Journal of Institute of Control, Robotics and Systems (2025) 31(5):566-572
https://doi.org/10.5302/J.ICROS.2025.25.0063 ISSN:1976-5622 eISSN:2233-4335

금속 SEM 이미지 분할을 위한 멀티 스케일 크로스-어텐션 및 SegFormer 기반 시멘틱 분할 모델 개발

Development for a Semantic Segmentation Model for Metal SEM Images Based on Multi-scale Cross-attention and SegFormer

변 영 훈, 윤 준 석, 이 상 아, 김 민 수, 원 홍 인, 윤 종 필*
(Younghun Byeon[1,2], Jun Seok Yun[2], Sanga Lee[2], Min Su Kim[2], Hong-In Won[2], and Jong Pil Yun[1,2,*])
[1]Department of Convergence Manufacturing System Engineering, UST, Yuseong-gu, Daejeon, Korea
[2]Manufacturing AI Research Center, Korea Institute of Industrial Technology, Yeonsu-gu, Incheon, Korea

Abstract: This paper proposes a semantic segmentation model based on cross-attention and SegFormer for segmenting metal grains in scanning electron microscope (SEM) images of stainless-steel specimens manufactured through additive manufacturing. Unlike benchmark datasets commonly used for developing image segmentation models, metal grains in SEM images exhibit minimal color variation within individual grains and have diverse shapes among grains belonging to the same class, making segmentation challenging. The segmentation model often struggles to distinguish grains because of variations in grain size. Therefore, we enhance the performance of the vision transformer-based semantic segmentation model SegFormer by incorporating a cross-attention module. Furthermore, the model is trained with the backbone network of SegFormer initialized with ImageNet-pretrained weights. We demonstrate the superiority of the proposed model through comparative experiments between convolutional neural network- and transformer-based semantic segmentation models.

Keywords: semantic segmentation, multi-scale cross-attention, SEM image, metal additive manufacturing, deep learning

I. 서론

최근 딥러닝의 발전으로 인해 간단한 이미지 분류 문제를 넘어 픽셀 단위 예측을 수행하는 모델이 연구되고 있다. FCN (Fully Convolutional Networks)는 최초로 CNN (Convolutional Neural Network)을 활용하여 픽셀 단위의 예측을 가능하게 했으며, U-Net은 의료 영상에서의 분할을 위해 개발되어 encoder-decoder 구조를 통해 높은 성능을 보였다. 영상 분할(image segmentation) 중 대표적인 시멘틱 분할(semantic segmentation) 방법 DeepLab [1] 및 후속 모델들은 atrous pyramid pooling과 같은 기법을 도입하여 객체의 멀티 스케일(multi-scale) 정보를 효과적으로 처리하였다. 또한, 패치로 분할된 이미지를 Transformer [2]에 입력하여 전체적인 문맥 정보를 학습할 수 있는 Vision Transformer (ViT) [3]와 CNN을 동시에 활용한 SegFormer [4] 모델은 ViT의 전역적 특성과 CNN의 지역적 특성을 결합하여 우수한 성능을 달성하였다. 그 외에도 ViT 내 self-attention을 변형하거나 CNN 대신 ViT를 decoder에 활용한 시멘틱 분할 연구도 진행되고 있다 [5,6,7]. 최근에는 기존 encoder-decoder 구조에서 벗어나 다양한 데이터 유형과 새로운 모델 구조를 통해 높은 성능을 달성한 연구가 진행되고 있다. OneFormer [8]는 시멘틱 분할 뿐만 아니라 instance segmentation, panoptic segmentation을 포함한 범용 영상 분할을 위해 텍스트 정보를 활용해 단일 모델로 3가지 분할을 수행할 수 있는 구조를 제안하고, [9]은 프롬프트 기반 영상 분할 모델인 SAM [10]을 시멘틱 분할에 활용하기 위해 CAM (Class Activation Map)으로 프롬프트를 추출하는 방법을 제안한다.

제조 산업에서는 생산성 향상과 품질 개선을 위해 이미지 데이터를 결함 검사 등에 활용하고 있다. 머신 러닝 및 딥러닝 기반 이미지 분류, 검출, 분할 방법이 성능 및 자동화 면에서 효과적이기 때문에 식품, 제약, 항공, 철도, 반도체 등의 다양한 분야에 적용되고 있다[11]. 그러나, 제조 분야의 특성상 데이터 데이터 전처리, 보안 등의 문제와 이후 응용 시 실시간성을 고려해야 한다. 이러한 문제 해결을 위해 컨베이어 벨트 결함 검출을 위해 CNN과 Transformer를 조합하여 실시간 시멘틱 분할을 달성한 연구가 있으며[12], 연합 학습을 통해 대량의 민감한 데이터를 사용하여 학습하여 충분한 성능을 달성한 사례가 있다[13].

* Corresponding Author
Manuscript received March 14, 2025; revised March 23, 2025; accepted April 14, 2025
변영훈: 과학기술연합대학원대학교 대학원생, 한국생산기술연구원 학생연구원(qoraksehfl@kitech.re.kr, ORCID 0000-0002-4923-8672)
윤준석: 한국생산기술연구원 연구원(yunjs@kitech.re.kr, ORCID 0000-0002-8319-6437)
이상아: 한국생산기술연구원 수석연구원(ivory@kitech.re.kr, ORCID 0000-0002-6567-0502)
김민수: 한국생산기술연구원 선임연구원(kms6777@kitech.re.kr, ORCID 0000-0001-5702-1732)
원홍인: 한국생산기술연구원 수석연구원(luvhayym@kitech.re.kr, ORCID 0000-0003-1609-8447)
윤종필: 과학기술연합대학원대학교 교수, 한국생산기술연구원 수석연구원(rebirth@kitech.re.kr, ORCID 0000-0002-2802-9978)
※ 본 연구는 한국생산기술연구원 기업수요기반생산기술실용화사업(KITECH-JA250003) 및 한국산업기술기획평가원 로봇산업핵심기술개발사업(RS-2024-00507383)에 의하여 연구되었음.

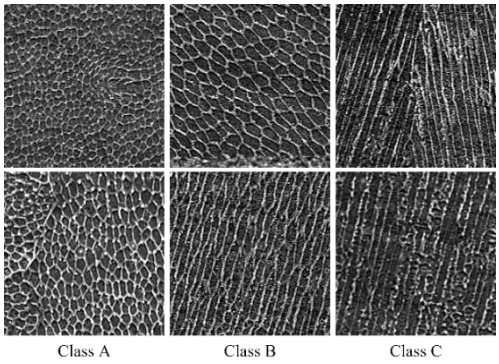

그림 1. 금속 SEM 이미지 데이터 예시.
Fig. 1. Examples for metal SEM image data.

금속 적층 제조는 3D 프린팅을 사용해 레이저 등으로 용융된 금속을 쌓아 금속 부품을 제조하는 분야다. 금속 부품을 제조할 때 사용된 금속 종류 및 공정 조건에 따라 물성이 달라지며 이를 측정하여 품질 향상에 활용한다. 물성 중 하나인 상분율은 금속 결정 구조에 따라 구분하는 상의 비율을 측정하는 것이다. 이를 통해 적층된 제품의 인장강도, 경도, 내구도 및 내열성과 같은 제품 품질을 측정할 수 있다. 상분율을 측정하기 위해 제품의 단면을 SEM (Scanning Electron Microscope)으로 촬영하고 각 상을 분류하는 과정이 필요하다. 이때 픽셀 단위 분류가 가능한 딥러닝 시멘틱 분할을 활용하여 금속 상의 비율을 측정할 수 있다. 이를 위해 해당 분야 적용성을 고려한 모델 사용성과 데이터 특성을 고려해야 한다. 최근 발표된 시멘틱 분할 연구들은 성능 향상을 위해 OneFormer [8]와 같이 텍스트 데이터 활용을 위한 네트워크를 추가하거나 프롬프트 사용이 가능한 SAM [10] 기반의 [9]과 같이 복잡한 구조의 모델을 개발하고 있다. 그러나, 제조 산업 현장에서 새로운 유형의 데이터를 취득하는 것이 어렵고 계산량이 많고 구조가 복잡한 모델을 사용하는 것 또한 어렵다. 한편, 금속 SEM 이미지 내 결정은 적층 시 사용한 원재료 및 공정 변수 등을 고려해 분류되는데, 크기, 방향, 분포에 따라 같은 종류임에도 다른 구조를 가질 수 있다. 이때 이미지 내 결정 크기, 방향, 분포는 이미지의 공간 해상도에 따라 달라지므로 멀티 스케일 정보를 학습할 수 있는 모델이 필요하다. 따라서 금속 SEM 이미지 내 금속 결정 분할을 위해 encoder-decoder로 구성된 단순한 구조의 시멘틱 분할 모델에 멀티 스케일 정보 학습 능력을 향상시킬 수 있는 방법이 필요하다.

이를 위해, 본 논문에서는 시멘틱 분할 모델 SegFormer에 멀티 스케일 크로스 어텐션(cross-attention)을 추가한 새로운 모델을 제안한다. SegFormer는CNN 및 ViT로 이루어진 단순한 encoder-decoder 구조로 효과적인 멀티 스케일 정보 추출이 가능하며 이를 활용해 픽셀 단위 클래스가 포함된 마스크를 예측할 수 있는 시멘틱 분할 모델이다. 제안하는 모델의 크로스 어텐션은 서로 다른 두 특징 사이의 유사도를 연산할 수 있는 기법으로 스케일 간 상관관계를 효과적으로 학습할 수 있다. 제안하는 방법을 검증하기 위해 자체 수집한 금속 SEM 이미지 데이터셋에서 기존 SegFormer 모델과 성능을 비교하고, 최적의 크로스 어텐션 구성을 알아내기 위해 수행한 실험도 제공한다.

II. BACKGROUND

1. 멀티 스케일 시멘틱 분할

시멘틱 분할(semantic segmentation)은 한 이미지에서 픽셀 단위 분류를 수행하는 문제이며, 이미지 내 영역의 스케일에 따른 상관관계 분석이 중요하다. 더 많은 네트워크 레이어를 거친 낮은 해상도의 특징 맵(feature map)의 경우 더 높은 수준의 의미론적 정보와 문맥 정보를 포함하고, 적은 네트워크 레이어를 거친 높은 해상도의 특징 맵은 이미지 내 형상과 위치 정보를 포함하고 있다[15]. 따라서 스케일마다 달라지는 여러 정보를 융합하는 것이 픽셀 단위 문제에서 중요하다. 이러한 사실로 인해 CNN 기반의 백본(backbone) 네트워크에서 공간 해상도를 달리하여 멀티 스케일 정보를 학습하는 방법이 활발히 연구되고 있다[1,15,16,17,18,21,22]. 또한, 최근 ViT의 발전에 따라 attention 메커니즘 및 ViT를 활용한 멀티 스케일 시멘틱 분할 연구도 활발히 진행되고 있다 [4, 5,6,14]. 최근 영상 분할 연구 중 하나인 [8]은 3가지 영상 분할인 semantic, instance, panoptic 분할을 단일 모델로 수행하기 위해 텍스트 정보를 각 분할에 맞게 사용하며 마스크 예측에 멀티 스케일 특징 맵을 활용한다. [9]은 픽셀 단위가 아닌 이미지 단위 라벨로 학습하는 weakly supervised learning을 사용해 시멘틱 분할을 수행하기 위해 프롬프트 기반 분할 모델인 SAM [10]에 CAM (Class Activation Map)을 접목한 방법을 제안하며 이미지에서 멀티 스케일 특징 맵을 추출하여 SAM으로 예측한 마스크와 같이 활용하여 시멘틱 분할을 달성한다. 위의 멀티 스케일 학습을 고려한 연구들은 모두 백본 네트워크로 멀티 스케일 특징 맵을 추출하여 마스크를 예측하는데 활용하는 방법에 초점을 두고 있다. 이는 convolution, attention 등을 통한 특징 추출 후 이들을 통합하여 마스크를 예측하는 과정으로 구성된다.

2. 금속 SEM 이미지 데이터

본 논문에서 사용하는 SEM 이미지는 스테인리스강의 적층 제조로 만들어진 시편의 단면을 촬영한 것으로 적층 시 금속 결정 형성 방향에 따라 3가지 상으로 분류된다. 그림 1의 데이터 예시에서 확인할 수 있듯이 결정 경계는 밝고 결정 내부는 어두운 색으로 구분된다. 또한, 동일한 상의 결정도 결정의 크기, 방향, 분포에 따른 패턴이 복잡하여 영역 분할이 어렵다. 그림 2에서 볼 수 있듯, 특징 맵의 스케일에 따라 포함하는 정보가 다르기 때문에 금속 결정 간 상관

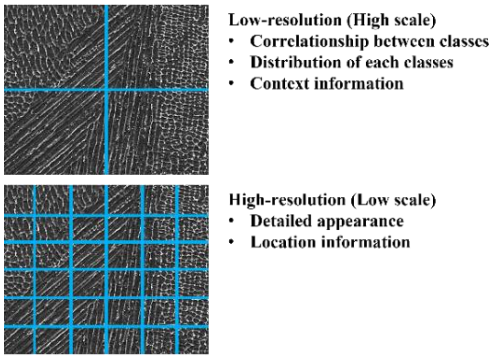

그림 2. 금속 SEM 이미지 내 multi-scale 특징 비교.
Fig. 2. Multi-scale features comparison for metal SEM image data.

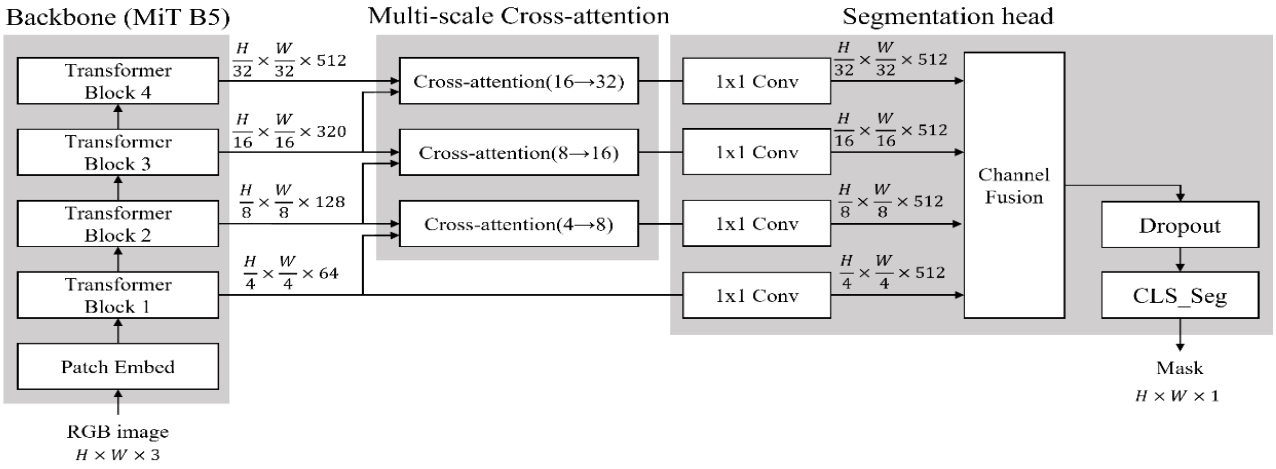

그림 3. 제안하는 모델 아키텍처 개요.

Fig. 3. Overview of an architecture of the proposed method.

관계 및 문맥 정보를 통해 경계면 분할에는 전역 정보가 포함된 저해상도, 세부 결정 클래스 분할에는 지역 정보가 포함된 고해상도 특징 맵 활용이 필요하다. 따라서 이미지 내 전역 정보와 지역 정보를 활용한 결정 크기 및 분포에 따라 달라지는 멀티 스케일 기반 시멘틱 분할이 요구된다. 멀티 스케일 학습 능력을 강화할 수 있는 모델을 개발하는 한편, 금속 적층 제조 분야에 적용하기 위해 능력 강화를 위한 복잡한 모델 혹은 새로운 데이터 유형을 추가하지 않는 단일 이미지 기반 모델을 개발하는 것이 중요하다.

III. 제안하는 모델

본 논문은 ViT 기반의 SegFormer [4] 모델에 멀티 스케일 크로스 어텐션 모듈을 추가한 방법을 제안한다. SegFormer 모델은 입력 RGB 이미지 $I \in \mathbb{R}^{H \times W \times 3}$가 주어질 때, 백본 네트워크인 MiT (Mix Transformer encoder)는 4개의 Transformer Block으로 구성되며 각 블록에는 효율적인 연산을 위한 Efficient Self-Attention과 MLP (Multi-Layer Perceptron)으로 구성된 Mix-FFN (Mix-Fead-Forward Network), 멀티 스케일을 부여하는 Overlapped Patch Merging으로 구성된다. MiT의 각 단계 별 레이어 수 및 채널 수에 따라 구분되며, 본 논문에서 사용한 B0는 (2, 2, 2, 2) 및 (32, 64, 160, 256), B5는 (3, 6, 40, 3), (64, 128, 320, 512)로 레이어수와 채널 수에 차이가 있다. 두 모델은 파라미터 수에서 약 24배, 벤치마크 데이터셋 성능에서 약 1.8배의 차이를 보인다.

백본 네트워크에서 출력한 특징 맵들은 공간 해상도를 각각 1/4, 1/8, 1/16, 1/32배 축소한 $F_4 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_4}$, $F_8 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_8}$, $F_{16} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_{16}}$, $F_{32} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_{32}}$ 으로 구성된다. 각 출력 특징 맵은 멀티 스케일 정보를 포함하고 있으며, 이후 segmentation head 입력 채널 수 C_{embed}에 맞춰 해상도를 재설정한 뒤 convolution을 통해 융합하여 마스크를 예측하는데 활용된다.

본 논문에서 제안하는 멀티 스케일 크로스 어텐션 모듈은 그림 2와 같이 4개의 특징 맵 중 인접한 스케일(1/4 및 1/8, 1/8 및 1/16, 1/16 및 1/32) 사이의 상관관계 학습하기 위해 총 3개의 크로스 어텐션을 사용한다. 크로스 어텐션의 연산 부하를 고려해 4가지 멀티 스케일 특징 맵의 모든 조합이 아닌 인접한 스케일의 특징 맵 사이에서만 수행한다. 크로스 어텐션으로 계산한 인접 스케일 별 결과 $F_{4\leftarrow 8}, F_{8\leftarrow 16}, F_{16\leftarrow 32}$는 백본 네트워크의 forward 순서대로 수행되며, 이를 통해 특징 맵이 추출되는 동안 발생하는 상관관계를 학습할 수 있다. 인접한 두 특징 맵의 해상도와 채널 수가 다르기 때문에 해상도가 더 작은 특징 맵을 더 큰 해상도에 맞춰 업샘플링(upsampling)하고 kernel 크기가 1인 convolution layer를 통해 채널 수를 조정한다. 조정된 두 특징 맵 사이의 크로스 어텐션은 다음과 같이 연산된다.

$$\text{Attention}(Q_i, K_j, V_j) = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j \qquad (1)$$

표 1. SEM 이미지 시멘틱 분할 모델 성능 비교.

Table 1. Comparison on semantic segmentation models for SEM images.

	mIoU	Class ratio[%p]
Mask2Former [5]	0.328	16.07
U-Net (MobileViT [20])	0.411	16.87
SegFormer B0 [4]	0.430	14.37
SegFormer B5 [4]	0.451	10.39
Proposed method	**0.478**	**9.97**

표 2. Cross-attention 방향에 따른 성능 비교.

Table 2. Comparison based on cross-attention direction.

	mIoU	Class ratio[%p]
Backward	0.476	10.03
Forward (Proposed)	**0.478**	**9.97**

표 3. Cross-attention 결합 방법에 따른 성능 비교.

Table 3. Comparison based on integration methods for cross-attention.

	mIoU	Class ratio[%p]
Addition	0.412	10.28
Concatenation (Proposed)	**0.478**	**9.97**

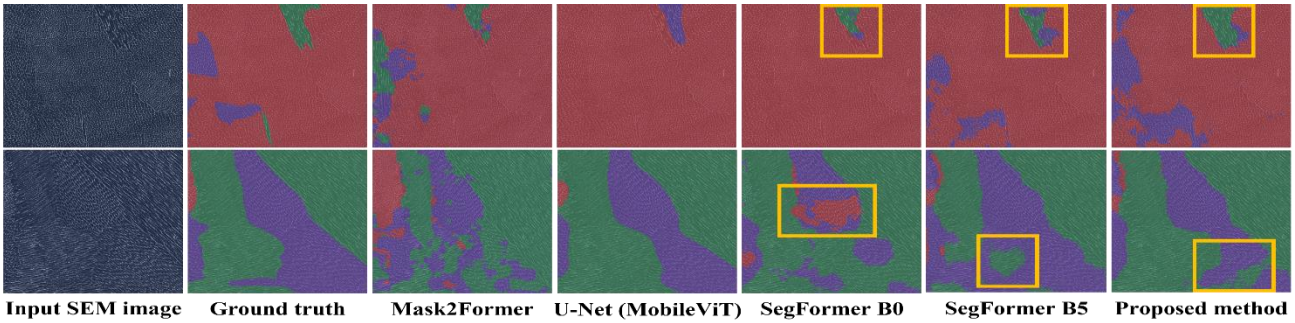

그림 4. 제안하는 모델 및 Mask2Former, U-Net (MobileViT), SegFormer B0, SegFormer B5 정성적 성능 비교 (클래스 A, B, C는 각각 빨강, 파랑, 초록으로 표시).

Fig. 4. Qualitative comparison on the proposed method, Mask2Former, U-Net (MobileViT), SegFormer B0, and SegFormer B5 (Class A, B, and C are described in red, blue, and green, respectively).

여기서 $Q_i = W_Q F_i, K_i = W_K F_i, V_i = W_V F_i$는 각각 attention 계산을 위한 query, key, value이고 W_Q, W_K, W_V는 학습 가능한 가중치이며, d_k는 key 벡터의 차원 수이다. 그림 1에서 볼 수 있듯, 연산된 각 특징 맵은 segmentation head에 입력되어 C_{embed} 채널로 kernel 크기가 1인 convolution layer를 이용해 조정된 다음, 채널 차원으로 concatenate하여 융합된다. 융합된 특징 맵은 segmentation layer에 의해 픽셀 단위 클래스를 예측하는데 활용된다.

IV. 실험

1. 데이터셋 및 학습 하이퍼파라미터

본 논문은 스테인리스강의 적층 제조로 만들어진 금속 시편의 절단면을 촬영한 SEM 이미지와 결정에 따라 분류된 픽셀 단위 ground-truth 마스크로 구성된다. 학습 데이터 및 평가 데이터는 각각 147, 96장으로 구성된다. 원본 이미지 해상도 1280x960에서 640x640으로 resize한 다음 480x480 크기의 무작위 crop을 생성하여 학습한다. 그 외, RandomGaussianBlur를 통해 데이터 증강을 수행하며 ImageNet [19] 학습 시 사용되는 정규화를 적용한다. 이후 정성적 평가에서 클래스 A는 빨강, B는 파랑, C는 초록으로 표시한다.

제안하는 방법 및 기존 시멘틱 분할 모델 학습 및 평가는 동일한 조건에서 수행한다. 학습은 mini-batch 크기로 150 epoch 동안 진행하며, momentum 0.9로 설정한 AdamW optimizer를 학습률 0.0005로 지정하여 Polynomial Learning Rate Scheduler를 이용해 학습률을 아래 식 2와 같이 감소시켜 학습 안정성을 높인다.

$$LR_t = LR_0 \times \left(1 - \frac{t}{T}\right)^p \tag{2}$$

식 2의 감소율 p는 0.0005이고 T는 총 학습 단계 수, t는 현재 학습 단계, LR_0은 초기 학습률, LR_t는 현재 단계 t에서의 학습률이다. 손실 함수는 일반적인 cross-entropy 기반 함수를 사용하며 식은 아래와 같다.

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3}$$

식 3에서 C는 클래스 개수, y_i는 ground-truth 마스크, $\hat{y}_i$은 모델 예측 확률 값이 포함된 마스크이다. 모델 성능 평가를 위해 mIoU (mean Intersection-Over-Union)과 클래스 비율 오차 Class ratio를 사용한다. mIoU는 분할한 마스크의 겹침 비율을 계산하여 분할 정확도를 평가하는 방법이고, Class ratio는 한 이미지에서 클래스 별 비율을 계산하여 클래스 분포를 평가하는 방법이다. 특히, Class ratio는 금속 적층 제조 산업에서 SEM 이미지로 측정하는 물성 중 하나인 상분율과 유사한 방식으로 계산되기 때문에 Class ratio의 수치가 높을수록 상분율 예측 능력이 우수함을 검증할 수 있다. mIoU는 클래스별 confusion matrix를 계산한 뒤 예측한 영역과 ground-truth 영역을 합친 영역 대비 겹친 영역의 비율을 계산한다. Class ratio는 모든 평가 데이터에 예측한 마스크와 ground-truth 마스크의 픽셀 별 클래스 비율 오차의 평균 값을 계산한다.

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{4}$$

$$\text{Class ratio} = \frac{1}{N}\sum_{c=1}^{N} e_c, \tag{5}$$

$$e_c = \left| \frac{1}{N_p}\sum_{i=1}^{N_p} p_i - \frac{1}{N_p}\sum_{j=1}^{N_p} g_j \right| \times 100 \tag{6}$$

여기서 TP_i는 클래스 i에 대해 올바르게 예측한 픽셀 수, FP_i는 클래스 i가 아닌 픽셀을 잘못 예측한 픽셀 수, FN_i은 클래스 i를 예측하지 못한 픽셀 수, N은 클래스 개수, p_i, g_j는 각각 예측한 마스크와 ground-truth 마스크의 픽셀 값, N_p는 마스크의 총 픽셀 수, e_c는 클래스 c의 예측 및 ground-truth 마스크 간 비율 오차를 의미한다.

제안하는 모델에서 사용한 SegFormer는 B5 모델을 사용하며 이는 B0부터 B5의 모델 중 가장 크고 백본 네트워크인 MiT B5는 총 4단계로 구성되며 각 단계의 출력은 각각 64, 128, 320, 512의 은닉층을 가진다. 위 출력들은 각각 크로스 어텐션 연산에 사용된 특징 맵 F_4, F_8, F_{16}, F_{32}이다. 그리고 segmentation head의 입력 채널 수 C_{embed}는 512로 설정한다.

2. 실험 결과

본 논문에서 제안하는 방법의 우수성을 검증하기 위해 SegFormer 외에 Mask2Former와 MobileViT [20]를 백본 네트워크로 사용한 U-Net (이하 U-Net (MobileViT))을 동일한 데이터로 학습하여 표 1에서 모델 간 성능을 비교한다. 제안하는 모델은 모델 크기가 작은 B0보다 0.048 더 높은 mIoU를,

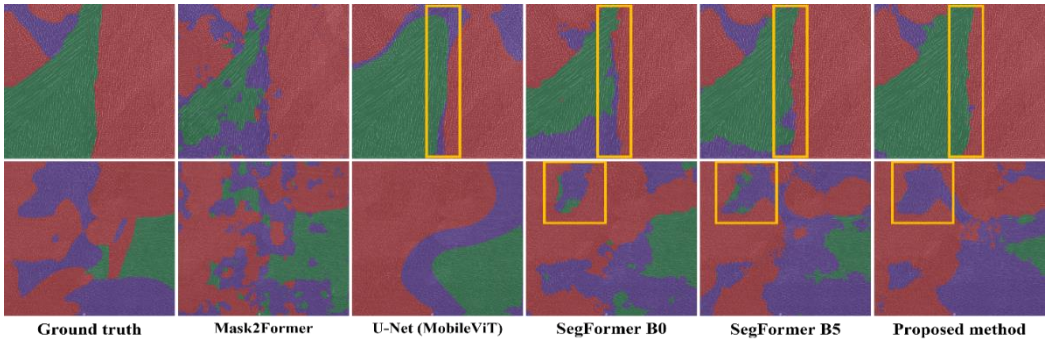

그림 5. 클래스 경계면 구분 성능 비교

Fig. 5. Comparison on class discrimination in boundaries

B5보다 0.027 더 높은 수치를 달성하였다. 또한 Mask2 Former 및 U-Net (MobileViT)와 비교해 각각 0.150과 0.067의 mIoU 향상을 확인하였다. 시멘틱 분할 모델 성능 평가에 주로 활용되는 mIoU 비교를 통해 기존 방법 대비 제안하는 방법의 우수함을 검증할 수 있고, 본 논문에서 사용하는 데이터에 멀티 스케일 크로스 어텐션을 적용하는 것이 ViT 기반 모델의 성능을 향상시킬 수 있음을 알 수 있다. 또한, Class ratio 성능 향상을 통해 멀티 스케일 학습으로 금속 결정 간 분리 능력이 강화되었음을 예측할 수 있고, 이를 통해 제안하는 방법이 향상된 멀티 스케일 구분 능력으로 금속 SEM 이미지에서 상분율을 측정하기에 적합함을 알 수 있다.

표 2는 크로스 어텐션 계산 시 특징 맵 방향에 따른 성능을 비교한다. 표에서 Backward는 백본 네트워크의 계산 방향과 반대, 즉 $F_{8\leftarrow4}, F_{16\leftarrow8}, F_{32\leftarrow16}$의 특징들을 계산하는 것을 의미한다. 두 방향은 0.002 mIoU의 차이를 보여 성능 차이가 거의 없으나, backward의 경우 모델 학습 시 불안정한 loss 값 변화를 보이며 오히려 상승하는 경우도 있어 forward를 제안하는 모델에 적용하였다. 표 3은 크로스 어텐션 출력 특징 맵과 기존 멀티 스케일 특징 맵 사이의 결합 방법에 따른 성능을 비교한다. 크로스 어텐션 결과로 나온 특징 맵을 결합할 때 요소별 덧셈(element-wise addition) 연산과 채널별 concatenation(channel-wise concatenation) 연산을 적용해 실험했다. 실험 결과 concatenation 연산을 적용했을 때 0.066 만큼 더 높은 mIoU를 보여 단순한 합이 아닌 concatenation 후 가중합을 통한 방식이 더 효과적임을 증명한다.

한편, 제안하는 모델의 정성적 성능을 비교하는 그림 4의 주황색 박스에서, SegFormer B0 및 SegFormer B5와 비교하여 기존 모델에서 잘못 예측한 부분을 올바르게 예측하고 ground-truth의 경향을 비교적 반영하는 모습을 보인다. 또한, 정확한 클래스 예측이 어려운 Mask2Former 및 U-Net (MobileViT)와 비교해 SegFormer와 제안하는 방법의 성능이 더 좋음을 알 수 있다. 그러나 SegFormer에 사용된 Transformer의 패치 분할로 인해 클래스 경계 부분이 뚜렷하지 않은 부분이 존재하지만, 기존 모델에 비해 구분 성능이 조금 더 향상된 모습을 보여 제안하는 멀티 스케일 크로스 어텐션의 효과를 확인할 수 있다. 그림 5는 모델 별 클래스 경계면 구분 능력을 검증하기 위한 정성적 비교 결과를 제시한다. 주황색 박스로 표시된 부분에서 볼 수 있듯, 기존 SegFormer B0 및 B5의 경우 클래스 C와 A의 경계에서 클래스 B로 잘못 예측한 부분이 발생하는 경우가 있지만, 제안하는 모델의 경우 이러한 클래스 혼동이 적음을 알 수 있다. 게다가, 기존 모델과 비교해 제안하는 모델이 경계면을 비교적 일정하게 구분하는 것을 보여 멀티 스케일 크로스 어텐션의 효과를 입증할 수 있다. 또한, ViT 기반 모델인 Mask2Former와 U-Net (MobileViT)을 비교할 때, 정확한 클래스 경계면 예측이 어려운 두 모델과 달리 제안하는 방법의 경우 보다 정확한 경계면을 예측하여 ViT 기반 모델의 단점을 제안하는 방법으로 완화할 수 있음을 알 수 있다.

V. 결론

본 논문은 ViT 및 CNN을 활용한 시멘틱 분할 모델인 SegFormer에 멀티 스케일 크로스 어텐션을 도입하여 금속 SEM 이미지 분할 문제를 해결하였다. 제안하는 모델은 ViT 기반 백본 네트워크에서 출력된 멀티 스케일 특징 맵 사이의 계층적 관계를 학습할 수 있는 크로스 어텐션을 활용해 스케일 간 상관관계 학습 능력을 강화하였다. 제안하는 모델의 성능을 자체 수집한 금속 SEM 이미지 데이터셋을 활용해 학습하고 기존 모델인 SegFormer와 비교하여 성능 향상을 검증하였다. 향후 연구에서 ViT 내 패치 분할로 인한 불안정한 경계면 구분 능력을 강화하고 새로운 데이터 전처리 기법을 도입하여 학습 안정성을 높일 것이다.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Proc. of the European Conference on Computer Vision (ECCV),* pp. 801-818, 2018.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic

segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077-12090, 2021.
[5] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290-1299, 2022.
[6] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7262-7272, 2021.
[7] T. Li, Z. Cui, and H. Zhang, "Semantic segmentation feature fusion network based on transformer," *Scientific Reports*, vol. 15, p. 6110, 2025.
[8] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One transformer to rule universal image segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2989-2998, 2023.
[9] H. Kweon and K.-J. Yoon, "From SAM to CAMs: Exploring segment anything model for weakly supervised semantic segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19499-19509, 2024.
[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015-4026, 2023.
[11] L. Zhou, L. Zhang, and N. Konz, "Computer vision techniques in manufacturing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no.1, pp. 105-117, 2022.
[12] Y. Shen and X. Ma, "Real-time semantic segmentation model combining CNN and transformer to detect belt runout," *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, pp. 451-456, 2024.
[13] M. Mehta and C. Shao, "Federated learning-based semantic segmentation for pixel-wise defect detection in additive manufacturing," *Journal of Manufacturing Systems*, vol. 64, pp. 197-210, 2022.
[14] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.
[15] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 603-619, 2018.
[16] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3562-3572, 2019.
[17] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, pp. 555–570, 2019.
[18] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 2612-2626, 2019.
[19] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.
[20] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," arXiv:2110.02178, 2021.
[21] S.-H. Kim, "Landing pad recognition technology based on semantic image segmentation using virtual reality dataset," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 2, pp. 68-73, 2024.
[22] Y.-G. Jung and T.-H. Park, "Image segmentation system of soft capsules based on deep learning", *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 6, pp. 607-613, 2024.

변 영 훈
2021년 한국해양대학교 제어자동화공학부(공학사). 2023년 한국해양대학교 제어계측공학과 석사(공학석사). 2023년~현재 UST KITECH 스쿨 융합제조시스템공학과 박사과정 재학중. 관심분야는 영상분할, 결함탐지.

윤 준 석
2023년 전남대학교 인공지능융합학과 석사(공학석사). 2023년~현재 한국생산기술연구원 제조AI연구센터 연구원. 관심 분야는 이미지 초해상화, 영상분할, 결함탐지, 이상탐지.

이 상 아
2010년 서울대학교 기계항공공학부 공학사. 2012년 동 대학원 공학석사. 2017년 동 대학원 공학박사. 2019년~ 현재 한국생산기술연구원 수석연구원(보). 관심분야는 AI-Simulation.

김 민 수
2014, 2016, 2022년 포항공과대학 전자공학과 공학사. 2016년 동 대학원 공학석사. 2017년 동 대학원 공학박사. 2024년~현재 한국생산기술연구원 선임연구원. 관심분야는 데이터 기반 결함 분석, 결함 탐지, 신호 처리, 딥러닝 신경망.

원 홍 인
2017년 한양대학교 기계설계공학과 박사. 2017~현재 한국생산기술연구원 수석연구원. 관심 분야는 인공지능, 데이터 마이닝, 디지털 트윈, 제조 시스템 자동제어.

윤 종 필

2003년 경북대학교 전자전기공학부 졸업. 2009년 포항공과대학교 전자전기공학과 박사. 2009~2016년 포스코 기술연구원 책임연구원. 2016~현재 한국생산기술연구원 수석연구원. 관심분야는 인공지능, 머신비전, 결함검사, 설비진단, 의료데이터분석.