Journal of Institute of Control, Robotics and Systems (2025) 31(5):517-525
https://doi.org/10.5302/J.ICROS.2025.25.0020 ISSN:1976-5622 eISSN:2233-4335

SSL4STR: Self-Supervised Learning 기반의 MAE 를 활용한 한글 장면 문자 인식 성능 향상

SSL4STR: Enhancing Korean Scene Text Recognition Performance Using Self-supervised Learning With Masked Auto-encoder

유 준 혁*, 송 현 석
(Joonhyuk Yoo[1,*] and Hyun-Seok Song[1])
[1]College of Information and Communication Engineering, Daegu University

Abstract: This paper proposes a novel approach to enhance Korean scene text recognition (STR) performance using self-supervised learning (SSL) with masked auto-encoder (MAE), called by SSL4STR. This research integrates MAE with state-of-the-art deep learning methods, such as vision transformer (ViT), MGPSTR and DBNet++, to create the integrated Korean Scene Text Recognition System. The SSL4STR leverages large-scale unlabeled datasets for pre-training and combines synthetic and real-world labeled data for fine-tuning. The proposed model demonstrates robust performance under challenging conditions, including degraded and occluded images, achieving significant improvements over existing models. The applicability of SSL4STR is validated in real-world scenarios such as coastal surveillance and vessel identification, highlighting its potential in various Korean language applications.

Keywords: self-supervised learning, masked auto-encoder, Korean scene text recognition, vision transformer

I. 서론

최근 해양 환경에서 자동화된 해안 감시 시스템에 대한 필요성이 증가하고 있다. 우리나라 영해에서는 그림 1과 같이 외국 및 우리 어선의 불법 조업, 북한 어선이나 목선의 NLL (북방한계선) 침범 같은 사건들이 빈번히 발생하며, 어민의 안전과 국가 안보를 위협하고 있다. 이러한 문제를 예방하고 적시에 대응하기 위해서는 해상에서 운항 중인 선박을 특정하고 추적할 수 있는 시스템이 필수적이다. 특히, 카메라 영상을 통해 선박을 정확히 식별할 수 있는 기술은 불법 선박의 식별, 귀순 의도 선박의 판별, 해양 사고 발생 시 해경이나 육군의 신속한 대처에 중요한 역할을 할 것이다 [11,28].

이러한 지능형 해안 감시 시스템을 구축하기 위해서 vessel re-identification [38] 등의 다양한 연구가 이루어지고 있고, 본 논문은 그 중에서 선박명을 식별할 수 있는 장면 문자 인식 (scene text recognition, STR) 기술을 다루고 있다. 딥러닝 기술의 발전으로 인해 STR은 다양한 응용 분야에서 중요한 역할을 차지하고 있다. 대부분의 STR 모델들[9, 15, 16, 20, 35-42]은 주로 영어와 같은 단순한 알파벳 문자 구조에 기반을 두고 설계되었다. 하지만 한국어는 초성, 중성, 종성의 조합으로 이루어진 복잡한 음운 구조를 가지며, 이론적으로 11,172가지 이상의 글자 형태가 가능하여 기존 STR 모델의 한계를 드러내고 있다. 특히, 열악한 해상 환경에서의 선박명 인식과 같은 실세계 응용 분야에서는 이런 문제들이 더욱 두드러진다.

그림 1. 2023년 대한민국 영해 불법 활동 건수와 해안 감시 환경에서의 선박명 인식.

Fig. 1. Number of illegal activities in South Korean territorial sea in 2023 and identification of vessel name in coastal surveillance.

기존 STR 연구는 주로 대규모 합성 데이터셋(MJSynth [1], SynthText [2])과 영어 기반 벤치마크(IIIT [3], IC13 [4], SVT [5], IC15 [6], SVTP [7], CUTE [8])를 활용하여 모델 성능을 평가하는데 초점을 맞추고 있다. 이러한 평가 방식을 통해 상기한 6개의 벤치마크에서 성능은 더 이상 크게 개선되지 않고 있으며, SOTA (State-of-the-Art) 성능 향상 속도가 둔화되고 있다. 이러한 접근법은 현실 세계에서 발생하는 다양한 문자 환경을 충분히 반영하지 못하며, 특히 비-영어권 국가에서 한국어와 같은 문자에 대한 적용 가능성에도 한계를 보인다. 따

* Corresponding Author
Manuscript received February 3, 2025; revised March 28, 2025; accepted April 7, 2025
유준혁: 대구대학교 컴퓨터정보공학부 교수(joonhyuk@daegu.ac.kr, ORCID 0000-0002-8311-5342)
송현석: 대구대학교 AI학과 학부생(21827863@daegu.ac.kr, ORCID 0009-0008-8270-9094)
※ 본 연구는 국방부 재원으로 정보통신기획평가원 지원을 받아 수행된 "이동형 모바일 환경 인공지능을 활용한 경계 감시 시스템 기술개발(A)" 연구 결과 중 일부입니다.
※ 본 연구에 사용된 데이터는 ㈜엠지티의 지원을 받아 수집되었습니다.

라서 본 논문에서는 기존 STR 모델의 일반화 성능과 한국어 특화 성능을 개선하기 위한 새로운 접근법을 제시한다.

Wang et al [9]은 기존 벤치마크를 활용한 모델 성능 평가의 제한성을 강조하며, 대규모 데이터셋 Union14M을 제안하여 현실 세계 STR 문제를 해결하고자 했다. 이 연구는 학습 데이터셋의 다양성과 품질 향상이 STR모델의 일반화 성능 개선에 중요한 역할을 한다는 점을 입증하였지만, 한국어를 포함한 영어 이외의 외국어에 대한 구체적인 해결책을 제시하지는 못했다. 한편, MAE (Masked Auto-Encoder) [10]를 활용한 SSL (Self-Supervised Learning) 학습 기법은 대규모 Unlabeled 데이터에서 유의미한 특징 표현을 학습할 수 있는 잠재력을 보여주고 있으며, 특히 열화와 폐색된 이미지에서도 강인한 성능을 보이고 있다. 그러나 이러한 SSL 기술이 실세계 환경에서의 한글 장면 문자 인식에 직접 적용된 사례는 아직까지 제시된 바 없다.

현재의 해양 감시 시스템은 해안 감시 레이더에서 해안 경계지 내의 모든 선박을 탐색하는데, AIS (Automatic Identification System) 또는 V-Pass (vessel Pass) 등의 선박 통신 장치가 있는 경우에는 선박 식별이 가능하다. 그러나, 식별되지 않는 선박의 경우에는 병사들이 직접 CCTV를 활용하여 24시간 연속으로 모니터링하면서 선박을 수동으로 식별해야 하기 때문에 이를 운용하기 위한 병력도 많이 필요하고 그 과정에서 사람의 실수도 존재할 가능성이 있다. 이를 개선하기 위해 본 연구에서는 AI 기반의 자동화된 해안 감시 시스템 구축의 핵심 요소로서, 해상에서 운항 중인 선박에 인쇄된 한글 선박명 검출과 인식의 성능을 향상시키고 열화와 폐색 환경에서도 문자열 인식의 강인함을 보장하기 위해, Masked Auto-Encoder 기반의 Self-Supervised Learning 기법을 활용한 새로운 STR 모델인 SSL4STR (Self-Supervised Learning for Scene Text Recognition) 방법을 제안한다. 본 연구에서 개발 중인 SSL4STR 시스템은 해안 경계 부대의 실질적 운용을 목적으로 설계되었다. 특히 실제 운용 단계에서는 선박을 자동으로 검출하고 문자 인식 결과를 운영자에게 제공하여 신속한 판단을 지원하는 시스템으로 활용된다.

본 논문은 다음과 같이 구성된다. II장에서는 기존 STR 모델과 MAE 및 SSL 등의 관련 연구를 소개하고, III장에서는 본 논문에서 제안하는 SSL4STR 방법을 제시한다. IV장에서는 실제 해안 감시 도메인에서 제안된 모델에 대한 실험과 기존 STR 모델과의 비교 평가 및 분석을 진행한다. 마지막으로, V장에서는 결론 및 향후 연구 방향을 기술한다.

II. 관련 연구

1. STR 모델의 개요

CRNN [12]과 같은 초기의 STR (Scene Text Recognition) 모델들은 CNN (Convolutional Neural Network)와 RNN (Recurrent Neural Network)를 결합하여 이미지 특징을 추출하고 시퀀스 형태로 문자를 인식하였다. 이러한 모델들은 영어와 같이 단순한 알파벳 구조의 문자 인식에는 효과적이었지만, 한국 산업 규격으로 지정된 한국어 문자 집합 2,350자(KS X 1001 [24])와 같이 복잡하고 다양한 자모 조합을 가진 한글 인식에는 한계가 있었다. 그 이후에 ASTER [13], SAR [14]과 같이 Attention 메커니즘을 도입하여 문맥 정보를 활용한 문자 인식 성능을 향상시키는 방법들이 등장하였다. 하지만 이러한 모델들도 한글의 복잡한 음운 구조와 복합 글자 형태를 충분히 반영하지 못하여 영어에 비해 한글의 인식률은 여전히 낮은 편이다.

2. 기존 한글 STR 연구

한글은 초성, 중성, 종성의 조합으로 이루어져 이론적으로 11,172가지 이상의 글자 형태가 존재하며, 한 글자 내에 여러 개의 자모가 결합되어 있어 단순한 CNN 필터로는 세부 특징을 추출하기 어렵다. 또한, 기존의 딥러닝 기반 STR 모델은 대부분 영어 데이터로 학습되어 있어 한글과 문자 구조가 복잡한 문자에 대한 일반화 능력이 낮다. 이를 향상시키기 위해서는 한글 데이터를 학습하는 것이 효과적이지만, 영어에 비해 한글로 구성된 대규모의 Labeled 공공 이미지 데이터셋이 부족하여 한글 STR 모델의 학습에 한계가 있다. 최근 발표된 한국어 STR 모델이 ICDAR2017 한국어 벤치마크 20,000장에 대해 테스트한 결과 종단간 인식 성능이 50% 미만일 정도로 매우 낮은 정확도를 보여 실세계에 적용이 불가능한 상황이다[31-32].

3. Self-Supervised Learning 개요

SSL (Self-Supervised Learning)은 레이블링된 데이터가 부족한 상황에서도 유용한 특징을 학습할 수 있는 접근법으로, 특히 대규모 비지도 데이터셋을 활용하는 데 강점을 가지고 있다. ViT (Vision Transformer) [17] 기반의 딥러닝 모델은 사전 정의된 패턴을 탐지하는 CNN 기반의 딥러닝 모델에 비해 데이터 간의 문맥적 관계를 학습하는 특성 때문에 모델이 높은 성능을 내기 위해서는 대규모의 데이터 학습이 선행되어야 한다. 하지만 현실적으로 대규모의 Labeled 데이터셋을 확보하는 것은 비용과 시간이 많이 소요되는 작업이다. 특히 의료 영상이나 위성 사진과 같은 전문적인 데이터 도메인에서는 레이블링 작업에 전문가의 참여가 필요하기 때문에 그 어려움이 가중된다. SSL은 레이블링된 데이터가 부족한 상황에서도 데이터의 구조적 특성과 문맥적 관계를 학습함으로써 유의미한 Representation을 학습할 수 있으며, 다양한 downstream 작업을 가능하게 한다. 결과적으로 이러한 학습 방식은 ViT 모델이 데이터 효율성을 극대화하고, 대규모의 Labeled 데이터 부족 문제를 극복하며, 더 넓은 범위의 응용 사례에 적용할 수 있는 기반을 제공한다.

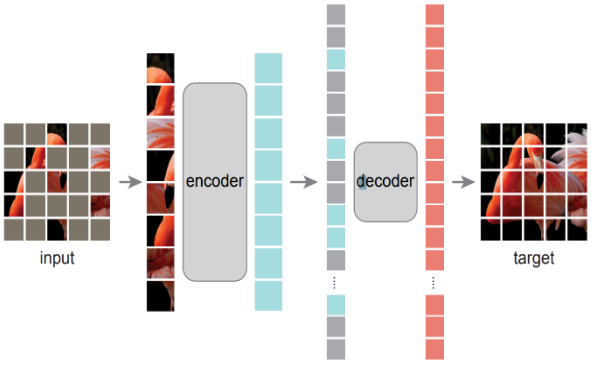


그림 2. Masked auto-encoder의 구조[10].
Fig. 2. Architecture of masked auto-encoder [10].

4. Masked Auto-Encoder 개요

MAE (Masked Auto-Encoder) [10]는 그림 2와 같이 ViT (Vision Transformer) 기반의 인코더-디코더 구조를 가지고 있는 생성형 AI 모델이다. MAE는 이미지의 일부를 마스킹한 후에 이를 복원하는 과제를 통해 시각적 특징을 학습하는데, 이미지에 마스킹 처리된 부분을 Auto-Encoder를 통해 생성하여 원본과 유사한 이미지를 생성하는 기술이다. MAE가 보유한 이러한 특성은 특히 열화와 폐색된 이미지에서의 시각적 특징 복원에 유용하며, 이는 장면 문자 인식 분야에서도 잠재적인 응용 가능성을 시사한다.

이러한 MAE를 장면 문자 인식에 적용하려는 연구들이 본 논문과 비슷한 동기로 최근 등장한 바 있다[15,16]. 이들은 MAE를 활용하여 열악한 환경에서 촬영된 이미지의 품질을 향상시키고, 복원된 이미지를 기반으로 문자 인식 성능을 개선하는 방법을 탐구하고 있다. 그러나 MAE를 직접적으로 한글 STR에 적용한 사례는 아직 수행된 바 없다. 한글의 복잡한 자모 구조와 다양한 글자 형태를 고려할 때, STR 분야에서 MAE의 적용은 장면 이미지에서 한글 인식 성능 향상에 큰 잠재력이 있다. 본 논문에서는 해상 환경에서 열화와 폐색된 이미지에서도 높은 한글 인식률을 달성하기 위해 MAE를 한글 선박명 인식에 적용하였다.

III. MAE를 활용한 SSL4STR 시스템

본 논문에서는 MAE [10]의 인코더(Encoder) 특성을 활용하여 ViT 기반의 한글 장면 문자 인식 모델인 KR-MGPSTR [11]의 인코더 모듈을 MAE의 인코더 모듈로 대체하였다. 다시 말하면, 제안된 SSL4STR (Self-Supervised Learning for Scene Text Recognition) 시스템은 MAE로 학습된 인코더를 한글 STR 모델인 KR-MGPSTR 모델의 인코더에 통합하여 열화와 폐색된 한글 문자 이미지에서도 높은 인식 성능을 달성하였다.

기존의 MGPSTR [23] 모델은 STR 분야에서 SOTA를 달성하여 높은 정확도를 보여주는 ViT 기반의 장면 문자 인식 모델이다. 이 모델은 ViT 인코더의 최종 출력 벡터를 특정 문자에 해당하는 하나의 출력 토큰으로 통합하기 위해 A3 (Adaptive Addressing and Aggregation) 모듈을 제안하였다. 각 문자 영역에서 필요한 토큰에 대한 가중치를 주어 일종의 어텐션(attention) 역할을 하는 A3 마스크를 만들어서 다중 입도 정확도(multi-granularity precision)를 향상시켰다.

본 연구에서는 한국어 STR의 성능을 향상시키기 위해 기존의 영어 인식 중심의 MGPSTR 모델을 한글 인식에 특화한 버전인 KR-MGPSTR [11] 모델이 우선 설계되었다. 이를 위해, 기존 MGPSTR 모델에서는 WordPiece, BPE, Character 토크나이저를 모두 사용하였으나, 영어와 달리 한글 적용 시 WordPiece 및 BPE 토크나이저를 사용할 경우 연산량이 증가하고 외부 모델(GPT-2, KR-BERT 기반) 의존성 증가와 비효율적인 fusion 알고리즘 등의 문제가 발생하였다. 실험을 통해 이러한 복합적 토크나이저 사용으로 얻는 정확도 향상이 약 0.6%로 미미함을 확인하여 본 연구에서는 그림 3에 도시한 바와 같이 Character 토크나이저만을 활용하여 한글 데이터를 추가 학습하였다. 출력 클래스 개수는 한국 산업 규격(KS X 1001 [24])에 따라 지정된 한글 문자 집합 2,350개로

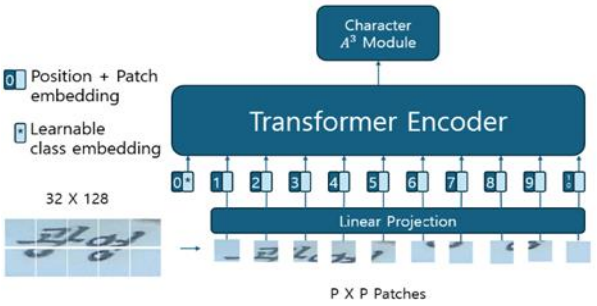


그림 3. ViT 기반의 KR-MGPSTR의 구조[11].
Fig. 3. Architecture of ViT-based KR-MGPSTR [11].

구성하여 한글의 모든 완성형 글자를 포함시켰다. KR-MGPSTR 모델은 Character A3 모듈을 통해 입력 이미지의 특징 맵에서 각 한글 문자의 위치와 형태를 정확히 파악하고, 해당 문자에 대한 특징 벡터를 생성한다. 이를 통해 한글의 음운 구조와 자모 결합의 다양성을 충분히 반영하여 한글의 복잡한 자모 결합 구조를 효과적으로 인식할 수 있다.

본 논문에서는 MAE의 강점을 활용하여 열화와 폐색된 한글 문자 이미지에서도 강인한 인식 성능을 달성할 수 있도록 한글 문자 인식 모델인 KR-MGPSTR의 Encoder로 사용하였다. MAE는 대조 학습(contrastive learning)과 달리 이미지의 상대적 차이를 학습하는 것이 아니라, 마스킹된 이미지의 일부를 복원하는 자기 복원(Self-Reconstruction) 과제를 통해 이미지 자체의 시각적 패턴과 구조를 학습한다. 이를 통해 이미지의 세부적인 정보와 전체적인 문맥을 동시에 이해하게 하여 열화와 폐색된 이미지에서도 강인한 특징 표현(Feature Representation)을 추출할 수 있다. 따라서 MAE는 제한된 정보로부터 이미지의 중요한 시각적 특징을 효과적으로 학습하고 이미지의 전반적인 구조와 패턴을 이해함으로써 Wei et al [21]과 Chen et al [22] 등의 연구와 같이 다양한 downstream 작업에서 우수한 성능을 보인 바 있다.

MAE의 사전 학습을 위해 한글 문자 이미지에 다양한 확률(예를 들면, 30%, 50%, 75%, 90%)로 무작위로 마스킹(Masking) 처리를 하였다. 이 중 마스킹되지 않은 이미지 패치들을 MAE의 인코더로 입력하여 이미지 패치를 선형 임베딩하고 포지션 임베딩을 추가하여 MAE 모델을 사전 학습시켰다. MAE의 인코더를 통해 사전 학습된 모델 중 문자 이미 50% 마스킹을 수행한 모델이 가장 좋은 문자 복원 및 인식 성능을 보여 random masking ratio는 50%로 고정하여 MAE를 사전 학습하였다.

제안된 SSL4STR 시스템은 그림 4에 도시한 바와 같이 KR-MGPSTR 모델에 MAE로 사전 학습된 MAE 인코더를 사용하였다. 그림 4에서 MAE 인코더/디코더를 사전 학습할 때 자기감독 학습(Self-Supervised Learning)이 적용되는데, 대규모의 Unlabeled 이미지 데이터로부터 일반화된 특징을 학 습할 수 있어서 모델이 데이터의 다양한 시각적 패턴과 구조를 이해하도록 도와주기 때문에, 열화와 폐색된 이미지에서도 강인한 특징 표현을 추출할 수 있게 한다. 더불어 SSL4STR은 실해역 선박명 문자 이미지로 파인 튜닝(Fine-Tuning)할 때 자기감독 학습으로 사전 학습된 MAE 인코더를 사용함으로써 한글 STR 모델의 학습 효율성을 높일 수 있다. 초기 가중치를

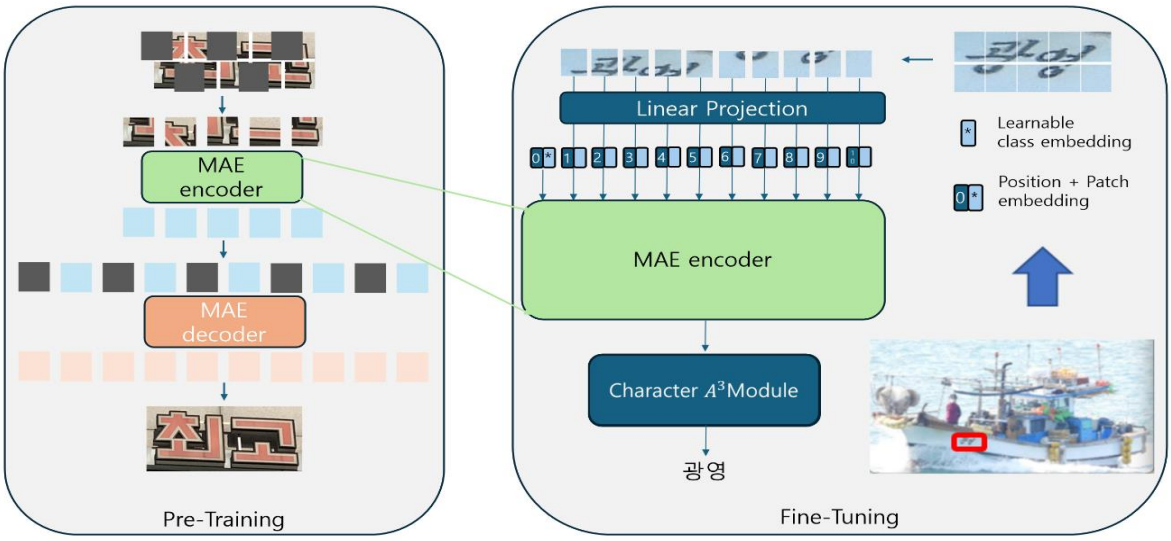


그림 4. 제안된 SSL4STR시스템의 구조와 자기감독 학습 방법: 50% 랜덤 마스킹된 문자 이미지의 visible patch에서 특징을 추출하도록 사전학습이 진행되며 사전학습된 인코더는 실해역 선박명 문자 이미지로 fine-tuning을 수행함.

Fig. 4. Architecture and Self-Supervised Learning Method of the Proposed SSL4STR: Pre-training is performed to extract features from visible patches of 50% randomly masked character images, and the pre-trained encoder is fine-tuned using the real-world character images of the vessels' name.

보다 풍부한 특징 표현으로 설정할 수 있어 학습 과정의 수렴 속도를 높이고, 파인튜닝 시 적은 양의 Labeled 데이터로도 높은 성능을 달성할 수 있다. 이처럼 SSL4STR은 SSL을 활용한 Pre-Training 과정과 실세계 Labeled 데이터를 활용한 Fine-Tuning 과정을 수행하여 데이터 라벨링 비용과 시간을 절감하면서도 열화와 폐색된 한글 장면 문자 이미지에서도 높은 성능을 유지할 수 있다.

SSL4STR은 MAE의 사전 학습을 위한 자기지도 학습 데이터셋으로 그림 5와 같은 대규모의 Union14M-U (Unlabeled) 데이터 1천만 장과 그림 6과 같이 AIHub (www.aihub.or.kr)에서

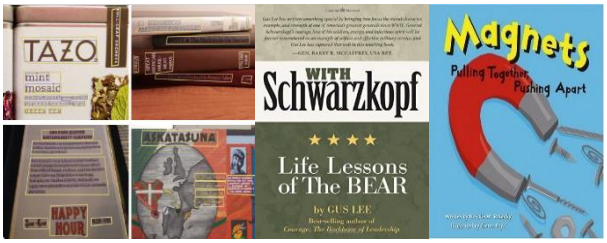


그림 5. Union14M-U (Unlabeled) 데이터 예시.

Fig. 5. Samples of Union14M-U (unlabeled) data.

그림 6. AIHub의 가로형 한글 간판 이미지 데이터.

Fig. 6. Data of Korean horizontal signboard images at AIHub.

공개한 야외 실제 촬영 한글 이미지의 가로형 간판 이미지 10만 장을 활용하였다. 이러한 사전 학습은 모델이 다양한 환경에서의 문자 이미지 패턴과 구조를 깊이 있게 이해하도록 돕는다. 특히, 사전 학습에 사용되는 대규모의 Union 14M-U (Unlabeled)와 실제 간판 이미지는 한글 문자의 복잡한 형태와 다양한 배경 조건을 포함하고 있는 Book32 [25], CC (Conceptual Captions) [26], OpenImages [27] 등의 데이터셋들을 사용하고 있기 때문에 제안된 SSL4STR 모델의 일반화 능력을 향상시키는 데 중요한 역할을 한다.

이와 더불어 파인 튜닝 단계에서의 학습 데이터셋으로 인공 문자 이미지 생성기인 SynthTIGER [30]를 통해 생성된 영어, 숫자, 한글을 포함한 1,000,000장의 합성 이미지와 실제 해상에서 촬영된 선박 이미지 13,000장을 사용하였다. SynthTIGER로 생성된 합성 이미지는 다양한 글꼴, 배경, 노이즈 등을 포함하여 SSL4STR 모델이 다양한 환경 조건에서의 문자 인식에 적응할 수 있도록 해준다. 또한, 그림 7과 같이 실제 해상에서 촬영된 선박 이미지는 해수의 염분으로 인해 열화와 폐색된 선박명 문자 이미지가 많기 때문에, SSL4STR 모델이 실해역 테스트 환경에서 발생하는 어려운 선박명 인식 상황에서도 강인할 수 있도록 학습시킬 수 있다.

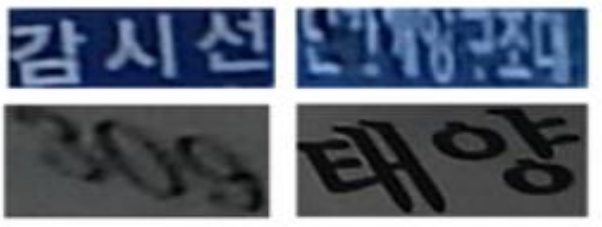


그림 7. 실제 해상에서 촬영된 선박명 이미지.

Fig. 7. Images of vessels' name photographed in real sea.

상기한 학습 데이터셋 구축 전략을 통해 SSL4STR 은 한글 장면 문자 인식에서 활용 가능한 레이블링된 데이터의 부족 문제를 해결하고, 제안된 모델이 실제 환경에서의 다양한 변형과 열악한 조건에서도 강인한 인식 성능을 발휘하도록 설계되었다. 특히, 대규모의 Unlabeled 데이터와 실제 이미지를 활용한 사전 학습은 레이블링에 필요한 비용과 시간을 절감하면서도, 제안된 모델의 일반화 능력을 크게 향상시킬 수 있다.

IV. 실험 결과

제안된 SSL4STR 모델은 사전 학습 시 대규모의 Unlabeled 데이터와 파인 튜닝 시 소량의 Labeled 데이터를 활용하여 학습한다. 표 1은 모델의 학습 시에 사용된 데이터셋과 하이퍼-파라미터들(Hyper-Parameters)을 정리한 것이다.

SSL4STR의 모델 학습 과정은 그림 4와 표 1에 도시한 바와 같이 크게 두 단계로 구분할 수 있다. 첫 번째는 Pre-Training 단계로서, MAE가 자기지도 학습을 수행할 때 Union14M-U (Unlabeled) 데이터셋 1천만 장과 AIHub 한글 간판 이미지 10만 장을 입력 받아 일부 패치를 무작위로 50%

표 1. SSL4STR 학습 시 적용된 Hyper-parameters.
Table 1. Hyper-parameters applied during training of SSL4STR.

Hyper-Parameters	Pre-Training	Fine-Tuning
Training Dataset	Union14M-U 10M, AIHub 간판 이미지 100K	SynthTIGER 합성 이미지 1M, 실제 해상 촬영 선박 이미지13K
Input Size (H x W)	32 x 128	32 x 128
Batch Size	64	50
Epochs	2	54
Optimizer	AdamW (betas: 0.9, 0.95, weight decay: 0.05)	Ada-delta
Initial Learning Rate	0.0025	1.0
Loss Function	Mean Squared Error	Cross Entropy

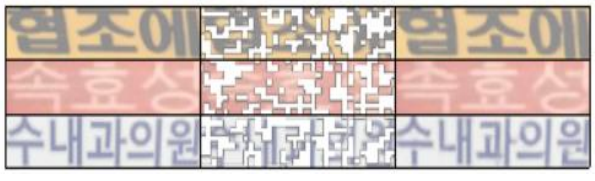


그림 8. MAE 사전 학습 과정 중 이미지 복원 샘플: 원본 이미지(좌), 마스킹 이미지(중), 복원 이미지(우).
Fig. 8. Samples of reconstructed images during pre-training of MAE: Original image (left), masked image (middle), reconstructed image (right).

마스킹해서 MAE의 인코더로 전달한다. MAE는 마스킹된 이미지 패치로부터 문자의 특징을 ViT 인코더/디코더를 통해 학습한다. 학습 시 AdamW를 Optimizer로 사용함으로써 급격한 업데이트로 인한 불안정한 학습을 피하고, 과적합을 억제하며 보다 안정적인 학습을 유도하였다. 또한, 기존 MAE [10] 연구의 실험적 결과에 따라 Initial Learning Rate를 0.0025로 설정하였다. 이를 통해 사전 학습된 MAE 인코더는 이미지의 구조적 패턴과 시각적 특징 등의 전반적인 표현을 학습하며, MAE 디코더를 사용해 복원한 결과를 통해 MAE의 사전 학습이 잘 이루어졌는지 훈련 상태를 평가하였다. 그림 8은 MAE 사전 학습 과정 중 AIHub 한글 간판 이미지를 복원한 결과들 중 일부 사례들이고, 원본 이미지를 무작위로 50% 마스킹한 상태에서도 MAE가 원본 이미지에 가깝게 이미지 복원 능력을 보여 준다는 것을 확인할 수 있다.

두 번째 모델 학습 과정은 Fine-Tuning 단계로서, 우선 한글에 특화된 STR 모델인 KR-MGPSTR의 인코더 부분을 첫 번째 과정에서 사전 훈련된 MAE의 인코더로 대체한다. SynthTIGER [30]를 통해 생성된 영어, 숫자, 한글을 포함한 1백만 장의 합성 이미지와 실제 해상에서 촬영된 선박 이미지 1만3천 장을 사용하였다. Fine-Tuning 시에는 손실 함수로 Cross Entropy 함수를 사용하였고, Optimizer로 Ada-delta [30]를 적용하며 이의 특성에 따라 Initial Learning Rate를 1.0으로 설정하여 보다 빠르게 수렴하도록 하였다.

기존의 KR-MGPSTR 방법에 더해 제안된 SSL4STR 모델은 마스킹 등 MAE 과정이 추가되었기 때문에 사전학습 단계에서는 계산 시간 증가가 있지만, 파인튜닝 단계에서는 기존 KR-MGPSTR 모델과 유사한 학습 시간이 소요되었다.

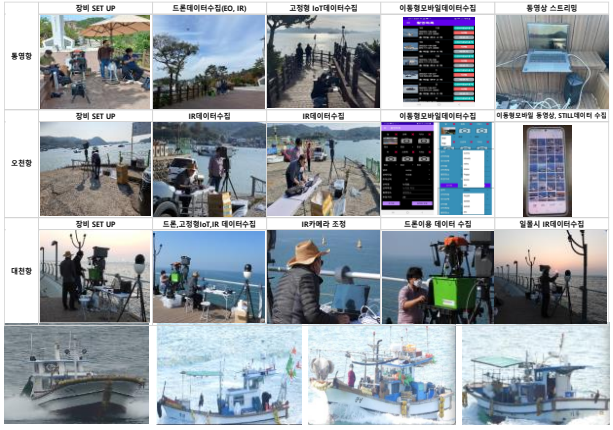


그림 9. 데이터 수집 과정과 수집된 선박 데이터.
Fig. 9. Process of collecting data and the collected vessel data.

표 2. MAE 사전 학습 시 마스킹 확률에 따른 성능 비교.
Table 2. Performance comparison according to random masking ratio during pre-training of MAE.

Random Masking Ratio	Test Accuracy
MAE with 30%	71.62
MAE with 50%	**73.31**
MAE with 70%	70.61
MAE with 90%	72.64

특히, 추론 단계에서는 기존 모델 대비 특별한 연산 부담의 증가 없이 거의 유사한 처리 속도를 유지하였기 때문에, 초기 사전학습에 따른 연산 부담은 실제 서비스 환경에서는 크게 문제가 되지 않았다.

성능 검증을 위한 테스트 데이터셋은 실제 해상에서 촬영된 선박명 이미지 1,023장으로 구성하였다. 이 데이터는 다양한 실해역 환경에서 촬영된 것으로 열화 및 폐색을 포함하여 실제 해안 감시 분야에서 마주할 수 있는 복잡한 상황을 반영하고 있다. 그림 9는 고정형 CCTV와 드론을 활용한 실제 해상 선박 데이터 수집 과정과 이 과정을 통해 수집된 선박 영상 데이터를 나타낸 것이다. 해당 데이터셋은 선박명 또는 선박에 부착된 주요 문자열을 포함하고 있으며, 실세계에서 제안된 SSL4STR 모델의 성능을 검증하기 위한 중요한 평가 지표로 사용되었다.

평가 단계에서는 실해역에서 촬영된 선박명 이미지 데이터 일부를 사용하였고, 제안된 모델의 성능을 평가하기 위하여 모델의 예측 결과를 정답 값과 비교하여 전체 테스트 데이터 중 예측 결과와 정답 값이 일치한 데이터 수의 비율로 정확도를 계산하는 방식으로 평가하였다.

표 2에 제시한 바와 같이 MAE의 인코더를 통해 사전 학습된 모델 중 텍스트 이미지에 50% 마스킹을 수행한 모델이 가장 좋은 텍스트 인식 성능을 보여 Random Masking Ratio는 50%로 고정하였다.

SSL4STR은 Fine-Tuning 단계에서부터 실제 해상에서 촬영된 선박 문자 이미지 데이터로 학습이 수행되고 있다. 그림 10에 도시한 바와 같이 실제 해역에서 수집된 선박명 문자 이미지에 대해서도 원본 이미지를 무작위로 50% 마스킹한 상태에서 MAE 디코더가 원본 이미지에 가깝게 이미지 복원 능력을 보여 준다는 것을 확인할 수 있다. 이와 같이 제안된 SSL4STR 모델은 Random Masking으로 사전 훈련된 MAE 인코더를 STR에 적용함으로써 해상 환경에서의 열화와 폐색된 선박명 텍스트 이미지에서도 강인한 인식 성능을 달성함을 볼 수 있다.

본 논문에서 제안된 SSL4STR 모델과 기존에 발표된 KR-MGPSTR [11] 모델의 성능을 표 3에 비교 평가하였다. Pre-Training 단계와 Fine-Tuning 단계에서 사전 학습 유무와 합성 데이터(SynthTIGER)와 실제 선박 데이터(vessel)의 학습 포함 여부 등 다양한 학습 데이터 조합이 실제 해상에서 선박명 인식 성능에 미치는 영향을 분석하였다.

그림 10. 실해역에서 수집된 선박명 문자 이미지(좌), 50% 마스킹된 이미지(중), MAE를 통해 복원한 이미지(우).

Fig. 10. Vessles' Name Text Image Collected from the Actual Sea (left), 50% Masked Image (middle), Image Reconstructed by MAE (right).

표 3. 실해역 선박명 인식 테스트에서 KR-MGPSTR 모델과 제안된 SSL4STR 모델의 성능 비교.

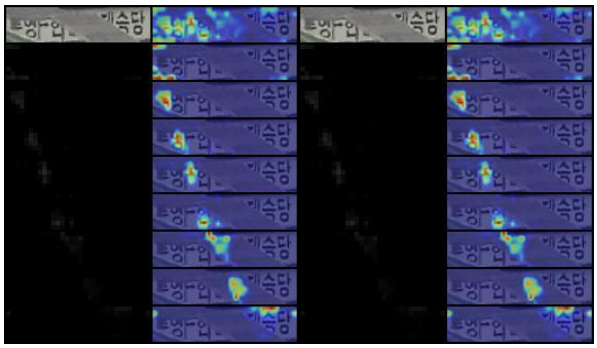
Table 3. Performance comparison between KR-MGPSTR and the proposed SSL4STR in real-world vessel name identification test.

Method	Unlabeled Pre-Training Data	Labeled Fine-Tuning Data	Test Accuracy
KR-MGPSTR [11]	N/A	Vessel	52.52
		SynthTIGER	65.66
		SynthTIGER+ Vessel	70.27
SSL4STR	N/A	Vessel	5.42
		SynthTIGER	72.97
		SynthTIGER+ Vessel	73.31
	Union 14M-U & AIHub	Vessel	3.04
		SynthTIGER	68.24
		SynthTIGER+ Vessel	**75.34**

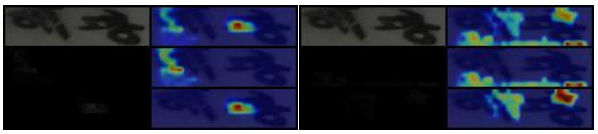
표 3의 실험 결과를 분석하면 다음과 같다. 우선 MAE 사전 학습 단계가 없는 KR-MGPSTR 모델의 경우에는, 실제 선박 데이터(vessel)만으로 학습했을 때 52.52%의 정확도를 기록하였고, 합성 데이터(SynthTIGER)만 학습하였을 때의 정확도는 65.66%로 약 13.14% 향상되었다. 실제 선박 데이터보다 합성 데이터로 성능이 개선된 원인은 ViT 기반의 딥러닝 모델 특성상 대량의 데이터를 학습하는 것이 일반화 성능을 향상하는데 더 많은 기여를 하게 되고 합성 데이터의 양이 실제 선박 데이터보다 약 77배 많기 때문인 것으로 분석된다. 하지만 합성 데이터와 실제 선박 데이터를 통합하여 Fine-Tuning 학습한 경우에는 70.27%의 정확도로 합성 데이터만 학습하였을 때보다 약 4.61%의 성능 향상을 보였다.

이와 비교해서 SSL4STR 모델은 사전 학습 여부와 관계없이 실제 선박 데이터만으로 Fine-Tuning이 진행되었을 때 현저히 낮은 정확도를 보여주었다. 이는 소량의 실세계 데이터만으로 모델을 Fine-Tuning 하는 것은 성능 향상을 거의 기대할 수 없다는 것을 나타낸다. 사전 학습 단계가 없는 SSL4STR 모델의 경우에는, 합성 데이터와 실제 선박 데이터를 함께 Fine-Tuning 하였을 때 73.31%의 정확도로 기존 KR-MGPSTR 모델과 비교하면 3.04%의 성능 향상을 달성하였다.

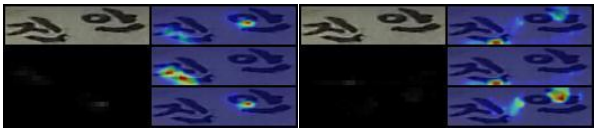
사전 학습의 유무는 제안된 SSL4STR 모델의 성능에 중요한 영향을 미쳤다. 사전 학습 단계가 없이 학습한 SSL4STR 모델은 73.31%의 정확도를 기록한 반면에, 레이블링 되지 않은 Union14M-U와 AIHub 한국어 간판 이미지 데이터를 자기감독 학습(SSL)을 통해 MAE를 사전 훈련시킨 모델은 합성 데이터와 실제 선박 데이터를 함께 Fine-Tuning 하였을 때 정확도가 75.34%로 증가하였고, 이는 사전 학습 단계가 없는 경우 대비 2.03%의 성능 향상 효과가 있었음을 확인하였다. 본 실험 결과를 통해 자기감독 학습을 통한 MAE 사전 훈련 방법이 제안된 SSL4STR 모델의 일반화 성능을 향상시킨다는 점을 명확하게 보여주었다.



(GT): 통영한산도제승당.
(좌): 통영 R 영도제제당당. (우): 통영한산도제승당



(GT): 해광
(좌): 해성. (우): 해광



(GT): 진안
(좌): 진양. (우): 진안

(GT): Ground Truth. (좌): KR-MGPSTR 의 예측 문자열. (우): SSL4STR 의 예측 문자열.

그림 11. 열화와 폐색된 선박 텍스트 이미지에서 제안된 SSL4STR 모델의 개선된 예측 값과 어텐션 맵.
Fig. 11. Improved prediction results in the degraded and occluded vessels' text images and attention maps of the proposed SSL4STR model.

그러나 표 3에서 실제 해상 선박(vessel) 데이터만으로 fine-tuning 시에는 5.42%와 3.04% 등 큰 성능 저하가 확인되었는데, 이에 대한 원인은 주로 도메인 차이, 즉 데이터 부족과 데이터 품질 문제 때문인 것으로 분석된다. 특히 사전학습 과정에서 사용된 일반 문자 데이터와 극히 열악한 환경의 실제 선박 이미지 간의 도메인 갭(domain gap)이 매우 커서, MAE 기반 자기지도 학습으로 학습된 특성들이 fine-tuning 과정에서 제대로 전이되지 못했을 가능성이 가장 크다. 본 연구에서 사용한 약 1,000,000장의 합성 데이터와 13,000장의 실해역 데이터로는 한국어의 복잡한 자모 결합 구조와 해상 환경의 특수성을 충분히 반영하는 데 한계가 있다. 따라서 향후 연구에서는 도메인 적응 기술을 적용하여 데이터 품질과 다양성을 높이고, 더욱 다양한 형태의 한글 데이터셋을 추가 확보하며, 라벨링 품질 향상 및 최적의 하이퍼파라미터 탐색을 통해 모델의 일반화 성능을 개선할 계획이다.

ICDAR 한국어 벤치마킹 중 지금까지 가장 높은 한국어 STR 정확도를 보인 것은 49.09% 정도에 불과하였고[31,32], 실제 Vessel 데이터 기준으로는 정확도가 더 낮은 상황에서 KR-MGPSTR 모델[11]은 70.27%로 한국어 인식 성능을 향상시켰고, 제안된 SSL4STR 모델은 추가적으로 5% 이상 더 개선하였다. 실험에서 나타난 약 25%의 오인식 사례의 대부분은 이미지 품질 저하와 심한 폐색 및 열화로 인한 것이었다. 이를 개선하기 위해 MAE 방법을 추가 도입한 것이고, 향후 추가적으로 데이터 증강 기법과 함께 이미지 복원(Image Restoration) 기술을 전처리 단계에서 도입하여 인식률을 더욱 향상시킬 방안을 고려 중에 있다.

그림 11은 실제 해상에서 촬영된 열화와 폐색된 선박 텍스트 이미지에서 제안된 SSL4STR 모델의 개선된 예측 결과 사례를 보여준다. 그림 11에서 (GT)는 정답 레이블 값을 나타내고, (좌)는 KR-MGPSTR 모델의 예측 값을 나타내고, (우)는 SSL4STR 모델의 예측 값을 나타낸다. 기존 KR-MGPSTR 모델이 열화와 폐색 상황에서 선박명을 잘못 예측하는 데 반해, 제안된 SSL4STR 모델은 정답 레이블과 동일한 결과를 예측하였다. 이를 통해, SSL4STR 모델이 열화와 폐색 상황에서도 강인한 한글 STR 모델임을 입증하였다.

V. 결론

본 논문에서는 SSL (Self-Supervised Learning) 기반 MAE (Masked Auto-Encoder)를 활용하여 한국어 장면 문자 인식의 성능을 향상시키는 SSL4STR (Self-Supervised Learning for Scene Text Recognition) 모델을 제안하였다. 제안된 모델은 한국어의 복잡한 자모 결합 구조를 효과적으로 학습하기 위해 대규모 Unlabeled 데이터로 MAE를 사전 훈련시켜 한글 STR 모델인 KR-MGPSTR 구조에 통합하였으며, Fine-Tuning 단계에서 실제 해상에서 수집된 선박 데이터와 합성 데이터를 결합하여 학습하였다. 본 논문의 주요 기여점은 다음과 같이 요약할 수 있다.

- Masked Auto-Encoder 기반의 한국어 특화 모델 설계: MAE를 기존에 저자들이 제안했던 KR-MGPSTR [11] 구조와 통합한 새로운 SSL4STR 모델을 통해 한국어의 복잡한 자모 결합 구조를 효과적으로 학습하였다.
- 대규모 Unlabeled 학습 데이터 활용: Union14M [9]과 AIHub (www.aihub.or.kr)에 있는 한글 데이터셋을 사용하여 Self-Supervised Learning을 수행함으로써, 대량의 Labeled 데이터셋 구축을 위한 비용과 시간 투입 없이 Labeled 데이터 부족 문제를 해결하고 한국어 STR 모델의 일반화 성능을 개선하였다.
- 실세계 응용 가능성 입증: 제안된 SSL4STR 모델을 해상에서 고정형 CCTV와 이동형 드론을 통해 촬영된 선박 이미지에 적용하여 실제 응용 가능성을 평가하였다. 실제 환경에서 선박명을 자동으로 식별하고 해상에서의 열화와 폐색된 문자열 환경에서도 강인한 선박명 인식 성능을 입증하였다. 실험 결과, 제안된 모델은 기존 KR-MGPSTR 모델 대비 5.07%의 선박명 인식 성능 향상을 보였으며, 열화와 폐색된 문자에서도 안정적으로 선박명을 식별하여 해상 감시, 선박 식별 및 다양한 한국어 응용 분야에서 활용될 가능성을 보여주었다.

본 연구는 한국어에 특화된 STR 모델을 제안하고 있지만, 이러한 접근법은 향후 연구에서 한글, 영어, 숫자를 포함하여 중국어, 일본어와 같이 복잡한 문자 구조를 가진 외국어 등에도 적용이 가능해 다국어(multi-lingual) STR 모델 연구로 확장할 수 있다.

REFERENCES

[1] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," arXiv preprint arXiv: 1406.2227, June 2014.
doi: https://doi.org/10.48550/arXiv.1406.2227

[2] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315~2324. April 2016.
doi: https://doi.org/10.1109/cvpr.2016.254

[3] M. Mathew, M. Jain, and C.V. Jawahar, "Benchmarking scene text recognition in devanagari, telugu and malayalam," *In International Conference on Document Analysis and Recognition*, vol. 7, pp. 42~46, Nov. 2017.
doi: https://doi.org/10.1109/icdar.2017.364

[4] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez I Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. Almazan, and L. Heras, "ICDAR 2013 robust reading competition," *In International Conference on Document Analysis and Recognition*, pp. 1484~1493, Aug. 2013.
doi: https://doi.org/10.1109/ICDAR.2013.221

[5] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," *In International Conference on Computer Vision*, pages 1457~1464, November 2011.
doi: https://doi.org/10.1109/iccv.2011.6126402

[6] D. Karatzas, L. Gomez, A. Nicolaou, S. Ghost, A. Bagdanov, and M. Iwamura, "ICDAR 2015 competition on robust reading competition," *In International Conference on Document Analysis and Recognition*, pp. 1156~1160, Aug. 2015.
doi: https://doi.org/10.1109/ICDAR.2015.7333942

[7] T. Q. Phan, P. Shibakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," *Proc. of the IEEE International Conference on Computer Vision*, pp. 569~576, March 2014.
doi: https://ieeexplore.ieee.org/document/6751180

[8] A. Risnumawan, P. Shivakumara, C. Chan, and C. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, pp. 8027~8048, 2014.
doi: https://doi.org/10.1016/j.eswa.2014.07.008

[9] Q. Jiang, J. Wang, D. Peng, C. Liu, and L. Jin, "Revisiting scene text recognition a data perspective," *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 20542-20554, Jan. 2024.
doi: https://doi.org/10.1109/iccv51070.2023.01878

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000-16009, June 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01553

[11] H. Song and J. Yoo, "ViT-Based vessel name identification system to improve korean recognition performance," *IEMEK Symposium on Embedded Technology (in Korean)*, May 2024.

[12] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, Dec. 2016.
doi: https://doi.org/10.48550/arXiv.1507.05717

[13] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 2035-2048, Jun. 2018.
doi: https://ieeexplore.ieee.org/document/8395027

[14] H. Li, P. Wang, C. Shen and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8610-8617, July 2019.
doi: https://doi.org/10.1609/aaai.v33i01.33018610

[15] P. Lyu, C. Zhang, S. Liu, M. Qiao, Y. Xu, L. Wu, K. Yao, J. Han, E. Ding, and J. Wang, "MaskOCR: Text recognition with masked encoder-decoder pretraining," arXiv preprint arXiv:2206.00311, Jun 2022.
doi: https://doi.org/10.48550/arXiv.2206.00311

[16] J. Wu, Y. Peng, S. Zhang, W. Qi, and J. Zhang, "Masked vision-language transformers for scene text recognition." arXiv preprint arXiv:2211.04785, Nov 2022.
doi: https://doi.org/10.48550/arXiv.2211.04785

[17] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representation*, pp. 1-21, May 2021.
doi: https://doi.org/10.48550/arXiv.2010.11929

[18] J. Lee, S. Park, J. Baek, S. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Wokrshops*, pp. 546~547, Oct 2019.
doi: https://doi.org/10.1109/cvprw50498.2020.00281

[19] H. Liu, B. Wang, Z. Bao, M. Xue, S. Kang, D. Jiang, and B. Ren, "Perceiving stroke-semantic context: hierarchical contrastive learning for robust scene text recognition," *Proc. of the AAAI Conference on Artificial Intelligence*, vol.36, no.2, pp.1702~1710, Jun. 2022.
doi: https://doi.org/10.1609/aaai.v36i2.20062

[20] M. Yang, M. Liao, P. Lu, J. Wang, S. Zhu, H. Luo, Q. Tian, and X. Bai, "Reading and Writing: Discriminative and generative modeling for self-supervised text recognition", *In ACM Multimedia*, pp. 4214~4223, October 2022.
doi: https://doi.org/10.1145/3503161.3547784

[21] C. Wei, H. Fan, S. Xie, C. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668~14678, December 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01426

[22] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 9640~9649, Apr. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.00950

[23] P. Wang, C. Da, and C. Yao, "Multi-Granularity Prediction for Scene Text Recognition," *European Conference on Computer Vision*, pp. 339~355, October 2022.
doi: https://doi.org/10.1007/978-3-031-19815-1_20

[24] KS X 1001, "Code for Information Interchange (Hangul and Hanja), https://standard.go.kr/KSCI/standardIntro/getStandardSearchView.do?ksNo=KSX1001.

[25] B. Iwana, S. T. R. Rizvi, S. Ahmed, A. Dengel, and S. Uchida, "Judging a book by its cover," arXiv preprint arXiv: 1610.09204,

October 2016.
doi: https://doi.org/10.5040/9781641899543.ch-009

[26] S. Piyush, D. Nan, G. Sebastian, and S. Radu, "Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *In Annual Meeting of the Association for Computational Linguistics*, pp. 2556~2565, Jul. 2018.
doi: https://doi.org/10.18653/v1/p18-1238

[27] K. Alina, Rom. Hassan, A. Neil, U. Jasper, K. Ivan, P. T. Jordi, K. Shahab, P. Stefan, M. Matteo, K. Alexander, D. Tom, and F. Vittorio, "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no.7, pp. 1956~1981, Nov. 2018.
doi: https://doi.org/10.1007/s11263-020-01316-z

[28] S. Ham, M. Kang, S. Jeong, and J. Yoo, "Deep-learning-based water shield automation system by predicting river overflow and vehicle flooding possibility," *IEMEK Journal of Embedded Systems and Applications*, vol. 18, no. 3, pp. 133-139, June 2023.

[29] S. Jeong and J. Yoo, "TextReID: Transformer-based text re-identification to supplement vessel identification," *IEMEK Symposium on Embedded Technology (in Korean)*, May 2023.

[30] M. Yim, Y. Kim, H. Cho, and S. Park, "SynthTIGER: Synthetic text image generator towards better text recognition models," *International Conference on Document Analysis and Recognition*, pp. 109~124, September 2021.
doi: https://doi.org/10.1007/978-3-030-86337-1_8

[31] J. Kim and S. Kim, "Deep learning network-based end-to-end scene text spotter for Korean characters," *Journal of the Korean Institute of Industrial Engineers (in Koream)*, vol. 48, no. 4, pp. 398-408, August 2022.

[32] S. Kim and S. Kim, "Korean scene text recognition using semi-supervised learning with character-level consistency regularization," *Journal of the Korean Institute of Industrial Engineers (in Koream)*, vol. 49, no. 3, pp. 258-266, June 2023.

[33] C. Baek and K. Kong, "Deep learning-based face recognition through low-light enhancement", *IEMEK Journal of Embedded Systems and Applications (in Koream)*, vol. 19, no. 5, pp. 243-250, October 2024.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representation*, May 2019.
doi: https://arxiv.org/pdf/1711.05101v3

[35] M. Lee, S. Lee, and T. Kim, "Performance evaluation of efficient vision transformers on embedded edge platforms," *IEMEK Journal of Embedded Systems and Applications (in Koream)*, vol. 18, no. 3, pp. 89-100, Jun. 2023.

[36] Zeiler and D. Matthew, "AdaDelta: an adaptive learning rate method," arXiv preprint arXiv: 1212.5801, December 2012.
doi: https://arxiv.org/pdf/1212.5701v1

[37] Y. Lee, W. Shin, J. Yun, and M. Joo, "Image segmentation for microstructure based on semi-supervised learning," *IEMEK Journal of Embedded Systems and Applications (in Koream)*, vol. 19, no. 6, pp. 307-312, December 2024.

[38] G. Zeng, W. Yu, R. Wang, and A. Lin, "A transfer learning-based approach to marine vessel re-identification," arXiv preprint arXiv:2207.14500, July 2022.

[39] S. Jung, E. Kim, and J. Yoo "CAPS: Autonomous child abuse prediction system based on deep learning with CCTV video," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol.27, no. 12, pp. 1029-1037, Dec. 2021.

[40] W. Jung, H. Choi, B. Kim, H. Chang, D. Lee, and S. Kim, "Enhancing text recognition using masking segmentation network and domain adaptation," *Proc. of 2022 37th ICROS Annual Conference (ICROS 2022)*, pp. 68-69, Jun. 2022.

[41] S. Son and H. Oh, "SIMCAP: Similarity-based image captioning," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol.30, no. 12, pp. 1380-1388, Dec. 2024.

[42] Y. Wang and J. Ha, "Scene text recognition with dual encoders," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol.29, no. 12, pp. 973-979, Dec. 2023.

유 준 혁

1993년 포항공과대학교 전자전기공학과 학사. 1995년 동 대학원 석사. 2007년 미국 매릴랜드대학교 컴퓨터공학 박사. 2009년~현재 대구대학교 컴퓨터정보공학부 교수. 관심분야는 온-디바이스 AI, 컴퓨터비전, 분산형 기계학습, Physical AI.

송 현 석

2018년~현재 대구대학교 인공지능전공 재학 중. 관심분야는 딥러닝, 컴퓨터 비전, 장면 문자 인식, 이미지 복원.