, Lee, & Toutanova, 2019; Radford & Narasimhan, 2018; Vaswani et al., 2017), transformers are anchored by a self-attention mechanism. This mechanism is tailored to capture long-range dependencies, a trait that is invaluable in the context of hyperspectral imaging. Given the spectral richness of each pixel in such imaging, understanding the inter-band relationships is crucial. Transformers, with their self-attention mechanism, are well-equipped to cater to this analytical demand, as supported by emerging research (Dosovitskiy et al., 2021; Hatamizadeh, Yang, Roth, & Xu, 2021; Liu et al., 2021).

Capitalizing on this foundation, our study introduces INSTRAS, a novel deep learning architecture. INSTRAS seamlessly integrates the virtues of transformers with the tried-andtested capabilities of the U-Net structures. Our aim with this innovation is to elevate the efficacy of multi-class IR medical image segmentation. Preliminary empirical evaluation showcases the prowess of INSTRAS, indicating its superior performance over established models, such as U-Net and attention U-Net. This positions INSTRAS as a potential foundation in the future landscape of IR data analysis within medical imaging.

2. Related work

2.1. CNN-based models

The evolution of image segmentation techniques has been marked by the increasing prominence of fully convolutional networks (FCNs). Unlike their predecessors, which were rooted in manual data handling, FCNs are equipped to adapt to a myriad of images and tasks. This adaptability is manifest in the development of architectures like VGG (Simonyan & Zisserman, 2015), ResNet (He, Zhang, Ren, & Sun, 2015), and Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017). By integrating CNNs with fully connected layers, advancements in segmentation have been observed across diverse medical imaging areas such as skin (Esteva et al., 2017; Li & Shen, 2018), colon (Chen et al., 2017; Graham et al., 2019), and breast (Hatipoglu & Bilgin, 2017). U-Net's success (Ronneberger et al., 2015) further underscores the remarkable feature extraction capabilities of CNNs, leading to subsequent models employing the U-Net's encoder–decoder architecture (Alom, Yakopcic, Hasan, Taha, & Asari, 2019; Isensee et al., 2018; Zhou, Siddiquee, Tajbakhsh, & Liang, 2018; Zhuang, 2018). However, with the technological advancements in image acquisition producing increasingly high-resolution images, CNNs confront challenges regarding their receptive field. Various architectures attempt to expand the receptive field either through larger kernel sizes or by implementing dilated convolution (Yu & Koltun, 2015). Yet, capturing long-range information throughout an image remains an area that needs further exploration. In our current study, we enhance the U-Net architecture by integrating the transformer layer to supplant traditional convolution.

2.2. Transformer models

The impact of transformers, initially observed within the realm of NLP (Devlin et al., 2019; Radford & Narasimhan, 2018; Vaswani et al., 2017), has found its place within the visual processing domain. One of the most salient attributes of transformers is their self-attention mechanism, which addresses the locality constraints endemic to conventional CNN encoders. The Vision Transformer (ViT) (Dosovitskiy et al., 2021) approach emerged as a significant milestone in this arena. By extracting non-overlapping image patches and repurposing them as sequential inputs to the transformer encoder, it was able to integrate and process spatial information effectively. This was further enhanced with the introduction of positional embeddings, ensuring that spatial relationships between patches were maintained. The overarching principle underscoring this design is the transformer's ability to discern intricate, high-level features, which becomes especially invaluable when confronted with complex datasets like IR images. This pioneering work with ViT catalyzed the exploration of diverse architectures tailored for distinct vision tasks.

The DETR model (Carion et al., 2020), for instance, presented a unique fusion of the self-attention mechanism with CNN encoders. This amalgamation not only streamlined the object detection process but also eliminated the need for multiple hand-crafted components, rendering the process more end-to-end. The SWIN Transformer (Liu et al., 2021) brought in another evolution, introducing a hierarchical structure reminiscent of pyramid networks. By leveraging shifted windows and strategic partitioning of self-attention computation, SWIN balanced the need for both global and local information extraction, while ensuring computational efficiency. Recent innovations in medical imaging have underscored the potential of hybrid models. The TransUNet (Chen, Lu et al., 2021), for example, employs a hybrid encoder interweaving both CNNs and vision transformers. This encoder, when combined with a U-Net style decoder supplemented with skip connections, has proven effective at medical image segmentation tasks. Similarly, the UNETR model (Hatamizadeh et al., 2021) takes this hybrid approach into the realm of 3D imaging. Here, non-overlapping 3D patches extracted from input volumes are transformed using linear embedding and positional encoding. Once processed through the transformer blocks, these features are subsequently decoded using a U-Net inspired 3D structure, illustrating the versatility of transformer-based approaches across imaging modalities. In summary, transformer architectures, initially conceptualized for text, have shown immense promise in visual domains, from general object detection to specialized medical imaging tasks, highlighting their adaptability and potential for future explorations.

2.3. Deep learning for IR imaging

The fusion of IR imaging with deep learning has opened up new avenues for advancing scientific research. Historically, IR imaging has been valued for its capability to probe the molecular composition of samples, delivering insights without the requirement for external labels. This unique advantage is even more pronounced when combined with deep learning techniques, given their capacity to extract patterns from complex datasets (Pradhan, Guo, Ryabchykov, Popp, & Bocklitz, 2020). Recent years have seen a noteworthy upswing in the application of deep learning models to IR imaging. These methods, while previously constrained to conventional imaging modalities, have now demonstrated immense promise in the IR spectrum (Keogan et al., 2021; Muniz, Baffa, Garcia, Bachmann, & Felipe, 2023). Conventional convolutional neural networks, for instance, have been employed to segment and distinguish varying regions of interest within IR images, achieving a level of precision previously unattained (Berisha et al., 2019; Falahkheirkhah, Yeh, Mittal, Pfister, & Bhargava, 2021; Tiwari, Falahkheirkhah, Cheng, & Bhargava, 2022). Furthermore, these deep learning techniques have not just been limited to analysis but have also made strides in enhancing the raw IR data quality. Advanced algorithms have been developed to refine the resolution, suppress noise, and improve the overall reliability of the collected IR data, further elevating the value and accuracy of subsequent analyses (Falahkheirkhah, Yeh, Confer, & Bhargava, 2022; Magnussen et al., 2020).

3. Proposed method

3.1. Overview

In this section, we present our model architecture and loss function. An overview of the proposed method, which is composed of seven main modules, is shown in Fig. 1. First, a $P \times P$ convolution layer with stride size P is utilized for linear projection and patch embedding, where P is the patch size. Second, a multi-head transformer layer is designed for image feature encoding. Third, a patch merging layer is used after each transformer block and aims to down-sampling features and increase the channel dimension. Fourth, several sequences 3 × 3 convolution layer followed by batch normalization and a Rectified Linear Unit (RELU) activation serves as a decoder for feature extraction. Fifth, an attention gate with three 1 × 1 convolution layers is applied to help the network to focus on relevant regions. Sixth, a 2 × 2 transposed convolution layer is implemented for feature extraction and upsampling. Finally, a 1 × 1 Convolution layer is used as the output layer to generate the final segmentation result.

The motivation behind this design is to fully utilize the U-Net shape structure to compress multidimensional spatial information. Simultaneously, the self-attention mechanism allows the model to focus on the long-range relevant patterns. The combination ensures coherent feature propagation and optimizes the model's capability to process the multidimensional nature of IR data.

3.2. INSTRAS architecture

Given an input image $I \in R^{H \times W \times C}$ and a patch size of P, we uniformly extract non-overlapping patches each with the size of $P \times P$ to form a new shape of $x_p \in R^{N \times P^2 \times C}$, where $N = \frac{H}{P} \times \frac{W}{P}$.

We then project flattened patches x_p into E dimensional space with a linear layer. To maintain the spatial information, we apply a learnable positional embedding to x_p, which can be represented as:

$$z_0 = \left[x_p^1 \mathbf{E}; x_p^2; \ldots; x_p^N \mathbf{E}\right] + \mathbf{E}_{pos}, \tag{1}$$

where the $\mathbf{E} \in R^{\left(P^2 \cdot C\right) \times E}$ and $\mathbf{E}_{pos} \in R^{N \times E}$ denote the embedding projection, and the positional embedding respectively. Notice that unlike the transformer used in NLP and ViT, we do not have the $[x_{Class}]$ since our task is about image segmentation but not classification.

The embedded patches are fed into a transformer encoder, consisting of four transformer blocks and four patch merging blocks. Each transformer block consists of L layers of multiheaded self-attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Thus each layer of the transformer encoder can be written as:

$$z_l' = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1}, \tag{2}$$

$$z_l = \mathrm{MLP}\left(\mathrm{LN}\left(z_l'\right)\right) + z_l', \tag{3}$$

where LN is the layer normalization and $l \in [1, L]$ denotes the layer number.

Inspired by the U-Net (Ronneberger et al., 2015) and Swin-Unet (Liu et al., 2021) architectures, we extract four transformer block outputs and merge them back to the original image size. Convolution and transpose convolution layers are used for decoding and upsampling features, respectively. We use an attention gate (Oktay et al., 2018) before each concatenation operation to selectively weight feature maps. For each pixel $x_i^n \in R^{F_n}$ at skip-connection level n with F_n size of the feature map, and its corresponding gating signal $g_i \in R^{F_g}$ taken from its previous level, the attention gate can be represented as:

$$q_{att}^n = \psi^T\left(\sigma_1\left(W_x^T x_i^n + W_g^T g_i + b_g\right)\right) + b_\psi, \tag{4}$$

$$\sigma_2(x) = \frac{1}{1 + exp(-x)}, \tag{5}$$

$$\alpha_i^n = \sigma_2(q_{att}^n(x_i^n, g_i; \Theta_{att})), \tag{6}$$

where $W_x \in R^{F_n \times k}$, $W_g \in R^{F_n \times k}$, and $\psi \in R^{k \times 1}$ are linear projections implemented by using 1 × 1 2D Convolutions. We typically set the hidden dimension of k as half of the input dimension F_n. The $b_g \in R^k$ and $b_\psi \in R$ are added as bias terms. The σ_1 is an activation layer that RELU implements. The σ_2 is a sigmoid function that maps the attention into the range between 0 and 1. The q_{att}^n represents the process of the attention calculation for level n with the attention gate parameters Θ_{att}. After getting the attention coefficient α_i^n, we time it to the input vector x^n to get the attention results:

$$\hat{x}^n = \alpha^n x^n, \tag{7}$$

where the $\hat{x}^n$ is the output of the attention gate, whose spatial regions are selected by the information from its lower skip-connections.

We utilize the dense layer to implement the patch merging blocks. These merging blocks play a crucial role in dividing the input feature into four parts and subsequently concatenating them together. Afterward, the dense layer is responsible for transforming the resulting four-dimensional feature into a two-dimensional representation.

3.3. Loss function

To train our model, we use a combination of the Soft Dice (Milletari, Navab, & Ahmadi, 2016) loss function and SoftMax cross entropy. The Soft Dice can be represented as:

$$\mathscr{L}_{DICE}(P, Y) = 1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N} P_{ic}Y_{ic} + \epsilon}{\sum_{c=1}^{C}\sum_{i=1}^{N} P_{ic}^2 + \sum_{c=1}^{C}\sum_{i=1}^{N} Y_{ic}^2 + \epsilon}, \tag{8}$$

where P and Y are the predicted output probability and the target label for all C classes, respectively, and N is the number of pixels in the image. A small positive value ϵ is added to avoid division by zero.

The SoftMax cross entropy loss calculates the difference between the predicted class probability processed by a softmax activation function and the target labels. The cross entropy loss is defined as following:

$$\mathscr{L}_{CE}(P, Y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} Y_{ic}\log(P_{ic}), \tag{9}$$

and the final loss is a combination of $\mathscr{L}_{DICE}$ and $\mathscr{L}_{CE}$:

$$\mathscr{L}(P, Y) = (1-\lambda)\mathscr{L}_{DICE}(P, Y) + \lambda\mathscr{L}_{CE}(P, Y), \tag{10}$$

where λ is a constant that balances the weight between dice loss and cross entropy loss. Experimentally, we set λ as 0.98 for all trainings in this study.

4. Experiments and results

4.1. Dataset

A breast Tissue MicroArray (TMA) known as BR1003, sourced from US Biomax Inc., comprising 101 cores with a 1 mm diameter obtained from 47 individuals was used in this study. The TMA was constructed from formalin fixed paraffin-embedded tissue and was processed as per typically protocols prior to imaging (Fernandez et al., 2005). We used Hematoxylin and Eosin (H&E) and smooth muscle actin to stain two of the sections, which were then viewed under a light microscope. An unstained section of the same TMA, 5 μm in thickness, was positioned on a BaF2 salt plate for FT-IR imaging, utilizing transmission mode. The FT-IR imaging data were recorded using a 680-IR spectrometer that was paired with a 620-IR imaging microscope from Agilent Technologies. This setup had a numerical aperture of 0.62 and employed a Mercury Cadmium Telluride (MCT) 128 × 128 focal plane array detector, which was cooled by liquid nitrogen. More details about the dataset has been published previously (Mittal et al., 2018). We constructed INSTRAS utilizing weakly-supervised methods for the purpose of classifying breast tissue into key histological categories. The model leverages a previously designed random forest classifier for training label creation (Mittal et al., 2018). Essentially, pathologists annotate the FT-IR images – obtained from the FT-IR device – based on corresponding images from H&E-stained glass slides. This forms the basis or ‘‘ground truth’’ for the classification process. Subsequently, manually created metrics like peak height or area ratios serve as the input features for the network. The random forest classifier then uses these metrics and labels to establish a pixel-level classification model. This resulting classified imagery is then used to train INSTRAS. It is worth noting that this label generation procedure is a one-time training process used to train the model. Once complete, the model can segment the tissue with no additional intervention needed. Using weak labels to train IR images has been published previously (Falahkheirkhah et al., 2021).

This study employs the following six histological categories: dense stroma, loose stroma, reactive stroma, benign, epithelium, and malignant epithelium. We also include other cell types, grouped under the category ‘‘others’’ as previously defined. Based on biological significance, 9 IR bands were selected as the image channels: 1545, 3288, 1238, 2956, 1454, 2848, 1084, 1404, 1655 cm^{-1}.

4.2. Metrics

We use the Receiver Operating Characteristic (ROC) curve, Accuracy, and the Area Under the Curve (AUC) as the performance metrics for evaluation of the effectiveness of segmentation.

The ROC curve is a graph that shows the model performance for all different classification thresholds, represented by the true positive rate (Recall) against the false positive rate (Fall-Out). The true positive rate and the false positive rate can be written as:

$$TPR = \frac{TP}{TP + FN}, \tag{11}$$

$$FPR = \frac{FP}{FP + TN}, \tag{12}$$

where the TPR and FPR are the true positive rate and false positive rate, respectively. TP, FP, FN are a number of true positives, false positives, and false negatives. All TP, FP, and FP are calculated with the one-vs-all method, which treats each class as positive and other classes as negative.

The AUC is defined by the total area under the ROC curve, which measures the model's ability to distinguish the positive and negative samples. We can define the AUC as:

$$AUC = \int_0^1 TPR(x), dx, \tag{13}$$

where the domain of x is the false positive rate.

4.3. Implementation details

The networks in this study are implemented using the PyTorch deep learning library and executed on NVIDIA DGX A100 40 GB system with CUDA 11.0. To optimize the network performance, we used AdamW (Loshchilov & Hutter, 2017) as the optimizer with a learning rate of 0.001 and a weight decay rate of 0.001 for 1500 epochs. With a batch size of 32, the average training duration is around 21 h.

The training images have varying widths ranging from 900 to 1280 pixels and heights ranging from 900 to 1230 pixels. To ensure a fixed input image size, we randomly crop ten 224 × 224 pixel patches from each image for each iteration. Since the size of our data set is limited, we augment our data by flipping each patch along the vertical axis (with a 0.5 probability), rotating each patch by 90, 180, and 270 degrees (with a 0.25 probability for each angle), and adding a uniform noise (between 0 to 0.2) to each pixel (with a 0.5 probability).

We use the Embedding size E as 48, the patch size P as 2, and transformer layers number L for each transformer block as 3. All the transformer layers have 12 heads with a dropout rate of 0.1.

4.4. Comparison with pure convolution methods

To evaluate our proposed method and demonstrate that the IR images can be well encoded by involving a transformer, we compare the performance of INSTRAS with the widely used U-Net model and the attention U-Net. We train all the models using 6 and 9 IR bands. The IR bands we used follow the order mentioned in Section 4.1, which means if we train a 6-channel model, we use 1545, 3288, 1238, 2956, 1454, and 2848 cm^{-1} bands.

The ROC curves of both INSTRAS and U-Net trained on 9 IR bands are illustrated in Fig. 2. We use the validation set of 32 image patches from two external validation samples. By evaluating the area under the ROC curves, we observed in Table 1 that the AUC scores across various histological components reflect a clear advantage of our proposed method over other models for both 6 and 9 band analyses. Specifically, INSTRAS achieved mean AUC scores of 0.9595 and 0.9788 for 6 and 9 IR bands respectively, indicating an improvement in segmentation prediction over both U-Net and Attention U-Net. A detailed inspection of the AUC scores for each histological component further emphasizes the superiority of our proposed method, notably within noncancerous epithelium and reactive stroma.

Table 2 displays the performance accuracy of each model, with our proposed method, INSTRAS, outperforming all others across all evaluated cases. In addition to these quantitative evaluations, we also demonstrated the qualitative aspect by comparing the segmentation masks produced by INSTRAS and the U-Net models. Fig. 2 shows each model's prediction under different IR band configurations against the target mask. We employ an orange dashed box to underscore the pronounced differences between the outputs. The segmentation mask generated by our method retains more intricate information, whereas the U-Net versions tend to overlook some of the histological units. For instance, the segmentation offered by both the U-Net and Attention U-Net models inaccurately represents the shape of the malignant epithelium in the center of the second sample.

4.5. Entire image inference

As our model employs a transformer encoder designed to process a fixed number of patches of a specific size, it is incapable of directly predicting the entirety of a high-resolution IR image in a single inference. To surmount this limitation, we introduce an overlapping image patch-based prediction technique.

Our inference method divides the input image into overlapping patches, each adhering to the size requirements of the model's input. This scheme ensures efficient image processing while preserving the consistent input shape of our model. By significantly overlapping each patch with its neighbors, we aim to harness broader contextual data, thereby minimizing the potential discrepancies at the patch boundaries. For every image pixel, the prediction probabilities from all overlapping patch inferences are aggregated. Once all patches undergo processing, the class garnering the highest cumulative value gets designated as the ultimate prediction for that particular pixel.

Table 3 elucidates the interplay between the average inference time, the prediction accuracy of INSTRAS_9, and the overlapping coefficient δ. This coefficient represents the overlap percentage between adjacent patches. It is imperative to highlight that a larger overlap area can yield more accurate predictions, albeit at the expense of a more prolonged inference duration. However, as the δ increases, the gain in accuracy brought by additional inference time diminishes. By configuring δ to be less than 50%, we can attain real-time inference on a 1k × 1k high-resolution IR image without compromising the quality of the results.

5. Discussion

Based on our experimental results, we have demonstrated that INSTRAS outperforms traditional, purely convolutional models in the task of IR image segmentation. This superiority becomes particularly evident when the IR image encompasses a large number of bands. The advent of discrete frequency IR (DF-IR) imaging (Mittal et al., 2018) indicates a significant acceleration in data acquisition speed, primarily by recording only those bands pertinent to specific downstream tasks. It is also noteworthy that, unlike traditional visible light images which primarily rely on three channels (RGB), IR data is hyperspectral. This multi-channel property poses a unique advantage for transformers in IR segmentation. The ability to discern complex interrelationships across spectral bands and efficiently process high-dimensional data enhances scalability and effectiveness.

While prior studies have demonstrated advancements in achieving accurate medical image segmentation, it is crucial to note that transformer-based machine learning techniques generally require more computational resources compared to traditional convolutional approaches (Chen, Yang et al., 2021; Lu et al., 2023; Tang, Nan, Walsh, & Yang, 2023). As a result, transformer models inherently have longer inference times. Empirical data showcases that while the UNet model averages an inference time of 0.399 s on the test dataset and the Attention UNet completes the inference in approximately 0.554 s, INSTRAS requires about 2 s using default settings. But, with the self-attention mechanism intrinsic to transformers, there is a distinct advantage: the ability to model long-range dependencies across images and enable parallel sequence processing. With the monumental advancements in computational hardware in recent years, transformers can exploit these resources more effectively. Their proficiency in feature extraction and inherent scalability designates transformers as a promising avenue for future medical imaging tasks. The flexibility of INSTRAS, characterized by its adaptability to various data types and the customizable number of transformer layers in the encoder, emphasizes its potential utility in a myriad of medical imaging applications. The model's design allows for seamless retraining on new datasets, even with a divergent number of segmentation features. While transformer-based models offer numerous benefits, they also present several limitations. For instance, these models typically require large datasets, making the provision of extensive datasets for biomedical imaging both resource-intensive and challenging. Furthermore, transformer models tend to have longer training times due to their complex architectures, potentially prolonging the development cycle of machine learning projects in the biomedical field. Another significant drawback is their longer inference time, which poses challenges for real-time analysis and applications requiring immediate results.

As shown in Table 4, our model, in its default configuration, boasts trainable parameters comparable to conventional convolutional models. Nonetheless, its self-attention mechanism necessitates an extensive dataset of labeled samples. Securing high-quality IR biomedical images is both cost-intensive and intricate, making it difficult to obtain very large sets of IR images annotated with the ground truth. Predominantly, research in this domain hinges on expert evaluations and is supplemented by statistical methodologies, such as leveraging random forests (Breiman, 2001). In our study, the dataset labels, though comprehensive, are not flawless, hinting at potential annotation inconsistencies that could influence the training phase.

6. Conclusion

In this study, we introduced a novel U-Net-like deep learning architecture that incorporates a transformer encoder. Our findings indicate that this architecture outperforms both U-Net and Attention U-Net in the multi-class IR medical image segmentation task. The results underscore the transformer's superior capability to extract features from IR images in comparison to conventional convolution encoders. Moreover, it retains finer details of the microenvironment throughout the segmentation process. Future research directions can aim to enhance the explainability of such transformer models. This effort will enable their better integration into the medical imaging domain, potentially further improving their performance (Yang, Ye, & Xia, 2022). Overall, the proposed method provides a significant edge in harnessing the rich information embedded in hyperspectral images.

Acknowledgments

This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. The authors would like to thank Matthew Krafczyk and Shirui Luo for the helpful insights and guidance. This study was supported in part by the National Institutes of Health, United States via grants R01EB009745 and R01CA260830.

Data availability

Data will be made available on request.

References

Alom Z, Yakopcic C, Hasan M, Taha TM, & Asari VK (2019). Recurrent residual U-net for medical image segmentation. Journal of Medical Imaging.

Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. (2014). Using Fourier transform IR spectroscopy to analyze biological materials. Nature protocols, 9(8), 1771–1791. [PubMed: 24992094]

Berisha S, Lotfollahi M, Jahanipour J, Gurcan I, Walsh M, Bhargava R, et al. (2019). Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks. Analyst, 144(5), 1642–1653. [PubMed: 30644947]

Bhargava R (2023). Digital histopathology by infrared spectroscopic imaging. Annual Review of Analytical Chemistry, 16(1), 205–230,

Breiman L (2001). Random forests. Machine Learning, 45(1), 5–32.

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, & Zagoruyko S (2020). End-to-end object detection with transformers. arXiv: Computer Vision and Pattern Recognition.

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. Cornell University - arXiv.

Chen L-C, Papandreou G, Kokkinos I, Murphy K, & Yuille AL (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848. [PubMed: 28463186]

Chen H, Qi X, Yu L, Dou Q, Qin J, & Heng P-A (2017). DCAN: Deep contour-aware networks for object instance segmentation from histology images. Medical Image Analysis.

Chen J, Yang G, Khan H, Zhang H, Zhang Y, Zhao S, et al. (2021). JAS-GAN: generative adversarial network based joint atrium and scar segmentations on unbalanced atrial targets. IEEE Journal of Biomedical and Health Informatics, 26(1), 103–114.

Devlin J, Chang M-W, Lee K, & Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv abs/1810.04805.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In 9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3–7, 2021. OpenReview.net.

Esteva A, Kuprel B, Novoa RA, Ko JM, Swetter SM, Blau HM, et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature.

Falahkheirkhah K, Yeh K, Confer MP, & Bhargava R (2022). DRB-net: Dilated residual block network for infrared image restoration. In Advances in visual computing: 17th international symposium, ISVC 2022, san diego, CA, USA, October 3–5, 2022, proceedings, part II (pp. 104–115). Springer.

Falahkheirkhah K, Yeh K, Mittal S, Pfister L, & Bhargava R (2021). Deep learning-based protocols to enhance infrared imaging systems. Chemometrics and Intelligent Laboratory Systems, 217, Article 104390.

Fernandez DC, Bhargava R, Hewitt SM, & Levin IW (2005). Infrared spectroscopic imaging for histopathologic recognition. Nature biotechnology, 23(4), 469–474.

Geinguenaud F, Militello V, & Arluison V (2020). Application of FTIR spectroscopy to analyze RNA structure. RNA Spectroscopy: Methods and Protocols, 119–133.

Graham S, Chen H, Gamper J, Dou Q, Heng P-A, Snead D, et al. (2019). MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Medical Image Analysis.

Hatamizadeh A, Yang D, Roth HR, & Xu D (2021). UNETR: Transformers for 3D medical image segmentation. In 2022 IEEE/CVF winter conference on applications of computer vision (WACV), (pp. 1748–1758).

Hatipoglu N, & Bilgin G (2017). Cell segmentation in histopathological images with deep learning algorithms by utilizing spatial relationships. Medical & Biological Engineering & Computing.

He K, Gkioxari G, Dollár P, & Girshick R (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961–2969).

He K, Zhang X, Ren S, & Sun J (2015). Deep Residual Learning for Image Recognition. Cornell University - arXiv.

Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl SAA, et al. (2018). Nnu-net: Self-adapting framework for U-net-based medical image segmentation. arXiv: Computer Vision and Pattern Recognition.

Kedzierski M, Falcou-Préfol M, Kerros ME, Henry M, Pedrotti ML, & Bruzaud S (2019). A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the mediterranean sea. Chemosphere, 234, 242–251. [PubMed: 31226506]

Keogan A, Nguyen TNQ, Phelan JJ, O'Farrell N, Lynam-Lennon N, Doyle B, et al. (2021). Chemical imaging and machine learning for sub-classification of oesophageal tissue histology. Translational Biophotonics, 3(4), Article e202100004.

Lasch P, Diem M, Hänsch W, & Naumann D (2006). Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. Journal of Chemometrics: A Journal of the Chemometrics Society, 20(5), 209–220.

Li Y, & Shen L (2018). Skin lesion analysis towards melanoma detection using deep learning network. Sensors.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9992–10002).

Loshchilov I, & Hutter F (2017). Decoupled weight decay regularization. In International conference on learning representations.

Lu S, Zhang Z, Yan Z, Wang Y, Cheng T, Zhou R, et al. (2023). Mutually aided uncertainty incorporated dual consistency regularization with pseudo label for semi-supervised medical image segmentation. Neurocomputing, 548, Article 126411.

Magnussen EA, Solheim JH, Blazhko U, Tafintseva V, Tøndel K, Liland KH, et al. (2020). Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. Journal of Biophotonics, 13(12), Article e202000204.

Milletari F, Navab N, & Ahmadi S-A (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565–571).

Mittal S, Yeh K, Leslie LS, Kenkel S, Kajdacsy-Balla A, & Bhargava R (2018). Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology. Proceedings of the National Academy of Sciences, 115(25), E5651–E5660.

Muniz FB, Baffa M. d. O., Garcia SB, Bachmann L, & Felipe JC (2023). Histopathological diagnosis of colon cancer using micro-FTIR hyperspectral imaging and deep learning. Computer Methods and Programs in Biomedicine, 231, Article 107388.

Oktay O, Schlemper J, Folgoc LL, Lee MCH, Heinrich MP, Misawa K, et al. (2018). Attention U-net: Learning where to look for the pancreas. arXiv: Computer Vision and Pattern Recognition.

Pang S, Du A, Orgun M, & Yu Z (2018). A novel fused convolutional neural network for biomedical image classification. Medical & Biological Engineering & Computing, 57.

Pradhan P, Guo S, Ryabchykov O, Popp J, & Bocklitz TW (2020). Deep learning a boon for biophotonics? Journal of Biophotonics, 13(6), Article e201960186.

Radford A, & Narasimhan K (2018). Improving language understanding by generative pre-training.

Ronneberger O, Fischer P, & Brox T (2015). U-net: Convolutional networks for biomedical image segmentation. cite arxiv:1505.04597Comment: conditionally accepted at MICCAI 2015.

Simonyan K, & Zisserman A (2015). Very deep convolutional networks for large-scale image recognition. In International conference on learning representations.

Tang Z, Nan Y, Walsh S, & Yang G (2023). Adversarial transformer for repairing human airway segmentation. IEEE Journal of Biomedical and Health Informatics.

Tiwari S, Falahkheirkhah K, Cheng G, & Bhargava R (2022). Colon cancer grading using infrared spectroscopic imaging-based deep learning. Applied Spectroscopy, 76(4), 475–484. [PubMed: 35332784]

Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017). Attention is all you need. ArXiv abs/1706.03762.

Yang G, Ye Q, & Xia J (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77, 29–52. [PubMed: 34980946]

Yu F, & Koltun V (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Zhang CC, Gao X, & Yilmaz B (2020). Development of FTIR spectroscopy methodology for characterization of boron species in FCC catalysts. Catalysts, 10(11), 1327.

Zhou Z, Siddiquee MR, Tajbakhsh N, & Liang J (2018). Unet++: A nested u-net architecture for medical image segmentation. Springer International Publishing eBooks.

Zhuang J (2018). LadderNet: Multi-path networks based on U-net for medical image segmentation. arXiv: Computer Vision and Pattern Recognition.

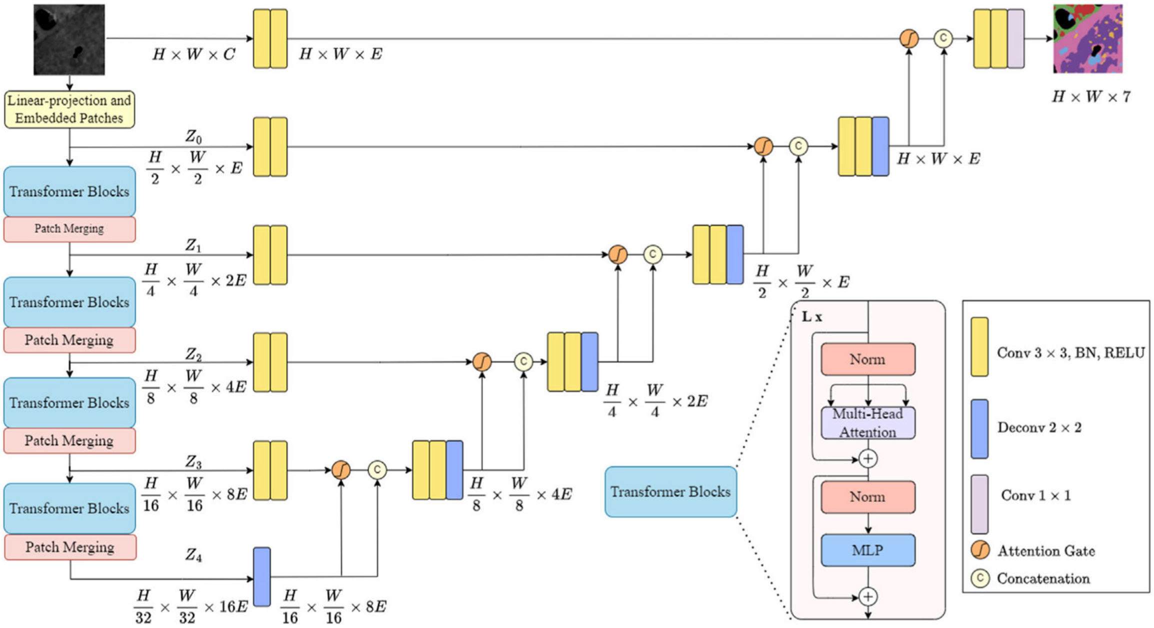
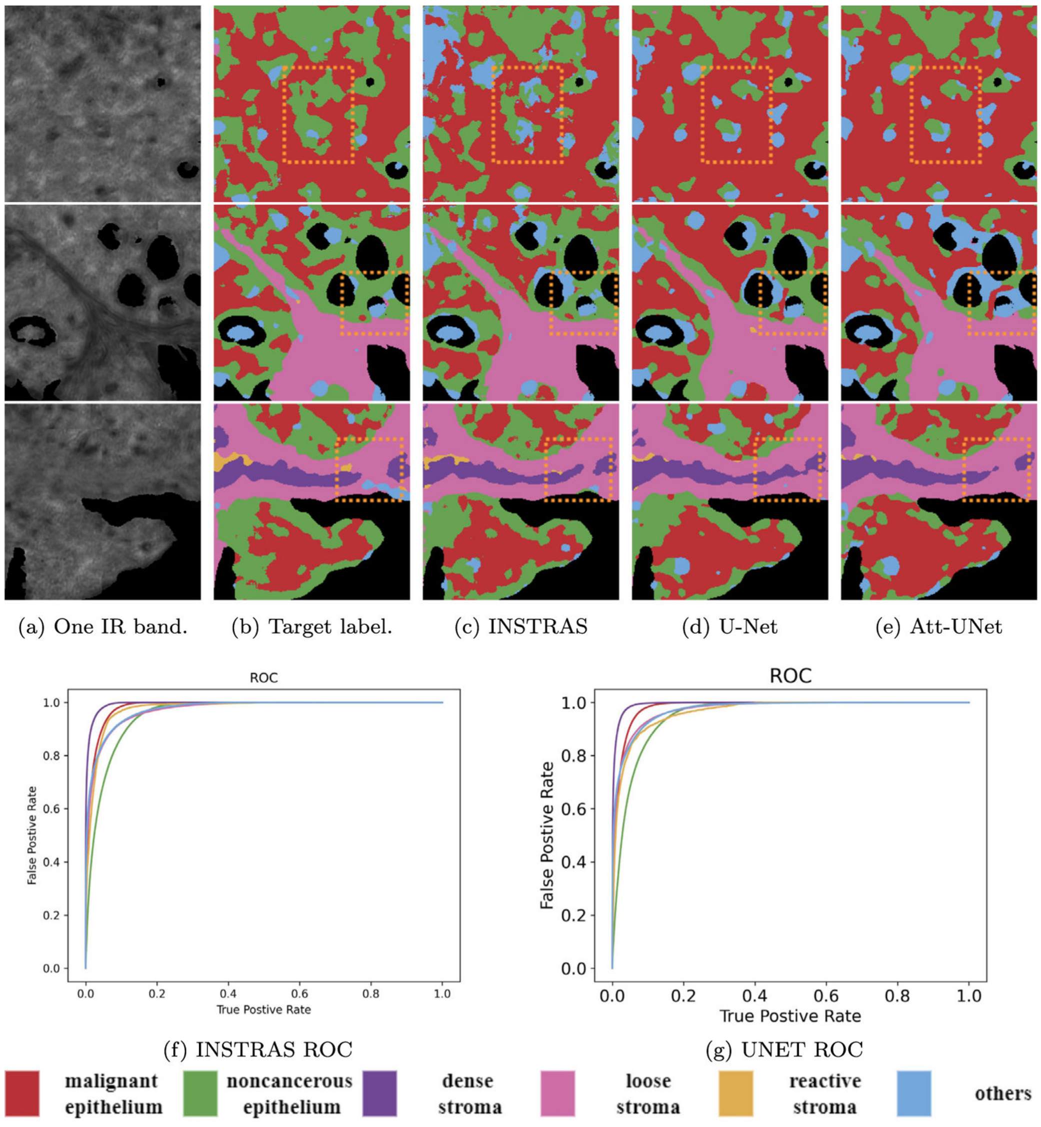

$H \times W \times C$
$H \times W \times E$
$H \times W \times 7$
Linear-projection and Embedded Patches
Z_0
$\frac{H}{2} \times \frac{W}{2} \times E$
Transformer Blocks
Patch Merging
Z_1
$\frac{H}{4} \times \frac{W}{4} \times 2E$
Transformer Blocks
Patch Merging
Z_2
$\frac{H}{8} \times \frac{W}{8} \times 4E$
Transformer Blocks
Patch Merging
Z_3
$\frac{H}{16} \times \frac{W}{16} \times 8E$
Transformer Blocks
Patch Merging
Z_4
$\frac{H}{32} \times \frac{W}{32} \times 16E$
$\frac{H}{16} \times \frac{W}{16} \times 8E$
$\frac{H}{8} \times \frac{W}{8} \times 4E$
$\frac{H}{4} \times \frac{W}{4} \times 2E$
$\frac{H}{2} \times \frac{W}{2} \times E$
$H \times W \times E$
Transformer Blocks
L x
Norm
Multi-Head Attention
Norm
MLP
Conv 3 × 3, BN, RELU
Deconv 2 × 2
Conv 1 × 1
Attention Gate
Concatenation

Fig. 1.

The overview of INSTRAS architecture. Input IR images (C is up to 9 in our data set) are divided into equal size patches and fed to a linear embedding layer with positional encoding. The embedded patches are then passed to L multi-head attention blocks, and four intermediate output features Z_1 to Z_4 with E embedding size are extracted and merged by the convolution decoder with attention gates.



(a) One IR band. (b) Target label. (c) INSTRAS (d) U-Net (e) Att-UNet

(f) INSTRAS ROC (g) UNET ROC

(h) Color labels

Fig. 2.

Visualization of results for INSTRAS, U-Net, and Attention U-Net under the same training settings. The images contain both models' predictions trained on 9 IR bands. We use red dashed boxes to highlight the major difference between each prediction. (f) and (g) show the ROC curves of INSTRAS and UNET respectively. The meaning of the pixel color is shown in (h). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1

The AUC↑ value of INSTRAS (proposed), U-Net, and Attention U-Net models on the validation data set. We use six and nine IR bands to train all networks and record the AUC value of each histological component. The Mean AUC is the average value over the 6 classes. The **best** results and the **second best** results are set in red and bold, respectively.

Models	Malignant epithelium	Noncancerous epithelium	Dense stroma	Loss stroma	Reactive stroma	Others	Mean
U-Net$_6$	0.9782	**0.9432**	**0.9869**	**0.9537**	**0.8807**	0.9596	**0.9504**
Attention U-Net$_6$	**0.9787**	**0.9461**	**0.9851**	**0.9517**	0.8906	**0.9653**	0.9529
INSTRAS$_6$	**0.9809**	0.9422	0.9834	0.9386	**0.9451**	**0.9670**	**0.9595**
U-Net$_9$	**0.9836**	**0.9500**	**0.9951**	0.9750	0.9709	**0.9731**	**0.9746**
Attention U-Net$_9$	0.9803	0.9465	0.9947	**0.9750**	**0.9739**	0.9725	0.9738
INSTRAS$_9$	**0.9855**	**0.9594**	**0.9953**	**0.9754**	**0.9773**	**0.9797**	**0.9788**

Table 2

The accuracy↑ of INSTRAS (proposed), U-Net, and Attention U-Net model on the validation data set. We train all networks by six, and nine IR bands. The accuracy is the average value over the 6 classes. The **best** results and the **second best** results are set in red and bold, respectively.

Models	**Accuracy↑**
U-Net$_6$	**81.419**
Attention U-Net$_6$	81.276
INSTRAS$_6$	**81.674**
U-Net$_9$	**84.836**
Attention U-Net$_9$	83.758
INSTRAS$_9$	**85.811**

Table 3

The average $\mathrm{INSTRAS}_9$ inference time over entire IR image using different patch overlapping coefficient δ, and the accuracy of prediction over the entire image.

δ	**Inference time (s)**	**Accuracy↑**
0	0.7573	85.76
25	1.181	85.80
50%	2.069	85.81
75%	9.040	85.90
87.5%	33.45	85.93

Table 4

Trainable parameters number of each model taking 9 IR bands.

Model	Trainable parameters
U-Net$_9$	34530k
Attention U-Net$_9$	34882k
INSTRAS$_9$	33882k