Materials Characterization 229 (2025) 115645

Contents lists available at ScienceDirect

Materials Characterization

journal homepage: www.elsevier.com/locate/matchar

Accelerated quantification of reinforcement degradation in additively manufactured Ni-WC metal matrix composites via SEM and vision transformers

Mutahar Safdar [a,b], Bashir Kazimi [c], Karina Ruzaeva [c], Gentry Wood [d], Max Zimmermann [e], Guy Lamouche [b], Priti Wanjara [b], Stefan Sandfeld [c,f], Yaoyao Fiona Zhao [a,*]

[a] Department of Mechanical Engineering, McGill University, Montreal, QC H3A 0C3, Canada
[b] National Research Council Canada, Montreal, QC H3T 1J4, Canada
[c] Institute for Advanced Simulation – Materials Data Science and Informatics (IAS-9), Forschungszentrum Jülich GmbH, Jülich 52425, Germany
[d] Apollo-Clad Laser Cladding, a division of Apollo Machine and Welding Ltd., Edmonton, AB T6E 5V2, Canada
[e] Fraunhofer Institute for Laser Technology ILT, Aachen 52074, Germany
[f] Chair of Materials Data Science and Materials Informatics, Faculty 5 – Georesources and Materials Engineering, RWTH Aachen University, Aachen 52056, Germany

ARTICLE INFO

Keywords:
Additive manufacturing
Scanning electron microscopy (SEM) images
Metal matrix composites
Carbide damage
Semantic segmentation
Vision transformers

ABSTRACT

Machine learning (ML) applications have shown potential in analyzing complex patterns in additively manufactured (AMed) structures. Metal matrix composites (MMC) offer the potential to enhance functional parts through a metal matrix and reinforcement particles. However, their processing can induce several co-existing anomalies in the microstructure, which are difficult to analyze through optical metallography. Scanning electron microscopy (SEM) can better highlight the degradation of reinforcement particles, but the analysis can be labor-intensive, time-consuming, and highly dependent on expert knowledge. Deep learning-based semantic segmentation has the potential to expedite the analysis of SEM images and hence support their characterization in the industry. This capability is particularly desired for rapid and precise quantification of defect features from the SEM images. In this study, key state-of-the-art semantic segmentation methods from self-attention-based vision transformers (ViTs) are investigated for their segmentation performance on SEM images with a focus on segmenting defect pixels. Specifically, SegFormer, MaskFormer, Mask2Former, UPerNet, DPT, Segmenter, and SETR models were evaluated. A reference fully convolutional model, DeepLabV3+, widely used on semantic segmentation tasks, is also included in the comparison. A SEM dataset representing AMed MMCs was generated through extensive experimentation and is made available in this work. Our comparison shows that several transformer-based models perform better than the reference CNN model with UPerNet (94.33 % carbide dilution accuracy) and SegFormer (93.46 % carbide dilution accuracy) consistently outperformed the other models in segmenting damage to the carbide particles in the SEM images. The findings on the validation and test sets highlight the most frequent misclassification errors at the boundaries of defective and defect-free pixels. The models were also evaluated based on their prediction confidence as a practical measure to support decision-making and model selection. As a result, the UPerNet model with the Swin backbone is recommended for segmenting SEM images from AMed MMCs in scenarios where accuracy and robustness are desired whereas the SegFormer model is recommended for its lighter design and competitive performance. In the future, the analysis can be extended by including higher capacity as well as smaller models in the comparison. Similarly, variations in specific hyperparameters can be investigated to reinforce the rationale of selecting a specific configuration.

1. Introduction

Additive manufacturing (AM) or 3D printing is a layer-based fabrication technique [1]. Metal AM can realize fully dense metallic structures, thus offering the potential to compete with conventional manufacturing in industry. Some AM techniques, such as directed

* Corresponding author.
E-mail address: yaoyao.zhao@mcgill.ca (Y.F. Zhao).

https://doi.org/10.1016/j.matchar.2025.115645
Received 8 July 2025; Received in revised form 26 September 2025; Accepted 7 October 2025
Available online 10 October 2025

energy deposition (DED), offer freedom in depositing materials, thereby enabling the repair of components and the manufacture of hybrid products, as well as the enhancement of functional parts with changing chemistry and/or mechanical properties [2]. The capability to enhance parts has wide-ranging applications in the industry. For instance, DED processing of metal matrix composites (MMCs) can lead to wear-resistant overlays on functional parts and tools [3]. MMCs are engineered composite materials with a metal matrix integrated with one or more reinforcement particles [4,5]. These composites leverage the advantageous characteristics of metals, such as toughness and ductility, alongside the reinforcing properties of the added materials, such as stiffness and strength. MMCs are highly valued for their exceptional strength-to-weight ratio, wear resistance, and performance at elevated temperatures. However, the production of MMCs, such as nickel tungsten carbide (Ni-WC) overlays on steel substrates, poses significant challenges [6]. While these composites can significantly enhance the wear resistance of components used in several sectors (e.g., agriculture, automotive, mining, and aerospace), their development requires intensive optimization of the key process parameters and iterative adjustment to meet the specific needs of part geometries [7].

Metallographic characterization is an essential tool to process development as it can reveal the constituents in materials' structure, enabling the evaluation of their properties [8,9]. The individual constituents in the microstructure of DED-based Ni-WC MMCs complement each other by reducing material wastage, while allowing reasonable plastic deformation. In addition to the metal matrix and reinforcement carbides, the deposited structures can also contain defects related to the processing conditions (e.g., porosities, cracks, dissolved carbides, poor carbide distribution) [10]. Therefore, it is essential to quantitatively evaluate printed structures and characterize parameter adjustments needed for the numerous key process variables. This evaluation is usually accomplished in the industry through optical metallography, as it highlights the distribution of carbide particles in the matrix and certain processing defects (e.g., porosity). High thermal conditions also lead to degradation of the reinforcement carbide particles, which is usually reflected by partially diluted carbides and reprecipitated hard phases in the matrix.

Dilution band represents the rim region around a carbide particle in which elevated thermal exposure partially dissolves the original spherical WC particle, producing a contrast-visible zone that reflects reinforcement degradation. Reprecipitated carbides represent fine secondary carbide phases that form within the matrix during cooling/solidification after partial dissolution of the original WC particles. Reprecipitated carbides can appear as dispersed hard particles locally or across the matrix. Optical microscopy is limited in distinguishing these phases of degraded carbides. On the other hand, scanning electron microscopy (SEM) is better suited to analyze these constituents owing to the shorter wavelength of electrons enabling higher magnifications and resolutions as well as highlighting compositional changes [11]. However, the characterization of SEM-generated samples can be slow, labor-intensive, and dependent on expert knowledge, which makes it challenging to meet the industrial requirements of high throughput quantitative and accurate metallography.

Additively manufactured (AMed) microstructures can be evaluated and analyzed using various methods [12]. Traditional manual techniques for characterizing and identifying the constituents of these microstructures are prevalent in industrial settings. Although such techniques can produce detailed analyses, these methods are notably labor-intensive and time-consuming. Alternatively, computer-assisted or semi-automated techniques can accelerate the evaluation process, but demand consistent effort for each new analysis [13]. Consequently, there has been a shift toward employing image-processing or computer vision (CV) techniques to fully automate microstructural quantification. These approaches span a range from basic to highly complex, based on the algorithms' sophistication. Despite their utility, CV methods often struggle with generalizing across different contexts and often fail to accommodate the characteristic variability in AMed micrographs [14].

Semantic segmentation, also known as dense or pixel-level prediction, is a process for classifying image pixels and is widely utilized in various fields, such as autonomous vehicles, medical diagnostics, robotics, and object identification [15]. Traditionally, machine learning (ML) models employed for this task have relied on established convolutional neural network (CNN) based frameworks and architectures. These models, including architectures like U-Nets, are recognized for accurately delineating distinct regions within images by clustering at the pixel level [16]. Although U-Net architectures are prevalent, the ML community has recently explored the application of cutting-edge vision transformers (ViTs) to semantic segmentation, achieving comparably robust performance [17]. This progression underscores the imperative to enhance precision, tailor models to specific domains, broaden generalization capabilities and create high-quality annotated datasets.

This work aims to evaluate state-of-the-art ViTs for their capability to segment SEM images from AMed Ni-WC MMC structures, while also comparing the ViTs against a reference CNN architecture. Although CNNs and ViTs have been shown to be effective for a range of segmentation tasks, these models also suffer from domain-specific challenges. Engineering applications of artificial intelligence (AI) face issues of data imbalance and quality. Industrial deployment requires that ML models be lightweight and robust to changing data distributions. To support the requirement of high throughput quantitative metallography for SEM images generated through the analysis of AMed Ni-WC MMCs, this work contributes to the following:

- An open-source annotated SEM dataset from AMed Ni-WC MMC metallographs, as well as an open-source codebase for reproducibility
- An investigation of seven state-of-the-art ViT architectures to quantitatively analyze SEM images from AMed Ni-WC MMC using semantic segmentation
- A comparison between a reference CNN architecture and evaluated ViTs for insights on complementary or superior segmentation capabilities of each category
- Findings on the impact of encoder variations, model category, model size, as well as pixel categories on segmentation performance
- Recommendations for industrial deployment grounded in practical considerations of model size and prediction confidence

The rest of this paper is arranged into background discussions on semantic segmentation and their AM applications (Section 2), dataset introduction (Section 3), segmentation architectures with ViT backbones (Section 4), training experiments (Section 5), discussions with findings (Section 6), and conclusions including suggestions for future works (Section 7).

2. Semantic segmentation and AM applications

Semantic segmentation models work at the pixel level by predicting the labels and grouping them to highlight the regions of interest [18]. Standard CV algorithms can also accomplish this task reasonably, but these models struggle with generalization. This section is constrained to ML-based semantic segmentation and consists of two parts.

The first part provides a concise overview of the development and advancement of ML-based semantic segmentation algorithms, beginning with the basic fully convolutional networks and U-Net models and concluding with the most current state of the broader domain. The subsequent part emphasizes the applications of semantic segmentation in AM by addressing recent advancements, as well as by identifying ongoing research efforts.

2.1. ML-based semantic segmentation

The capacity to delineate information within images on a pixel-by-

pixel basis has been incorporated into ML models through various approaches. Fully convolutional networks (FCNs) marked a significant shift in developing ML-driven semantic segmentation models [19]. These networks adapted conventional CNN architectures to execute the task of semantic segmentation. In FCNs, fully connected layers were completely substituted by convolutional layers to preserve spatial relations, facilitating seamless end-to-end segmentation across images of any size. A notable model within this category was introduced by Long et al. [19], which featured a skip architecture that effectively restored information lost during the down-sampling process. Upon its introduction, their model demonstrated remarkable performance on the PASCAL VOC dataset, surpassing existing benchmarks [20].

The subsequent phase in developing semantic segmentation models was marked by the introduction of encoder-decoder structures, exemplified by the U-Net series. These models were designed to retain detailed information through skip connections. Ronneberger et al. [16] provided a seminal architecture in this category. The encoder was designed for contraction, capturing the contextual details of the pixels, whereas the decoder focused on expansion, allowing for precise localization. Additionally, the architecture utilized skip connections linking the contractive and expansive components, enhancing the model's accuracy in pixel segmentation. Initially developed for medical imaging, this model has since become a standard in various other fields due to its effectiveness and remains widely adopted.

Building on the success of deep convolutional models for semantic segmentation, Chen et al. [21] combined a CNN trained for capturing a certain level of feature representation with a fully connected conditional random field (CRF) for refining the segmentation results from the CNN. The addition of CRF considered the dependencies and context of pixels that helped refine the results from CNN. The architecture is referred to as DeepLab, and improvements through research have led to several variants of the architecture (e.g., versions v1, v2, v3, v3+). At the time of their work, the hybrid approach of combining CNN with fully connected CRF achieved state-of-the-art results on semantic segmentation tasks.

Sultana et al. [22] conducted a comprehensive review and analysis of semantic segmentation models that relied on convolution-based techniques before ViTs emerged for image segmentation tasks. They categorized CNN-based semantic segmentation models into five principal architectural frameworks: (i) networks composed entirely of convolutional layers, (ii) networks incorporating dilated/atrous convolutions, (iii) networks employing a top-down/bottom-up strategy, (iv) networks that integrate global contextual information, and (v) networks that enhance the receptive field and incorporate multi-scale context. Notable examples within these categories include (i) FCN, (ii) DilatedNet and DeepLab, (iii) Deconvnet, U-Net, SegNet, FC-DenseNet, (iv) ParseNet, GCN, and EncNet, and (v) DeepLabV2, DeepLabV3, PSPNet, and Gated-SCNN. Their analysis offers an extensive overview of the field by providing a detailed comparison of these architectures and their respective advantages and limitations. Readers are encouraged to refer to their review for a more detailed examination [22].

The introduction of transformers and their extension to vision data have paved the way for their application to image semantic segmentation tasks. Transformer models are built upon the self-attention mechanism, which enables parallel data processing and the capture of long-range dependencies. This key attention feature allows transformers to selectively focus on relevant information, marking a substantial shift from the sequential processing typical of earlier natural language processing (NLP) models. Dosovitskiy et al. [23] extended the self-attention mechanism to visual data by representing an image as a series of tokens from 2D patches. This adaptation has recently prompted interest in applying self-attention mechanisms to vision-related tasks. The initial implementation of ViTs in semantic segmentation was introduced by Zheng et al. [24] in their segmentation transformer (SETR) model, which employs a pure transformer instead of the traditional encoder, systematically reducing the spatial resolution of inputs. Since then, several ViT architectures for semantic segmentation have been developed. Thisanke et al. [17] provide a detailed review and comparison of significant ViT architectures used in semantic segmentation. Their review offers insights into the increasing applications of ViTs in this field.

The latest research in CV for semantic segmentation tasks continues to be guided by the needs of specific domains, efficient architectures, and data challenges. The following section discusses the applications of semantic segmentation in AM, the domain of interest for the current work.

2.2. Applications of semantic segmentation in additive manufacturing

AM technology is rapidly progressing toward industrial maturity. The freedom in deposition mechanisms and material compositions enables the development of new materials whose microstructure brings unique challenges (e.g., segmentation of processing defects). Integrating AI theory and ML applications with AM has expedited its development. These applications are well documented in the open literature [25]. The AM contexts to ML solutions cover a broad spectrum of data modalities and information representations [26]. Among other applications, semantic segmentation is of interest to the data-driven AM community for in-process and post-process evaluation of quality for the deposited material by investigating macro and microstructures. As a result, several applications of image semantic segmentation models already exist in AM, both for 2D and 3D data representation captured at different stages (e.g., in-situ, ex-situ) of the AM process flow.

Applications of semantic segmentation in AM are predominantly focused on in-process datasets (e.g., monitoring) and post-process datasets (e.g., evaluation). In-process applications aim to swiftly identify defects and irregularities during manufacturing by segmenting the relevant pixels. Scime et al. [27] designed a CNN-based semantic segmentation architecture named dynamic segmentation CNN (DSCNN). This model was designed for real-time segmentation at the native resolution of both visible-light and infrared imaging systems and is adaptable across different machines, process technologies, and sensing systems. Moreover, recent efforts in semantic segmentation within AM have also concentrated on anomaly detection of powder beds in selective laser sintering processes [28]. Other in-process applications involve image segmentation under varying printing conditions of fused filament fabrication processes [29] and segmenting areas of interest from in-situ sensing representations [30]. A recent development in semantic segmentation for in-situ defect detection addresses the challenge of imbalanced datasets. Wang [31] introduced a class-aware semantic contrast and attention amalgamation model tailored explicitly for semantic segmentation. The proposed model demonstrated effective performance in scenarios with data imbalance.

Semantic segmentation techniques are also gaining traction for post-process datasets in AM as they facilitate rapid structure evaluation. These techniques are of particular interest within the data-driven AM community, as they enable the quantification of multi-scale features (e.g., macro, *meso*, micro), which are crucial for quality assessments. Scott et al. [11] integrated SEM images with synthetic thermal tomography images using a common U-Net encoder. This network segmented defects in the usual manner and classified their parameters by leveraging the encoded features for a subsequent fully connected network. Their approach enhanced the segmentation of thermal tomography results by incorporating SEM images. It led to superior performance compared to traditional methods. However, it did not address data scarcity and class imbalance challenges.

Similarly, Rose et al. [13] implemented a convolution-based semantic segmentation model to automate the segmentation of NiCrBSi-WC MMC metallographic images. This model specifically segmented carbide particles whilst designating the matrix as the background. Although their research utilizes the same MMC as the current work, their approach was limited to binary segmentation. It did not tackle class imbalances in AMed optical metallographs, which was addressed in a

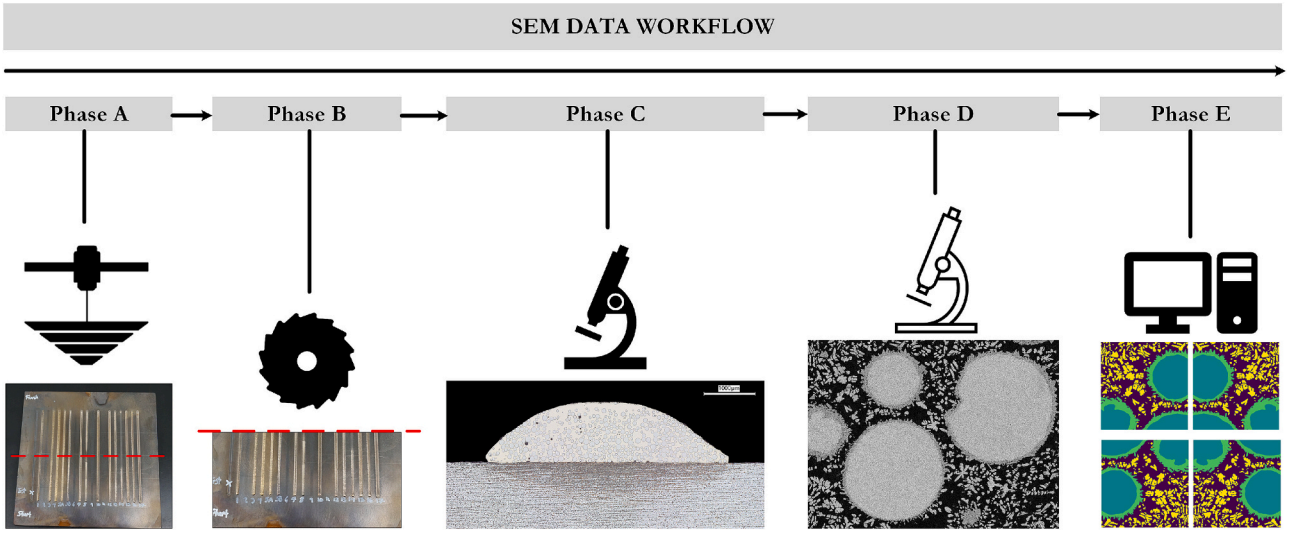

Fig. 1. Steps to prepare SEM images for additively deposited Ni-WC MMC powder on steel substrate.

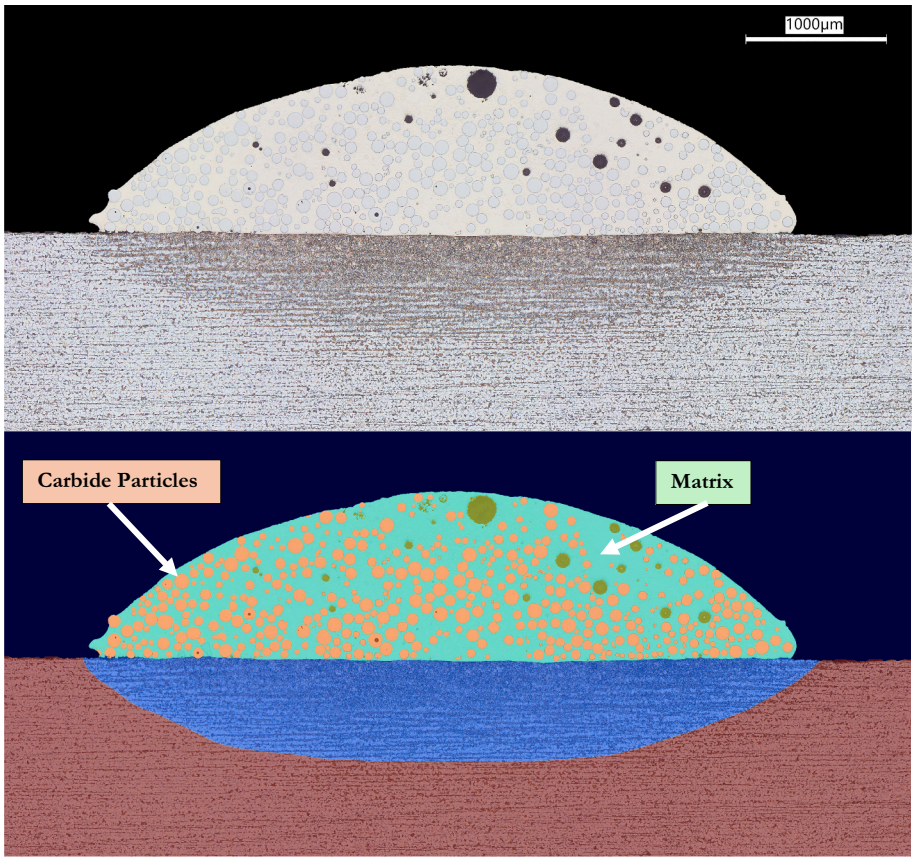

Fig. 2. Stitched panoramic sample cross-section based on optical microscopy. The lower portion of the figure shows an overlaid ground truth mask. Carbide particles and the surrounding matrix are highlighted owing to their relevance to this analysis. Optical microscopy does not effectively highlight the damage to reinforcement carbide particles from higher thermal conditions during processing.

recent work [10]. To the best of the authors' knowledge, no approach in the open literature has yet focused on addressing challenges associated with SEM images from Ni-WC MMCs, such as segmenting diluted and reprecipitated carbides, representing anomalies of the processing conditions [32].

Recent advancements in the segmentation of AMed metallographic images have seen the adoption of ensemble methods to address the complexities of segmenting multi-phase materials. Luengo et al. [33] conducted detailed investigations using CNN-based architectures on their publicly available MetalDAM SEM dataset, introduced alongside their study. They developed an ensemble model tailored for semantic segmentation tasks, where their stacking-based approach demonstrated superior performance compared to the individual models. In parallel, Biswas et al. [34] developed an ensemble comprising three dilated, attention-guided U-Net models. The outputs of these models were combined pixel-by-pixel to construct the final segmentation mask. While

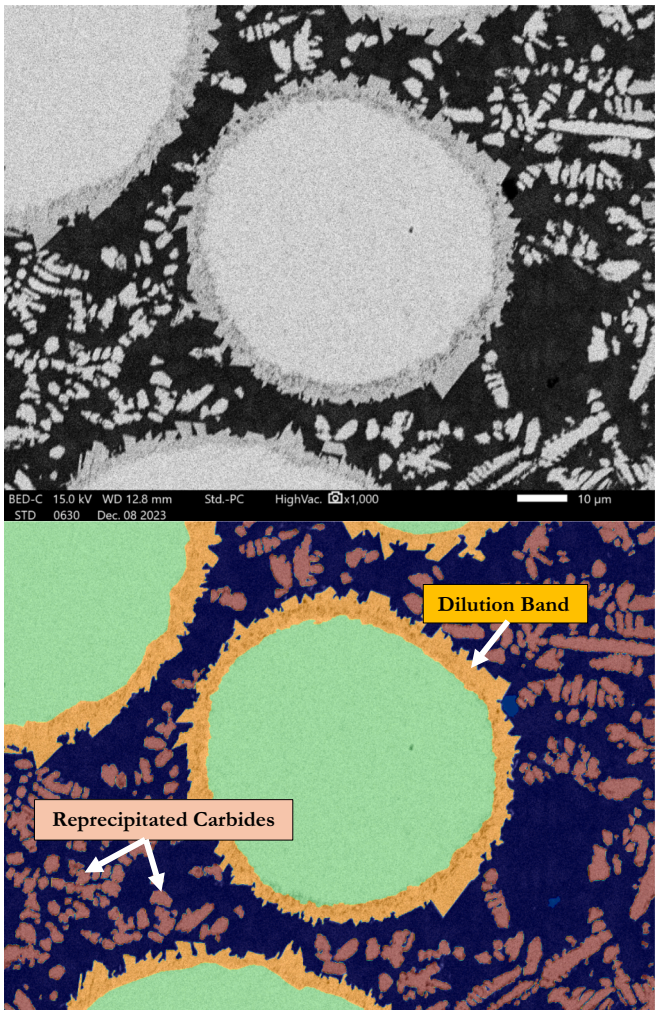

Fig. 3. A sample image from the SEM analysis captured with a BED-C signal at 1000× magnification. At the bottom, a labeled ground truth mask highlights the dilution band and reprecipitated carbide particles. These two phases highlight damage to the reinforcing carbide particles, and their quantification can support decision-making during the process development.

both ensemble strategies yielded encouraging results, these did not include the minority and challenging-to-segment precipitate class within their predictions. These applications were also aimed at optical microscopy images of AMed structures.

The applications of semantic segmentation in AM are mostly limited to fully CNNs, while transformer architectures are gaining traction due to their global context modeling and flexibility, necessitating a systematic investigation of key ViT architectures and associated methods. Moreover, there is a need to generate material specific datasets to accelerate engineering applications of AI in AM. The current work covers these gaps through segmenting SEM images and quantitatively analyzing anomalies of AMed MMC structures to optimize the processing parameters by limiting damage to reinforcement particles (diluted carbides and reprecipitated hard phases in SEM).

3. SEM dataset generation

The SEM images presented in this study were obtained through DED-based processing of Ni-WC MMC powders. Fig. 1 shows the major phases of the processing pipeline for SEM dataset generation, starting with the deposition process (Phase A).

The DED system consisted of a 6-kW fiber laser, a twin-disk powder feeder, and a modified milling platform for substrate positioning. Before conducting the experiments, the laser was calibrated using a beam profiler to attain the optimal spot size (approximately 5.5 mm) suitable for wear-resistant overlay applications of Ni-WC MMCs. Additionally, the powder feed rate was precisely adjusted to within a tolerance of ±1 g/min by modifying the hopper's rotational speed. Following the calibration phase, the system was prepared for the experimental trials. Steel substrates were first milled and marked with reference lines aligned with the deposition direction at 0, 25 %, 50 %, 75 %, and 100 % along the deposition length (e.g., 200 mm). These reference lines facilitated the correlation of sensor data with metallographic evaluations. Subsequently, the test plates underwent cleaning with acetone and were preheated using a propane torch to temperatures exceeding 260 °C. After preheating, the plates were allowed to cool to approximately 260 °C to enhance deposition quality. Any oxides that formed on the plate surfaces during preheating were removed using a wire wheel. The deposition head was then positioned at the starting point, and the appropriate nozzle spacing was established before depositing the beads on the steel substrates.

For the metallographic examination of individual beads within the dataset, the beads were sectioned perpendicular to the deposition direction (Phase B in Fig. 1). The sectioned samples were subsequently embedded in Bakelite to facilitate the metallographic analysis. Post-embedding, the samples underwent a series of grinding steps using progressively finer SiC papers (80 grit, 120 grit, 220 grit, 500 grit, and 1200 grit) to achieve a flat, smooth surface and minimize the grinding artifacts. The grinding and polishing of the samples were performed using the Saphir 550 grinding and polishing system manufactured by ATM Qness GmbH. This equipment maintained a consistent applied force of 200 N for each abrasive grain over two minutes. Following the grinding process, the samples were polished for eight minutes under a force of 25 N to eliminate any residual grinding marks. To make the heat-affected zone (HAZ) in the cross-section more noticeable, the samples were etched with Nital (approximately 3 % nitric acid in ethanol) for ten seconds and subsequently cleaned with ethanol to remove any etching residues. The prepared samples were then examined using an Olympus BX53M microscope at a magnification of 10× (Phase C in Fig. 1). Optical microscopic images were automatically captured using Olympus Stream Motion software, which employed an autofocus and stitching technique to achieve a resolution of 0.97 µm per pixel. A typical resulting cross-section is shown in Fig. 2 alongside the labeled ground truth.

A subset of samples was identified for SEM analysis, and the analysis was carried out on JEOL-JCM 7000 NeoScope Benchtop SEM (Phase D in Fig. 1). Specifically, during optical microscopy, the samples with premise for having diluted and reprecipitated carbide particles were selected for further investigation with SEM. Each selected sample was first individually mounted on the stage of the SEM, and an electrically conductive, non-porous carbon tape was used to act as a conducting path to prevent electric charge buildup on the sample surface (which would lower the image quality/resolution). The SEM chamber was then evacuated to establish high vacuum conditions. When applicable, the view was adjusted to visualize the samples in the correct orientation with the bead reinforcement area set upside. Manual adjustment for visual conditions was preferred over automatic focus as it provided better results at the expense of effort. The low-angle detector C of backscattered electrons (BED—C) was found to work best for obtaining compositional images of the cross-section under focus, though different signals were tested. The landing voltage and working distance were fixed at 15.0 kV and 12.8 mm, respectively. All samples were analyzed across a range of magnification levels (270×, 500×, 600×, 700×, 800×, 1000×), and subsequently, the images were recorded for ML modeling. Out of the analyzed cross-sections, sample #29 led to the most representative carbide defects and was selected for the subsequent modeling. The main objective of ML is to assist with high throughput and accurate quantitative metallography. Therefore, labeled higher magnification images (1000×, 800×, 700×) were used for training while validating and

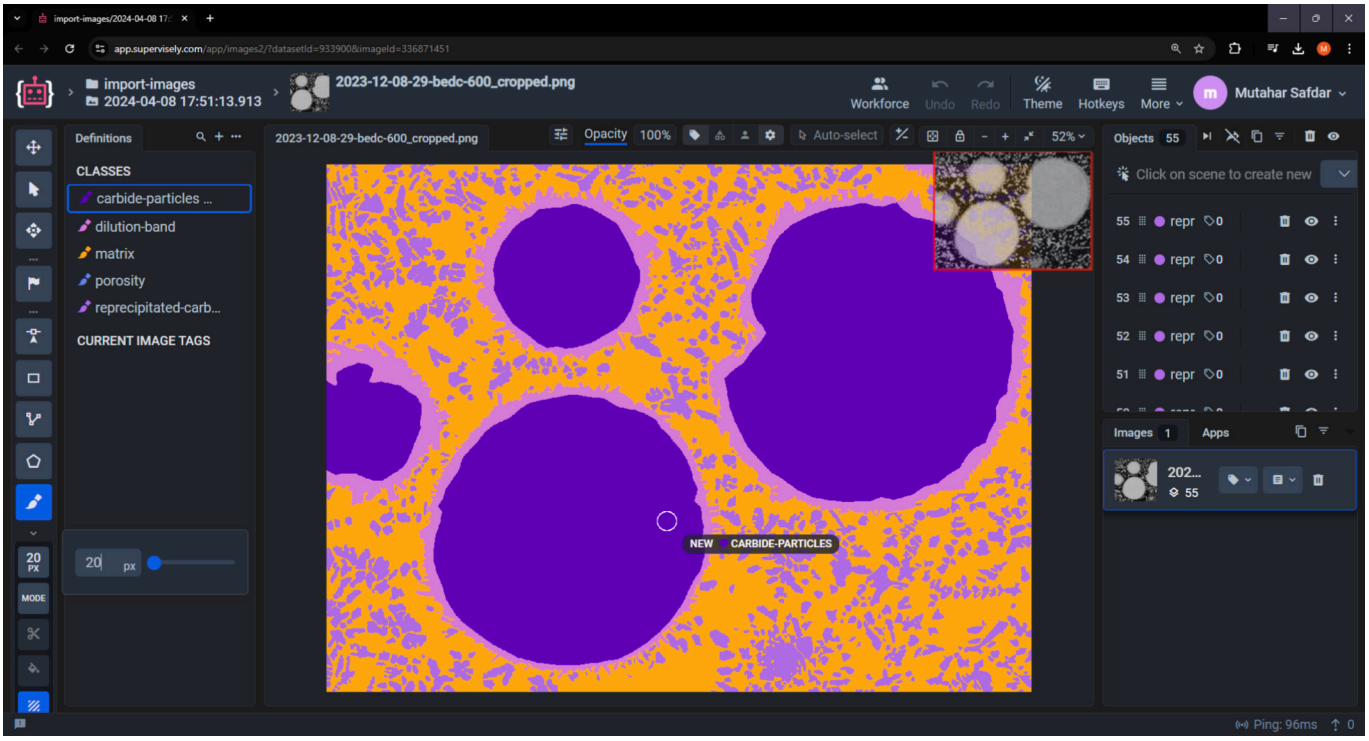

Fig. 4. Web-based Supervisely interface. Among other features, the interface allows adjustment to enable per-pixel labeling through the brush tool.

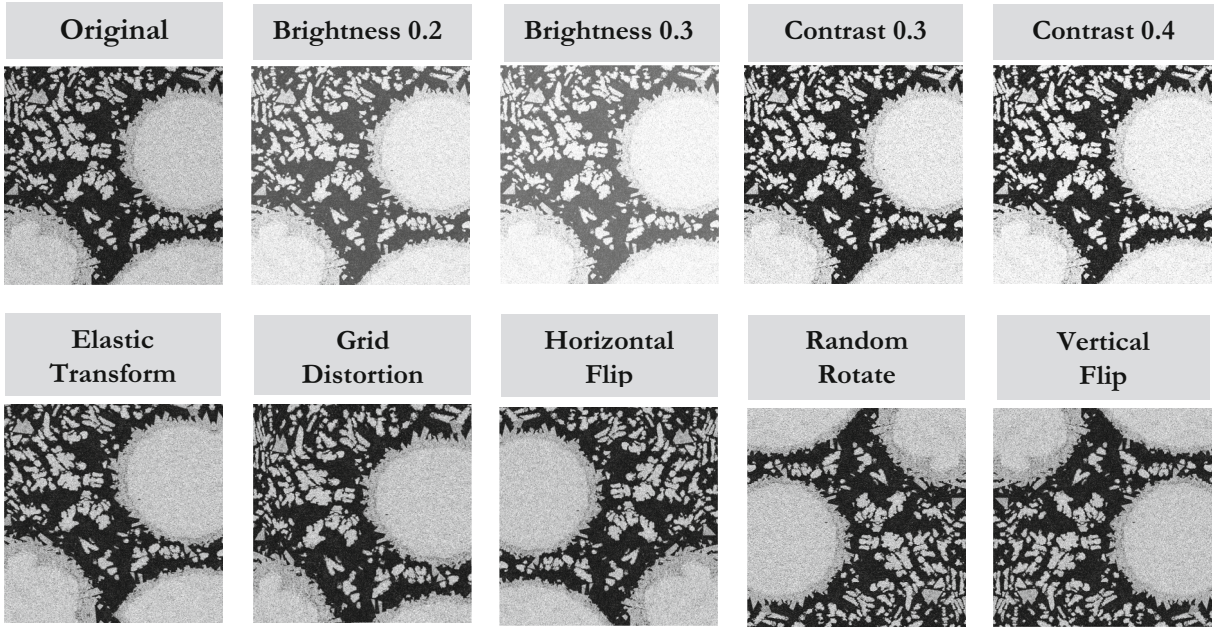

Fig. 5. Original and augmented crops on a sample input image from the SEM dataset.

testing the models at lower magnification (600×) to mimic real-world SEM analysis.

Fig. 2 shows an optical microscopy image for a sample cross-section. The phases of interest are labeled in the sample and highlighted in the overlaid mask. Among these, carbide presence is of particular interest as it determines the quality of overlayed deposition for functional parts. While optical microscopy has significance for efficiently evaluating the quality of depositions for practical applications, it is difficult to visualize the carbide degradations and quantify their presence. Fig. 3 shows a sample SEM image from a deposited cross-section. The unique phases are labeled and represented by the overlaid mask. The carbide dilution band and reprecipitated carbides are highlighted, as these two are closely related to the anomalies of the processing parameters at higher thermal conditions.

The melting and subsequent re-solidification of the Ni-Cr-B-Si matrix, which contains substantial quantities of tungsten and carbon, induce significant modifications in the carbide morphology. These structural changes result in a considerable reduction in the abrasion resistance as well as the impact toughness. Since tungsten carbide is the primary wear-resistant component within the MMC, it is critical to

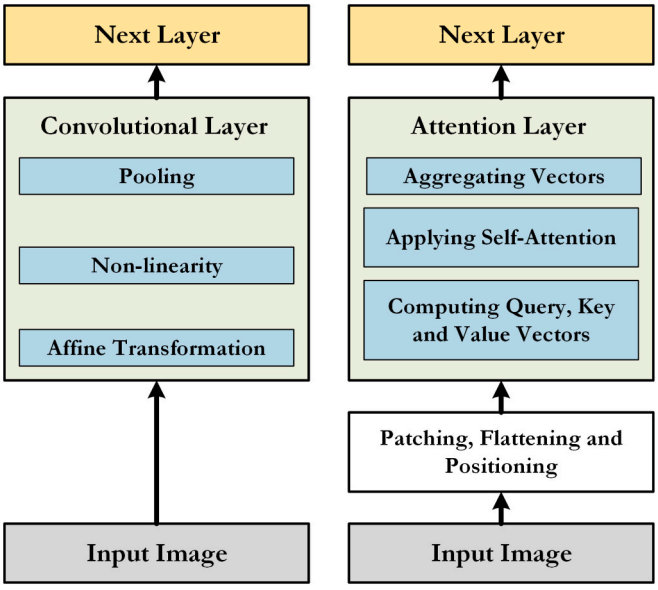

Fig. 6. Comparison between the sequence of operations in a typical convolutional layer (adapted from Deep Learning book [40]) and a typical attention layer.

minimize any thermal degradation of this carbide. The proportion of undamaged tungsten carbide in the deposit is the most critical factor in determining the relative performance of the deposition after porosity. Since porosity is easily detected optically, the SEM analysis focuses on segmenting the two types of reinforcement degradation namely dilution band and reprecipitated carbides [10]. These degradation phases are clearly visible in SEM images, which enable their segmentation and quantification to support process characterization.

The images were labeled (Fig. 1 Phase E) in the web-based software tool Supervisely [35], whose interface is shown in Fig. 4. A brushing tool was used at pixel resolution (20 pixels) to label the SEM phases of interest manually. To optimize the time-consuming labeling effort, the labeling order was carefully selected, starting with the matrix where the entire image was masked as the matrix class. This was possible through the overlay feature, which allows multiple pixel labels with preference given to the most recent value. Following this, entire carbide particles (including non-diluted phases) were labeled as the dilution band, which was subsequently updated by drawing an inner enclosure of the carbide mask on the existing dilution mask. This strategy made separating the dilution band from carbide particles easier by eliminating any repetition of effort when labeling intricate boundaries of dilution bands. Finally, the reprecipitated carbides were labeled, which represented the main portion of the labeling effort. Porosities did not contribute to the decision-making for the labeling sequence due to their highly sparse presence and, as such, these were eventually removed from the subsequent comparison study to keep the focus on the carbide defects. Moreover, porosities can be quantified through optical microscopy to an acceptable degree, eliminating the need to include them in SEM segmentation task. The "Export with Masks" option was used for the data export process. An older version (e.g., 2.0.6) was chosen to ensure that the machine mask order corresponded with the interface's default display. This selection was crucial because newer versions of the application exhibit a tendency for smaller area objects to overwrite larger ones, which could compromise the integrity of the annotations.

A total of four SEM images (each 1280 pixels by 960 pixels) were labeled at magnifications 1000×, 800×, 700×, and 600×, producing 45 crops (50 % x-y overlap) in total, with each crop size of 512 pixels by 512 pixels. Subsequently, this labeled dataset was then augmented (Fig. 1 Phase E) using flips (horizontal and vertical), rotates (random 90 degrees), elastic transform (alpha = 1.0, sigma = 50.0, linear interpolation), grid distortion (num_steps = 5, distort_limit = ±0.3, linear interpolation), contrast (0.3, 0.4), and brightness (0.2, 0.3). The specific contrast and brightness levels were based on visual validation, as higher values lead to invalid augmentations (e.g., leading to no difference between the dilution band and the carbide body). All augmentations were implemented using the Albumentations library [36]. This process resulted in 405 crops for the comparison process. Fig. 5 illustrates the applied augmentations against the original sample crop through visualizations.

4. Segmentation operations and architectures

This section discusses the segmentation operations and architectures.

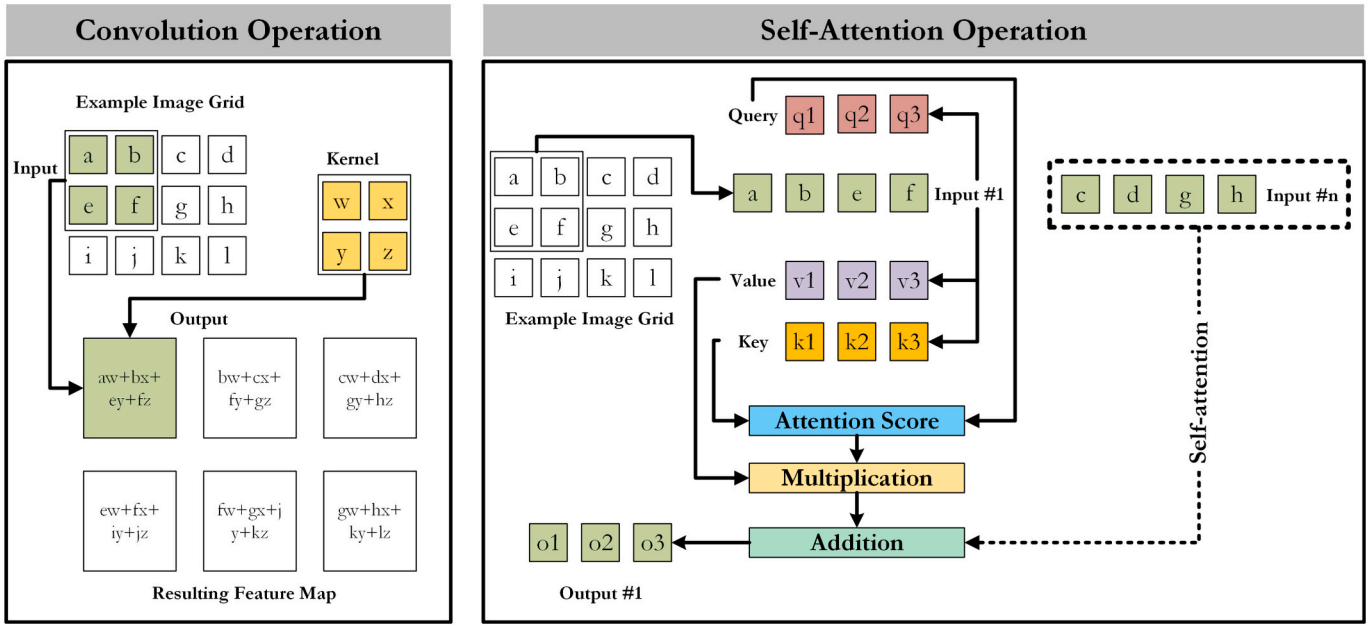

Fig. 7. Pictorial comparison between a convolution operation (left) and a self-attention operation (right). Unlike convolution, which applies the same kernel across the entire image grid to generate feature maps, the self-attention operation considers all relevant input components when generating output for each input element.

Table 1
Segmentation architectures under comparison. In the table, 'L'' refers to layers in the architecture, 'P'' refers to patch size of the input used in the transformer backbone, 'W'' specifies the attention window size for Swin-based transformer backbones. 'b0' represents the SegFormer variant used in the comparison.

Architecture #	Encoder/ backbone	Method/ decoder	Parameters	Reference
Reference CNN	ResNet (50 L)	DeepLabV3+	43,655,648	[39]
Transformer 1	Swin (Base P4 W7)	UPerNet	121,238,503	[43]
Transformer 2	ViT (Large P16)	SETR	309,351,977	[24]
Transformer 3	ViT (Base P16)	DPT	109,674,708	[44]
Transformer 4	MiT (b0 P4)	SegFormer	3,719,018	[45]
Transformer 5	ViT (Base P16)	Segmenter	102,385,162	[46]
Transformer 6	Swin (Small P4 W7)	MaskFormer	63,095,272	[47]
Transformer 7	Swin (Small P4 W7)	Mask2Former	68,777,896	[48]

The comparison is being done between fully convolutional network (reference CNN) and ViTs. ViT architectures can include convolutional and fully connected layers in addition to self-attention blocks [37]. This discussion on architectures follows basic convolution and self-attention operations. An architecture contains both the encoder (or backbone), which usually extracts the features from input images, as well as the specific segmentation method (or segmentation head or decoder), which converts the extracted features into final predicted masks. These discussions are limited to transformer-based architectures, whereas the details on the reference CNN architecture can be found in the original ResNet backbone [38] and segmentation method papers [39].

4.1. Convolution and self-attention

Fig. 6 compares a typical convolutional layer with an attention layer. The first step of a convolutional layer is a kernel-sliding operation that computes dot products between the kernel values and the covered image grid region. The results of this linear transformation go through a non-linear activation function, which enables an architecture employing convolutions to learn non-linear relationships between inputs and outputs. Pooling (max-pooling or average-pooling) is often applied to the resulting grids and helps to reduce the spatial dimensions. Finally, before the resulting feature maps are fed to the next convolutional layer, normalization (e.g., batch, instance, layer) is typically used to stabilize the learning process by normalizing the output of activation functions. Eq. 1 represents the convolution operation.

$$y[i,j,c] = \sum_{m=-k}^{k} \sum_{n=-k}^{k} \sum_{C'} w[m,n,c'].x[i+m,j+n,c'] + b[c] \tag{1}$$

Where:

$y[i,j,c]$: The output feature map at spatial position (i,j) and channel c.

$w[m,n,c']$: Convolution filter weights of kernel size $k \times k$

$x[i+m,j+n,c']$: Input feature map at position $(i+m,j+n)$ and channel c'

$b[c]$: Bias term for channel c

C': Number of input channels

Unlike the fixed convolutional filters, which apply the same transformations across an entire image, the self-attention operation enables dynamic computation of the relevance for an input element based on all other elements in the entire input sequence. The attention layer begins by transforming the input sequence into queries (Q), keys (K), and values (V) through learned linear projections. Next, the dot product between Q and transposed K is calculated to determine how much each input relates to the others. The attention scores are scaled by dividing with the square root of the dimensionality of K to ensure numerical stability. These scaled scores are then passed through a Softmax function to normalize them into attention weights. The attention weights are finally multiplied with the corresponding V and summed to produce the output for each input. The exact sequence of steps is repeated for all input sequence elements. As illustrated in Fig. 7, while convolutional filters focus on local spatial patterns with fixed receptive fields, self-attention considers the entire input grid (e.g., Input #n contributing to output #1), allowing it to capture long-range dependencies across an image. For vision data, this capability makes self-attention particularly effective in modeling global relationships, but it also requires flattening the image into patch embeddings, adding encodings to retain positional information, and handling higher computational costs compared to convolution.

Eqs. 2–5 represent the aforementioned self-attention operation.

$$\text{Score}(q_i,k_i) = q_i.k_i = \sum_{d=1}^{d_k} Q[i,d].K[j,d] \tag{2}$$

$$\text{Scaled Score}(q_i,k_i) = \frac{\text{Score}(q_i,k_i)}{\sqrt{d_k}} \tag{3}$$

$$\alpha_{ij} = \text{softmax}\left(\frac{\text{Score}(q_i,k_i)}{\sqrt{d_k}}\right) \tag{4}$$

$$\text{Output}(i) = \sum_{j=1}^{N} \alpha_{ij}.v_j \tag{5}$$

Where:

q_i and k_i are query and key vectors for input positions i and j

d_k represents the dimensionality of the key vectors

α_{ij} represents the attention weight between positions i and j

v_j is the value vector at position j and N represents the input elements in the sequence

4.2. Vision transformer encoders or backbones

Table 1 lists the reference CNN and candidate transformer architectures alongside their encoders and decoders. A diverse set of encoders was selected to evaluate the varying capabilities in feature extraction for SEM image segmentation. The reference encoder (ResNet-50) represents a widely used convolution-based backbone and is often regarded as a benchmark for semantic segmentation. Among the transformer encoders, the ViT backbone (Base and Large variants) extracts fixed resolution features through its global context modeling whereas the shifted window (Swin) backbone (Base and Small variants) extracts hierarchical multi-scale features through local attention in shift windows. The mix transformer or MiT backbone (b0 variant) also extracts multi-scale features but balances the performance with efficiency through its lightweight design. This functionality could be ideal for real-time or industrial applications. In the future, large encoders like BERT pre-training of images transformers (BEiT [41]) can be considered for comparison employing pertained encoder checkpoints. Moreover, depending on the findings, encoders smaller than MiT such as the MobileViT encoder (Extra Small variant with 2.3 million parameters [42]) can also be investigated. As a result, the current selection provides representative capabilities to investigate transformer backbones for segmenting SEM images of AMed MMCs.

The transformer backbones considered in this study, ViT, Swin, and MiT, represent distinct approaches to encoding features for vision tasks. The ViT backbone uses a consistent patch size of 16 × 16 to generate non-overlapping patches and applies linear embedding to produce tokenized representations. ViT computes self-attention globally across all patches, maintaining a fixed resolution throughout its layers, as illustrated in Fig. 8. In contrast, the Swin backbone employs a hierarchical encoder with a patch size of 4 × 4 and applies shifted windows to compute attention locally. It incorporates patch merging in deeper

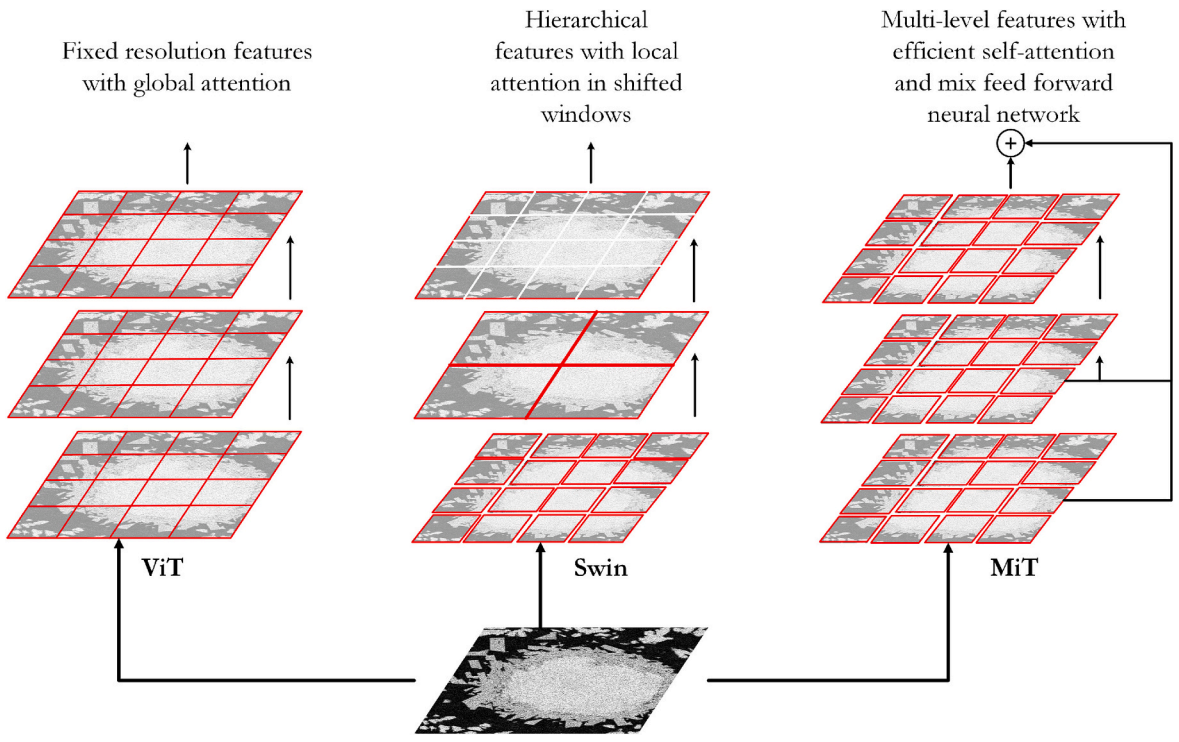

Fig. 8. Transformer-based backbones employed in the semantic segmentation architectures of this study. The red and grey patches symbolically highlight the extraction of features with global and local attention as well as at varying scales. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

layers, progressively reducing spatial resolution by 4, 8, and 16 factors, enabling multi-scale feature extraction. Finally, the MiT backbone builds on these principles with overlapping patches for tokenization and a hierarchical encoder comprising four transformer blocks. These blocks produce multi-level feature maps at scales 1/4, 1/8, 1/16, and 1/32 of the original image dimensions. To maintain feature consistency with non-overlapping patches, MiT employs overlapped patch merging. This design ensures both spatial continuity and efficient self-attention across varying resolutions, as symbolically depicted in Fig. 8.

4.3. Segmentation methods

Similar to the transformer encoders, the decoders in Table 1 represent a diverse selection with methods that have been proposed to work specifically with transformer encoders. DeepLabV3+ is based on atrous convolutions that expand the receptive field and capture multi-scale context and thus serves as a strong CNN reference for comparison. Unified perceptual parsing network (UPerNet) can integrate hierarchical feature maps from Swin transformer backbone to accomplish multi-scale fusion. ViT for dense prediction (DPT) uses a transformer-based architecture to improve global context aggregation and has been proposed for dense prediction tasks like semantic segmentation and depth estimation. SegFormer contains a lightweight multi-layer perceptron (MLP) decoder to process features from MiT encoder. This design has been shown to maintain performance while offering simplicity. Segmenter extends pure transformer architectures through mask embeddings for dense pixel-level classification and can demonstrate the potential of attention mechanism for SEM segmentation. SETR is another pure transformer-based decoder and has been used in conjunction with a ViT encoder for semantic segmentation. Finally, MaskFormer and its extension Mask2Former have been proposed for universal segmentation tasks (combining semantic, panoptic and instance segmentation) and represent advancements in decoder designs. This diverse selection of decoders can be extended in the future by adding more transformer-based methods for semantic segmentation (e.g., HRFormer [49], Data2VecVision [50]).

Vision transformer for dense prediction or DPT introduces an approach for dense or pixel-level prediction tasks (e.g., semantic segmentation, depth estimation) by replacing traditional convolutional backbones with transformer-based architectures. As depicted in Fig. 9 (A), the DPT framework begins by transforming an input image into a sequence of patches. These patches are generated using a linear embedding process or derived from a ResNet-50-based hybrid feature extractor, with positional embeddings added for spatial context. The transformer backbone processes these tokens at a uniform and high resolution across multiple stages, enabling a global receptive field that captures fine-grained details and broader spatial relationships. To construct full-resolution predictions, the architecture reassembles tokens into feature maps at varying scales (e.g., 1/32, 1/16, 1/8, and 1/4 of the original image resolution) through hierarchical fusion modules. These modules (shown in green between transformer and fusion modules in Fig. 9 (A)) progressively refine the features using convolutional units and up-sampling. DPT uses a mix of hierarchical features and convolution-based decoding to produce precise and globally consistent outputs. This makes it very useful for dense vision tasks (e.g., segmentation).

The segmentation transformer or SETR presents an alternative approach to semantic segmentation by treating it as a sequence-to-sequence prediction task, departing from the conventional encoder-decoder FCN framework. Instead of relying on progressive down-sampling and convolutions to capture semantic context, SETR utilizes a pure transformer-based encoder that processes an image as a sequence of fixed-size patches, while maintaining the original spatial resolution. This design allows for global context modeling at every layer of the transformer by improving the ability to capture fine-grained details and large receptive fields. Fig. 9 (B) shows that the architecture begins by embedding non-overlapping image patches into tokens, which are later enhanced with positional embeddings. These tokens are passed through multiple transformer layers to generate comprehensive feature representations. For pixel-wise segmentation, the architecture employs

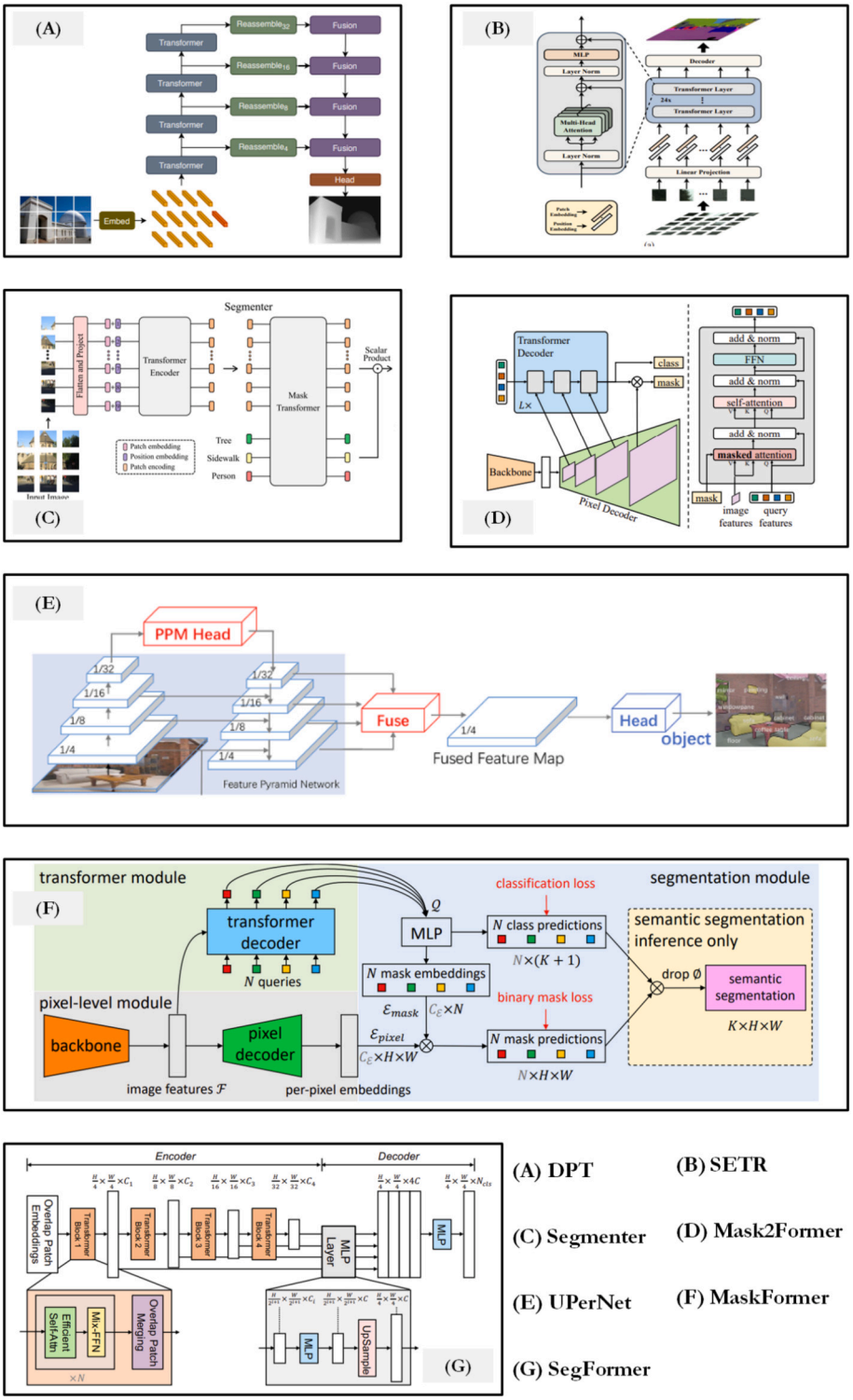

Fig. 9. Segmentation methods employed in the comparison study. (A) Dense Vision Transformer. (B) Segmentation Transformer. (C) Transformer for Semantic Segmentation. (D) Masked-attention Mask Transformer. (E) Unified Perceptual Parsing. (F) MaskFormer. (G) SegFormer. Figures taken or adapted from the original papers referenced in Table 1.

decoder designs, such as progressive up-sampling (SETR-PUP), which reshapes and incrementally upscales feature maps to restore the original image resolution, and multi-level feature aggregation (SETR-MLA), which integrates features from different transformer layers to enhance spatial accuracy. This framework eliminates the dependence on convolutional layers, achieving context-aware segmentation through the transformer's global attention mechanism.

Transformer for semantic segmentation or Segmenter model introduces a transformer-based approach to semantic segmentation by leveraging global context at each network layer. Unlike traditional convolutional methods, Segmenter builds on the ViT architecture, adapting it for segmentation tasks by treating image patches as tokens and projecting them into embeddings. As illustrated in Fig. 9 (C), the model consists of an encoder that transforms input images into patches,

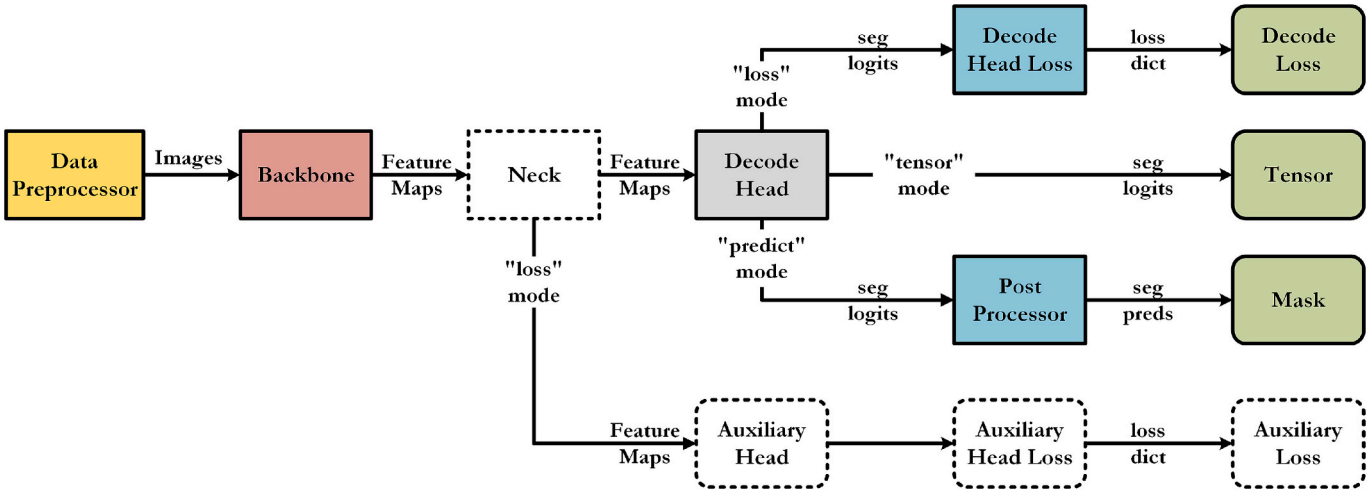

Fig. 10. Components of MMSegmentation forward function implemented for a model instance—figure adapted from official MMSegmentation documentation.

followed by the addition of positional information. These embeddings are passed through a transformer encoder to capture global contextual information. The output embeddings are then processed by a mask transformer decoder, which generates class-specific masks via scalar products, resulting in the final segmentation map. While a linear decoder provides robust baseline performance, the mask transformer decoder further enhances segmentation accuracy by generating detailed class masks. This architecture demonstrates the potential of transformer-based models in semantic segmentation by enabling fine-grained and globally coherent predictions.

The Masked-attention mask transformer or Mask2Former is a universal architecture designed to handle a broad spectrum of image segmentation tasks, including panoptic, instance, and semantic segmentation. Unlike traditional approaches that necessitate specialized architectures for each segmentation task, Mask2Former unifies these under a single framework. The model architecture in Fig. 9 (D) comprises three key components: a backbone, a pixel decoder, and a transformer decoder. Masked attention within the transformer decoder leads to localized features by limiting cross-attention to predicted mask regions. This functionality improves the efficiency of feature extraction and enhances the handling of small objects. The pixel decoder processes multi-scale features, feeding them to the transformer decoder in a layer by layer fashion. By removing redundant computation through learnable query features and reordered self- and cross-attention layers, Mask2Former achieves superior performance across various segmentation benchmarks.

The Unified perceptual parsing network or UPerNet provides a versatile multi-task architecture capable of recognizing various visual elements within an image, including scene contexts, objects, materials, and textures. UPerNet effectively captures hierarchical multi-scale features by integrating a feature pyramid network (FPN) and a pyramid pooling module (PPM). The architecture presented in Fig. 9 (E) processes fused feature maps at different scales and directs them to distinct output heads for scene and object-level segmentation. In the context of semantic segmentation, this fusion mechanism enhances spatial coherence and captures fine-grained details. By focusing exclusively on semantic segmentation, this framework takes advantage of UPerNet's ability to combine features hierarchically, which enables precise and reliable segmentation results.

MaskFormer introduces a unified framework for tackling semantic and panoptic segmentation tasks by leveraging a mask classification approach rather than the traditional per-pixel classification paradigm. As depicted in Fig. 9 (F), the architecture utilizes a backbone to extract image features F, which are processed by a pixel decoder to generate per-pixel embeddings. A transformer decoder utilizes these features and produces per-segment embeddings Q that correspond to a set of binary mask predictions and their associated global class labels. This unified approach simplifies the segmentation pipeline by treating each binary mask as a standalone prediction. It allows the same model and training procedure to address semantic and instance-level segmentation tasks. By integrating segmentation inference through a dot product of mask and pixel embeddings, MaskFormer achieves strong empirical results in scenarios involving many classes.

SegFormer offers a simple and efficient framework for semantic segmentation. It combines a hierarchical transformer encoder with a lightweight MLP decoder. As illustrated in Fig. 9 (G), the transformer encoder extracts multi-scale features using overlapping patch embeddings and a series of hierarchical transformer blocks. These blocks progressively reduce the spatial resolution from $\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{32} \times \frac{W}{32}$ (H represents image height, W represents image width). Notably, SegFormer eliminates the need for positional encodings, which helps avoid interpolation issues when the testing resolution differs from the training resolution. The lightweight MLP decoder directly fuses features across multiple levels using feed-forward layers and up-sampling operations. This approach integrates local and global attention mechanisms. The framework provides variants from MiT-b0 (3.7 million parameters) to MiT-b5 (82.0 million parameters), which successively demonstrate superior efficiency and accuracy compared to prior methods. This combination of simplicity and performance makes SegFormer a robust choice for semantic segmentation tasks.

5. Training experiments

The dataset was split into 75 % training samples, 15 % validation samples, and 15 % test samples for the experiments. The test and train splits were composed of only the crops coming from the lowest magnification (e.g., 600×). This was done to evaluate the potential of ML for high-throughput segmentation, since a montage-based data extraction can support fast SEM analysis of printed bead cross-sections at lower magnifications, while also reasonably capturing the carbide defects.

The MMSegmentation framework of OpenMMLab was used to conduct the experiments [51–53]. The MMSegmentation framework provides a modular and flexible pipeline for semantic segmentation, as illustrated in Fig. 10. The data preprocessor first processes input images to generate normalized inputs for the backbone. Subsequently, the backbone extracts hierarchical feature maps. Optional neck modules further refine these feature maps before being passed to the decode head. The head is responsible for producing segmentation logits. The logits serve as the raw predictions of the model and provide valuable information on the confidence of each class prediction. In "loss" mode,

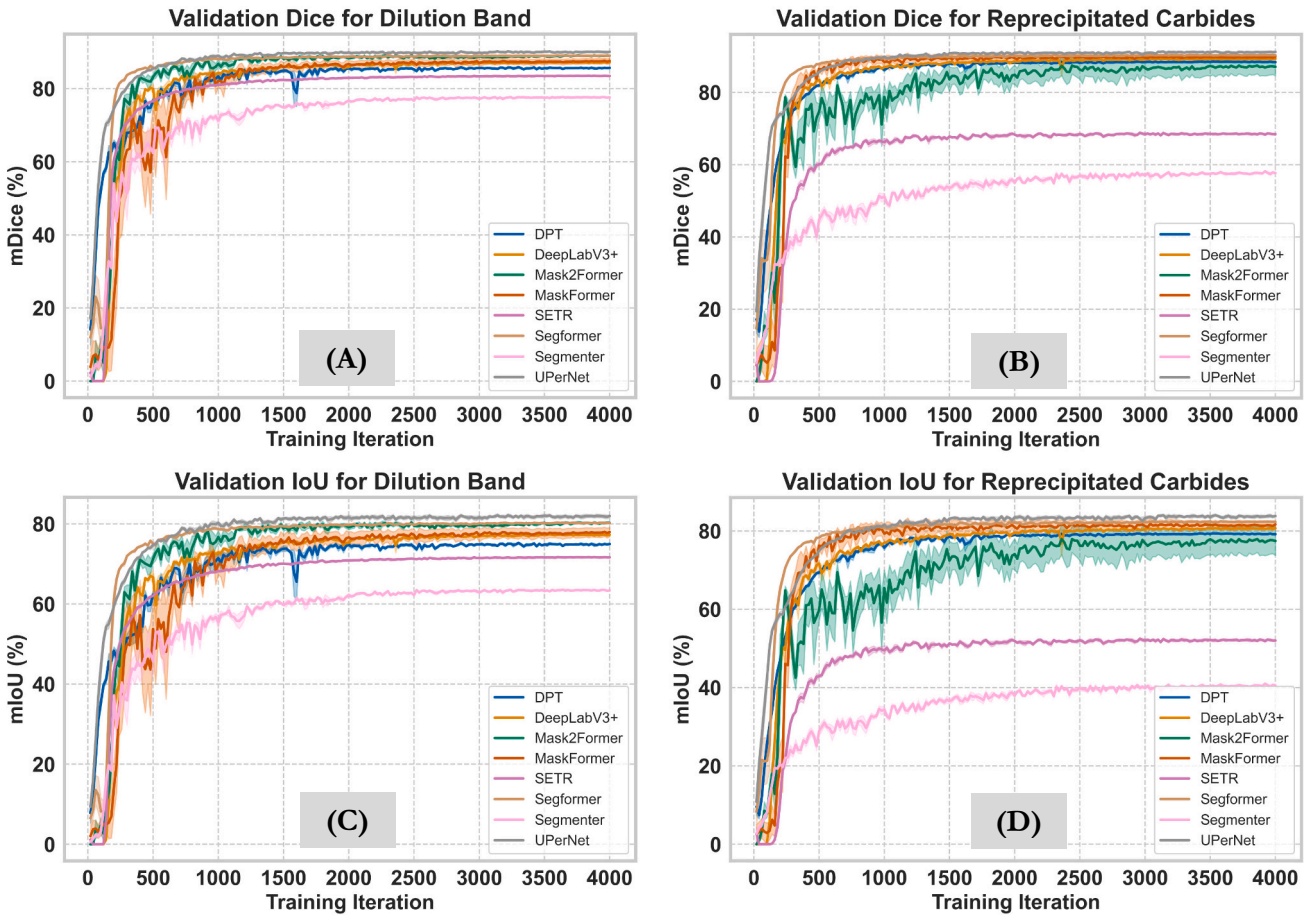

Fig. 11. Validation curves for percentage mean Dice Coefficient and percentage mean IoU on dilution band and reprecipitated carbides across the models under comparison.

the decode head computes the primary segmentation loss. The logits from the decode head are post-processed in "predict" mode to produce final segmentation masks. When adapting models for specific tasks, decode and auxiliary heads are modified by adjusting the number of classes or by introducing task-specific loss functions. This modularity allows for efficient customization and performance optimization across diverse segmentation datasets and applications.

All MMSegmentation models were implemented using Python programming language and the PyTorch deep learning framework. For each model, the default configurations were adapted to meet the needs of our dataset, which included setting the crop size to 512 by 512 pixels and updating the preprocessor to handle the mean and standard deviation of the dataset. The decoder head was updated by defining five custom classes (e.g., matrix, porosity, carbide particles, dilution band, and reprecipitated carbides) for SEM image segmentation. The class label values were zero-indexed to meet the requirements of the library. As mentioned in Section 3, the porosity class (index 1) was ignored in the training process due to its sparse presence and irrelevance to the SEM image analysis. A batch size of 4 was used across all experiments. All models were trained for 5000 iterations with a 20-step validation interval leading to 250 sets of validation metrics during training. One model checkpoint with updated weights was saved at the end (e.g., 5000th iteration), while the checkpoints with the best mean metrics were also saved. Instead of using pre-trained weights from MMSegmentation checkpoints, these were randomly initialized to mitigate the impact of different pre-trained configurations (different datasets, objectives, and normalization conventions) on comparison results. These differences could advantage particular backbones for reasons unrelated to performance on SEM images. Across all experiments, five different seed values (e.g., 1, 2, 3, 4, 5) were used to compare the effect of different weight initialization and dataset shuffling, while keeping everything the same in the model configuration.

During training, a stochastic gradient descent (SGD) optimizer was used to lower the training error in the process of iterative improvement. A learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0005 were used during the optimization process. A combination of schedulers was used to dynamically adjust the value of the learning rate, which started with a linear warm-up (iteration $\leq$ 1000) followed by a cosine annealing schedule (iteration >1000). This combination supported stable initial training, followed by a gradual reduction in the learning rate to enable smooth convergence and to avoid any local minima.

Three different evaluation metrics were used to compare the performance of models during training and on the test set: accuracy, intersection over union (IoU), and f-score or Dice Coefficient. Accuracy measures the proportion of accurately classified pixels against the entire image or dataset. The IoU metric, also known as the Jaccard Index, is a class-sensitive measure and less dependent on the total number of pixels. The Dice Coefficient, also known as the f-score, quantifies the extent of overlap between two distinct sets. In semantic segmentation tasks, the Dice Coefficient measures the similarity between the predicted segmentation results and ground truth. This metric is highly responsive to the degree of overlap between the prediction and the ground truth mask, making it especially effective for semantic segmentation where precise overlap is essential. While mean IoU and f-score were recorded for all models, special attention was given to models' performance on segmenting the dilution band from the carbide particles and reprecipitated phases from the matrix by recording per-class metrics. We also used a

Table 2
Mean test metrics for **DeepLabV3+** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	91.87 ± 0.01	95.76 ± 0.01	95.76 ± 0.05
Carbide particles	97.69 ± 0.02	98.83 ± 0.01	98.58 ± 0.05
Dilution band	77.47 ± 0.16	87.31 ± 0.10	90.11 ± 0.25
Reprecipitated carbides	80.56 ± 0.13	89.24 ± 0.08	87.95 ± 0.16

Table 3
Mean test metrics for **SegFormer** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	92.60 ± 0.01	96.16 ± 0.00	95.95 ± 0.08
Carbide particles	98.09 ± 0.02	99.04 ± 0.01	98.80 ± 0.05
Dilution band	80.19 ± 0.06	89.00 ± 0.04	93.21 ± 0.10
Reprecipitated carbides	82.05 ± 0.04	90.14 ± 0.02	88.18 ± 0.13

Table 4
Mean test metrics for **Mask2Former** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	89.94 ± 1.13	94.69 ± 0.63	98.68 ± 0.30
Carbide particles	98.07 ± 0.02	99.03 ± 0.01	98.69 ± 0.04
Dilution band	80.31 ± 0.18	89.08 ± 0.11	87.84 ± 0.39
Reprecipitated carbides	77.12 ± 3.57	86.89 ± 2.40	79.89 ± 4.24

Table 5
Mean test metrics for **MaskFormer** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	90.86 ± 0.71	95.21 ± 0.39	97.63 ± 0.59
Carbide particles	97.97 ± 0.04	98.98 ± 0.02	98.39 ± 0.07
Dilution band	78.12 ± 1.13	87.70 ± 0.71	87.10 ± 1.88
Reprecipitated carbides	81.48 ± 1.34	89.77 ± 0.83	86.65 ± 2.34

Table 6
Mean test metrics for **UPerNet** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	92.75 ± 0.01	96.24 ± 0.01	96.10 ± 0.02
Carbide particles	98.02 ± 0.01	99.00 ± 0.01	98.55 ± 0.02
Dilution band	82.04 ± 0.43	90.13 ± 0.26	94.29 ± 0.08
Reprecipitated carbides	83.92 ± 0.37	91.25 ± 0.22	89.97 ± 0.44

Table 7
Mean test metrics for **DPT** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	91.57 ± 0.02	95.60 ± 0.01	95.57 ± 0.03
Carbide particles	97.00 ± 0.03	98.48 ± 0.02	97.47 ± 0.04
Dilution band	74.57 ± 0.18	85.43 ± 0.12	91.47 ± 0.09
Reprecipitated carbides	79.13 ± 0.08	88.35 ± 0.05	87.25 ± 0.07

confusion matrix to visualize and compare the performance of models across different classes, as it highlights which classes were consistently confused for other classes by the models. Eqs. 6–8 show basic formulas of these metrics with parametric explanations.

Table 8
Mean test metrics for **Segmenter** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	74.76 ± 0.02	85.56 ± 0.01	89.44 ± 0.11
Carbide particles	97.00 ± 0.02	98.48 ± 0.01	98.39 ± 0.02
Dilution band	63.64 ± 0.14	77.78 ± 0.10	77.25 ± 0.15
Reprecipitated carbides	41.31 ± 0.10	58.47 ± 0.10	52.22 ± 0.29

Table 9
Mean test metrics for **SETR** with standard error across all five runs.

Metrics	mIoU ± SE	mDice ± SE	mAcc ± SE
Class			
Matrix	75.96 ± 0.04	86.34 ± 0.03	87.41 ± 0.28
Carbide particles	97.51 ± 0.03	98.74 ± 0.01	98.37 ± 0.05
Dilution band	70.78 ± 0.23	82.89 ± 0.16	87.23 ± 0.25
Reprecipitated carbides	51.58 ± 0.36	68.05 ± 0.31	64.16 ± 0.84

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Where:

TP = True Positives are the number of pixels correctly predicted as belonging to the target class.

TN = True Negatives are the number of pixels correctly predicted as not belonging to the target class.

FP = False Positives are the number of pixels incorrectly predicted as belonging to the target class.

FN = False Negatives the number of pixels incorrectly predicted as not belonging to the target class.

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP + FP + FN} \tag{7}$$

Where:

P and G are the predicted and ground truth sets respectively.

$P \cap G$ is the intersection (overlap) between prediction and ground truth.

$P \cup G$ is the union of the prediction and ground truth.

$$\text{Dice} = \frac{2\,|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

6. Discussions and findings

Fig. 11 (A-D) shows the performance of models on the validation set during training. The plots are limited to 4000 iterations to highlight key regions of performance improvement, as the order of performance for all models remained the same for the rest of the training. The plots highlight the overall mean Dice (mDice) and mean IoU (mIoU) metrics for carbide defect categories, namely dilution band and reprecipitated carbides, as the overall mean may not be representative of performance on these two categories of interest. UPerNet, with a Swin backbone, performs the best in segmenting both categories. For dilution band segmentation on the validation set, mIoU and mDice show that SegFormer and Mask2Former perform similarly, as the second-best segmentation models. These are followed by MaskFormer and DeepLabV3+, where the former model performs slightly better. DPT, SETR, and Segmenter maintain the last three positions, with Segmenter being the lowest-performing model.

Notably, the SegFormer model, while being much smaller than the other models, achieves top performance in the first 500 iterations. However, its capacity may be restricted for learning complex patterns; this could have limited its performance against the UPerNet model in the following iterations. For segmentation of reprecipitated carbides on the

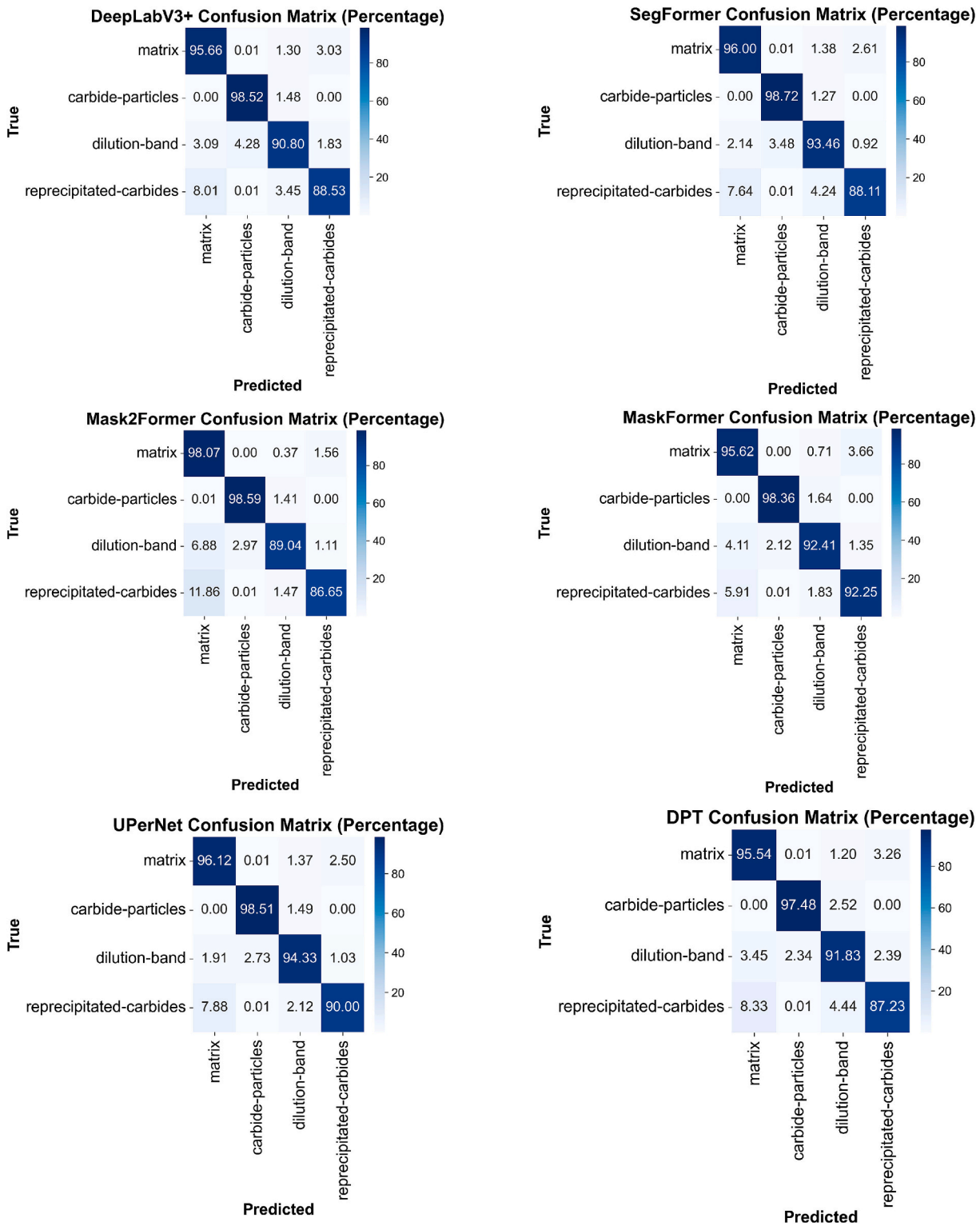

Fig. 12. Confusion matrices on the test set for each model. Presented results correspond to best performing model on dilution band out of the five runs. DeepLabV3+ (Seed 3, 5000 iteration) SegFormer (Seed 5, 5000 iteration), Mask2Former (Seed 1, 5000 iteration), MaskFormer (Seed 4, 5000 iteration), UPerNet (Seed 4, 5000 iteration), DPT (Seed 5, 5000 iteration), Segmenter (Seed 3, 5000 iteration), SETR (Seed 5, 5000 iteration).

validation set, mIoU and mDice show that while SegFormer is still the second-best model after UPerNet, it is not the case for Mask2Former, as it struggles to learn and segment reprecipitated carbides throughout the training. The SegFormer model is followed by MaskFormer and DeepLabV3+, and the performance of these three models is close to each other. The fifth model in the performance is DPT, where Mask2Former model performs lower and takes the 6th spot in segmenting reprecipitated carbides on the validation set. The last two performing models remain the same as the dilution band. However, their performance is significantly lower on reprecipitated carbides (77.46 mDice dilution

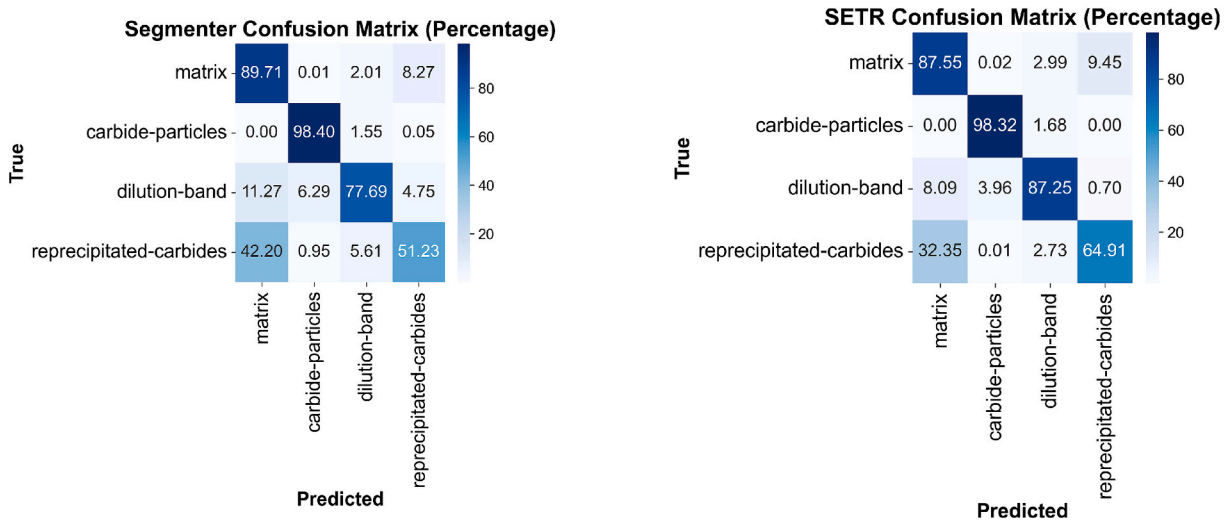

Fig. 12. (*continued*).

band versus 57.71 mDice reprecipitated carbides for Segmenter, 83.53 mDice dilution band versus 68.51 mDice reprecipitated carbides for SETR).

6.1. Results

Tables 2–9 present predictions for each class using mean IoU, mean Dice, and mean Accuracy over the entire test set. The standard error or SE, as in [54], highlights the variations in the performance across the five runs for each model. In the case of the dilution band, while UPerNet maintains top performance also on the test set, Mask2Former performs slightly better than SegFormer, followed by Mask2Former. The CNN-based DeepLabV3+ model is ranked 5th on the test set for its performance in segmenting the dilution band followed by DPT, SETR, and Segmenter. Notably, the two masked transformers (MaskFormer & Mask2Former) have much higher standard errors than any other model. In the case of reprecipitated carbides, SegFormer performs better than both masked transformers after the UPerNet model. The CNN-based DeepLabV3+ performs better than one of the masked transformers and is ranked 3rd. This performance could be explained based on the ability of the CNN models to better capture intricate shape features, as in the case of reprecipitated carbides. The variability of masked transformers across the five runs makes their performance less robust, giving an edge to UPerNet and SegFormer as the two top-performing transformer models.

Fig. 12 shows the confusion matrices for the best-performing run on the test set across all models. The values are averaged over the entire test set and presented as a percentage of the total pixels in each class. Notably, the lower performance for all models can be attributed to classification errors related to the dilution band and reprecipitated carbide pixels. Between these two classes, the misclassification of reprecipitated carbide pixels as matrix pixels represents the highest misclassification error across all models. While MaskFormer has the lowest misclassification (5.91 %) error in this regard, the results could be subjected to high-performing run as indicated earlier by higher SEs. This makes SegFormer (7.64 % misclassified reprecipitated carbides) and UPerNet (7.88 % misclassified reprecipitated carbides) the two top-performing models in this regard, with UPerNet still performing better overall on reprecipitated carbides (90 % correctly classified pixels). While the dilution band has the second highest misclassification errors after the reprecipitated carbides; these are more distributed between the matrix and carbide particles, suggesting that all models struggle similarly when classifying dilution band pixels as either matrix or carbide particles. On dilution band pixels, UPerNet has the highest performance (94.33 %), with the SegFormer model as a close second (93.46 %).

6.2. Observations

Figs. 13 and 14 show the iterative improvement on the validation set during the training of the best (UPerNet) and worst (Segmenter) performing models on a sample of the SEM image dataset from the AMed Ni-WC MMCs. In this study, the interval selected contained the most improvement in performance across all training iterations. Fig. 13 illustrates how the Segmenter model fails to precisely capture the context of the dilution band, leading to misclassification of dilution band pixels with the matrix, carbide particles, or even with reprecipitated carbides. It also struggles to learn and segment reprecipitated carbides, as several of these are misclassified for matrix. Lastly, some of the reprecipitated carbides were consistently misclassified as carbide particles, indicating that overall, the model struggles to learn the class-wise fine intensity gradients, leading to poor contextualization and the subsequent localization.

Fig. 14, on the other hand, highlights how the top-performing UPerNet model promptly captures the context of each class and refines the precise localization of dilution band and reprecipitated carbide pixels over the remaining training iterations. One notable similarity between the two models in these visualizations is their struggle to consistently segment the partial dilution band on the left (e.g., without the associated carbide body). The Segmenter model consistently segments only a portion of the dilution band across the training iterations, whereas the UPerNet model randomly captures it, while missing the band altogether for most iterations. This observation could highlight the dependence of the models on the neighborhood context but requires focused investigation before conclusion.

For practical evaluation, the models were also qualitatively tested on a larger crop (e.g., 1280 pixels by 960 pixels) at lower magnification (e.g., ×600) comprising validation and test sets. Fig. 15 presents the results for the UPerNet model alongside the confusion matrix highlighting the quantified pixels. The original predicted mask was updated to replace the misclassified pixels with red before being laid on the input image. Most of the misclassified labels lie at the carbide-dilution and reprecipitate-matrix boundaries, highlighting the source of segmentation errors. Interestingly, the small, diluted carbide at the bottom was partially misclassified due to its relatively dense structure resembling carbide particles.

Fig. 16 shows the results for the best-performing run of the model with overall low performance. Like the UPerNet mask, the Segmenter mask was updated to replace the misclassified pixels with red before

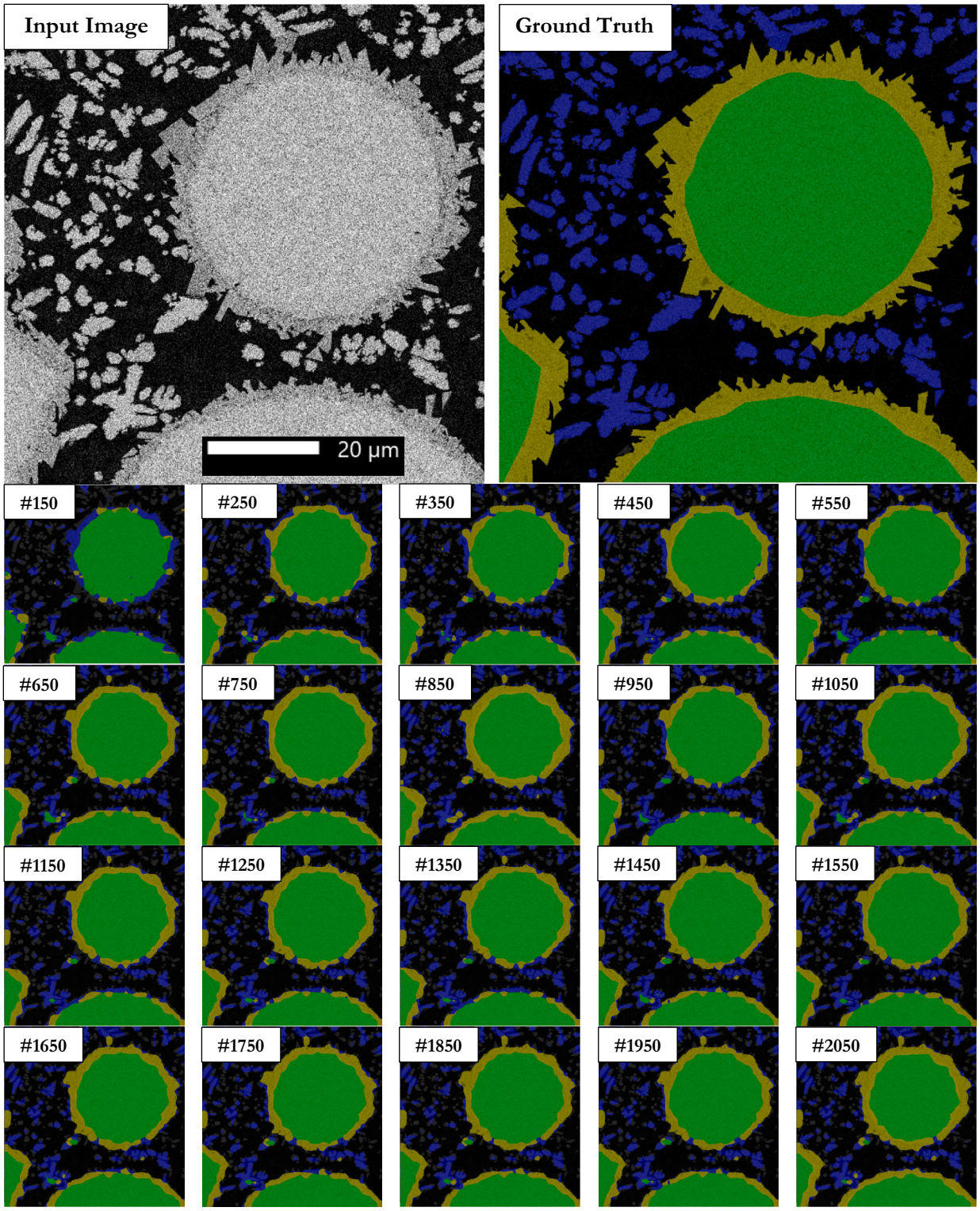

Fig. 13. Iteration-wise visual validation of predictions from **Segmenter** with ViT backbone during the lowest performing run (seed 5) from iterantion#150 to iteration#2050 in steps of 100 on a random validation crop (e.g., sem600_x256_y0_HorizontalFlip).

being laid on the original input image. In addition to poor performance on challenges discussed earlier (e.g., separating phase boundaries), the Segmenter model also struggled to accurately segment the dilution band and matrix pixels. Most notably, 39,877 matrix pixels in the original image were misclassified as reprecipitated carbide, whereas 78,166 reprecipitated carbide pixels were incorrectly predicted as matrix. These misclassifications led to an overall lower performance of the model. Similar to the UPerNet model, the Segmenter model was also able to identify partially appearing dilution bands, indicating its ability to learn without depending on the neighborhood context of the pixels.

The misclassification errors at phase boundaries can be attributed to two main factors: (i) annotation ambiguity, and (ii) model limitations. Since phase boundaries in SEM images often exhibit gradual intensity transitions, making precise pixel-level delineation can be subjective

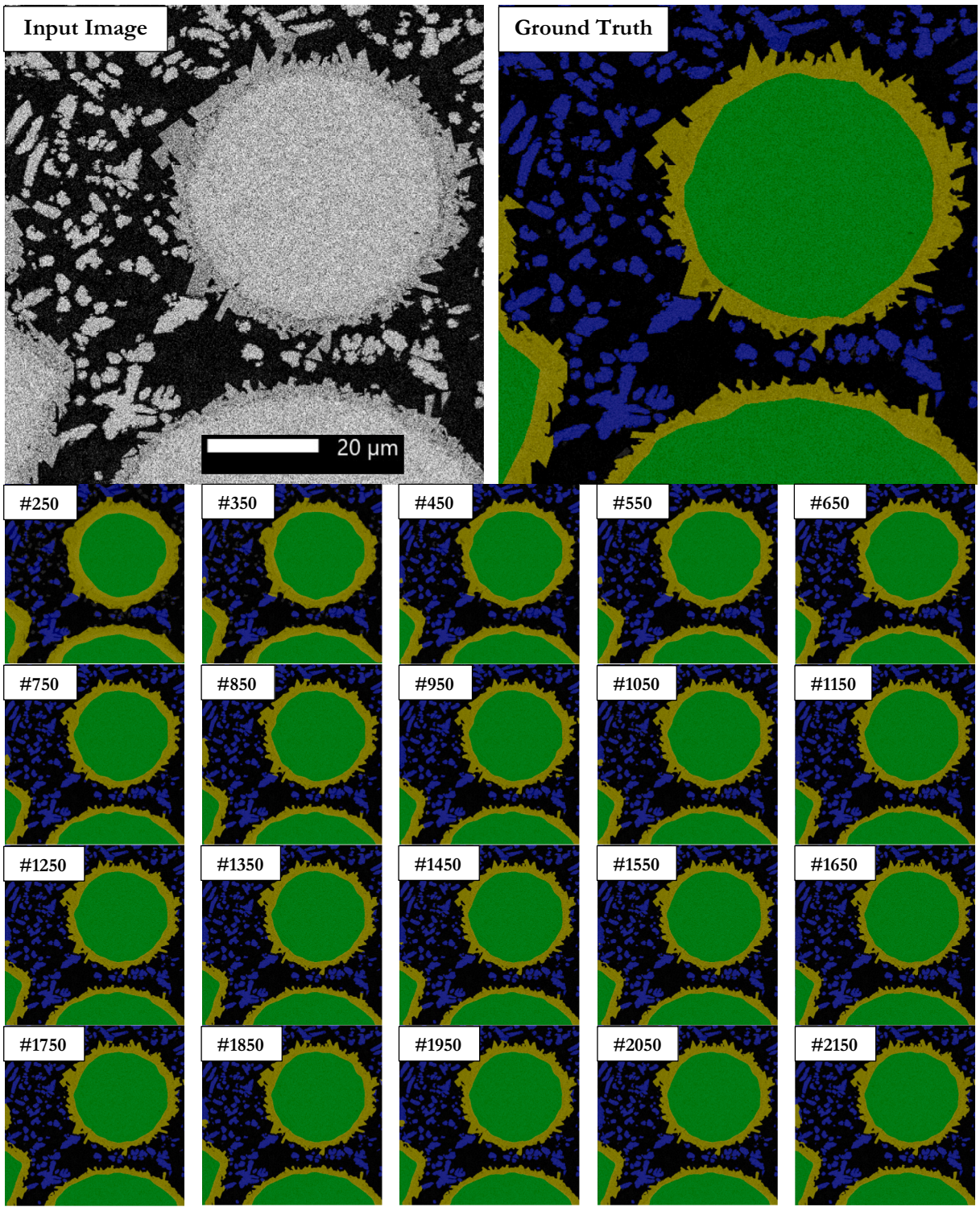

Fig. 14. Iteration-wise visual validation of predictions from **UPerNet** with Swin backbone during the highest performing run (seed 4) from iterantion#250 to iteration#2150 in steps of 100 on a random validation crop (e.g., sem600_x256_y0_HorizontalFlip).

even for experts leading to some degree of annotation ambiguity. Moreover, certain model architectures (e.g., transformers), despite their global context awareness, may still struggle to capture fine-grained gradients at small scales. The future work could focus on boundary-aware loss functions or multi-annotator labeling to reduce such ambiguities leading to reduced misclassifications at phase boundaries.

6.3. Practical considerations and recommendations

The confidence in the prediction of a model when segmenting classes of interest can be highlighted using logits. A logit is a raw model output from its final layer before it is normalized through an activation function. Fig. 17 (A) presents the logits across all classes to compare the confidence in models' predictions under comparison for segmenting SEM images. The models are significantly more confident when

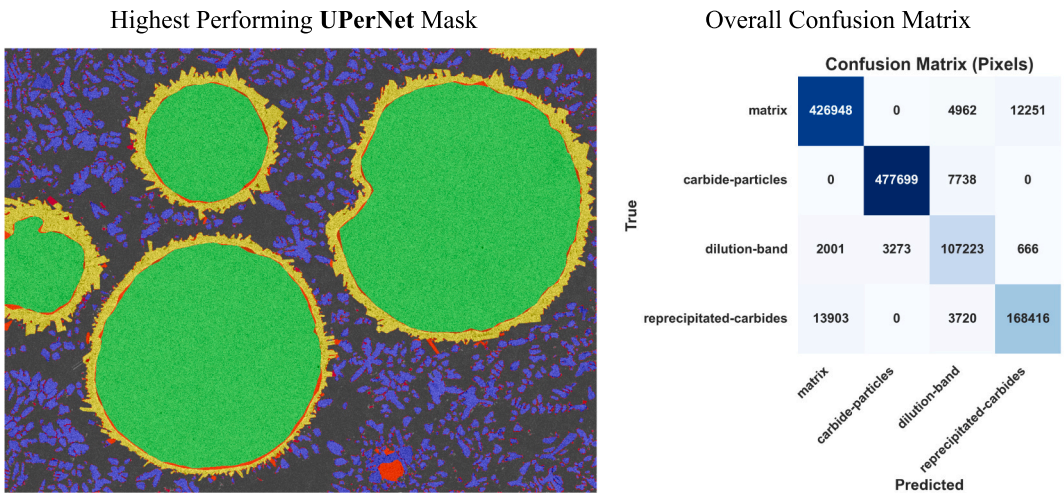

Fig. 15. UPerNet checkpoint from 5000th iteration taken from best performing seed for evaluation on a larger image (e.g., 1280 pixels by 960 pixels). Misclassified pixels are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

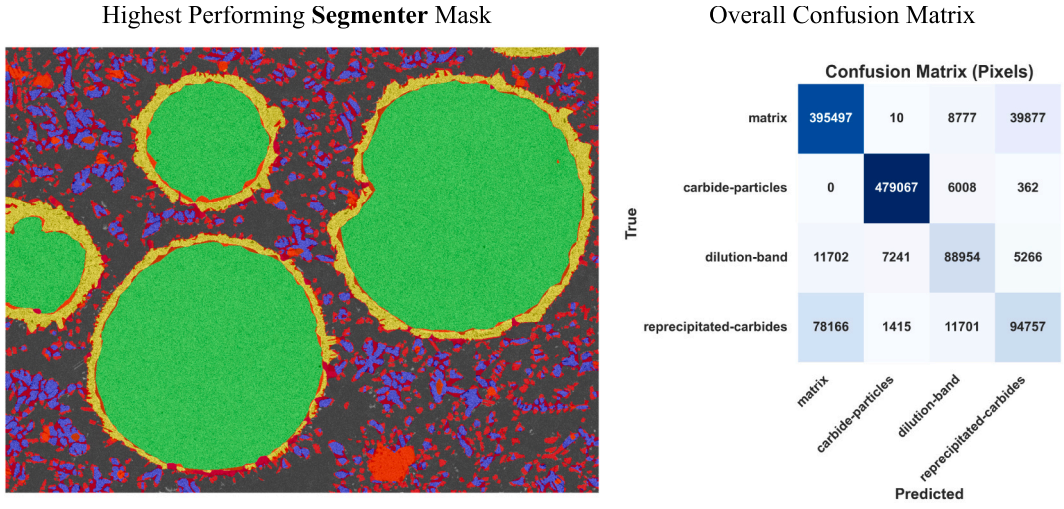

Fig. 16. Segmenter checkpoint from 5000th iteration taken from the best-performing seed for evaluation on a larger image (e.g., 1280 pixels by 960 pixels). Misclassified pixels are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predicting the pixels for the carbide particles and matrix compared to the dilution band and reprecipitated carbides with the lowest values corresponding to the dilution band pixels.

The plots in Fig. 17 (B) present logit values as a percentage of the global maximum. It can be seen that masked transformers are more confident on "easy-to-predict" matrix and carbide particle pixels, but their confidence fluctuates on the dilution band and reprecipitated carbide pixels. This behavior could also be the reason for the high SE in the predictions of MaskFormer and Mask2Former on the test set. Fig. 17 (C) shows the logit values as a percentage of the maximum value within each pixel category or class. DPT, UPerNet, and SegFormer models have the highest values for logits when predicting the dilution band pixels. Out of these three models, the UPerNet and SegFormer models performed the best in predicting the dilution band on the test set. On the reprecipitated carbides, these two models have similar logit values. Notably, the CNN-based DeepLabV3+ has slightly higher logit values and its performance in predicting the reprecipitated carbide pixels on the test set is also comparable to UPerNet and SegFormer. This can be attributed to the inherent characteristics of CNNs, as these models are especially effective when detecting patterns, edges, and shapes, features that are representative of the reprecipitated carbide pixels.

We compared single-image (using larger micrograph of Fig. 15 and Fig. 16) inference runtimes on a workstation with an NVIDIA RTX 4090 GPU using PyTorch and MMSegmentation library. We report median per-image latency (milliseconds) and the corresponding single-image equivalent throughput in frames per second (FPS = 1000/median latency in milliseconds) and include the 90th-percentile latency (p90), the time by which 90 % of runs complete, to characterize tail-latency variability. SegFormer achieved a median 42.6 milliseconds per image (23.5 frames per second) with p90 = 51.4 milliseconds, whereas SETR required 396.0 milliseconds per image (2.5 frames per second) with p90 = 453.6 milliseconds, corresponding to an approximately 9.3 times speedup on GPU (8.8 times by p90). On CPU, SegFormer ran in 1.18 s per image (0.85 frames per second; p90 = 1.23 s) versus 23.19 s (0.04 frames per second; p90 = 23.29 s) for SETR. This reflected a 19.7 times speedup (18.9 times by p90). These measurements reflect repeated inference on the same single image and frames per second is reported as the single-image equivalent (1000 divided by median latency in milliseconds).

Based on the findings, the UPerNet method with the Swin backbone is recommended for segmenting SEM images from AMed MMCs in scenarios where accuracy and robustness are desired, such as lab-scale or research and development setups. This recommendation is supported by the performance of the model on challenging features of dilution bands

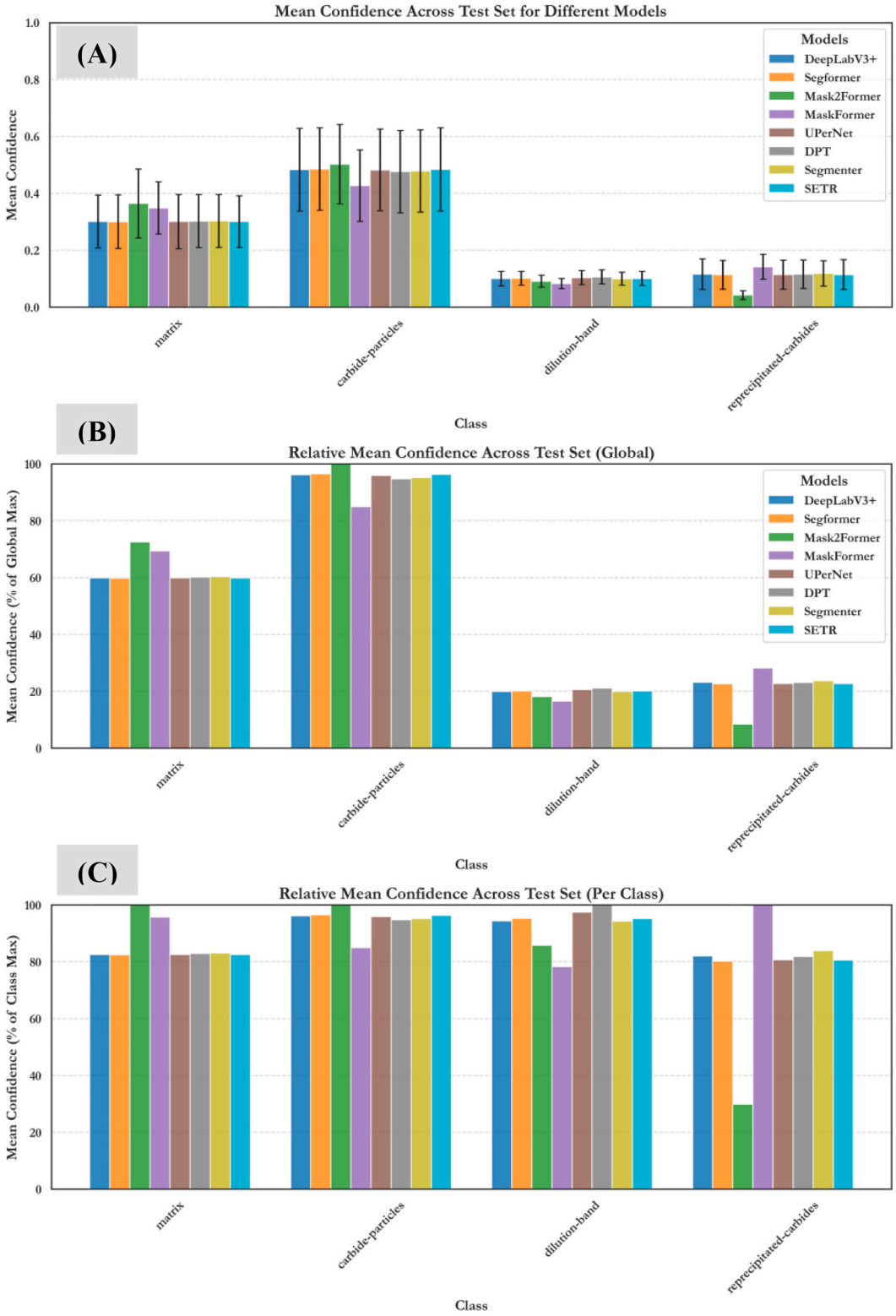

Fig. 17. Logits from the final layer of each model. (A) average raw values, (B) percentage of global maximum, and (C) percentage of class-wise maximum.

and reprecipitated carbides. For industrial applications where computational resources and productivity are critical, the SegFormer model is recommended as a promising alternative due to its lighter design and competitive performance. Moreover, future works are recommended on hyperparameter variations, as well as lighter and high-capacity segmentation methods to build upon the findings of the current work.

7. Conclusions and future works

In conclusion, widely used semantic segmentation methods and encoders based on transformers were evaluated and compared for segmenting damage to the carbide particles at higher thermal conditions in the case of AMed Ni-WC MMCs. The SEM images were used as the inputs and augmentations were applied to enhance the dataset and make the

models more robust to changing data distributions. Three transformer backbones, namely ViT, MiT, and Swin, with different patch sizes and depth variants, were used to extract the features from the SEM images during training. These features were fed to different methods for segmenting the input images into pixels of the matrix, carbide particles, dilution band, and reprecipitated carbides. Specifically, SegFormer, MaskFormer, Mask2Formaer, UPerNet, DPT, Segmenter, and SETR based methods were used for the semantic segmentation task. A reference CNN architecture, DeepLabV3+, with ResNet-50 backbone was also included in the comparison. During training, all models, except SETR and Segmenter, were found to reasonably learn the pixels of carbide anomalies. The masked transformers, MaskFormer and Mask2Formaer, were found to fluctuate significantly across the five runs during training. UPerNet and SegFormer were found to perform best on dilution band (94.33 % and 93.46 % overall accuracy on test set, respectively) and reprecipitated carbide (90.97 % and 88.52 % overall accuracy on test set, respectively) classes. The low-performing models SETR and Segmenter were found to struggle with the precise localization of the dilution band features. The models were also tested on a large industrial image mimicking high throughput analysis for process characterization. The primary source of misclassification was found to be the segmentation errors arising from poor separation between the carbide-dilution and matrix-reprecipitate boundaries. For practical considerations, the models were also compared in terms of predicted raw logits from the last layer before the post processing step for generating final mask. Much higher logit values were observed for all models on carbide particles and matrix pixels with the lowest logit values reported for the dilution band pixels. The top-performing UPerNet and SegFormer models had comparable logit values across all four classes. For industrial deployment, SegFormer can be preferred over UPerNet owing to its much smaller size; however, from pure performance considerations, UPerNet, with its higher capacity, could be more suited to handle changing data distributions.

In the future, the current comparison can be made more robust by extending through the following considerations:

- Within each selected model, more variations of key hyperparameters (e.g., batch size, learning rate, optimizer) can be considered
- The effect of specific components (e.g., multi-feature backbones) can be investigated in detail to evaluate their impact on the segmentation task. Moreover, different available variants (small, base, large, and extra large) of the existing encoders can be tested
- While the current study considered representative categories, the list can be expanded to include more transformer-based architectures (e.g., BEiT, Data2VecVision, HRFormer)
- The developed models can also be blind tested on Ni-WC images from other SEM setups to evaluate their generalizability in segmenting carbide damages
- To achieve absolute performance limits, more SEM data can be generated through new experiments or relevant augmentations. The enhanced dataset can improve the performance of selected models by providing robust model generalization

CRediT authorship contribution statement

Mutahar Safdar: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis. **Bashir Kazimi:** Writing – review & editing, Supervision. **Karina Ruzaeva:** Writing – review & editing, Supervision. **Gentry Wood:** Writing – review & editing, Resources, Data curation. **Max Zimmermann:** Writing – review & editing, Resources, Data curation. **Guy Lamouche:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Priti Wanjara:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Stefan Sandfeld:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yaoyao Fiona Zhao:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

M.S. acknowledges the Helmholtz Information & Data Science Academy (HIDA) for providing financial support enabling a short-term research stay at the Institute for Advanced Simulation – Materials Data Science and Informatics (IAS-9) of Forschungszentrum Jülich in Germany where a portion of the current work was conducted. M.S. also received financial support from the National Research Council Canada (Grant# NRC INT-015-1).

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.

Acknowledgements

The authors are grateful to Digital Research Alliance of Canada (RRG# 4294) for providing computational resources to support this research. We would like to acknowledge the Artificial Intelligence Enhancement of Process Sensing for Adaptive Laser Additive Manufacturing (AI-SLAM) consortium between Canada and Germany for their support that made this work possible. We are also thankful to Xavier Pelletier, Connor MacDowell, and Dr. Sheida Sarafan at the National Research Council of Canada for technical support with metallography and SEM.

Data availability

The annotated SEM dataset used in this work is open-source and has been publicly released at Zenodo: https://zenodo.org/records/17315241 The codebase is made available to support reproducibility of the investigation: https://github.com/Mutahar-Safdar/SEM_Segmentation_ViTs.git

References

[1] E. Pei, et al., Springer Handbook of Additive Manufacturing, Springer Nature, 2023.
[2] Additive Manufacturing — General Principles — Fundamentals and Vocabulary, ASTM, 2022.
[3] M. Zhang, C. Wang, G. Mi, C. Zhai, J. Li, Dispersion of reinforcing micro-particles in the laser welding of metal matrix composites: high-fidelity modeling with experimental characterization, Mater. Charact. 207 (2024) 113561.
[4] P.F. Mendez, et al., Welding processes for wear resistant overlays, J. Manuf. Process. 16 (1) (2014) 4–25.
[5] A. Pariyar, C.S. Perugu, K. Dash, S.V. Kailas, Microstructure and mechanical behavior of high toughness Al-based metal matrix composite reinforced with in-situ formed nickel aluminides, Mater. Charact. 171 (2021) 110776.
[6] M. Xu, et al., Ni-based superalloy fabricated by wire arc additive manufacturing exhibits excellent mechanical properties at 650° C by combining ultrasonic-assisted and solution treatment, Mater. Charact. 224 (2025) 115032.
[7] G. Wood, P. Mendez, Disaggregated metal and carbide catchment efficiencies in laser cladding of nickel-tungsten carbide, Weld. J. 94 (11) (2015) 343–350.
[8] L.E. Murr, A metallographic review of 3D printing/additive manufacturing of metal and alloy products and components, Metallogr. Microstruct. Anal. 7 (2018) 103–132.
[9] Z.-H. Qiu, N. Xu, Q.-N. Song, C. Zhong, Y.-F. Bao, Synergistic enhancement of mechanical properties and corrosion resistance of ultralight carbon nano-onion reinforced LA141 metal matrix composites fabricated by friction-stir processing, Mater. Charact. 223 (2025) 114940.
[10] M. Safdar, et al., Accelerated semantic segmentation of additively manufactured metal matrix composites: generating datasets, evaluating convolutional and transformer models, and developing the MicroSegQ+ tool, Expert Syst. Appl. 251 (2024) 123974.
[11] S. Scott, W.-Y. Chen, A. Heifetz, Multi-task learning of scanning electron microscopy and synthetic thermal tomography images for detection of defects in additively manufactured metals, Sensors 23 (20) (2023) 8462.

[12] X. Zhang, et al., Exceptional strength and ductility in 18Ni300/316 L heterogeneous bionic structures through laser additive manufacturing, Mater. Charact. 227 (2025) 115296.
[13] D. Rose, J. Forth, H. Henein, T. Wolfe, A.J. Qureshi, Automated semantic segmentation of NiCrBSi-WC optical microscopy images using convolutional neural networks, Comput. Mater. Sci. 210 (2022) 111391.
[14] N. O'Mahony, et al., Deep learning vs. traditional computer vision, in: Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC) Volume 1, Springer, 2020, pp. 128–144.
[15] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, A review of semantic segmentation using deep neural networks, Int. J. Multimed. Inf. Retr. 7 (2018) 87–93.
[16] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[17] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, D. Herath, Semantic segmentation using vision transformers: a survey, Eng. Appl. Artif. Intell. 126 (2023) 106669.
[18] J. Zhao, C. Shen, M. Huang, Y. Qi, Y. Chai, S. Zheng, Deep learning accelerated micrograph-based porosity defect quantification in additively manufactured steels for uncovering a generic process-defect-properties relation, Mater. Charact. 225 (2025) 115094.
[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[20] D. Hoiem, S.K. Divvala, J.H. Hays, Pascal VOC 2008 challenge, World Literature Today 24 (1) (2009).
[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[22] F. Sultana, A. Sufian, P. Dutta, Evolution of image segmentation using deep convolutional neural network: a survey, Knowl.-Based Syst. 201 (2020) 106062.
[23] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[24] S. Zheng, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[25] M. Safdar, G. Lamouche, P.P. Paul, G. Wood, Y.F. Zhao, Applications in data-driven additive manufacturing, in: Engineering of Additive Manufacturing Features for Data-Driven Solutions: Sources, Techniques, Pipelines, and Applications, Springer Nature Switzerland, Cham, 2023, pp. 45–121.
[26] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, Y.F. Zhao, A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management, J. Intell. Manuf. (2022) 1–36, https://doi.org/10.1007/s10845-022-02017-9.
[27] L. Scime, D. Siddel, S. Baird, V. Paquit, Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: a machine-agnostic algorithm for real-time pixel-wise semantic segmentation, Addit. Manuf. 36 (2020) 101453.
[28] A.-M. Schmitt, C. Sauer, D. Höfflin, A. Schiffler, Powder bed monitoring using semantic image segmentation to detect failures during 3D metal printing, Sensors 23 (9) (2023) 4183.
[29] Z. Jin, Z. Zhang, J. Ott, G.X. Gu, Precise localization and semantic segmentation detection of printing conditions in fused filament fabrication technologies using machine learning, Addit. Manuf. 37 (2021) 101696.
[30] J. Zhang, et al., Image segmentation for defect analysis in laser powder bed fusion: deep data mining of X-ray photography from recent literature, Integr. Mater. Manuf. Innov. 11 (3) (2022) 418–432.
[31] K. Wang, Contrastive learning-based semantic segmentation for in-situ stratified defect detection in additive manufacturing, J. Manuf. Syst. 68 (2023) 465–476.
[32] M. Safdar, G. Wood, M. Zimmermann, G. Lamouche, P. Wanjara, Y.F. Zhao, Detecting the extent of co-existing anomalies in additively manufactured metal matrix composites through explainable selection and fusion of multi-camera deep learning features, Virtual Phys. Prototyp. 20 (1) (2025) e2515240.
[33] J. Luengo, et al., A tutorial on the segmentation of metallographic images: taxonomy, new MetalDAM dataset, deep learning-based ensemble model, experimental analysis and challenges, Inform. Fusion 78 (2022) 232–253.
[34] M. Biswas, R. Pramanik, S. Sen, A. Sinitca, D. Kaplun, R. Sarkar, Microstructural segmentation using a union of attention guided U-net models with different color transformed images, Sci. Rep. 13 (1) (2023) 5737.
[35] Q. Liu, et al., PseudoClick: Interactive image segmentation with click imitation, in: European Conference on Computer Vision, Springer, 2022, pp. 728–745.
[36] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: fast and flexible image augmentations, Information 11 (2) (2020) 125.
[37] J. Minnema, M. van Eijnatten, W. Kouw, F. Diblen, A. Mendrik, J. Wolff, CT image segmentation of bone for medical additive manufacturing using a convolutional neural network, Comput. Biol. Med. 103 (2018) 130–139.
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[40] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016.
[41] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254, 2021.
[42] S. Mehta, M. Rastegari, Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, arXiv preprint arXiv:2110.02178, 2021.
[43] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 418–434.
[44] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188.
[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Proces. Syst. 34 (2021) 12077–12090.
[46] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272.
[47] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Proces. Syst. 34 (2021) 17864–17875.
[48] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1290–1299.
[49] Y. Yuan, et al., Hrformer: High-resolution transformer for dense prediction. arXiv 2021, arXiv preprint arXiv:2110.09408 vol. 19, 2021.
[50] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: A general framework for self-supervised learning in speech, vision and language, in: International Conference on Machine Learning, PMLR, 2022, pp. 1298–1312.
[51] M. Contributors, "MMCV: OpenMMLab Computer Vision Foundation," Ed, 2018.
[52] M. Contributors, MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020, 2023.
[53] K. Chen, et al., MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155, 2019.
[54] B. Kazimi, S. Sandfeld, Enhancing semantic segmentation in high-resolution TEM images: a comparative study of batch normalization and instance normalization, Microsc. Microanal. 31 (2024) ozae093.