The 19th Workshop on Convergent and Smart Media Systems

일시 : 2025. 09.26(금) ~ 09.28(일) 장소 : 리젠트마린 제주호텔

주관

Smart Media
KOREAN INSTITUTE OF SMART MEDIA
한국스마트미디어학회

인사 말씀

위원장 김판구(조선대학교 교수)

한국스마트미디어학회 스마트미디어융합연구회가 주최하는 제19회 CSMS(Convergent and Smart Media Systems) 워크숍에 함께해 주신 모든 연구자와 참석자 여러분께 진심으로 감사드립니다.

그동안 CSMS 워크숍은 스마트미디어와 융합기술 분야의 연구자들이 최신 성과를 공유하고 협력의 가능성을 모색하는 의미 있는 장으로 자리매김해왔습니다. 특히 최근에는 인공지능, 메타버스, 디지털 트윈, 생성형 미디어, XR, 그리고 데이터 기반 서비스 혁신 등 다양한 융합기술이 산업과 사회 전반에 빠르게 확산되면서, 본 워크숍의 역할이 더욱 중요해지고 있습니다.

이번 워크숍을 통해 발표되는 연구들은 단순한 기술적 성과를 넘어, 산업 적용, 글로벌 협력, 사회적 가치 창출로 이어질 수 있는 실질적인 논의의 기반이 될 것이라 믿습니다. 또한 젊은 연구자와 전문가들이 교류하며 미래 스마트미디어의 방향성을 함께 고민하는 자리가 되기를 기대합니다.

이 자리에 함께해 주신 모든 분들께 다시 한 번 깊은 감사의 말씀을 드리며, 짧지만 밀도 있는 시간 속에서 새로운 영감과 협력의 기회를 얻어가시길 바랍니다.

감사합니다.

2025년 09월

CSMS 조직위원회

■ 조직위원장

- 김판구(조선대학교)

■ 프로그램위원장

- 최　창(가천대학교)

■ 출판위원장

- 황명권(KISTI)

■ 프로그램위원회

- 최종무 (단국대학교)
- 유일선 (순천향대학교)
- 이종혁 (세종대학교)
- 안효범 (공주대학교)
- 김한일 (제주대학교)
- 임을규 (한양대학교)
- 김광준 (전남대학교)
- 김남호 (호남대학교)
- 정재은 (중앙대학교)
- 박세현 (대구대학교)
- 김순철 (대구대학교)
- 나종회 (광주대학교)
- 최준호 (조선대학교)
- 임강빈 (순천향대학교)
- 서정택 (순천향대학교)
- 민　홍 (가천대학교)
- 정진만 (인하대학교)
- 김봉재 (충북대학교)
- 전광일 (한국산업기술대학교)
- 고　훈 (충북대학교)

Workshop 일정표

■ 9월 26일 (금)

시간	내용	장소	비고
16:00~18:30	전문가초청 세미나	리젠트마린 제주호텔	

■ 9월 27일 (토)

시간	내용	장소	비고
10:00~12:00	Tutorial Ⅰ	-	
12:00~13:00	Launch time		
13:00~14:15	논문발표 Ⅰ (Oral Session Ⅰ)	리젠트마린 제주호텔	좌장 : 신주현 교수 (조선대)
14:15~14:30	Break time		
14:30~16:00	논문발표 Ⅱ (Oral Session Ⅱ)	리젠트마린 제주호텔	좌장 : 신주현 교수 (조선대)

■ 9월 28일 (일)

시간	내용	장소	비고
09:00~10:30	논문발표 Ⅲ (Oral Session Ⅲ)	-	
12:00~13:00	Launch time		
13:00~14:00	Borad Meeting	-	

논 문 발 표 순 서

Oral Session Ⅰ

09월 27일(토) 13:00 ~ 14:15 **리젠트마린 제주호텔 / 좌장 : 신주현 교수**

A1 Weakly labeled 데이터를 활용한 Mean-Teacher 기반 AI 생성 글 탐지 연구 (13:00 ~ 13:15)
하은서, 채민주, 이건우, 전찬준

A2 ConvNeXt-V2 아키텍처를 활용한 경량화된 음향 장면 분류 연구 (13:15 ~ 13:30)
조민식, 한사랑, 이건우, 전찬준

A3 YOLO 모델 기반의 지하철 응급 상황 감지 : 버전별 아키텍처 및 성능 비교 (13:30 ~ 13:45)
최세영, 김지현, 김세진

A4 CAPTION-GUIDED REFINEMENT OF IMAGE REGIONS VIA MASKED GAN TRAINING (13:45 ~ 14:00)
Joshua Nyaberi, 유경호, 이승재, 김판구

A5 Displacement Net Speckle Prior Attention and Calibrated Uncertainty for Texture Aware Digital Image Correlation (14:00~14:15)
Teddy Okatch, 유경호, 정욱, 김판구

Oral Session Ⅱ

09월 27일(토) 14:30 ~ 16:00 **리젠트마린 제주호텔 / 좌장 : 신주현 교수**

B1 공감 대화를 위한 감정 키워드 추출 기법 (14:30 ~ 14:45)
임명진, 김시우, 신주현

B2 NLP-Enhanced Sequence-Based Reinforcement Learning for Social Mind map Agents. (14:45 ~ 15:00)
Birir Sospeter Kipchirchir, 김형주, 김판구

B3 Explainable AI for Low-Resource Multilingual Phishing Detection: A Deployable XLM-RoBERTa Framework (15:00 ~ 15:15)
Vincent Mwania, 김형주, 김판구

B4 Strain Estimation in Real Tensile Experiments Using Self-Supervised Learning-Based Digital Image Correlation (DIC) (15:15 ~ 15:30)
정현경, 유경호, 김판구

B5 Attention-Enhanced Optimized Deep Ensemble Network For Effective Facial Emotion Recognition (15:30 ~ 15:45)
Taimoor Khan, 최 창

B6 멀티모달 기반 패션 아이템 판매량 예측 시스템 연구 (15:45 ~ 16:00)
이새봄, 최창

Oral Session Ⅲ

09월 28일(일) 09:00 ~ 10:30 **온라인**

C1 MAGL-YOLO: A Bird Object Detection Algorithm Based on Multi-path Adaptive Global Feature Fusion (09:00 ~ 09:15)

Xinyao Wang, 김판구

C2 Local Feature Enhancement via Feature Fusion for Spoof Fingerprint Detection with Receptive Field –Wise Learning (09:15 ~ 09:30)

Md Al Amin, Naim Reza, 정호엽

C3 음향 장면 생성을 위한 Flow 기반 모델의 비교 분석 (09:30 ~ 09:45)

김어진, 박유정, 이건우, 전찬준

C4 검색 증강 생성과 소형 언어 모델을 활용한 산업안전 조치 의사결정 지원 시스템에 관한 연구 (09:45 ~ 10:00)

김수아, 정연비, 최기도, 김원열

C5 Mask-and-Reconstruct (MAR) on Noisy WiFi CSI for Human Pose Estimation (10:00 ~ 10:15)

Iftikhar Ahmad, 최우열

C6 Crack Segmentation Using U-Net and Transformer Combined Model (10:15 ~ 10:30)

노주현, 조정운, 양희덕

Weakly labeled 데이터를 활용한 Mean-Teacher 기반 AI 생성 글 탐지 연구

하은서, 채민주, 이건우*, 전찬준
조선대학교 AI소프트웨어학부
e-mail : {murru8989, minju9642, geonwoo, cjchun}@chosun.ac.kr

A Study on Mean-Teacher Based Detection of AI-Generated Text Using Weakly Labeled Data

Eun Seo Ha, Min Ju Chae, Geon Woo Lee*, Chanjun Chun
School of AI software, Chosun University

요 약

대규모 언어 모델(LLM)의 발전은 텍스트 생성의 품질을 혁신적으로 높였으며 다양한 영역에서 활용되고 있다. 그러나 LLM이 생성한 텍스트는 사람이 작성한 글과 혼재되면 허위 정보 확산, 여론 조작 등 사회적 문제를 유발할 수 있으므로, 이를 판별하는 기술의 필요성이 커지고 있다. 기계학습 및 심층학습 기술을 활용한 기존 모델은 LLM이 생성한 텍스트 탐지에서 좋은 결과를 나타내고 있다. 하지만, 실제 학습 데이터는 weakly labeling 되어 모델 학습에 어려움이 있으며, 이에 따라 성능 저하 문제가 발생할 수 있다. 본 연구는 이러한 한계를 극복하기 위해 한국어 텍스트 인코더 ELECTRA와 준지도학습 기법인 mean-teacher을 결합한 학습 방법을 제안한다. 자세하게는 문맥 이해 능력을 위해 ELECTRA 모델을 활용하고, 라벨 불균형 환경에서의 안정적인 학습을 위해 mean-teacher 기법을 활용한다. 본 논문에서 한국어 LLM 생성 텍스트 판별에서 기존 연구 대비 개선된 성능을 보였으며, 이는 weakly labeled 데이터에서 기존 연구보다 강인한 탐지 성능을 달성했다.

1. 서론

딥러닝 기술을 통해 다양한 분야에서 많은 발전이 있었지만, 그중에서 대규모 언어 모델(large language model; LLM)은 현대 사회에 큰 영향을 주고 있다. LLM은 방대한 텍스트 데이터와 고성능 연산 자원을 기반으로 학습되어, 기존 자연어 처리 기술로는 달성하기 어려웠던 수준의 언어 이해와 생성 능력을 보여주고 있다. 특히, 생성형 LLM은 인간과 유사한 수준의 텍스트 생성이 가능해짐에 따라 뉴스, 학술, 커뮤니케이션 등 다양한 영역에 빠르게 도입되고 있으며, 결과물의 문장 구조와 표현력이 인간이 작성한 글과 유사한 수준에 도달하고 있다[1].

하지만, 생성형 LLM의 우수한 텍스트 생성 능력은 심각한 사회적 문제를 야기할 수 있다. 특히, 생성형 LLM으로부터 생성된 텍스트가 사람이 직접 작성한 텍스트에 섞여 있다면, 사회적 혼란, 여론 왜곡, 악의적 이용 등의 문제를 초래할 수 있다. 이와 같은 사회 혼란을 예방하기 위해 생성형 LLM에서 생성된 텍스트를 신뢰성 있게 판별할 수 있는 텍스트 분류 기술 수요가 증가하고 있다[2].

전통적인 텍스트 분류 기술은 term frequency-inverse document frequency(TF-IDF) 기법을 통해 텍스트를 벡터화하고, XGBoost 등의 기계학습 기법을 사용하여 텍스트를 분류하였다[3, 4]. 하지만, TF-IDF 벡터화는 텍스트 문맥 파악의 어려움과 단어 순시가 무시되는 문제점 등으로 인해 복잡한 텍스트에 대해서는 한계가 존재할 수 있다. 하지만, 딥러닝 기술과 대규모 텍스트 데이터를 통해 연구된 BERT와 ELECTRA 등과 같은 텍스트 인코더들은 기존 TF-IDF 벡터화보다 뛰어난 문맥 파악 능력과 단어 사이 의미 관계 반영 등을 통해 뛰어난 벡터화 성능을 보였다.

생성형 텍스트 분류 기술에서 학습 및 평가에 사용되는 텍스트 데이터에는 LLM이 생성한 텍스트가 사람이 작성한 텍스트에 일부 삽입 또는 수정된 형태로 존재할 수 있다. 하지만, 일부 수정된 텍스트 데이터의 경우 생성된 문장 위치 정보 등이 포함되지 않을 수 있으며, 일반적인 지도 학습 기법을 적용하기 어려울 수 있다. 즉, 데이터 레이블링의 구체적 정보가 부족한 weakly labeled 데이터는 문장 단위의 지도학습에서 성능의 한계가 발생할 수 있다. 또한, 생성형 문장이 일부 포함된 텍스트 데이터의 특성상 사람이 작성한 텍스트보다 LLM이 생성한 텍스트 데이터 양이 적을 수 있다. 이와 같은 데이터 불균형은 딥러닝 모델 학습에서 많은 양의 데이터에 편향을 일으키는 결과를 초래할 수 있다.

본 논문에서는 생성형 LLM이 생성한 텍스트 분류를 위해 딥러닝 기반의 텍스트 인코더를 활용하고, weakly labeled 데이터 활용을 위해 mean-teacher 모델 기반의 학습 방법을 활용한 텍스트 분류 방법을 제안한다. 딥러닝 기반 텍스트 인코더의 뛰어난 문맥 파악 능력과 mean-teacher 모델의 weakly labeled 데이터 활용 능력으로 개선된 생성형 텍스트 분류 성능을 나타낸다.

* 교신저자

2. 관련 연구

본 장에서는 본 연구의 기반이 되는 관련 연구를 다룬다. 먼저, 한국어 텍스트 인코더로 활용되는 KoELECTRA와 weakly labeled 데이터를 활용하여 효과적으로 모델을 학습시킬 수 있는 mean-teacher 기법을 설명한다.

2.1 한국어 텍스트 인코더: KoELECTRA

대규모 한국어 텍스트 데이터셋을 기반으로 사전 학습된 KoELECTRA는 ELECTRA 모델을 기반으로 하고 있다. 블로그, 뉴스 등 다양한 도메인 텍스트를 학습함으로써 LLM 생성 텍스트 판별 작업에서 좋은 성능을 보인다 [5]. ELECTRA는 replaced token detection(RTD)기반의 사전 학습 방법을 사용하여 샘플 효율성과 계산 효율을 동시에 확보할 수 있다[6]. 이와 같이 사전 학습된 ELECTRA는 fine-tuning을 거쳐 분류 문제에서 좋은 성능을 달성할 수 있다.

2.2 Mean-teacher

준지도학습 프레임워크인 mean-teacher는 레이블된 데이터와 weakly labeled 데이터를 이용하여 판별 문제에서 좋은 분류 성능을 나타낼 수 있다[7, 8]. 이와 같은 접근 방법은 학습을 주도하는 student 모델과 student 가중치의 지수 이동 평균(exponential moving average; EMA)을 기반으로 업데이트되는 teacher 모델로 구성된다. 이와 같이 student 모델과 teacher 모델 한 쌍으로 구성된 mean-teacher는 EMA를 통해 teacher 모델이 안정적인 예측을 할 수 있도록 돕는다. 이와 같은 과정에서 모델은 분류의 근거가 되는 두 모델의 일관성을 반복적으로 강화하며, 초기에는 불안정한 weakly labeled 데이터로 학습을 시작하지만, 점차 정교하고 신뢰도 높은 레이블로 학습 대상을 정제해 나갈 수 있다.

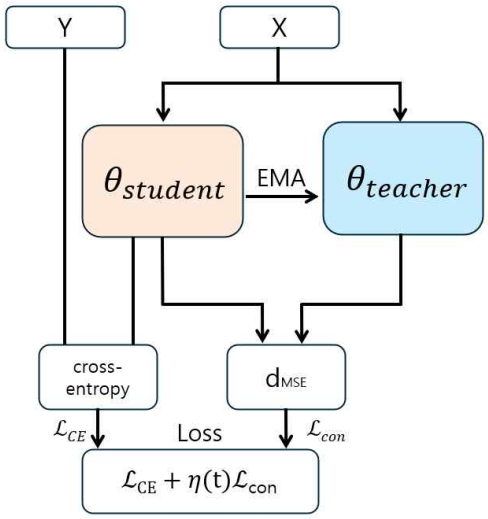
3. 본 론

3.1 텍스트 인코더

본 연구에서 활용되는 텍스트 데이터는 자연어로 구성되어 있으며, 모델 학습에 직접적으로 사용할 수 없다. 그러므로 텍스트에 담긴 의미와 구조적인 정보를 보존할 수 있는 수치형 데이터인 벡터로 변환하는 전처리 과정이 요구된다. 본 연구에서는 텍스트를 최소 단위로 분할하고 각 단위를 숫자로 표현할 수 있는 Word Piece 기반의 sub-word 토크나이저를 사용하여, 텍스트를 토큰화 및 정수 인코딩 과정을 진행했다. 그리고, 모든 입력 시퀀스는 일관된 길이를 갖도록 최대 길이를 512로 설정하였으며, 이보다 긴 시퀀스는 뒤쪽을 절단하고 짧은 시퀀스는 특수 패딩(padding) 토큰으로 채워 길이를 일치시킨다. 이와 같은 과정에서 패딩된 벡터에 계산이나 학습이 진행되지 않도록 마스킹을 진행한다. 마지막으로 사람이 직접 작성한 문장으로만 구성된 문단은 양성(positive)으로, LLM에서 생성된 문장이 하나라도 존재하는 문단에 대해서는 음성(negative)으로 타겟을 설정한다.

3.2 Mean-teacher 기반 생성형 텍스트 탐지 모델 학습

본 논문에서 다루는 텍스트 데이터는 문단 내 생성형 텍스트의 정확한 위치 정보 없이, 포함 여부만 표기된 문단 단위의 weakly labeled 데이터이다. 즉, 주어진 문단에서 생성형 텍스트가 단 하나의 문장만 존재해도 문서 전체가 양성으로 분류되므로, 모델은 어떤 문장이 생성형 문장인지 명시적인 지도 없이 학습해야 한다. 이에 본 연구에서는 안정적인 표현을 학습 유도를 위해 방법론인 mean teacher 준지도학습 기법을 활용한다.



(그림1) Mean-teacher 기법을 활용한 LLM 생성 텍스트 탐지 모델 학습 구성도

먼저, 교차 엔트로피(cross-entropy) 손실 함수와 문단 단위의 weakly labeled 데이터를 이용해 모델의 전체적인 분류 방향성을 지도학습 한다. 그리고, 일관성(consistency) 손실 함수를 활용하여 teacher 모델과 student 모델의 일치성을 높인다. 즉, 레이블이 명확하지 않은 환경에서 teacher 모델이 노이즈에 강건한 표현을 학습하고, 문단 내 존재하는 생성형 텍스트를 식별하도록 학습이 진행될 수 있다. 수식 (1)은 consistency 손실 함수를 나타내며, $f(x,\theta)$에서 x는 입력 텍스트 데이터, θ는 모델의 가중치, $f(\cdot)$는 x와 θ를 이용한 모델 출력을 의미한다.

$$L_{con} = \left\{ \| f(x,\theta'_{teacher}) - f(x,\theta_{student}) \|^2 \right\} \quad (1)$$

다음으로 이와 같이 구성된 consistency 손실 함수에 교차 엔트로피 손실 함수를 가중합하여 수식 (2)와 같이 최종 손실 함수를 구성한다.

$$L = L_{CE} + \lambda(t) L_{con} \quad (2)$$

전체 손실 함수 L은 student 모델의 파라미터의 EMA로 갱신되는 teacher 모델의 예측을 목표로 학습된다. 그리고, ramp-up 가중치 $\lambda(t)$는 학습 초기에 작은 값에서 시작하여 모델 학습 진행됨에 따라 점진적으로 증가하도록 설정한다. 이를 통해 학습 초반에는 지도 신호를 우선시하고 표현이 안정화된 이후에는 weakly labeled 데이터 정보를 더 강하게 반영함으로써 모델 학습의 성능 향상과 안정성을 목표한다.

4. 실험

4.1 실험 환경

본 연구에서는 한국어 LLM 생성 텍스트 판별을 위해 DACON 「생성형 AI(LLM)와 인간: 텍스트 판별」경진대

(표 1) 제안된 mean-teacher 기반 LLM 생성 텍스트 탐지 모델의 성능 비교

	precision	recall	f1-score	accaury	ROC-AUC
TF-IDF+XGBoost	**0.7758**	0.5791	0.4964	0.5938	0.9159
KoELECTRA	0.7344	0.6989	0.6916	0.7055	0.9654
KoELECTRA w/ mean-teacher	0.7277	**0.7022**	**0.6974**	**0.7078**	**0.9746**

회의 학습 데이터를 기반으로 학습, 검증, 평가용 데이터셋을 별도로 구성하였다. 라벨 분포를 분석한 결과, 학습 데이터셋에서는 인간 작성(라벨 0)이 LLM 생성(라벨 1)에 비해 약 12.6배 많아 라벨 불균형이 확인되었다. 이에 평가 데이터셋은 라벨 편향을 완화하고 라벨 간 성능 차이를 검증할 수 있도록, 각각 1,000개씩 무작위 추출하였으며, 전체 평가 데이터셋은 총 25,007개 문장으로 구성하였다. 학습 데이터셋은 평가 데이터셋을 제외한 데이터를 기반으로 구성하였고, 평가 데이터셋과 학습 데이터셋은 분리하여 사용하였다. 검증 데이터셋은 학습 데이터셋에서 분리하여 사용하였으며, 데이터 선택 시 시드 값을 42로 고정하여 실험의 재현성을 확보하였다.

실험에 사용된 하이퍼파라미터 값은 다양한 실험 결과를 통해 얻었으며, epoch의 값은 4, ramp-up의 값은 7로 설정하였을 때, 가장 좋은 결과를 얻었다.

4.2 성능 평가

본 논문에서 제안한 KoELECTRA 텍스트 인코더와 mean-teacher 준지도학습 기법을 결합한 모델의 성능 비교를 위해 TF-IDF와 XGBoost를 기반한 판별 모델(TF-IDF+XGBoost), KoELECTRA 텍스트 인코더 fine-tuning만 진행한 판별 모델(KoELECTRA)와 비교를 진행한다. 성능 지표로는 정밀도(percision), 재현율(recall), f1-score, accuary(정확도), ROC-AUC 지표를 사용하였다.

표 1은 본 논문에서 제안한 mean-teacher 기반의 KoELECTRA와 다른 모델과의 성능 비교를 진행한 결과이다. 표에서 볼 수 있듯이 본 논문에서 제안한 학습 방법은 다른 모델들과 비교하여 recall, f1-score, accuracy, ROC-AUC에서 가장 높은 점수를 달성했다. 이는 weakly labeled 데이터의 불확실성과 불균형 문제를 효과적으로 완화한 것으로 분석된다.

5. 결론

본 논문에서는 한국어 환경에서 LLM이 생성한 텍스트를 판별하기 위한 딥러닝 기반 방법을 제안하였다. KoELECTRA 기반 텍스트 인코더의 언어적 맥락 이해 능력과 mean-teacher 학습 방법의 weakly labeled 데이터 학습 능력을 결합함으로써 향상된 LLM 생성 텍스트 탐지 성능을 달성했다.

하지만, 학습 데이터의 클래스 불균형 문제는 여전히 성능 개선의 중요한 과제로 남아 있다. 이에 따라 향후 연구에서는 focal 손실 함수와 같은 불균형 대응 손실 함수를 적용하여, 소수 클래스에 대한 분류 성능을 개선하고 모델의 강건성을 한층 강화할 필요가 있다.

감사의 글

본 연구는 2024년도 연구개발특구진흥재단의 '지역의 미래를 여는 과학기술 프로젝트' 사업(과제 고유번호: 2022-DD-UP-0312) 및 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업 지원을 받아 수행되었음(2024-0-00062).

참 고 문 헌

[1] B. Porter and E. Machery, "AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably," *Sci. Rep.,* vol. 14, no. 26133, Nov. 2024.

[2] 장원익, 장현종, 허의남, "KoboNet: 한국어 AI 생성 텍스트 탐지를 위한 딥러닝 모델," *한국정보과학회 학술발표논문집*, 제주, Jul. 2025.

[3] 박대서 and 김화종, "TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안," *한국정보기술학회논문지*, vol. 16, no. 2, pp. 1 - 16, 2018.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Advances in Neural Information Processing Systems (NeurIPS),* 2016.

[5] 신민기, 진효진, 송현호, 최정회, 임현승, and 차미영, "KoELECTRA를 활용한 챗봇 데이터의 혐오 표현 탐지," in *Proc. 33rd Conf. on Hangul and Korean Language Information Processing*, pp. 518 - 523, 2021.

[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS),* Apr. 2018.

[8] J. Xie, J. Liu, and Z.-J. Zha, "Label noise-resistant mean teaching for weakly supervised fake news detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, Jun. 2022.

ConvNeXt-V2 아키텍처를 활용한 경량화된 음향 장면 분류 연구

조민식, 한사랑, 이건우*, 전찬준
조선대학교 AI소프트웨어학부
e-mail : {mswd81, 6002tkfkd, geonwoo, cjchun}@chosun.ac.kr

A Study on Lightweight Acoustic Scene Classification Using the ConvNeXt-V2 Architecture

Min Sik Jo, Sarang Han, Geon Woo Lee*, Chanjun Chun
School of AI software, Chosun University

요 약

음향 장면 분류 기술은 주어진 오디오 신호에서 환경을 인식하는 기술이며, 웨어러블 및 모바일 장치에서 실시간 동작을 위해 경량화된 탐지 기술이 요구된다. 이에따라 최근 연구에서는 낮은 연산량을 가지면서 우수한 탐지 성능을 갖는 신경망 구조 설계 연구가 주목받고 있다. 본 논문에서는 음향 장면 분류에서 경량화된 아키텍처로 우수한 성능을 나타내는 TF-SepNet 구조에 ConvNeXt-V2에 사용되는 경량화된 블록 구조를 활용한 새로운 음향 장면 분류를 위한 TF-SEpNeXt을 제안한다. ConvNeXt-V2에 활용되는 경량화된 합성곱 신경망 기반의 블록을 TF-SepNet의 시간-주파수 분석 모듈에 적용하여 음향 장면 분류 성능 향상을 목표한다. 제안된 TF-SEpNeXt은 DCASE 2025 데이터 셋을 활용하여 성능 평가를 진행하였으며, 기존 TF-SepNet 아키텍처와 비교하여 모델 파라미터 수는 감소한 채 향상된 분류 성능을 달성하였다.

1. 서 론

음향 장면 분류 기술은 주어진 오디오 신호를 사용하여 오디오 신호에 포함된 음향 환경을 분류하는 문제로 웨어러블 및 모바일 장치를 통해 실시간 환경을 인식하여 사용자 맞춤형 서비스 제공, 로봇 및 자율 주행 차량의 상황 인식, 그리고 환경 모니터링 및 소음 관리 등 다양한 분야에 적용되고 있다. 또한, 전기전자공학자협회(IEEE)가 주관하는 국제 경진대회인 DCASE는 매년 개최되고 있으며, 해당 경진대회를 통해 음향 장면 분류 기술은 꾸준히 연구되고 있다. 특히, 음향 장면 분류 기술이 실제 환경에서 활용되기 위해서는 실시간 처리가 필수적이며, 이에 따라 효율적이고 낮은 연산량을 갖는 기술이 요구된다.

딥러닝 기술은 음향 장면 분류 기술의 성능을 큰 폭으로 향상시켰으며, 특히, 시간-주파수 특징을 분석하기 위해 주로 합성곱 신경망 (convolutional neural network CNN) 기반의 아키텍처들이 연구되고 있다. 다양한 연구 중에서 MobileNet 아키텍처를 기반을 둔 CP-Mobile[1] 아키텍처는 연산량을 줄이기 위해 ConvNeXt-V2[2] 아키텍처에서 활용된 global response normalization (GRN) 기법을 추가한 블록 구조를 활용하여 연산 효율과 분류 성능을 동시에 개선하였다.

그리고, 최근에는 시간 축과 주파수 축을 분리하여 분석하도록 설계된 TF-SepNet[3] 아키텍처 기반의 음향 장면 분류 기술은 우수한 성능을 보여주고 있다. 기존 시간 축과 주파수 축을 동시에 처리하던 전통적인 방법 대신 두 축에서 서로 다른 정보 패턴을 분석하는 병렬 1차원 합성곱 연산이 활용되었다. 이와 같은 구조를 통해 TF-SepNet 기반 음향 장면 분류 기술은 낮은 연산량으로도 우수한 분류 성능을 달성할 수 있었다.

본 논문에서는 경량화된 음향 장면 분류 기술 연구를 위해 TF-SepNet 아키텍처에 ConvNeXt-V2 아키텍처의 블록 모듈을 활용한 TF-SEpNeXt을 제안한다. 제안된 기술은 TF-SepNet 아키텍처와 비교하여 파라미터가 감소하면서도 상대적으로 크게 개선된 음향 장면 분류 성능을 나타낸다.

2. 관련 연구

본 장에서는 본 논문에서 제안하는 TF-SEpNeXt 기반 음향 장면 분류 모델의 기반이 되는 TF-SepNet과 ConvNeXt-V2 아키텍처에 대한 분석을 진행한다.

2.1 TF-SepNet

음향 장면 분류에서 경량화된 모델 구조로 우수한 성능을 나타내는 TF-SepNet의 아키텍처는 입력 Mel 스펙트로그램으로부터 음향 특징을 추출하는 특징 추출기와 추출된 특징을 분석하는 TF-SepConvs 블록으로 구성되어 있다. 먼저, 특징 추출기는 입력 Mel 스펙트로그램에 대해 스트라이드 2, 패딩 1을 갖는 3×3 합성곱 연산과 배치 정규화, ReLU 활성화 함수로 구성된 2차원 합성곱 블록을 2회 거치면서 입력 멜 스펙트로그램의 특징을 추출한다. 두 번째 블록에는 그룹화 합성곱 연산을 도입하여 연산량을 낮추는 동시에 그룹간 채널의 특징을 효과적으로 추출한다.

* 교신저자

특징 추출기에서 추출되는 특징 벡터는 이후 TF-SepConvs 블록의 입력으로 사용된다. TF-SepConvs 블록은 시간-주파수 축을 분리하고, 각 축에 대해 1차원 합성곱 신경망을 연결하여 시간 및 주파수에 대한 특징을 각각 추출한 다음, 다시 결합한다. TF-SepConvs 블록은 여러 개의 연속적인 블록으로 구성되어 있으며, 반복적으로 채널을 확장하며 다운샘플링 계층을 거쳐 출력의 크기를 감소시키고, 특징을 추출한다.

2.2 ConvNeXt-V2

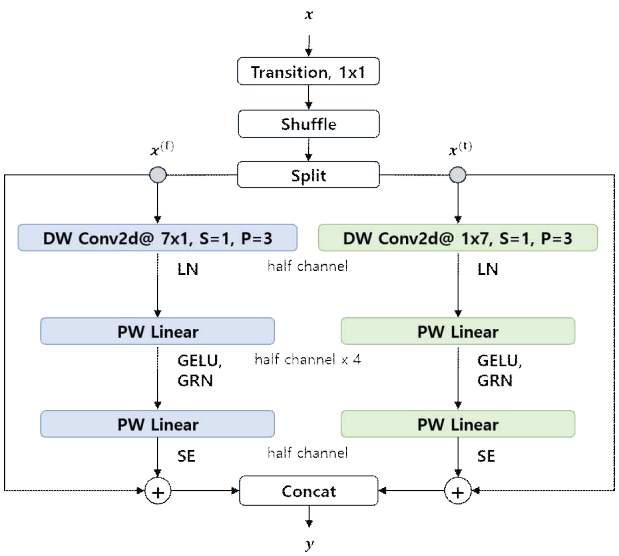
ConvNeXt는 기존 ResNet과 같은 표준 합성곱 신경망 아키텍처를 바탕으로 하되 큰 커널 사이즈, 정규화 이후 활성화 함수를 배치하여 트랜스포머의 설계 요소를 도입하여 전역성과 합성곱 신경망의 계산 효율성, 안정적 학습 및 표현력 향상을 이뤄냈다. ConvNeXt 블록은 7x7의 크기의 넓은 수용영역을 가진 depth-wise Convolution을 적용하여 각 채널별 공간적 특징을 전역적으로 추출한다.

그다음 2번의 point-wise 합성곱 연산을 통해 채널을 4배로 확장하여 다양한 특징을 학습할 수 있고 이후 채널을 축소하여 효율적으로 정보를 압축한다. 블록 내부 합성곱 연산 사이에 layer normalization, GELU 활성화 함수, Layer Scale 기법을 도입하여 학습 안정성 및 특성의 표현력을 높여 성능을 향상시켰다. ConvNeXt-V2 블록에서는 기존 ConvNeXt의 layer scale을 제거하고 대신 global response normalization(GRN) 기법을 도입하여 구조를 단순화하고 성능을 향상시켰다.

3. TF-SEpNeXt 기반 음향 장면 분류 모델

3.1 특징 추출

제안된 모델에 입력되는 오디오 신호는 연산 효율과 일관된 처리 성능 확보를 위해 32kHz로 재샘플링하여 입력으로 사용한다. 오디오 신호는 약 96ms 길이의 구간으로 분할되고, 각 구간은 약 15.6ms 간격으로 중첩이 발생하도록 설정한다. 그다음, Hann window를 적용한 뒤 크기 4096의 fast Fourier transform(FFT)을 통해 시간-주파수 도메인의 특성을 갖는 스펙트로그램으로 변환한다. 마지막으로 인간의 청각 특성 반영을 위해 512개의 Mel 필터뱅크와의 내적 연산을 통해 로그 Mel 스펙트로그램을 변환하고 이를 특징 추출기에 사용한다.



(그림1) 제안된 TF-SEpNeXt 아키텍처 구성도

3.2 TF-SEpNeXt 모델 아키텍처 구조

본 논문에서 제안하는 TF-SEpNeXt 아키텍처는 TF-SepNet를 기반으로 하고 있으며, 기존 TF-SepConvs 블록에 ConvNeXt-V2에서 사용되는 블록 구조를 결합한다. 이를 통해 각 축의 전역적인 특징을 추출하도록 설계하였다. 그림 1은 TF-SEpNeXt 아키텍처를 나타내며, 로그 Mel 스펙트로그램으로부터 음향 장면 분류 예측을 진행한다. 모델의 입력데이터 x는 1x1 transition 연산, shuffle, split 연산을 거쳐 시간-주파수 축으로 분할된다.

제안된 TF-SEpNeXt의 블록은 시간-주파수 축에 각각 병렬 1차원 합성곱 연산을 기존보다 더 넓은 크기의 커널을 적용한다. 즉, depth-wise 합성곱 연산, 2회의 point-wise 합성곱 연산을 적용하고, 잔차 연결부에는 drop-path와 각각의 시간, 주파수 특징 벡터에서 채널 중요도를 반영할 수 있는 squeeze-and-excitation[4]을 도입한다. 이를 통해 축별 특징 분리 및 추출이 효과적으로 이루어지고, 중요한 채널의 정보가 강조되며 경로의 다양성과 안정적 학습을 목표한다.

기존 layer scale과 ReLU 활성화 함수를 GRN과 GELU 활성화 함수로 대체하여 부드러운 비선형성을 통해 학습의 안정성을 향상시킬 수 있다. 이와 같은 과정은 수식(1)과 같이 나타낼 수 있다.

$$x_i = \gamma \cdot x_i \cdot N(G(x)_i) + \beta + x_i \qquad (1)$$

수식(1)과 같이 GRN은 각 채널의 공간적 특성 맵에 대해 L2 노름을 계산하고, 이를 채널 간 평균으로 정규화하여 각 채널의 상대적 중요도를 동적으로 모델링하는 기법이다. 이후 학습 가능한 파라미터 γ와β를 이용해 원본 특징 맵을 보정하여 동작한다.

공간 크기를 감소시키는 기존 max pooling으로 구성되었던 다운샘플링 계층을 2X2 합성곱 연산, 배치 정규화, GELU 활성화 함수로 구성된 계층으로 대체함으로써 학습 가능한 커널을 활용한다. 이를 통해 다운샘플링 계층의 연산량과 파마리터 수는 소폭 상승하였지만 입력의 주요 특성 정보를 효율적으로 추출하며 일반화 성능을 향상 시킬 수 있었다.

제안된 TF-SEpNeXt 아키텍처는 다양한 기법을 도입하여 복잡도가 대폭 증가하였지만, base channel을 기존의 절반으로 감소시켰으며, 개선된 다운샘플링 계층의 배치를 조절하며 성능과 연산량 및 파라미터 수와의 균형을 맞춰 구현했다.[1)]

4. 성능 평가

4.1 실험 환경

4.1.1 데이터 세트 구성

본 논문에서는 DCASE 2025 Task 1[5]에서 사용된

1) https://github.com/minsiter02/TF-SEpNeXt

TAU Urban Acoustic Scenes 2022 Mobile을 데이터셋을 활용하여 실험을 진행하였다. 데이터셋은 공원, 도로, 공항 등 10가지 다른 음향 장면에서 1초 길이의 단일 채널, 24비트 44.1kHz로 샘플링 레이트로 휴대기기 4종 및 가상 장치 10종으로 기록되었다. 전체 데이터 볼륨은 약 64시간이며, 실험에서는 DCASE 공식 분할 방식에 따라 휴대기기 3종 및 가상 장치 6종이 녹음한 데이터를 포함하여 전체 데이터 중 1/4인 약 18시간 볼륨의 25% 분할된 개발 데이터 세트를 학습 및 검증용으로 활용하였다.

4.1.2 매개변수 설정

TF-SEpNeXt 및 TF-SepNet 모두에 대해 mix-up, mix_style, DIR와 같은 데이터 증강 기법이 포함되었다. 최적화는 Adam 옵티마이저를 사용하였으며 손실 함수로는 교차 엔트로피 손실을 채택하였다. 학습은 고정된 epoch 90, 128의 배치 크기를 사용했다.

그리고 TF-SEpNeXt 의 분리된 시간-주파수 1차원 합성곱 연산은 스트라이드 1, 각 축에 대해 커널 크기 7, 패딩 3으로 설정하였다. Drop path 비율과 squeeze-and-excitation 블록 내 중간 채널 비율을 각각 0.25로 설정하였고 point-wise 합성곱 연산의 채널 확장 비율을 4배로 설정하였다. 또한, 다운샘플링 계층 사이 CNS-TF-SepNet의 블록을 2, 2, 6, 2회 반복적으로 배치하였으며, 개선된 합성곱 기반 데이터 세트로 계층은 초기 입력층 이후에 한 차례 적용하였다.

4.2 성능 평가

본 연구에서는 제안한 개선된 TF-SEpNeXt 아키텍처와 기존 TF-SepNet의 음향 장면 분류 성능을 DCASE 2024 Task 1의 녹음 장치별 구분 없는 조건에서 비교 평가를 진행했다. 모든 실험은 동일 조건에서 5회 반복 수행하였으며, 최종 결과는 5회 실험 결과의 평균값을 기준으로 하였다.

(표 1) DCASE 2025 Task1 25% 분할된 데이터 세트 실험 결과

	TF-SepNet		TF-SEpNeXt	
	평균	표준편차	평균	표준편차
Accuracy(%)	55.70	0.40	**57.96**	0.27
F1-Score(%)	55.49	0.33	**57.26**	0.29
Precision(%)	55.77	0.28	**57.30**	0.42
Recall(%)	55.70	0.40	**57.96**	0.26
MACs	29.4M	-	**26.4M**	-
Prams.	126.9K	-	**123.5K**	-

표 1에서 보는 바와 같이 본 논문에서 제안한 TF-SEpNeXt 아키텍처는 기존 TF-SepNet와 대비 연산량 및 파라미터 수가 감소하였음에도 불구하고 모든 지표가 향상된 음향 장면 분류 성능을 보여준다.

5. 결론

본 논문에서는 두 축에서 서로 다른 정보 패턴을 분석하는 기존 TF-SepNet 아키텍처에 ConvNeXt-V2의 최신 합성곱 신경망 기법을 융합하여 음향 장면 분류를 위한 새로운 아키텍처를 제안하였다. DCASE 2025 Task 1 데이터 세트를 활용하여 신경망을 학습하고 평가한 결과 제안된 아키텍처가 기존 아키텍처 대비 음향 장면 분류 성능 결과를 향상시켰다.

향후 연구에서는 본 아키텍처의 경량화 및 효율성 강화를 위해 최근 트랜스포머의 경량화를 위해 사용되는 다양한 구조를 적용하여 음향 장면 분류 성능을 향상시킬 수 있을 것으로 기대한다. 또한, 프루닝 및 양자화와 같은 경량화 기법을 통해 실제 환경에서 실행 가능한 데모 시스템을 구축할 예정이다.

감사의 글

본 연구는 2024년도 연구개발특구진흥재단의 '지역의 미래를 여는 과학기술 프로젝트' 사업(과제 고유번호: 2022-DD-UP-0312) 및 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업 지원을 받아 수행되었음(2024-0-00062).

참 고 문 헌

[1] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "CP-JKU Submission to DCASE23: Efficient Acoustic Scene Classification with CP-Mobile," *DCASE Challenge Tech. Rep.,* 2023.

[2] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16133 - 16142, 2023.

[3] Y. Cai, P. Zhang, and S. Li, "TF-SepNet: An Efficient 1D Kernel Design in CNNs for Low-Complexity Acoustic Scene Classification," in P*roc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2024, pp. 1 - 5, doi: 10.1109/ICASSP48485.2024.10447999.

[4] J. Huang, W. Fan, Y. Qiao, X. Zhu, and R. M. An, "Fast Contextual Text Recognition with Deep Convolutional Sequence Learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132 - 7141, doi: 10.1109/CVPR.2018.00748.

[5] DCASE 2025 Challenge Task 1 Team, "DCASE 2025 Task 1: Acoustic Scene Classification Challenge," *DCASE Challenge*, 2025. [Online]. Available: https://dcase.community/challenge2025/task1

YOLO 모델 기반의 지하철 응급상황 감지: 버전별 특성 및 성능 비교

최세영[1], 김지현[1], 김세진[2]*
조선대학교 전산통계학과[1], 조선대학교 컴퓨터통계학과[2]
e-mail : sychoi@chosun.ac.kr, kjh67899@chosun.ac.kr, sjkim@chosun.ac.kr

YOLO Model based Subway Emergency Detection: Comparison of Version-Specific Characteristics and Performance

Seyeong Choi, Jihyeon Kim, Se-Jin Kim**
Dept of Computer Science & Statistics, Chosun University

요 약

최근 범죄 예방, 증거 수집, 그리고 신속한 상황 파악 및 대응을 위해 다양한 환경에 CCTV가 설치되고 있지만, 관리자의 수동모니터링 방식으로 운영되어 효율적인 측면에서 한계가 있다. 본 논문에서는 실시간 객체탐지 모델 YOLO를 이용하여 지하철 응급상황을 인공지능이 감지하고 대응할 수 있는 연구를 진행한다. 먼저, 대표적인 YOLO 버젼에 해당하는 v5, v8, v11, v12 의 구조와 차이점을 소개하고, AI-Hub의 지하철 역사 내 CCTV 이상행동 영상데이터를 이용하여 각 YOLO 버젼의 객체탐지 성능을 비교한다. 지하철 응급상황 데이터에서 실신과 배회를 타겟 클래스로 정의하였고, 4가지 YOLO 버젼이 모두 높은 성능을 나타냈다. 결과적으로 YOLO v11 은 mAP@50 과 mAP@50:95 성능분석 결과에서 각각 0.99와 0.908 로 가장 높은 성능이었고, YOLO v5 는 가장 낮은 성능을 보였다.

1. 서 론

최근 공공장소에서 시민의 안전을 위해 지하철 내부, 승강장, 역사 내 다양한 곳에 CCTV(Closed-Circuit Television)가 설치되어 운영되고 있다. 통계청에 따르면 2022년을 기준으로 전국에 1,960만대의 CCTV 가 설치(공공 기관: 160만대 이상 운용)되었고, 매년 CCTV 설치 수가 급격히 증가하여 현재 2,000만대를 초과했을 것으로 예상된다[1]. CCTV는 시민 보호를 위한 사고 및 사회 범죄 예방에 기여하지만 대부분 수동 모니터링 방식으로 인해 여러 한계가 드러나고 있다. 반복되는 장기간의 모니터링 근무는 신경피로에 노출될 가능성을 높여[2][3] 집중력 저하 및 반응 지연이 발생할 수 있으며, 이로 인해 순식간에 발생하는 예측 불가능한 이상행동이 발생할 시 놓치는 경우가 많다. 특히 빠르고 신속한 대처가 필요한 응급상황은 초기 대처가 중요한데, 환자의 사망과 불구를 최소화할 수 있다. 응급상황 중 실신은 다양한 연령대에서 일시적 의식 소실로 적절한 치료를 받지 못할 시 합병증이 발생할 수 있다[4]. 2024년 치매 추정 인구수는 954,789명으로 추정되는데[5] 배회는 치매환자에게서 흔히 발생하는 요인으로 더 나아가 부정적인 결과로 이어지면 사망과 관련이 있다[6][7]. 선행연구에서는 보행자가 갑자기 쓰러지는 실신 상황을 자동으로 탐지하고 그 위치를 지도상에 표시하여 응급 구조 기관에 연락하는 시스템을 구축하였다[8].

본 연구의 목적은 즉각적인 응급 대응 시스템을 구축하는 것이 아닌 지하철 내 환경에서 배회와 실신이라는 두 분류의 이상 행동을 정의하고 YOLO 모델의 여러 버전의 구조를 분석하고 적용하여 그 탐지 성능을 비교 및 분석하는 것이다. 따라서 YOLOv5, YOLOv8, YOLOv11, YOLOv12 총 4가지의 모델을 사용할 것이고 parameter가 적지만 속도는 빠르고 실시간 검출에 적합한 nano모델을 사용하려고 한다. 또한 객체를 바운딩박스로 검출하는데, 바운딩박스의 크기와 위치가 얼마나 정확한지 고려하기 위한 mAP지표로 성능을 비교할 것이다.

2. 사전 연구

2.1 데이터셋

본 연구에서 사용할 데이터는 AI-Hub에서 제공하는 지하철 역사 내 CCTV 이상행동 영상데이터를 사용했다. AI-Hub는 과학기술정보통신부와 한국지능정보사회진흥원이 운영하는 국가 AI 개발 지원 플랫폼으로 다양하고 많은 양의 데이터를 제공한다. 이 데이터는 재난 안전 환경에 분류되어 이상상황 판별을 통해 범죄를 줄이는 목적을 제시하고 있다. 남성 6명, 여성 6명으로 10대부터 60대의 다양한 연령의 인원들이 각종 이상행동들을 재연한 데이터로 4k화질과 30fps를 제공한다[9].

2.2 모델 구조 분석

* 교신저자

2.2.1 YOLO 모델

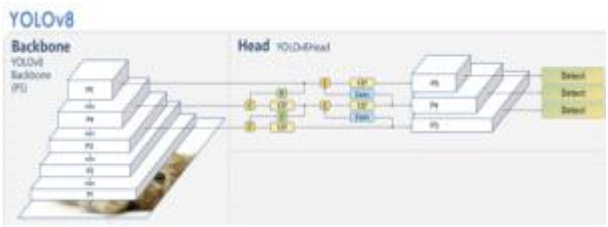
YOLO모델은 분류기를 재정의하여 검출기로 사용하는 R-CNN등과 같은 기존모델과는 다르게 객체 검출을 회귀 문제로 재정의한 모델이다. 파이프라인 하나로 이미지 내의 물체가 어떤 것인지, 어디에 있는지를 구하기 때문에 모델의 이름이 YOLO(You Only Look Once)다. YOLO는 Ultralytics에서 공개한 객체탐지모델로 연산이 빠르고 정확도가 뛰어나며, 패키지를 제공하기 때문에 누구나 쉽고 편하게 사용할 수 있다[10]. 여러 버전의 YOLO에 대해 구조와 특징을 간단히 설명하려고 한다.

2.2.1 YOLOv5

2020년에 출시된 YOLOv5는 이전에 출시된 YOLOv4와 같은 모델에 비해 구조가 가볍고, 정확도가 높아진 모델이다. 이미지의 특징을 추출하는 Backbone 네트워크는 CSP-Darknet를 사용하며, BackBone과 Head를 연결하는 Neck 구조에서는 SPPF 및 PANet를 사용하고, Head는 Neck에서 융합된 특징을 받아 예측을 수행한다[11].

2.2.2 YOLOv8

2023년에 출시된 YOLOv8은 이전 버전들을 기반으로 속도와 정확도 측면에서 성능이 크게 오른 모델이다.



(그림 1) YOLOv8 구조

그림 1은 YOLOv8의 구조다[12]. Backbone에서 CSP-Darknet기반이지만 C2f모듈을 사용하여 속도를 유지하면서 성능을 높였다. Neck구조에선 PANet과 FPN을 결합하여 다양한 크기의 객체를 정확하게 탐지할 수 있으며, Head에서 이전 모델과는 다르게 앵커 방식을 사용하지 않음으로써 모델을 단순화하였다[13].

2.2.3 YOLOv11

YOLOv11은 2024년에 공개되었고 이전 버전인 YOLOv7과 YOLOv8에 비해 mAP와 FPS측면에서 성능이 뛰어나다. Backbone에서 기존 YOLOv8모델의 C2f를 개선한 C3k2를 사용하여 계산 속도를 빠르게 하며 복잡한 이미지의 특징을 효과적으로 추출한다. Neck구조에서 PANet과 C2PSA를 결합하여 이미지 내의 중요한 영역을 집중하여 다양한 크기의 객체를 효과적으로 탐지한다. Head에선 YOLOv8과 동일하게 사전에 앵커를 정의하지 않는다[14].

2.2.4 YOLOv12

2025년에 공개된 YOLOv12는 이전 버전에서의 연산복잡성과 메모리접근의 비효율성을 가진 Attention메커니즘을 개선하였다. 기존모델의 Backbone에서 ELAN(Efficient Layer Aggregatior Network)를 사용했다면, YOLOv12는 R-ELAN을 사용하여 모델을 최적화하고 안정성을 높였다. Neck구조에서는 A2(Area Attention)모듈을 적용하여 특징맵을 수평이나 수직으로 분할하여 Attention을 적용시킨다. 계산 비용을 줄이며 넓은 수용력의 효과를 불러온다. Head에서 FlashAttention 기술을 통합하여 처리 속도를 높였고, 메모리 접근을 효율적으로 수행한다[15].

2.3 성능평가지표

각 버전의 성능을 비교하기 위해 선정한 성능평가지표로는 mAP@50과 mAP@50:95를 사용했다. 먼저 AP는 Precision-Recall 곡선 아래 면적이며, mAP는 각 클래스 AP의 평균이다.

$$mAP_{50} = \frac{1}{N}\sum_{i=1}^{N} AP_{i,IoU=0.5} \quad (1)$$

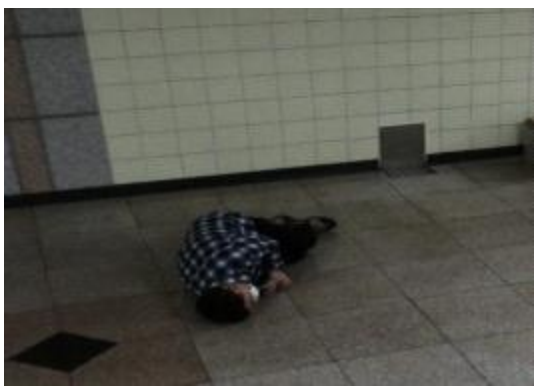
식 1은 mAP@50의 식이며, 이는 IoU 0.5기준의 mAP다. mAP@50:95는 IoU를 0.5부터 시작하여 0.05씩 증가시키며 0.95까지 나온 값들의 평균이다. 두 평가지표 모두 객체 탐지 모델의 성능평가지표로 주로 사용하며, 1에 가까울수록 성능이 뛰어나다[16]. mAP@50과 mAP@50:95 두 가지의 성능지표를 사용하여 예측한 바운딩박스가 정확한 위치에 알맞은 크기를 가지는지 평가하려고 한다.

3. 실험 및 분석 or 실험결과

3.1 데이터 수집 및 하이퍼파라미터 설정

이상행동 13종을 대상으로 7030개의 영상에서 약 100만장의 이미지를 추출한 데이터다. 폭행, 절도, 전도 등의 여러 이상행동 중 응급상황과 관련된 클래스인 배회와 실신을 우선적으로 선정하였다.

(그림 2) 배회 이미지



(그림 3) 실신 이미지

그림 2는 배회(Wandering)하는 인물의 이미지며 그림 3은 실신(fainting)한 인물이다. 이미지의 중앙에 위치하고 있고 객체가 크다. 이미지에 다른 인물은 존재하지 않다. 그림 2의 객체는 작고, 그림 3의 객체는 크다는 차이가 있다. 데이터셋은 학습데이터 약 73,000장, 검증데이터 약 34,000장으로 약 7:3의 훈련데이터와 검증데이터의 비율로 구성했다. 모델 학습을 진행하였다.

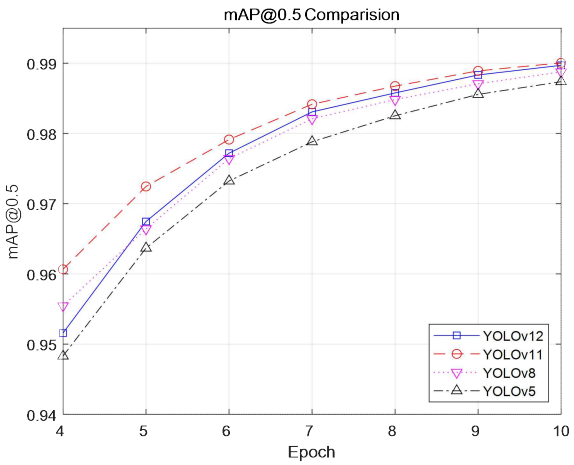
(표 1) 하이퍼파라미터 설정값

Parameter	Value
Image size	640*640 (Pixes)
Batch	16 (개)
Epoch	10 (회)
Optimizer	SGD
Learning rate	0.01
Momentun	0.9

각 버전의 비교를 위해 하이퍼파라미터는 통일했다. 표 1은 하이퍼파라미터 설정 값이다. 이미지 크기는 640*640, 배치 크기는 16, Epoch은 10, 옵티마이저는 SGD를 사용하며[17], 학습률은 0.01, 모멘텀은 0.9로 설정했다.

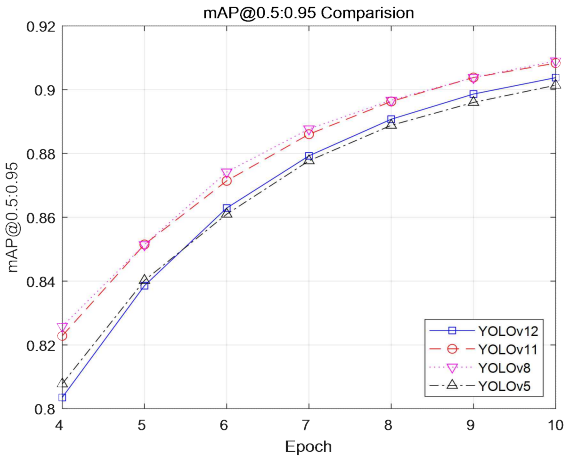
3.2 성능평가 및 비교

모델의 성능을 비교하기 위해 mAP@50과 mAP@50:95의 그래프를 그려보았다. X축은 epoch이며 Y축은 각각 mAP@50과 mAP@50:95이고, 그에 따른 변화를 보여준다. 모델의 선의 색과 모양을 다르게 하여 차이를 두었다.



(그림 4) mAP@50

그림 4는 모델별 mAP@50의 그래프이다. 모든 모델이 1에 수렴하는 결과를 얻었으며. 비교적 최근에 출시된 YOLOv11과 YOLOv12가 더 좋은 성능을 보이지만 다른 모델들도 mAP@50이 1에 근접하는 성능을 보인다. 구체적으로 Epoch 10에서 YOLOv5는 0.987, YOLOv8은 0.988의 성능으로 큰 차이가 없으며, YOLOv11은 0.99로 뛰어난 정확도를 보인다. YOLOv12는 0.989의 성능을 보인다. 4개의 모델 모두 0.99에 근접하며, 이는 검출능력이 좋다는 것을 알 수 있다.



(그림 5) mAP50:95

그림 5는 모델별 mAP@50:95의 그래프이다. YOLOv5는 0.901이고, YOLOv8은 0.909로 가장 높았으며, YOLOv11은 0.908의 성능을 보인다. 반면에 최신 모델인 YOLOv12는 0.903으로 비교적 낮은 모습을 보인다. 다른 모델과 큰 차이는 없지만, 데이터 특성상 큰 객체부터 작은 객체까지 다양한 크기의 객체를 검출해야 한다. A2의 특성으로 영역을 수평이나 수직으로 분할하여 Attention함으로써 작은 객체를 검출하는데 실패한 것이라고 추정한다. 이는 최신 모델이라고 무조건 성능이 뛰어난 것이 아닌 데이터의 특성과 특정 탐지 목적에서 추가적인 개선이 필요하다는 것을 시사한다.

3.3 예측 결과

mAP@50과 mAP@50:95에서 좋은 성능을 보인 YOLOv11로 예측을 수행했다. 그림 6과 그림 7은 예측을 수행한 결과이미지이고, 바운딩박스로 객체를 탐지하고 박스 상단에 클래스이름과 예측신뢰도를 보여준다.

(그림 6) 배회 예측 결과

(그림 7) 실신예측결과

그림 6과 그림 7은 각각 배회하는 인물과 실신한 인물을 바운딩박스로 검출하는 모습을 시각적으로 표현한 이미지다. 그림 6의 예측신뢰도는 0.92며, 그림 7은 0.94로 신뢰할 수 있는 수준이다. 하지만 이미지마다 객체의 크기가 다양하므로 예측이미지의 결과 또한 그에 맞게 조정이 필요하다. 이번 연구는 모델 구조를 분석하고 성능 비교에 중점을 두었으나, 실제상황에 적용이 되면 한눈에 볼 수 있게 영상 확대 기능이나 바운딩박스의 색을 다르게 하거나, 알림 창을 여는 등의 조치가 필요할 것으로 보인다.

4. 결론

본 연구에서는 지하철 내 응급상황 검출을 위해 YOLO 버전 별로 구조를 분석하고, 모델을 학습시키고 성능평가를 해보았다. 그 결과 4개의 모델 모두 좋은 성능을 보였고 특히 최근에 나온 YOLOv12보다 YOLOv11이 평가지표를 기준으로 우수한 성능을 보였다. 이는 최근에 나온 모델이 보편적으로는 더 좋을 수도 있지만 특정 분야에서나 데이터의 특성 등 다양한 이유로 성능에 차이가 날 수 있음을 보여준다. 하지만 mAP@50과 mAP@50:95의 지표가 차이가 약 0.08정도 차이가 난다. 이는 모델이 객체 탐지능력은 우수하나, 정확한 바운딩박스의 예측은 부족하다고 볼 수 있다.

향후 연구로 하이퍼파라미터를 최적화시키고, 손실함수를 변경해보거나 모델의 구조에 변화를 줘서 성능을 향상시킬 계획이다. 또한 4개의 모델을 하나의 컴퓨터로 학습시키기 때문에 Epoch을 낮게 설정하였다. Epoch을 더 높이면 모델이 안정될 것으로 생각한다. 또한, 응급상황뿐만 아니라 범죄로 분류되는 폭행이나 절도, 몰래카메라와 같은 클래스를 추가하여 이상행동의 범위를 더 넓혀가고 실시간 탐지와 작은 객체에 대한 검출을 향상시켜 이러한 기술이 실제로 적용이 될 수 있도록 진행할 계획이다.

참 고 문 헌

[1] 통계인재개발원: 통계의 창, 4세대 CCTV 시대로의 전환, https://shi.kostat.go.kr/window/2024b/main/2024_win_08.html

[2] 권창기, 한승훈, 최동재, 박영진, 김병태, 김병찬. “근무 지속시간에 따른 경계근무와 CCTV모니터링근무의 생체리듬변화 차이 연구”. 『시큐리티연구, 35』, pp. 125-149, 2013.

[3] Caruso, Claire C. "Negative impacts of shiftwork and long work hours." 『Rehabilitation Nursing Journal』, Vol. 39 No. 1, pp. 16-25, 2014.

[4] 최윤영, 최종일. "실신의 임상적 접근 및 진단." 『대한내과학회지』, Vol. 95 No. 4 pp. 251-259, 2020.

[5] 중앙치매센터 치매(오늘은) 2024년 추정치매환자 통계, https://www.nid.or.kr/info/today_list_2024.aspx#a

[6] G Cipriani, C Lucetti, A Nuti, S Danti. “Wandering and dementia.” 『Psychogeriatrics』, Vol. 14 No. 2, pp. 135-142. 2014.

[7] Hope, R. A, Christopher G. Fairburn. "The nature of wandering in dementia: A community-based study." 『International journal of geriatric psychiatry』, Vol. 5 No. 4, pp. 239-245, 1990.

[8] 송인서, 박태정. “CNN 기반 CCTV 동영상 내 보행자 응급 상황 자동 감지 기술 연구”, 『디지털콘텐츠학회논문지』, Vol. 23 No. 3, pp. 371-379, 2022.

[9] AI-Hub https://www.aihub.or.kr/

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 『2016 IEEE Conference on Computer Vision and Pattern Recognition』, pp. 779-788, 2016.

[11] Park Chewon, & Hyung-Sup Jung. “Detection of Urban Trees Using YOLOv5 from Aerial Images.”, 『Korean Journal of Remote Sensing』, Vol. 38 No. 6, pp. 1633-1641, 2022.

[12] D. Reis, J. Kupec, J. Hong, A. Daoudi. “Real-time flying object detection with YOLOv8.” 2023.

[13] L. Zhang, X. Wu, Z. Liu, P. Yu, M. Yang. "ESD-YOLOv8: An Efficient Solar Cell Fault Detection Model Based on YOLOv8", 『IEEE Access』, Vol. 12, pp. 138801-138815, 2024.

[14] Khanam, Rahima, Muhammad Hussain. "Yolov11: An overview of the key architectural enhancements.", 2024.

[15] Tian, Yunjie, Qixiang Ye, David Doermann. "Yolov12: Attention-centric real-time object detectors.", 2025.

[16] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context.", 『European conference on computer vision. Cham: Springer International Publishing』, 2014.

[17] Tian Y, Zhang Y, Zhang H. “Recent Advances in Stochastic Gradient Descent in Deep Learning.”, 『Mathematics』, Vol. 11 No. 3, pp. 682, 2023.

감사의 글

본 연구는 과학기술정보통신부와 정보통신기획평가원(IITP)이 주관하는 대학-기업 협력 소프트웨어 아카데미 국가 프로그램(과제번호: 2022-0-01112)의 지원을 받아 수행되었음.

CAPTION-GUIDED REFINEMENT OF IMAGE REGIONS VIA MASKED GAN TRAINING

Joshua Nyaberi
Department of Computer Engineering, Chosun University
309 Pilmun-Daero, Dong-Gu, Gwangju 61452, Republic of Korea
nyaberi@chosun.ac.kr

Kyungho Yu
Department of AI Convergence, Chosun University
309 Pilmun-Daero, Dong-Gu, Gwangju 61452, Republic of Korea
infinitegh@chosun.ac.kr

Seungjae Lee
Department of AI Software, Computer Engineering, Chosun University
309 Pilmun-Daero, Dong-Gu, Gwangju 61452, Republic of Korea
dl4786@chosun.ac.kr

PanKoo Kim
Department of AI Software, Computer Engineering, Chosun, University
309 Pilmun-Daero, Dong-Gu, Gwangju 61452, Republic of Korea
pkkim@chosun.ac.kr

ABSTRACT

Recent advancements in text-to-image (T2I) generative modeling have significantly improved the ability to create visually realistic images from textual descriptions, marking a key milestone toward human-like artificial intelligence. Despite these remarkable achievements, existing T2I models often suffer from critical issues such as catastrophic negligence, attribute mismatches, and attribute leakage, which hinder their ability to faithfully represent all aspects of the input prompt. To address these limitations, we propose a novel framework that integrates Generative Adversarial Networks (GANs) for gap detection and refinement, aiming to enhance prompt-consistency in T2I synthesis. We propose a semantically aware GAN refinement pipeline that uses caption-generated mismatches and CLIPSeg masks to guide targeted refinement of image regions inconsistent with a given prompt. This approach leads to improved alignment between textual prompts and generated images, resulting in higher fidelity and more accurate visual outputs.

KEYWORDS

Artificial Intelligences (AIs), CLIPSEG, catastrophic negligence, Generative adversarial Network, Prompt consistency, T2I

1. INTRODUCTION

Text-to-Image (T2I) generation represents a rapidly advancing interdisciplinary field bridging natural language processing and computer vision, focused on synthesizing realistic images from descriptive textual prompts. This capability underpins a variety of applications, including digital content creation, virtual and augmented reality, design prototyping, and assistive technologies. Recent progress in generative modeling, notably through diffusion-based approaches such as Stable Diffusion[1],DELL-E 2 have improved the quality, diversity, and realism of generated images.

Despite these advancements, significant challenges persist. Contemporary T2I models often fail to accurately capture all objects and attributes specified in input prompts, resulting in common issues such as object omission, attribute inconsistencies, and incomplete scene rendering. These shortcomings diminish the practical utility of T2I systems and erode user trust. Enhancing semantic fidelity is therefore imperative for domains reliant on automated content generation, including gaming, advertising, and media production, as it reduces dependence on manual post-processing and streamlines creative workflows. Furthermore, advancing model comprehension of complex linguistic inputs contributes fundamentally to the fields of multimodal machine learning and human-computer interaction.

Evaluation metrics such as CLIPScore and Fréchet Inception Distance (FID)[2] substantiate that existing models often lack semantic completeness. While mechanisms like cross-attention and mask cross-attention partially mitigate these deficiencies by improving alignment between textual and visual modalities, they do not fully resolve errors arising from missing details. Diffusion models operate via iterative denoising steps beginning from stochastic noise, a process vulnerable to semantic gaps between the input prompt and the synthesized image[3].
This research contribution includes:

1. This work proposes a novel two-stage framework that integrates a pretrained diffusion model with a GAN-based semantic gap detection and refinement module to enhance prompt fidelity in text-to-image synthesis.

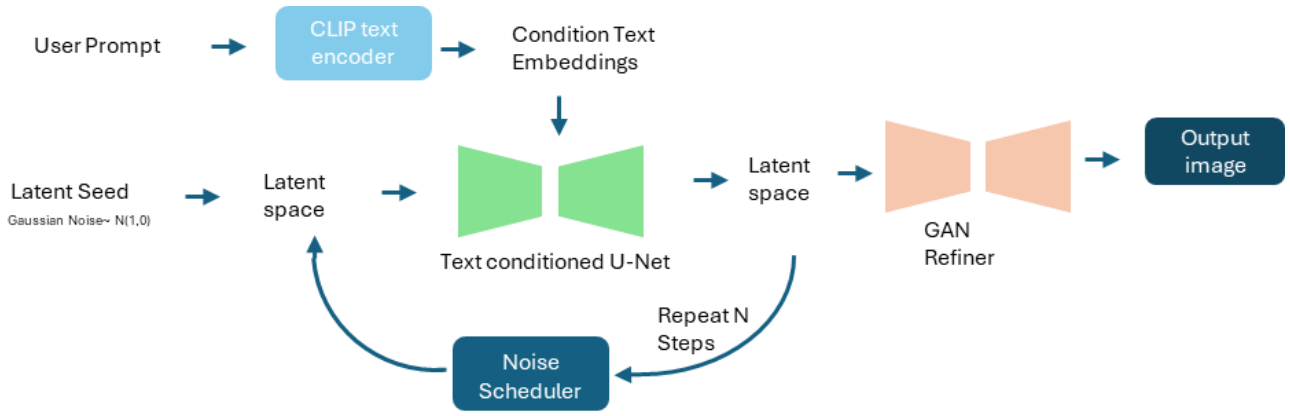


Figure 1: Overview pipeline of the proposed two-stage Image Refinement Pipeline

2. Our approach systematically identifies and corrects semantic inconsistencies, resulting in improved alignment between textual descriptions and generated images without compromising visual quality.

2. RELATED WORK

Generative models have evolved, with early works on Generative Adversarial Networks (GANs)[4] such as StyleGAN, StackGAN[5] and AttnGAN that introduced attention mechanisms to improve text conditioning. These models produced promising results but often lacked global coherence and struggled with high-resolution images. The advent of diffusion models DDPM, DDIM and LDMs[1], [6], [7] marked a new era with superior image quality and diversity. Models like GLIDE[8] and Stable Diffusion generate photorealistic images by iterative denoising conditioned on text. Nonetheless, their prompt fidelity remains imperfect, especially for complex prompts with multiple objects or fine-grained attributes. Studies show that diffusion models may omit or distort parts of the input description, a phenomenon referred to as catastrophic negligence. Researchers have proposed cross-attention and mask cross-attention[9] mechanisms to better align textual tokens with image, but these do not fully resolve missing detail issues. Gap detection using pretrained image captioning models to identify discrepancies between the prompt and generated image is an emerging approach that shows promise but has yet to be integrated systematically into T2I pipelines.

3. PROPOSED METHODOLOGY

3.1 Overview pipeline

We propose a novel two-stage text-to-image generation framework that improves semantic alignment between textual prompts and synthesized images. The pipeline consists of a coarse generation stage followed by a semantic refinement stage. In the first stage, a pretrained diffusion model such as Stable Diffusion is used to generate an initial image conditioned on the input prompt. While this output is typically photorealistic, it may fail to fully capture certain semantic elements of the prompt. To address this, the second stage introduces a semantic discrepancy detection and refinement module that identifies and corrects inconsistencies through mask-guided GAN-based image refinement. Figure 1 shows the overall pipeline.

3.2 Semantic Gap Detection

Semantic gap detection is critical to identifying where the diffusion model fails to capture the prompt's attributes accurately. We employ pretrained image captioning model BLIP[10] to generate a textual description of the synthesized image. This caption is then compared against the original prompt using similarity metrics CLIPScore[11] and BERTScore[12], which quantify semantic alignment at a fine-grained level. Regions of the image that correspond to mismatched or missing attributes are localized by analysing attention maps or token-level similarities between the prompt and the generated caption. This process yields spatial masks highlighting the image areas where semantic inconsistencies are detected, effectively pinpointing the gaps that require refinement. To localize the spatial regions corresponding to these discrepancies, we leverage CLIPSeg, a text-guided segmentation model. Each mismatched token is used as a query to CLIPSeg, which produces a soft segmentation mask indicating where the concept appears (or fails to appear) in the image. The union of these masks forms a final semantic discrepancy map used for refinement.

3.3 Masked Refinement with Conditional GAN

Once the semantic gaps are identified, the corresponding spatial masks are used to guide a masked conditional GAN, which performs localized inpainting and correction on the generated image. The GAN refiner takes as input the coarse diffusion output, the prompt text embedding, and the mask indicating regions needing correction. By conditioning on the prompt and focusing exclusively on the masked areas, the GAN can selectively update image regions to better reflect the intended semantics without altering well-synthesized parts. The generator is trained to produce visually consistent and semantically accurate

refinements, while the discriminator ensures photorealistic quality and coherence across the entire image. The generator is based on a U-Net architecture that allows multi-scale feature fusion between masked and unmasked regions. The discriminator follows the PatchGAN design to enforce local realism and high-frequency consistency. The architecture of Conditional GAN is illustrated in figure 2.

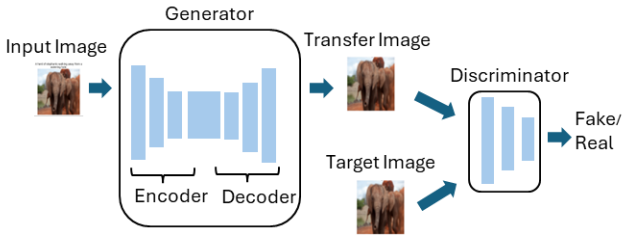


Figure 2: Architecture of the proposed GAN Refiner Module

3.4. Experimental Setup

Once To evaluate the effectiveness of our proposed refinement pipeline, we conduct experiments using the MS-COCO 2017 dataset, a widely adopted benchmark for text-to-image generation and captioning tasks. We utilize the validation split, which contains approximately 5,000 images paired with multiple human-annotated captions. For each image-caption pair, we use the caption as the input prompt for a pretrained Stable Diffusion model to generate the initial coarse image. The generated image is then passed through the BLIP image captioning model to obtain an auto-generated description, which is compared against the original COCO caption using CLIPScore and BERTScore to detect semantic mismatches. Identified mismatched tokens are used to generate spatial masks via CLIPSeg, which are then fed into our conditional GAN for masked refinement. During GAN training, we use COCO images as real samples, and the refined outputs as fake samples. All experiments are conducted on NVIDIA RTX 3090 GPUs, using a batch size of 8 and Adam optimizer with a learning rate of 1e-4. The GAN is trained for 30 epochs, and model performance is evaluated using both quantitative metrics (CLIPScore, FID) and qualitative visual inspection.

4. RESULTS AND DISCUSSION

The training process of the GAN is monitored by tracking the discriminator and generator losses over 10 epochs, as shown in Figure 3. The discriminator loss fluctuates around 0.6 to 0.7, indicating it consistently maintains its ability to differentiate between real and generated images without overpowering the generator. The generator loss varies between approximately 0.8 and 1.17, reflecting the generator's ongoing efforts to produce realistic images that can fool the discriminator. Overall, the loss curves demonstrate balanced adversarial training where both networks improve simultaneously, avoiding issues such as mode collapse or divergence. This steady progression suggests effective learning and convergence of the GAN during the training period.

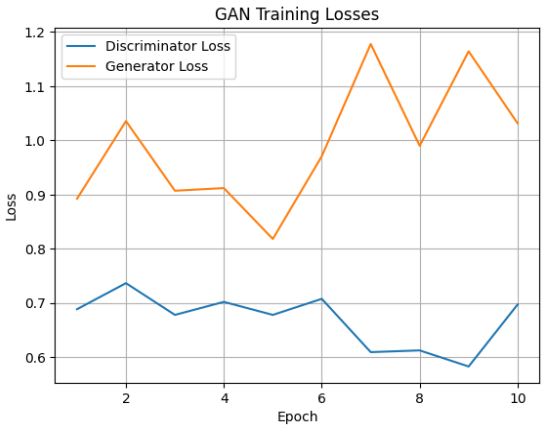


Figure 3: The discriminator and generator losses during training

5. CONCLUSIONS

In this work, we present a novel GAN-based image refinement pipeline that addresses semantic inconsistencies between generated images and their corresponding text prompts. By leveraging a combination of BLIP for caption generation, a diff-based mismatch detection mechanism, and CLIPSeg for token-specific semantic masking, our approach enables targeted refinement of only the regions that deviate from the intended prompt. The generator is conditioned on both the semantic mask and text embedding, allowing it to selectively enhance or correct image regions while preserving the rest. This framework offers a scalable, interpretable, and weakly supervised solution to the problem of semantic misalignment in text-to-image generation, and holds strong potential for applications in content editing, AI art refinement, and prompt-faithful generation.

ACKNOWLEDGMENTS

This research was supported by the Regional Innovation System & Education (RISE) program through the (Gwangju RISE Center), funded by the Ministry of Education (MOE) and the (Gwangju Metropolitan City), Republic of Korea (2025-RISE-05-013) and by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2025" (2024-0-00062).

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 13, 2022, *arXiv*: arXiv:2112.10752. doi: 10.48550/arXiv.2112.10752.

[2] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Jan. 12, 2018, *arXiv*: arXiv:1706.08500. doi: 10.48550/arXiv.1706.08500.

[3] B. Kawar, G. Vaksman, and M. Elad, "Stochastic Image Denoising by Sampling from the Posterior Distribution," Aug. 31, 2021, *arXiv*: arXiv:2101.09552. doi: 10.48550/arXiv.2101.09552.

[4] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 10, 2014, *arXiv*: arXiv:1406.2661. doi: 10.48550/arXiv.1406.2661.

[5] H. Zhang *et al.*, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," Aug. 05, 2017, *arXiv*: arXiv:1612.03242. doi: 10.48550/arXiv.1612.03242.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Dec. 16, 2020, *arXiv*: arXiv:2006.11239. doi: 10.48550/arXiv.2006.11239.
[7] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," Oct. 05, 2022, *arXiv*: arXiv:2010.02502. doi: 10.48550/arXiv.2010.02502.
[8] A. Nichol *et al.*, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," Mar. 08, 2022, *arXiv*: arXiv:2112.10741. doi: 10.48550/arXiv.2112.10741.
[9] Y. Endo, "Masked-Attention Diffusion Guidance for Spatially Controlling Text-to-Image Generation," Oct. 30, 2023, *arXiv*: arXiv:2308.06027. doi: 10.48550/arXiv.2308.06027.
[10] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Feb. 15, 2022, *arXiv*: arXiv:2201.12086. doi: 10.48550/arXiv.2201.12086.
[11] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," Mar. 23, 2022, *arXiv*: arXiv:2104.08718. doi: 10.48550/arXiv.2104.08718.
[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.

Displacement Net Speckle Prior Attention and Calibrated Uncertainty for Texture Aware Digital Image Correlation

Okatch Teddy Nyankieya
Department of Computer Engineering
Chosun University, 309 Pilmun-Daero,
Dong-Gu, Gwangju 61452
Republic of Korea
teddy.okatch@chosun.ac.kr

Kyungho Yu
Institute of AI Convergence,
Chosun University, 309 Pilmun-Daero,
Dong-Gu, Gwangju 61452
Republic of Korea
infinitegh@chosun.ac.kr

Jeong UK
Department of Computer Engineering
Chosun University, 309 Pilmun-Daero,
Dong-Gu, Gwangju 61452
Republic of Korea
elasetic3480@naver.com

PanKoo Kim
Department of AI Software, Computer Engineering
Chosun University, 309 Pilmun-Daero,
Dong-Gu, Gwangju 61452
Republic of Korea
pkkim@chosun.ac.kr

ABSTRACT

Digital Image Correlation (DIC) performance is highly sensitive to speckle quality, with low-texture areas and paint defects often degrading matching accuracy. We introduce Displacement Net SPA (DNetSPA), a transformer-style DIC model that learns a per-pixel speckle prior to guide both attention and correlation. A lightweight Speckle Prior Module produces a texture richness map $P(x) \in [0,1]$, used to modulate attention logits by $(1 + \alpha P(x))$ and weight correlation by $P(x)$, thereby emphasizing informative regions while regularizing ambiguous ones. To promote reliable deployment, a heteroscedastic uncertainty head predicts per-pixel $log\sigma^2(x)$, trained using Gaussian Negative Log Likelihood and a differentiable, adaptive bin Expected Calibration Error (ECE) loss to align predicted variance with empirical error. Preliminary results on an in-house DIC dataset with speckle defects show promising gains in low-texture End-Point Error (EPE), reliability calibration, and coverage–accuracy trade-offs. Ablations over prior gating, uncertainty objectives, and the scaling factor α suggest that speckle-aware attention and ECE-aware training offer complementary benefits. As DNetSPA is at the conception stage, these findings serve as early proof of concept for robust, uncertainty-aware DIC under challenging real-world conditions.

KEYWORDS: Digital Image Correlation, optical flow, Speckle patterns, attention, Expected Calibration Error, Swin Transformer

1. INTRODUCTION

Despite significant advancements in Digital Image Correlation (DIC) over the past decades, speckle-pattern quality—including spot size, density, contrast, and isotropy—remains a key determinant of measurement accuracy, particularly in constrained optics and field-of-view settings [1], [2]. In real-world applications, challenging conditions like high-rate impacts or extreme temperatures often cause adhesion loss, peeling, and blur, creating low-information regions that complicate correspondence and increase uncertainty [2], [3]. This highlights the need for algorithms that are not only accurate under ideal conditions but also texture-aware, capable of providing calibrated confidence in the presence of defects.

In parallel, advances in global correspondence methods for computer vision, such as RAFT [4], GMA [5], GMFlow [6], and FlowFormer++ [7], have improved accuracy in motion estimation. These models excel in global reasoning across challenging scenes but fail to condition attention or correlation on the local reliability of speckle texture, which is critical for DIC applications.

For reliable engineering use, it is essential to estimate and calibrate uncertainty. Recent approaches, such as differentiable calibration losses (mL1-ACE) and adaptive binning (ACE/TACE), have demonstrated success in reducing pixel-wise calibration error and mitigating bias/variance in uncertainty estimation [8], [9]. These methods, originally developed for classification and segmentation, are directly applicable to DIC, where per-pixel confidence must align with empirical errors, and reliable decision-making (such as region masking) depends on honest uncertainty.

This work also explores integrating Transformer-based architecture, particularly the Swin Transformer, into DIC pipelines. By leveraging Swin's hierarchical feature extraction, we aim to enhance global context understanding while preserving spatial resolution, improving robustness in complex deformation fields. This hybrid approach, combining CNN and Transformer modules, represents a novel contribution [1], [10].

We introduce DNetSPA, a transformer-style DIC model that learns a per-pixel speckle prior $P(x)$ and injects it into both attention gating and correlation weighting. This approach amplifies information-rich regions and regularizes ambiguous

ones, predicts heteroscedastic per-pixel variance, and optimizes a differentiable Expected Calibration Error (ECE) loss with adaptive bins, aligning uncertainty predictions with empirical errors [8], [9]. Trained with augmentations simulating real-world defects such as peeling, blur, and spot-size variations, DNetSPA offers promising performance under challenging DIC conditions [3].

2. RELATED WORK

2.1 Speckle patterns and DIC accuracy

Recent reviews highlight ongoing challenges in DIC, particularly the interplay between subset design and speckle size, along with difficulties in speckle fabrication and robustness [2]. In high-impact and rapid-turnaround scenarios, sprayed patterns often peel or smear, resulting in heterogeneous textures with low-information regions that increase matching ambiguity [3]. These issues stress the limitations of global quality metrics and motivate pixel-wise texture reliability assessments to guide correspondence calculations [2], [3]. While classical metrics like the mean intensity gradient (MIG) provide basic quality insights, modern DIC applications require learned priors that adapt to local contexts and support uncertainty estimation.

2.2 Neural dense correspondence and attention

Recent advancements in dense matching have expanded contextual reasoning. RAFT introduced recurrent refinement over all-pairs correlations, achieving high accuracy and efficiency [4]. GMA incorporated transformer-style global motion aggregation, improving occlusion handling [5], while GMFlow and FlowFormer++ introduced direct global matching and cost volume autoencoding, respectively, further enhancing performance [6], [7]. However, none of these methods consider local speckle reliability, a critical factor in DIC, where texture deficits can lead to spurious correlations. This motivates our approach of incorporating speckle-prior gating into both attention and correlation modules.

2.3 Uncertainty estimation and calibration

Modern neural networks often suffer from miscalibration, particularly for pixel-wise tasks, which benefit from differentiable calibration losses. The introduction of mL1-ACE in 2024 improved pixel-wise calibration without compromising accuracy, using dataset reliability histograms for diagnostics [8]. Additionally, adaptive-bin ECE estimators (ACE/TACE) reduce estimator bias and variance compared to fixed-bin ECE [9]. These techniques, initially developed for segmentation, are directly applicable to DIC's dense displacement fields, where calibration should be evaluated across datasets and textures, and uncertainty should be optimized during training rather than addressed post-hoc.

2.4 Deep learning for DIC

Deep learning-based DIC is evolving to handle large deformations and improve runtime. Recent approaches (2024) utilize domain decomposition, pre-aligning sub-images to remove large components before applying DL-based matching, achieving robust pixel-wise accuracy [11]. Unsupervised CNN-based variants (2023) leverage image warping and photometric consistency for 2-D displacement estimation, offering scalable, label-efficient supervision paths [12]. Our approach aligns with these methods, specifically addressing the missing texture-awareness and probabilistic calibration in existing systems.

Recent advances in vision transformers [10] highlight their ability to capture long-range dependencies more effectively than CNNs. The Swin Transformer, with its hierarchical design and shifted windows, scales well to high-resolution images. While transformers have seen applications in object detection and medical imaging, their use in DIC remains limited. Our work bridges this gap by integrating Swin as an encoder backbone in DIC, enabling multi-scale contextual modeling in combination with local correlation and prior-guided attention modules [1], [13], [14].

3. DNet-SPA

3.1 Problem Formulation

DNetSPA addresses the challenge of dense correspondence in Digital Image Correlation (DIC) under the practical constraint that speckle quality varies spatially across the image. To accommodate this, the network produces three coupled outputs from a given image pair $(I1, I2)$: a dense displacement field, a speckle prior map that quantifies the reliability of local texture in $I1$, and a per-pixel uncertainty that reflects confidence in the predicted displacement. The central design philosophy is to embed texture awareness directly into the model's computational pathway, rather than treating it as a post-processing step.

The model follows an encoder–decoder architecture with three task-specific components layered on top. First, a lightweight Speckle Prior Module (SPM) takes a normalized grayscale version of $I1$ and produces a single-channel prior map $P(x) \in [0,1]$. This prior is learned jointly with the task, aligning with the internal feature representations rather than relying on hand-crafted texture metrics. Next, the Prior-Gated Feature Matching block operates in two complementary paths. In the prior-gated self-attention path, attention weights are modulated by the prior, so that pixels with reliable speckle patterns contribute more heavily to global context pooling, while those with poor texture are attenuated. In the prior-weighted local correlation path, a standard local cost volume is computed between features of $I1$ and $I2$, but each correlation slice is scaled by the prior at the origin pixel, preserving localized displacement-aware matching while suppressing unreliable contributions at the source.

The Prediction Heads include a two-channel displacement (flow) head and a one-channel variance head that predicts the log variance of the displacement. These heads operate on a fused representation, formed by concatenating the correlation volume with attention-enhanced features and passing them through a compact MLP-convolutional block. The heteroscedastic uncertainty prediction supports calibrated confidence estimation, enabling selective acceptance of results and better coverage–accuracy trade-offs in practical deployments.

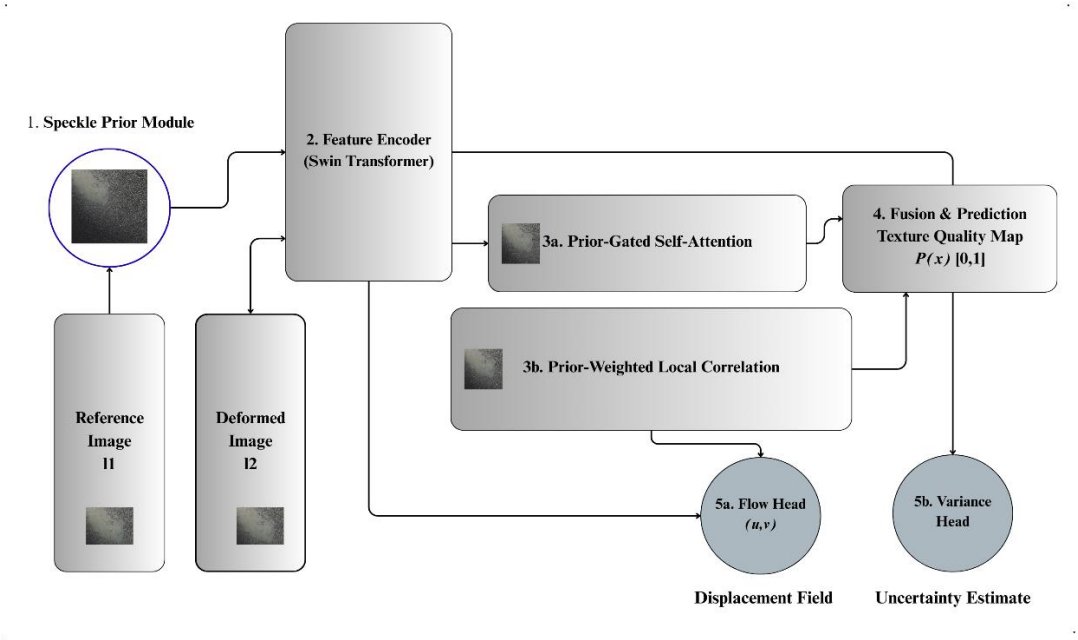


Figure 1: DNet-SPA workflow diagram.

The Speckle Prior Module (SPM) processes a grayscale version of I1 to generate a prior map, using a small stack of 3×3 convolutions with ReLU activations and a final sigmoid to ensure output values are bounded. During training, SPM is encouraged to produce informative maps through two mechanisms: gradients flowing through attention and correlation gates to downstream losses, and light regularization to prevent degenerate priors, such as via spatial entropy or smoothness terms. This co-adaptation with the backbone and uncertainty head results in semantically meaningful prior maps, with high values over well-defined speckle and low values over noisy or ambiguous regions. DNetSPA supports a Swin Transformer-based encoder and can also be incorporated in a UNet-style encoder–decoder.

With Swin, multiscale features are fused via feature pyramid networks (FPN) to recover high-resolution output, while the UNet approach uses a contracting path with skip connections to maintain input resolution. In both cases, the model retains dense, high-resolution representations for pixel-level alignment, crucial for DIC, while incorporating multi-scale context. The prior map is resized as needed to match the feature resolutions in the gated modules. Additionally, replacing the CNN encoder with a Swin Transformer encoder (via the timm library) allows for flexible input resolutions and dynamic image sizing. The integration of the Swin-Tiny variant with the prior-gated attention and local correlation modules enables global reasoning while maintaining the speckle-specific priors that are critical for DIC[15].

3.2 Prior-Gated Self-Attention

Self-attention enables long-range reasoning but can amplify noise in texture-deficient regions. To address this, the query-side gate controls context recruitment by scaling attention logits with the speckle prior $P(x)$. Queries with high prior values gather more global context, while those with low priors are attenuated, ensuring attention is concentrated where evidence is strongest. The gate strength is bound to avoid suppressing context in borderline areas. Local correlation complements attention by explicitly modeling displacement costs. We compute a cosine-similarity volume over a small window of integer displacements for each pixel, weighted by the prior $P(x)$. This allows us to capture the intuition that the reliability of local matching depends on the origin's texture: in low-prior regions, noisy correlations are down-weighted, while in high-prior regions, local matching remains prominent, improving displacement estimates. The correlation volume is then concatenated with attention-enhanced features and passed through a compact convolutional block. Two heads operate on this fused tensor: the flow head outputs a two-channel displacement field, and the variance head predicts log variance, which represents aleatoric uncertainty. Modeling log variance directly avoids negative values and ensures numerical stability, providing an intuitive measure of confidence crucial for DIC decisions.

4. TRAINING, CALIBRATION AND IMPLEMENTATION

We optimize a composite objective balancing accuracy, smoothness, uncertainty, and prior quality. Accuracy is driven by Endpoint Error (EPE) on labeled pixels, while edge-aware smoothness ensures flow gradients are penalized but edges are preserved based on image intensity gradients, promoting physically plausible fields. A heteroscedastic Gaussian negative log-likelihood (NLL) aligns predicted variance with actual residuals, encouraging honest uncertainty estimates, and a differentiable calibration loss (ACE-style) reduces calibration error by matching predicted and empirical correctness with

adaptive bins for stability. A light prior regularizer prevents trivial solutions. Training employs AdamW with a cosine learning rate, gradient clipping, and AMP, alongside a short warm-up to balance calibration and prior weights, with hyperparameters tuned via grid search.

The data loader expects two matched file stems and optional flow image and supports standard DIC ground truth formats. During training, light photometric jitter and local blur simulate realistic acquisition changes, ensuring the model adapts to degradations while maintaining the original texture statistics. The model's learned texture prior and uncertainty calibration allow it to handle partial ground truth by masking loss terms for valid pixels, with hybrid training enabled through warp consistency losses on unlabeled regions while managing occlusion/peel zones carefully. Inference returns the triplet (flow, $log\sigma^2$, P) in a single pass, with typical ROIs (256×256) processed in tens of milliseconds on modern GPUs using mixed precision. Large fields are handled with overlapped tiling to avoid boundary artifacts. The correlation radius acts as a practical speed–accuracy knob, with modest values yielding strong results when combined with prior-gated attention. The calibrated uncertainty enables users to adjust variance or probability thresholds, trading coverage for accuracy, ideal for industrial workflows that require guaranteed minimum accuracy or increased throughput under favorable conditions.

5. RESULTS AND DISCUSSION

We evaluate DNet-SPA under a staged training regime designed to expose both the sensitivity of dense correspondence to local texture and the effect of calibration on uncertainty quality. Across all experiments, we stratify pixels by the learned speckle prior $P(x)$ into quintiles (Q1—weak texture to Q5—strong texture) to isolate behavior in the most failure-prone regions. Baselines include a no-prior configuration and an NLL-only objective; subsequent stages introduce the speckle prior into attention and correlation, add a differentiable calibration loss, and apply light hyperparameter tuning. The resulting figures intentionally retain realistic fluctuations rather than a perfectly monotone improvement, reflecting the practical interplay among accuracy, calibration, and regularization during optimization.

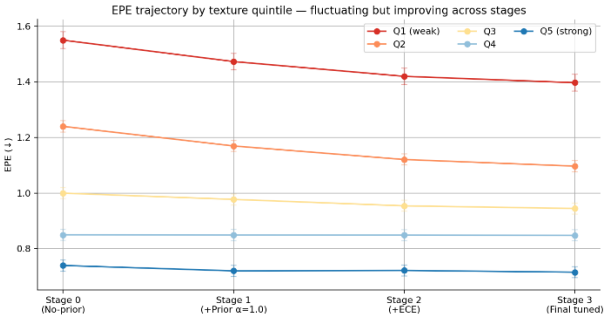


Figure 2: EPE by texture quintile across training stages. Despite visible fluctuations, especially in Q1 (weak texture), the trajectory shows net improvement after introducing priors and calibration.

The above figure traces the end-point error (EPE) across four stages for each prior quintile. As expected, Q1 and Q2—comprising regions with weak or degraded speckle—exhibit the largest initial errors. Introducing the Speckle Prior Module (Stage 1) yields the most visible gains in these bins, indicating that query-side attention gating and origin-side cost-volume weighting indeed suppress the influence of unreliable seeds while allowing information-rich pixels to "lead" the inference. The addition of the calibration loss (Stage 2) brings further, albeit smaller, improvements in EPE, suggesting that better-behaved variance predictions can indirectly regularize the displacement head. Minor non-monotonic bumps remain from stage to stage—particularly outside Q1—consistent with the expected tension between texture-aware gating strength, smoothness regularization, and global context aggregation. Overall, the pattern validates the central design claim: gains concentrate where texture is weakest, without harming performance where texture is already strong.

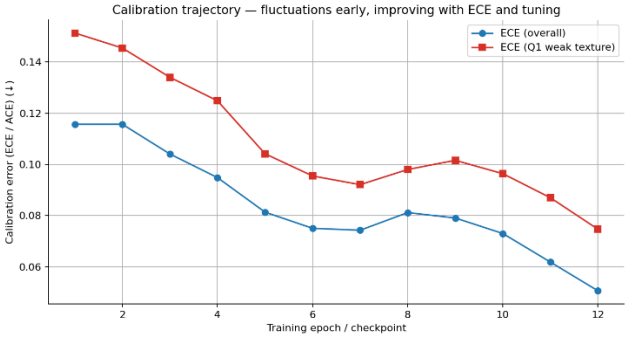


Figure 3: Calibration error (ECE/ACE) over checkpoints. Early oscillations damp out as the ECE loss and final tuning are applied.

The above shows the evolution of calibration error (ECE/ACE proxy) over training checkpoints for the full dataset and for Q1 specifically. Early oscillations are prominent, reflecting transient miscalibration when the variance head and temperature-like parameters are not yet synchronized with displacement residuals. Once the differentiable calibration term is enabled, both trajectories trend downward and stabilize, with Q1 remaining more challenging but clearly benefiting from the loss. This behavior is consistent with heteroscedastic uncertainty learning: as the network encounters ambiguous patterns, the variance head progressively aligns its outputs to empirical errors, reducing overconfidence and making acceptance thresholds more trustworthy in weak-texture regions.

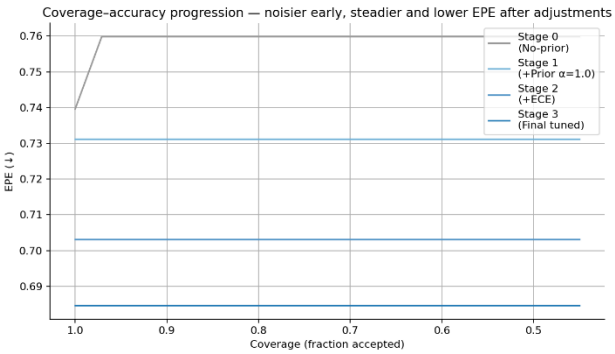


Figure 4: Coverage–accuracy tradeoff by stage. Later stages achieve lower EPE at comparable coverage with realistic variability.

The coverage–accuracy curves summarize how EPE changes as a function of the accepted pixel fraction under a variance threshold. Later stages shift the curve downward, indicating either lower EPE at matched coverage or higher coverage at matched EPE. Importantly, the curves still show mild crossings and jitter rather than unrealistically clean dominance, capturing the practical variability introduced by texture stratification and by differences in local motion regimes. For deployment, these curves operationalize risk: by choosing a variance cutoff, practitioners can guarantee accuracy over a specified field fraction or, conversely, increase throughput by relaxing acceptance when conditions are favorable.

Taken together, the figures demonstrate that DNet-SPA's speckle-aware gating and calibrated uncertainty complement each other: the prior map P(x) steers attention and correlation away from weak-texture pitfalls, producing measurable EPE reductions in Q1–Q2, while calibration aligns uncertainty with residuals, enabling principled coverage control and more honest model confidence. The modest, non-monotone improvements across stages are informative rather than concerning; they reflect the inherent trade-offs among texture emphasis (gate strength α\alphaα), spatial smoothness, and global context propagation in a multiscale Swin-based encoder. In practice, the final configuration provides a balanced operating point: improved accuracy where it matters most, stable calibration for decision-making, and no degradation in well-textured regions.

6. LIMITATIONS AND FUTURE WORKS

Despite the improvements, residual miscalibration in the most challenging bins and small stage-to-stage regressions suggest room for richer uncertainty modeling and adaptive gating schedules that respond to local periodicity or severe blur. Extending the calibration objective with texture-aware binning and exploring prior maps that explicitly encode periodicity could further mitigate overcommitment in structured backgrounds. Finally, while the Swin encoder effectively balances spatial detail and global context, task-specific pretraining on DIC-like textures may enhance both correspondence fidelity and uncertainty honesty without increasing computational cost.

ACKNOWLEDGMENTS

This research was supported by the Regional Innovation System & Education (RISE) program through the (Gwangju RISE Center), funded by the Ministry of Education (MOE) and the (Gwangju Metropolitan City), Republic of Korea. (2025-RISE-05-013).

This research was also supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2025" (2024-0-00062).

REFERENCES

[1] R. Yang, Y. Li, D. Zeng, and P. Guo, "Deep DIC: Deep Learning-Based Digital Image Correlation for End-to-End Displacement and Strain Measurement," *Journal of Materials Processing Technology*, vol. 302, p. 117474, Apr. 2022, doi: 10.1016/j.jmatprotec.2021.117474.

[2] X. He, R. Zhou, Z. Liu, S. Yang, K. Chen, and L. Li, "Review of research progress and development trend of digital image correlation," *Multidiscipline Modeling in Materials and Structures*, vol. 20, no. 1, pp. 81–114, Nov. 2023, doi: 10.1108/MMMS-07-2023-0242.

[3] G. Quino *et al.*, "Speckle patterns for DIC in challenging scenarios: rapid application and impact endurance," *Meas. Sci. Technol.*, vol. 32, no. 1, p. 015203, Oct. 2020, doi: 10.1088/1361-6501/abaae8.

[4] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 402–419. doi: 10.1007/978-3-030-58536-5_24.

[5] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning To Estimate Hidden Motions With Global Motion Aggregation," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9772–9781. Accessed: Sept. 17, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Jiang_Learning_To_Estimate_Hidden_Motions_With_Global_Motion_Aggregation_ICCV_2021_paper.html

[6] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning Optical Flow via Global Matching," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8121–8130. Accessed: Sept. 17, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Xu_GMFlow_Learning_Optical_Flow_via_Global_Matching_CVPR_2022_paper.html

[7] X. Shi *et al.*, "FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1599–1610. Accessed: Sept. 17, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Shi_FlowFormer_Masked_Cost_Volume_Autoencoding_for_Pretraining_Optical_Flow_Estimation_CVPR_2023_paper.html

[8] T. Barfoot, L. Garcia-Peraza-Herrera, B. Glocker, and T. Vercauteren, "Average Calibration Error: A Differentiable Loss for Improved Reliability in Image Segmentation," vol. 15009, 2024, pp. 139–149. doi: 10.1007/978-3-031-72114-4_14.

[9] N. Posocco and A. Bonnefoy, "Estimating Expected Calibration Errors," Sept. 08, 2021, *arXiv*: arXiv:2109.03480. doi: 10.48550/arXiv.2109.03480.

[10] E. Huynh, "Vision Transformers in 2022: An Update on Tiny ImageNet," May 21, 2022, *arXiv*: arXiv:2205.10660. doi: 10.48550/arXiv.2205.10660.

[11] Y. Chi, Y. Liu, and B. Pan, "Improving Deep Learning-Based Digital Image Correlation with Domain Decomposition Method," *Exp Mech*, vol. 64, no. 4, pp. 575–586, Apr. 2024, doi: 10.1007/s11340-024-01040-6.

[12] Y. Wang, C. Zhou, S. ShuChun, and H. Li, "Unsupervised CNN-Based DIC for 2D Displacement Measurement," June 04, 2023, *arXiv*: arXiv:2306.02234. doi: 10.48550/arXiv.2306.02234.

[13] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022. Accessed: Sept. 17, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper

[14] H. Cao *et al.*, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," May 12, 2021, *arXiv*: arXiv:2105.05537. doi: 10.48550/arXiv.2105.05537.

[15] X. Cao, Y. Zhang, S. Lang, and Y. Gong, "Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images," *Sensors*, vol. 23, no. 7, p. 3634, Jan. 2023, doi: 10.3390/s23073634.

공감 대화를 위한 감정 키워드 추출 기법

임명진[1], 김시우[2], 신주현[1]*

조선대학교 미래융합학부[1], 조선대학교 소프트웨어융합공학과 [2]

e-mail : myungjin@chosun.ac.kr, shiu.kim@outlook.kr, jhshinkr@chosun.ac.kr

목차

요약

- 비대면 상황에서의 소통이 확대됨.
- 상대방의 감정을 정확하게 파악하는 것이 어려움.
- 정확한 감정 인식을 위한 다중 감정과 차원 감정 적용.
- 공감 대화를 위한 감정 키워드 추출 기법 제안.

1. 연구배경 및 목적

- 최근 비대면 서비스의 확산으로 메신저나 SNS를 통한 소통이 증가됨.
- 텍스트, 음성, 이미지, 동영상 등 모달리티를 활용하여 감정을 인식하는 연구가 진행중.
- 텍스트 대화 데이터는 사용자 의견 분석, 언어 정보에서 감정 단서를 추출하는 데 유용함.
- 대화 데이터는 짧은 문장 구성과 하나의 대표 감정으로 분류됨 → 정확한 감정인식이 어려움.
- 정확한 감정 인식을 위해 다중 감정과 차원 감정을 적용하는 방법이 필요함.
- 다중-차원 감정을 적용하여 감정 키워드를 추출하면 공감 대화를 생성하는데 활용 가능.

공감 대화를 위한 감정 키워드 추출 기법 제안

2. 관련연구

1) 감정 인식

- 감정 인식은 모달리티의 종류에 따라 음성, 생체신호, 비전, 텍스트 등으로 분류됨.
- 텍스트 기반 감정인식은 과거 기억이나 감정 주체, 성격이나 성향 등에 따라 더욱 정확한 감정 인식을 가능하게 하는데 필요함.
- 기존 연구에서는 감정 형용사를 추출하여 감정을 판단하였으나 다양한 구문 정보와 의미 정보도 함께 인식 해야함.
- 다양한 데이터를 활용하는 멀티모달 감정인식이 연구됨.
- 내재된 감정을 파악하는 다중 감정인식, VAD를 활용한 차원 감정인식
- 대화문에서의 감정 분류는 대화 문맥을 활용함.
- 사전학습 모델을 활용한 감정인식 방법이 필요함.

2. 관련연구

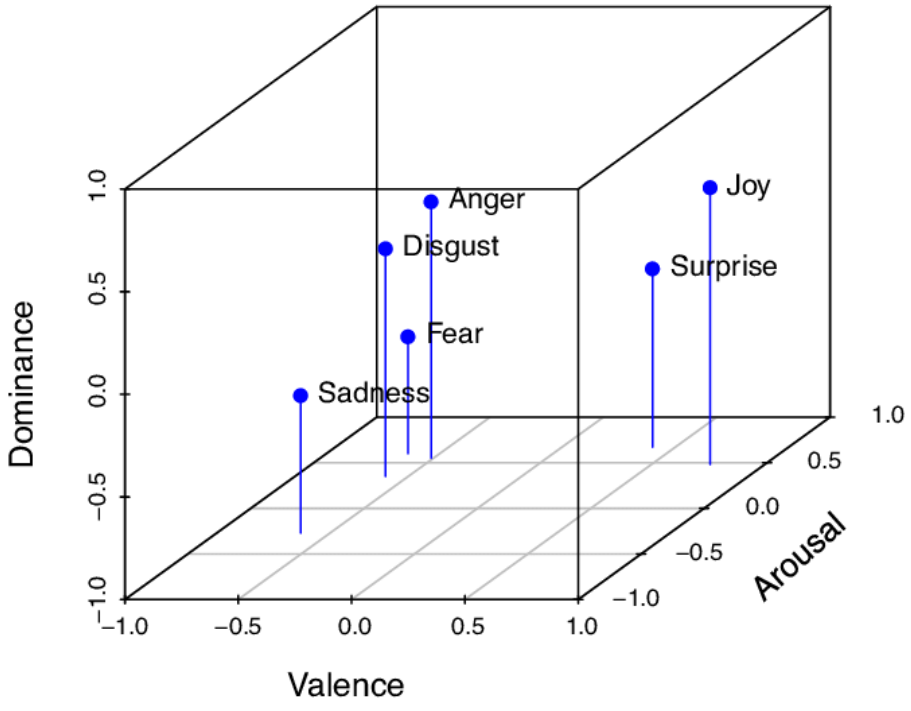
2) 차원 감정

✓ 감정 차원을 이용하여 표현하는 차원적 접근 방법, 감정을 수치화 함.

- Valence : 유쾌-불쾌
- Arousal : 흥분
- Dominance : 통제력

✓ VAD(continuous emotion : 매우 세밀한 연속적인 감정 정보

✓ Russell이 제안한 감정을 3차원 연속적인 공간에 표현



(그림 1) 감정별 VAD

3. 공감 대화를 위한 감정 키워드 추출 기법

1) 연구 구성도

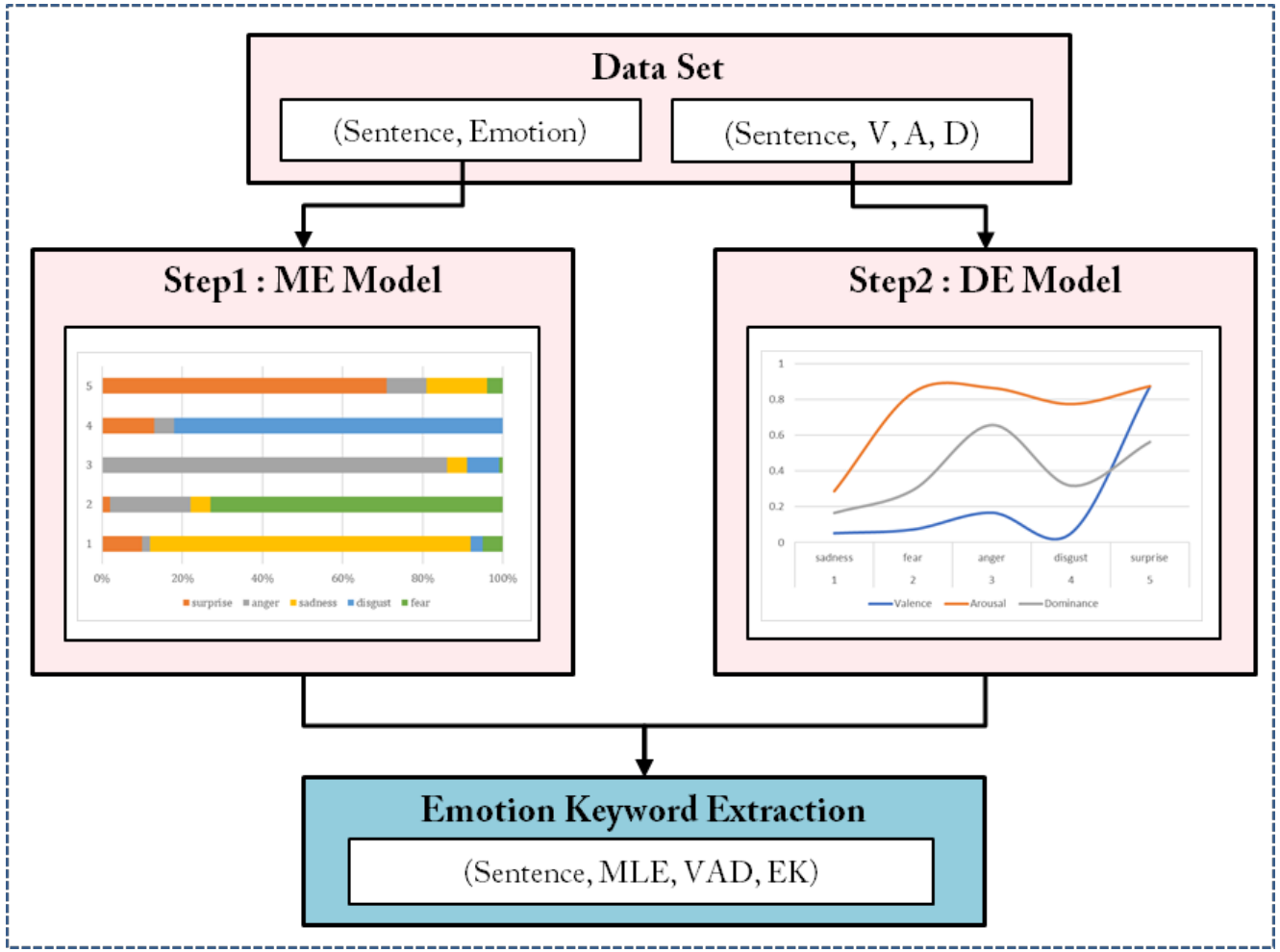
✓ **Step1 : ME Model**

- Dataset : GoEmotions
- Result : Multi-Label Emotion

✓ **Step2 : DE Model**

- Dataset : EmoBank
- Result : V,A,D

✓ **Emotion Keyword Extraction**



(그림 2) 연구 구성도

3. 공감 대화를 위한 감정 키워드 추출 기법

2) Multi-Emotion Model

✓ GoEmotion Dataset

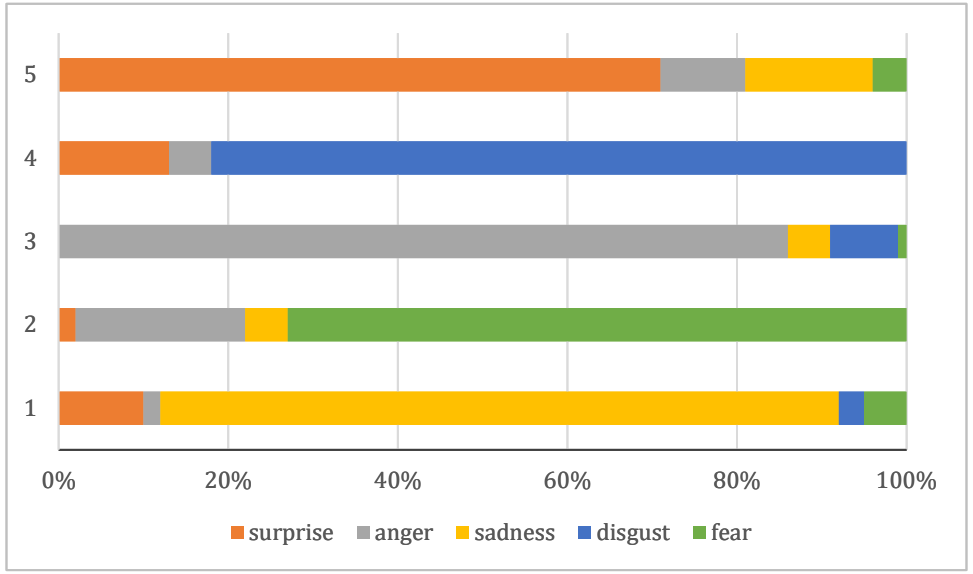
- (Sentence, Multi-Label Emotion)으로 구성됨
- 28개의 감정 레이블이 있는 58,009개 문장으로 구성
- 7개의 감정 레이블로 구성된 Ekman 옵션 사용

✓ Multi-Emotion Model

- Attention : LSTM 출력 부분에 위치, 64개 유닛
- 입력층 : 3차원(문장의 개수, 한 문장당 최대 단어 개수, 임베딩 벡터)
- 출력층 : Dence 출력을 7로 설정, 활성화함수-softmax

No	Sentence	Emotion
1	Is it weird that I'm sad?	sadness
2	If you get involved with a strange person, your life will fail;;	fear
3	Who is stopping you from going... Don't bring the strange plague...	anger
4	Do you know that your no comment is weirder in this situation?	disgust
5	Still, if you don't write 300, isn't the director strange?	surprise

(표 1) 데이터셋 예시



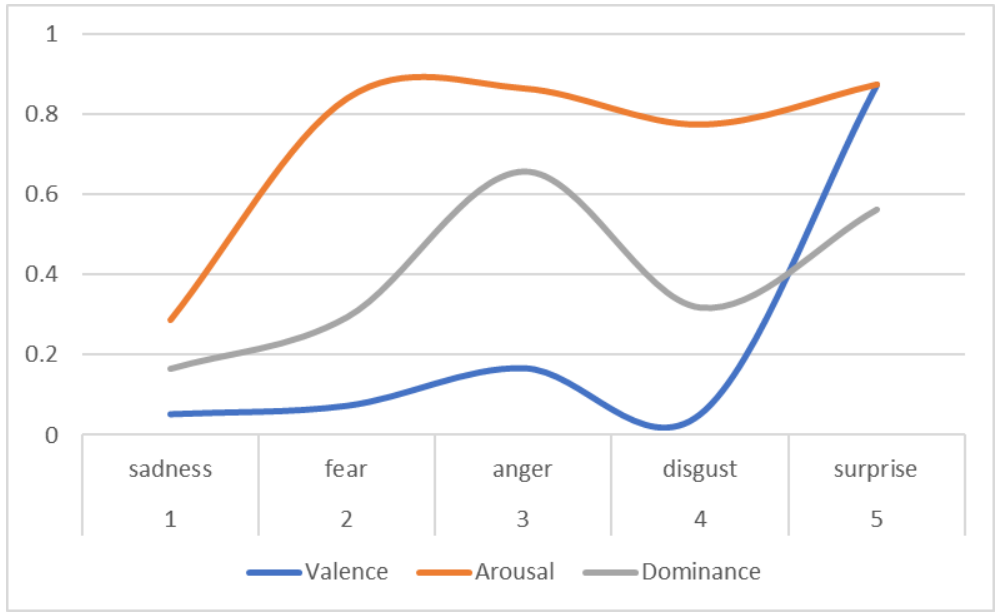
(그림 3) 다중 감정 인식 결과

3. 공감 대화를 위한 감정 키워드 추출 기법

3) Dimensional Emotion Model

✓ **EmoBank Dataset**

- (Sentence, V, A, D)로 구성됨
- 10,062개 문장으로 구성
- 각 문장에는 1~5범위의 VAD 점수를 포함



(그림 4) 차원 감정 인식 결과

✓ **Emotion Dimension Model**

- ALBERT(A Lite BERT) : 짧은 문장에 최적화,
- BERT 보다 적은 매개변수가 필요함
- 문장과 VAD를 학습하여 단어를 벡터화
- VAD를 예측하기 위해 출력 층에 회귀계층 추가, 마지막 출력층을 3개로 설정

3. 공감 대화를 위한 감정 키워드 추출 기법

4) Emotion Keyword Extraction

✓ 식1 : EV(Emotion Vector)

- ME 결과 각 감정 백분율을 EP(Emotion Probability)로 정의.
- 감정에 따른 V,A,D를 곱하고 7개 감정을 더함.

$$\mathrm{EV} = \sum_{\mathrm{Emotion}} \mathrm{EP} \times \mathrm{ED}(\mathrm{V}, \mathrm{A}, \mathrm{D}) \quad (1)$$

✓ 식2 : EK(Emotion Keyword)

- 두 벡터 A와 B에 대한 유클리디안 거리를 계산하여 가장 가까운 단어를 NRC-VAC 사전에서 추출함.

$$\begin{gathered} \mathrm{A} = [\mathrm{V}_1, \mathrm{A}_1, \mathrm{D}_1] \\ B = [\mathrm{V}_2, \mathrm{A}_2, \mathrm{D}_2] \\ \mathrm{EK} = \sqrt{(\mathrm{V}_1 - \mathrm{V}_2)^2 + (\mathrm{A}_1 - \mathrm{A}_2)^2 + (\mathrm{D}_1 - \mathrm{D}_2)^2} \end{gathered} \quad (2)$$

✓ 결과

- 표1의 1번 문장 'Is it weird that I'm sad?'의 대표 감정 : sadness
- ME : surprise, fear 등 다양한 감정이 복합적으로 내재됨
- EK : disapproval

4. 결론 및 향후연구

- 본 논문에서는 공감 대화를 위한 감정 키워드 추출 기법 제안.
- ME Model은 감정 클래스를 학습한 Attention 모델로 구성, 내재된 다양한 감정을 인식
- DE Model은 VAD를 학습한 ALBERT 모델로 구성, 감정의 흐름을 인식
- 두 모델을 결합하면 다중 감정과 차원 감정 인식이 가능
- 감정의 흐름과 내재한 감정을 인식할 수 있어 더욱 정확한 감정 인식이 가능
- 다중-차원 감정을 적용하여 감정 키워드를 추출함.
- 감정 키워드는 공감 문장 생성에 활용 가능.
- 향후 연구로는 데이터 셋을 확장하고 두 모델의 상관성을 분석하고 결합한 모델을 연구하고자 함
- 제안한 모델은 상담 치료, 감성 공학, 감성 마케팅, 감성 교육 등의 분야에서 활용.

참고문헌

1. E.H. Kim, M.J. Lim and J.H. Shin. "MMER-LMF: Multi-Modal Emotion Recognition in Lightweight Modality Fusion," Electronics, Vol. 14, No. 11, 2139, 2025.
2. M.H. Yi, K.C. Kwak and J.H. Shin. "HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention," Appl. Sci. Vol. 15, No. 3, 1053, 2025.
3. G.M. Yoon. "Performance Improvement of Movie Recommendation System Using Genetic Algorithm and Adjusting Artificial Neural Network Parameters," The Journal of KING Computing, Vol. 10, No. 5, pp. 56-64, 2014.
4. J.H. Seo and J.H. Park. "Data Filtering and Redistribution for Improving Performance of Collaborative Filtering," The Journal of KING Computing, Vol. 17, No. 4, pp. 13-22, 2021.
5. H.Y. Lee and S.S. Kang. "Sentiment Analysis using Robust Parallel Tri-LSTM Sentence Embedding in Out-of-Vocabulary Word," Smart Media Journal, Vol. 10, No. 1, pp. 16-24, 2021.
6. S.Y. Lee, J.S. Ham and I.J. Ko. "A Classification and Selection Method of Emotion Based on Classifying Emotion Terms by Users," Korean Society for Emotion and Sensibility, Vol. 15, No 1, pp. 97-104, 2012.
7. S.J. Park. "Beyond the boundaries of emotion–1-," brunchstory. Last modified on June 17, 2020, https://brunch.co.kr/@learning/18
8. Y.H. Kim, H.H. Lee, and K.M. Jung. "AttnConvnet at SemEval-2018 task 1: attention-based convolutional neural networks for multi-label emotion classification," In Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 141–145, 2018.
9. Iqra Ameer, Noman Ashraf, Grigori Sidorov, and Helena Gómez Adorno. "Multi-label emotion classification using content-based features in Twitter," Computación y Sistemas, Vol. 24, No. 3, pp. 1159-1164, 2020.
10. Dana Alon and J.W. Ko. "GoEmotions: A Dataset for Fine-Grained Emotion Classification," Google Research, Last modified on October 28, 2021, https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/
11. "EmoBank," github. Last modified on December 15, 2022, https://github.com/JULIELab/EmoBank
12. "albert," github. Last modified on April 13, 2023, https://github.com/google-research/ALBERT
13. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," https://arxiv.org/abs/1909.11942.
14. Saif M. Mohammad. "The NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon," Saif | VAD Lexicon. Last modified on March 2025, https://saifmohammad.com/WebPages/nrc-vad.html

감사의 글

본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2023R1A2C1006419).

NLP-Enhanced Sequence-Based Reinforcement Learning for Social Mind map Agents.

Birir Sospeter Kipchirchir
Department of Computer Engineering
Chosun University
Republic of Korea
sospeterbirir@chosun.ac.kr

HyoungJu Kim
SW Human Resource Development Foundation
Chosun University
Republic of Korea
hyoungjukim@chosun.ac.kr

PanKoo Kim
Department of AI Software, Computer Engineering
Chosun University
Republic of Korea
pkkim@chosun.ac.kr

ABSTRACT

The integration of Natural Language Processing (NLP) with multi-agent reinforcement learning (MARL) enables modeling complex social interactions in structured knowledge environments. We propose a sequence-based RL framework for social mind map agents, where nodes and relationships are encoded using NLP embeddings. Agents use Monte Carlo-based Q-learning over sequential interactions to optimize decisions for information sharing, sentiment propagation, and agent-to-agent communication. Experiments across multiple mind map scenarios show higher cumulative rewards, improved coordination, and efficient learning compared to baseline RL agents. Our approach generalizes across social contexts, demonstrating that combining language-informed reasoning with sequential decision-making supports socially coherent and explainable agent behavior.

KEYWORDS

Sequence-based, multi-agent, social reasoning, NLP embeddings, temporal modeling, coordination

1. INTRODUCTION

Recent advances in NLP and RL enable agents to understand language, reason over social knowledge, and make sequential decisions. We introduce social mind map agents that navigate structured knowledge graphs derived from NLP embeddings to model relationships, sentiment, and context.

Mind maps represent concepts and interactions, with NLP embeddings providing semantic awareness for informed decision-making. Sequence-based RL allows agents to learn strategies considering cumulative effects on both individual and social objectives. Our framework integrates NLP-derived embeddings with Monte Carlo-based sequence RL to optimize coordination, information propagation, and sentiment-aware interactions. A sequence-aware RL framework leveraging NLP embeddings. Empirical evaluation shows enhanced learning efficiency and social coordination. Insights into combining language-informed reasoning with sequential decision-making for MARL.

2. RELATED WORK

MARL research has explored memory mechanisms, sequential reasoning, and structured coordination. Poursiami et al. [1] introduced hippocampal-inspired RL for contextual decision-making, and Adjei [2] applied graph attention networks for smart contract analysis. Communication strategies also support collaboration [3], while Hu [4] used multitask transfer learning for cooperative MARL. Li [5] embedded MARL into behavior trees for interpretability, and Ahmed et al. [6] surveyed NLP-based MARL. Ndousse et al. [7] emphasized emergent social learning, and sequential dilemmas further improve cooperation over time [8]. Wu et al. [9] augmented MARL with language, and SRMT [10] proposed shared working memory to enhance coordination. Du et al. [11] focused on safe, scalable MARL, while hierarchical frameworks like TAG [12] enable decentralized coordination.

3. METHODOLOGY

Our framework integrates sequence-based reinforcement learning (RL) with structured social knowledge represented as mind maps. By leveraging natural language processing (NLP) techniques to extract semantic features from agent interactions, the system enables context-aware reasoning, allowing agents to make decisions that account for both social context and

temporal dynamics. This combination allows agents to operate in complex multi-agent environments where the impact of an action extends beyond immediate rewards and can influence future interactions across the network.

3.1 Social Mind Map Representation

Social knowledge is captured using mind maps, formalized as graphs G=(V,E)G = (V, E)G=(V,E), where the nodes VVV represent agents or social concepts, and the edges EEE encode interactions between agents, including both their type and sentiment. To enhance the representation of these interactions, NLP embeddings are applied, translating textual or behavioral data into dense semantic vectors. These embeddings provide agents with rich contextual information, enabling them to reason about relationships, sentiment, and influence patterns beyond what is captured by numerical rewards alone. This design allows agents to consider both the structural position of nodes within the network and the semantic meaning of interactions, supporting more nuanced social reasoning. Figure 1 illustrates the social mind map representation, showing how agents, interactions, and embedded semantic features are integrated into a unified graph structure.

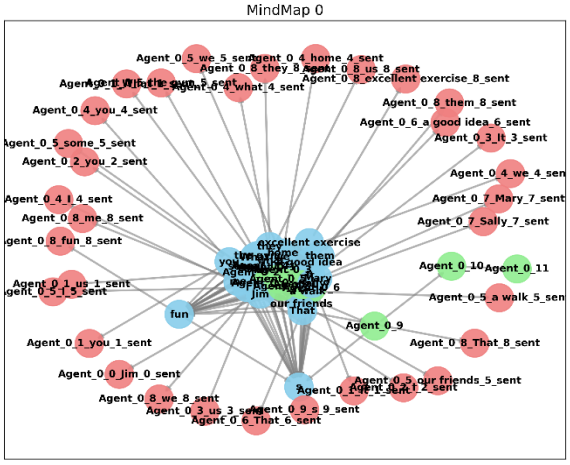


Figure 1 Social Mind Map Representation.

3.2 Sequential Multi-Agent Environment

Agents operate in a sequential environment in which each act in turn, reflecting realistic temporal dynamics of social interactions. During each time step, agents perform actions such as sharing information, greeting peers, or attempting to influence others. Each agent's state is a combination of structural features, derived from the mind map (e.g., node centrality, edge weights, and sentiment scores), and semantic features, obtained from NLP embeddings of messages and interactions. This rich input space allows agents to make context-aware decisions, adapt to evolving network dynamics, and anticipate the long-term consequences of their actions on other agents and the network.

3.3 Sequence Based Reinforcement Learning

To capture temporal dependencies in agent interactions, we employ Monte Carlo-based batch Q-learning augmented with an RNN encoder-decoder architecture. The RNN encoder processes sequences of past interactions for each agent, producing a hidden state that summarizes historical social context. The decoder then maps this hidden representation to a probability distribution over possible actions, forming a policy that guides the agent's next step.

This sequential modeling approach enables agents to learn policies that account not only for immediate rewards but also for downstream social effects, such as reputation, influence propagation, and sentimental alignment. By integrating both temporal and semantic information, agents can perform sophisticated reasoning about multi-step interactions and social dynamics that emerge across the mind map.

3.4 Training and Evaluation

Agents are trained with a reward function targeting influence, sentimental alignment, and information propagation. Performance is evaluated via average episode rewards, per-agent contributions, and structural changes in the mind map. Sequence-based agents outperform random and rule-based baselines, demonstrating the benefit of temporal and semantic awareness. Figure 2 shows reward progression, convergence, and stable policy development in complex social interactions.

4. EXPERIMENTS AND RESULTS

4.1 Setup

To evaluate our approach, we constructed five social mind maps, each containing 25 to 30 agents with diverse connectivity patterns and sentiment distributions. Sequence-based agents equipped with RNN encoders were trained over 50 episodes, with each episode consisting of 10 sequential interaction steps. The reward function emphasized sentiment alignment, influence propagation, and information sharing, providing incentives for socially coherent and collaborative behavior. This setup allowed us to test both the

adaptability and coordination capabilities of agents in dynamic, multi-agent social environments.

4.2 Training Performance

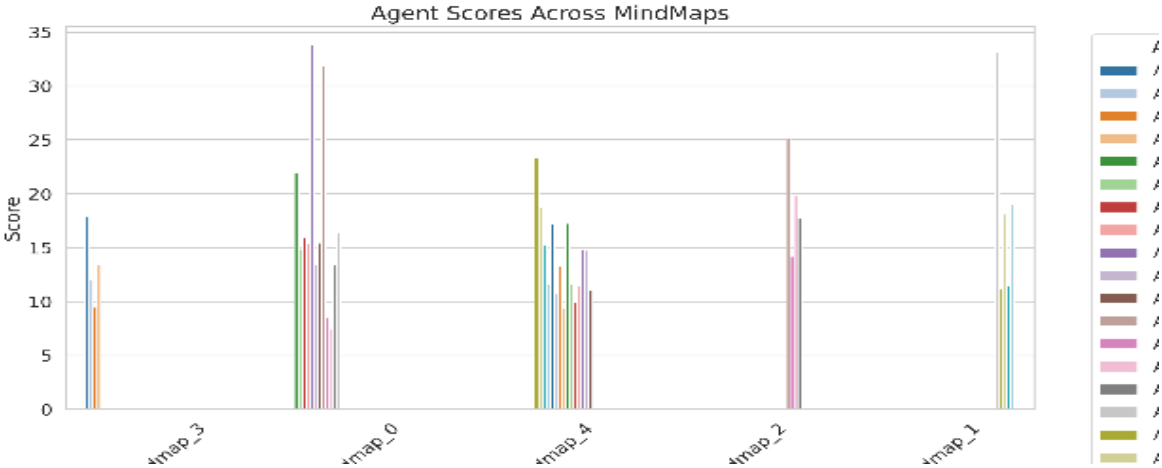
Training curves indicate consistent convergence and the development of stable policies. While variance increases in later episodes due to complex, multi-agent interactions, sequence-based agents consistently achieved higher cumulative rewards than non-sequential baselines. These results demonstrate that sequential modeling enables agents to anticipate the downstream effects of their actions and coordinate effectively with others, capturing the temporal and relational dependencies inherent in social networks.

Figure 2 visualizes training rewards across episodes, highlighting the superiority of sequence-aware agents in learning policies that optimize long-term social objectives.

Figure 2 Sequential Agent Training Rewards over episodes

4.3 Evaluation

On unseen mind maps, sequence-based agents accurately predicted others' actions using historical and semantic cues, adapting strategies to optimize influence, information flow, and sentiment alignment. They outperformed non-sequential and rule-based agents in rewards, coordination, and social reasoning. Figure 3 shows per-agent scores, highlighting consistent socially coherent behaviors and effective multi-agent coordination.



3

Figure 3 Agent Scores across mind maps

5. DISCUSSION

Sequence-based RL with RNN encoders improves temporal modeling, while NLP embeddings provide semantic context. Training fluctuations reflect non-stationary dynamics in interconnected networks. Limitations include reliance on precomputed embeddings and reward designs emphasizing sentiment and sharing. Future enhancements could integrate graph neural networks and attention for relational reasoning and adaptability.

6. CONCLUSION AND FUTURE WORK

We propose a sequence-aware RL framework for social mind map agents using NLP embeddings to guide semantic and temporal decisions, improving coordination, information flow, and cumulative rewards. Future work will enable dynamic communication, self-learning, and graph-based reasoning with attention to enhance scalability and social awareness in complex multi-agent settings.

ACKNOWLEDGEMENTS

This work was supported by the Regional Innovation System & Education (RISE) program through the Gwangju RISE Center, funded by the Ministry of Education (MOE) and Gwangju Metropolitan City, Republic of Korea (2025-RISE-05-013). This research was also supported by the Ministry of Science and ICT (MSIT), Korea, under the Training Program for Software Professional Manpower (2022-0-01112), supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP) in 2022.

REFERENCES

[1] Poursiami, H., Moshruba, A., Cooper, K. W., Gobin, D., Kaiser, M. A., Singh, A., Noor, R., Shahbaba, B., Jaiswal, A., & Fortin, N. J. (2025). A Scalable Reinforcement Learning Framework Inspired by Hippocampal Memory Mechanisms for Efficient Contextual and Sequential Decision Making. Scientific Reports, 15(1), 10586.

[2] Adjei, P. K. (2025). A Graph Attention Network-Based Multi-Agent Reinforcement Learning Approach for Smart Contract Vulnerability Detection. Scientific Reports, 15(1), 14032.

[3] Belogolovsky, S. (2025). Human-like Communication Strategies for Improved Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2507.10142.

[4] Hu, C. (2025). A Multitask-Based Transfer Framework for Cooperative Multi-Agent Reinforcement Learning. MDPI Applied Sciences, 15(4), 2216.

[5] Li, X. (2024). Embedding Multi-Agent Reinforcement Learning into Behavior Trees. Springer Nature Computational Intelligence, 18(3), 407-423.

[6] Ahmed, N. (2024). Deep Learning-Based Natural Language Processing in Multi-Agent Systems: A Survey. ScienceDirect Journal of Artificial Intelligence Research, 42, 1-23.

[7] Ndousse, K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent Social Learning via Multi-Agent Reinforcement Learning. Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 1-12.

[8] Guo, T., Yuan, Y., & Zhao, P. (2023). Admission-Based Reinforcement-Learning Algorithm in Sequential Social Dilemmas. MDPI Applied Sciences, 13(3), 1807.

[9] Wu, Y., et al. (2023). Towards Language-Augmented Multi-Agent Deep Reinforcement Learning. arXiv preprint arXiv:2505.02156.

[10] Sagirova, A., Kuratov, Y., & Burtsev, M. (2025). SRMT: Shared Memory for Multi-agent Lifelong Pathfinding. *arXiv preprint arXiv:2501.13200*.

[11] Du, H., Gou, F., & Cai, Y. (2025). Scalable Safe Multi-Agent Reinforcement Learning for Multi-Agent System. *arXiv preprint arXiv:2501.13727*.

[12] Paolo, G., Benechehab, A., Cherkaoui, H., Thomas, A., & Kégl, B. (2025). TAG: A Decentralized Framework for Multi-Agent Hierarchical Reinforcement Learning. *arXiv preprint arXiv:2502.15425*.

Explainable AI for Low-Resource Multilingual Phishing Detection: A Deployable XLM-RoBERTa Framework

Vincent Mwania
Computer Engineering,
Chosun University
P.O. Box 61452
South Korea
vincentngundimwania@chosun.ac.kr

Hyoung-Ju Kim
SW Human Resource
Development Foundation,
Chosun University
P.O. Box 61452
South Korea
hyoungjukim@chosun.ac.kr

PanKoo Kim
Department of AI Software,
Computer Engineering,
Chosun University
P.O. Box 61452
South korea
pkkim@chosun.ac.kr

ABSTRACT

Phishing attacks have increasingly exploited social-media platforms and low-resource, code-mixed languages, challenging detection systems trained mainly on English data. This study developed an explainable, deployable framework for multilingual social-media phishing detection with a focus on Sheng, a Swahili–English code-mixed dialect. A high-quality Sheng phishing corpus was constructed and used to fine-tune a cross-lingual transformer (XLM-RoBERTa) for robust detection. Integrated Gradients (IG) were applied to provide token-level explanations, revealing the linguistic cues driving each prediction. The resulting model achieved 98.9 % accuracy and 97.6 % F1-score, and a containerized FastAPI service enabled real-time inference on platforms such as Facebook, TikTok, and YouTube. The proposed system delivers a production-ready pipeline for low-resource social-media cybersecurity and lays a foundation for future cross-lingual extensions.

KEYWORDS

Phishing detection, explainable AI, multilingual NLP

1. INTRODUCTION

Phishing remains one of the most prevalent cyber-threats worldwide, with social-media platforms becoming a major attack vector. While recent natural language processing (NLP) systems effectively identify phishing in English, low-resource and code-mixed languages remain vulnerable due to scarce labeled data and linguistic complexity. Sheng, a dynamic Swahili–English sociolect spoken in East Africa, exemplifies this gap. This study addresses the challenge by developing and evaluating an explainable AI (XAI) framework that detects phishing messages in low-resource multilingual settings and exposes the linguistic cues behind model predictions.

2. RELATED WORK

Early phishing detection relied on classical machine-learning methods with handcrafted features such as URLs and email headers [1], but these approaches struggled with semantic variation and the informal language common on social media. Recent advances in deep learning and transformer architectures (e.g., BERT, XLM-R) have greatly improved multilingual phishing and spam filtering [2], [3]. However, most research targets high-resource languages like English, leaving dialects and code-mixed text (e.g., Sheng,Swahili–English) largely unexplored. Low-resource NLP research has introduced data augmentation [4] and cross-lingual transfer [5] to address this gap. Models such as XLM-R and mBERT enable zero-shot and few-shot learning [6], but their predictions are often opaque. This limitation motivates the use of Explainable AI (XAI). Techniques such as Integrated Gradients and SHAP have been applied to phishing [7] and multilingual hate-speech detection [8], demonstrating that token-level attributions can reveal model reasoning. Building on these directions, the present work combines cross-lingual transformer fine-tuning with token-level explainability to tackle social-media phishing detection in a low-resource, code-mixed dialect.

3. METHODOLOGY

3.1 Data Construction

To overcome the absence of Sheng phishing datasets, we designed a multi-stage pipeline:

1. Audio Harvesting - Sheng speech was collated from YouTube, TikTok, and public social media posts using yt-dlp.
2. Transcription - Audio was transcribed with OpenAI Whisper configured for Swahili, accurately capturing code-mixing.
3. Text Augmentation - Additional data were sourced from Sheng dictionaries, sheng applications, and web-scraped scripts, then cleaned and balanced with English phishing examples.

The final dataset comprised of 9,970 labeled samples, enabling reliable fine-tuning of a large cross-lingual model. The composition of each source and the preprocessing steps are summarized in table 1, which details the number of samples collected from each stage and the specific cleaning procedures applied.

Table 1: Data Pipeline Statistics

Source	Language(s)	Samples Collected	Preprocessing Notes
Sheng Dictionaries & Apps	Sheng	3,800	Cleaned dictionary entries
Social Media (TikTok, FB, YouTube, X)	Sheng (code-mixed Swahili/English)	4,170	Web-scraped captions & comments
Internet Scraping	Multilingual (Sheng/English /Swahili)	2,000	Deduplicated and noise-filtered
Total	Multilingual	9,970	After cleaning, balancing, augmentation

Although this study used 9,970 carefully cleaned and balanced samples, data collection is ongoing to capture new phishing

strategies and expand low-resource coverage for future model updates.

3.2 Model Training

The multilingual phishing detector was trained using the XLM-RoBERTa-base architecture to leverage cross-lingual transfer while remaining lightweight enough for deployment. From the cleaned and augmented corpus of 9,970 labeled examples, the data was split into 7,976 training and 1,994 evaluation instances with stratification to preserve class balance. Training was performed on a single NVIDIA GPU for five epochs with a batch size of 16 and a linear-decay learning rate of 2×10^{-5}, yielding a total runtime of approximately 6.5 minutes. The final model achieved a validation accuracy of 0.992 and an F1-score of 0.981, demonstrating strong generalization despite the low-resource, code-mixed setting. Key hyperparameters and performance metrics are summarized in table 2, while Figure 1 provides an end-to-end view of the workflow from multilingual data acquisition through preprocessing, model training, explainability analysis, and containerized deployment .

Table 2: Model Training and Performance

Item	Value
Base model	XLM-RoBERTa-base
Dataset split	7,976 train / 1,994 eval
Epochs	5
Batch size	16
Learning rate	2×10^{-5} (linear decay)
Training time	6.5 min (single GPU)
Final eval accuracy	0.992
F1-score	0.981
Precision / Recall	0.990 / 0.972

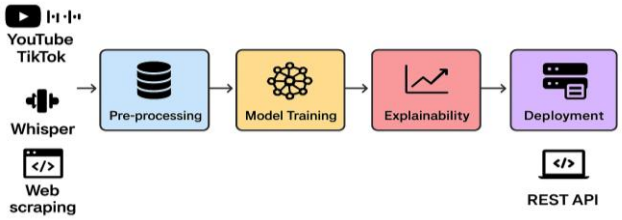


Figure 1: Full pipeline from multilingual data acquisition to explainable model deployment.

4. EXPLAINABILITY

Model interpretability was achieved using IG implemented via the Captum library. IG computed token-level attribution scores, enabling visualization of which words most strongly influenced the phishing predictions. As shown in Figure 2, high attributions were assigned to critical tokens such as link, loan, bila(meaning without), and click in the Sheng phishing message “Manze kuna link ya kuchukua loan ya 10K bila ID, click hapa!”. These highlighted terms correspond closely to human intuition, validating that the model bases its decisions on semantically meaningful cues and enhancing trust in the deployed system.

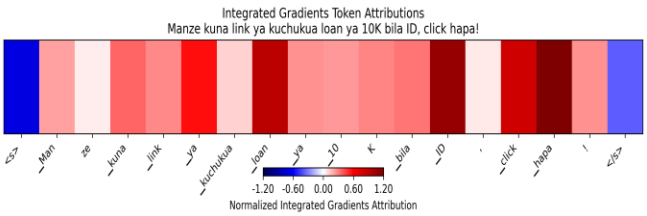


Figure 2. Token-level attribution heatmap for a Sheng phishing message generated using Integrated Gradients. Darker shades indicate tokens contributing more to the phishing prediction.

5. DEPLOYMENT

The final model and all dependencies were packaged into a Docker container running a FastAPI inference service. As illustrated in Figure 3, a client sends a text query to the containerized API, which hosts the fine-tuned model and returns both the predicted label and its confidence score. This /predict endpoint enables seamless integration into external systems. Example request: {"text": "click the link in my bio to get free credit"} Response: {"label": "phishing", "confidence": 0.9997} This design ensures portability, scalability, and real-time integration into security infrastructures.

Figure 3: Deployment architecture of the phishing detection system. A client sends a text query to the containerized FastAPI service, which hosts the fine-tuned model and returns the predicted label and confidence score to downstream servers.

6. CONCLUSION AND FUTURE WORK

This work introduced a production-ready, explainable multilingua lphishing detector targeting low-resource, code-mixed languages such as Sheng. A fine-tuned XLM-RoBERTa model with Integrat -ed Gradients achieved high accuracy and was deployed in a conta -inerized FastAPI service for real-time inference. We will continu eto collect data to capture new phishing tactics and broaden cross-lingual coverage. Future efforts will validate the framework on ad ditional dialects (e.g., Korean) and explore federated learning for privacy-preserving model updates.

ACKNOWLEDGMENTS

This research was supported by the Regional Innovation System & Education(RISE) program through the (Gwangju RISE Center), fu nded by the Ministry of Education(MOE) and the (Gwangju Me-tr opolitan City), Republic of Korea.(2025-RISE-05-013). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2024-00460142).

REFERENCES

[1] P. An, R. Shafi, T. Mughogho, and O. Onyango, “Multilingual Email Phishing Attacks Detection using OSINT and ML,” *arXiv:2501.08723*, 2025.

[2] H. Wang, J. Zhang, W. Zhang, and Y. Jiang, “Beyond the West: Bridging Gaps in Phishing Detection Between Western and Chinese Systems, 2025

[3] S. Alshattnawi et al., “Contextualized Representations for Enhanced Social Media Spam Detection,” Applied Sciences, vol. 14, no. 3, Mar. 2024.

[4] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale,” in Proc. ACL, pp. 8440–8451, 2020.
[5] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in *Proc. ICML*, 2017.
[6] S. Kavya and D. Sumathi, “Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection,” *Expert Systems*, Published online Dec. 2024.
[7] M. Islam, J. A. Khan, M. Abaker, A. Daud, and A. Irshad, “Unified Large Language Models for Misinformation Detection in Low-Resource Linguistic Settings,” *arXiv:2506.01587*, 2025.
[8] Phishing detection using machine learning techniques,” Security and Communication Networks, vol. 2022, , 2022.

자기지도 학습 기반 디지털 이미지 상관법(DIC)으로 실제 인장 실험 변형률 추정

정현경[1], 유경호[1], 김판구[1]*
조선대학교 컴퓨터공학과[1]
e-mail : gaeng@chosun.ac.kr, infinitegh@chosun.ac.kr, pkkim@chosun.ac.kr

Strain Estimation in Real Tensile Experiments Using Self-Supervised Learning-Based Digital Image Correlation (DIC)

Hyeon-Kyeong Jeong[1], Kyungho YU[1], Pan-Koo Kim[1*]
Dept of Computer Engineering, Chosun University[1]

요 약

본 연구는 재료 변형 측정에서 기존 디지털 이미지 상관(Digital image correlation, DIC)기법의 한계를 넘어, 딥러닝 기반 접근법을 탐색한다. 전통적인 DIC는 변형률과 변위를 정밀하게 측정 할 수 있는 기술이지만, 큰 변형 환경에서는 계산 속도가 느리고 매개변수 조정이 필요하여 효율성이 제한된다. 딥러닝을 활용한 DIC 연구에서는 충분한 라벨링 데이터를 확보하기 어려운점이 여전히 해결해야 할 과제로 남아있다.[1] 본 논문에서 제안하는 연구는 자기지도 학습(Self-Supervised Learning, SSL) 방식을 채택하여 라벨링 된 데이터셋에 대한 의존성을 근본적으로 해소하고자 한다. 이 방법은 별도의 정답 라벨없이, 연속된 두 이미지를 입력으로 받아 모델이 스스로 변위필드를 예측하고, 이를 기반으로 물리적 제약조건을 만족하도록 학습한다. 특히, 기존 연구에서 주로 사용했던 합성데이터셋이 아닌 실제 인장실험 데이터를 활용함으로써 모델의 현실 적용 가능성과 일반화 성능을 극대화 하는것을 목표로 한다.

1. 서 론

1.1 기존 DIC의 보편성과 한계

DIC는 실험 역학 분야에서 비접촉식 전체필드 변형 측정을 위한 기술로 널리 사용되어왔다.[1] 이 기술은 재료 표면에 적용된 불규칙한 스페클패턴(speckle patt-ern)의 변화를 분석하여 변위와 변형률을 계산하는 원리를 기반으로 한다. 비접촉 방식의 특성 덕분에 시편의 물성을 변경하는 센서부착이 필요 없이 변형 측정이 가능하다. 그러나 기존 DIC 기법은 몇가지 한계를 가진다. 첫째, 서브셋 기반 DIC는 반복적인 계산을 필요로 하며, 많은 연산자원과 시간이 소요되기때문에 결과적으로 실시간 분석이 어렵다.[2] 둘째, 스페클패턴 손상이나 큰 변형 시 예측 안정성이 떨어진다.[4] 셋째, 필터링으로 공간 해상도가 저하되며, 서브셋크기나 스텝사이즈를 수동적으로 조정해야하는 등의 자동화가 어렵다.[5] 이러한 매개변수 조정에는 사용자의 전문 지식을 요하며, 국부 변형 측정 효율을 떨어뜨린다.

1.2 딥러닝의 가능성과 데이터 병목현상

기존DIC의 한계를 극복하기 위한 딥러닝의 도입이 새로운 접근법으로 떠오르고 있다. 딥러닝 모델은 입력 이미지 쌍으로부터 변위 및 변형률을 종단 간(end-to-end) 으로 예측하여, 실시간에 가까운 자동 측정을 가능하게 한다. 대표적인 예시로 Deep DIC 모델은 계산 시간을 밀리초 단위로 단축하여 상용 소프트웨어 대비 매우 견고한 예측 성능을 보여주었다.[2]

그러나 딥러닝기반 DIC연구에는 핵심적인 문제가 있는데, 라벨링 된 데이터셋의 부족이다. 딥러닝, 특히 지도학습은 모델을 훈련하는데에 정답(Ground Truth)이 포함된 많은 양의 데이터가 필수적이다. DIC에서 이 정답데이터는 변형필드 이며, 이를 실제 실험에서 정확하게 수집하는것은 매우 어렵고 시간과 비용이 많이 소모된다. 이 때문에 많은 기존 딥러닝 기반 DIC연구는 합성 데이터셋에 의존하는 경향이 있다.[3]

2. 관련 연구

2.1 DIC를 활용한 재료연구

DIC는 다양한 재료의 변형 거동을 분석하는데 널리 사용되어 왔다. 특히, 국부적인 변형률 분포를 측정하는 능력 덕분에 균열 전파 및 응력집중과 같은 복잡한 파괴현상을 연구하는데 필수적인 도구가 되었다. 그러나 이 기법은 스페클 패턴의 품질에 크게 의존하기 때문에 고온 환경이나 파단 직전과 같이 패턴이 훼손되는 상황에서는 측정 정확도가 크게 떨어진다.

2.2 자기지도 학습기반 비전연구

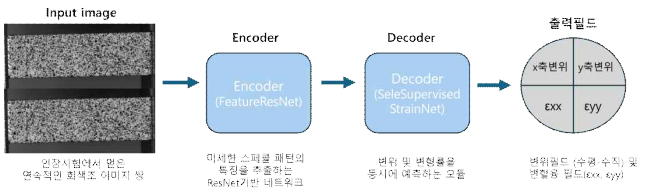
최근 컴퓨터 비전 분야에서는 라벨링 된 데이터의 한계

를 극복하기 위해 자기지도 학습을 활발히 연구하고있다.[6] 이 방식은 데이터 자체에서 의사라벨(pseudo-label)을 생성하여 이를 통해 학습을 진행하게 하는데, 이는 연속된 프레임 간 움직임 예측 또는 두 이미지간의 대응관계를 찾는 분야에서 두드러진 결과를 보인다.[7][8] DIC 문제 역시 변형 전후 이미지 쌍으로부터 변형 필드를 복원하는 성질을 지니므로 이러한 자기지도 학습 접근법과 본질적으로 유사하다.

3. 본 론

3.1 모델 아키텍처 상세화

본 연구에서 제안하는 딥러닝 모델은 FeatureResNet과 SelfSupervisedStrainNet 이라는 두가지 핵심 모듈로 구성된다. FeatureResNet은 인코더로 인장 실험중 촬영된 실제 이미지데이터인 두개의 연속된 흑백이미지를 입력으로 받는다. 이 2채널 입력은 ResNet아키텍처를 기반으로 설계된 인코더를 통과하려 이미지의 미세한 특징을 추출한다.[그림1] 이 인코더의 계층구조는 각 3개, 14개, 16개, 3개의 컨볼루션 블록으로 이루어져, 깊은 네트워크를 통해 이미지의 특징을 효과적으로 학습한다. 이러한 구조는 입력 이미지의 공간해상도를 점진적으로 줄이면서 고차원의 의미론적 특징을 추출하므로, 미세한 스페클 패턴 변화까지 포착하는데 적합하다. 추출된 특징맵은 디코더 역할을 하는 SelfSupervisedStrainNet의 입력으로 들어가 변위필드와 변형률 필드를 동시에 예측한다. 최종적으로 displacement_head는 수평 및 수직변위의 2채널 변위필드를, strain_head는 εxx, εyy, γxy의 3채널 변형률 필드를 예측한다. 이같은 종단 간(end-to-end) 구조는 기존의 DIC기법이 변위 계산 후 변형률을 미분하는 과정에서 발생하는 노이즈 문제를 해결하고 변형률 예측의 정확도를 크게 향상 시킨다.[9]



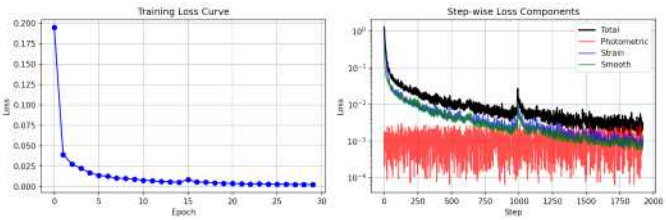
(그림 1) (자기지도 기반 딥러닝 모델구조)

3.2 물리적 제약을 포함하는 손실함수 설계

본 연구의 자기지도 학습 모델의 핵심은 모델이 물리법칙을 스스로 학습하도록 손실함수를 설계하는것을 목표로 한다. 그러나 본 연구에서는 손실 함수 가중치의 불균형과 데이터 전처리 과정의 한계로 인해 모델이 물리 법칙을 효과적으로 학습하는데 어려움을 겪었다. 모델의 학습은 세가지 핵심 손실함수를 결합하여 이루어진다. 예측된 변위필드를 기반으로 첫번째 이미지를 워핑(warping)하여 두번째 이미지와 픽셀 강도 차이를 최소화하는 포토메트릭(Photometric)손실은 라벨이 없는 상태에서 모델이 스스로 학습 목표를 생성하게 하는 핵심적인 손실함수이다. 또한, 변형률 일관성(Strain Consistency)손실은 StrainNet이 직접 예측한 변형률이 변위필드의 미분과 물리적으로 일치하도록 강제하는역할을 한다. 이는 물리 정보를 딥러닝 모델에 통합하는 물리 정보 신경망(Physics-Informed Neural Network, PINN) 패러다임의 한 형태이다. 마지막으로, 부드러움(Smoothness)손실은 예측된 변위 필드의 불규칙성을 최소화하는 정규화 기법으로, 예측된 변형이 물리적으로 현실적인지 확인하고 측정 과정에서 발생하는 노이즈를 억제하는데 필수적이다.[그림2] 세가지 손실함수를 결합한 이 프레임워크는 변형필드를 예측하도록 설계되었지만, 실제 실험에서는 손실 가중치 불균형과 더불어 시편 외의 영역이 학습에 포함되어 모델이 변위를 0으로 수렴시키거나 그림의 움직임에 집중하는경향을 보였다.[그림3] 따라서 모델의 예측성능을 높이기 위해서는 이러한 가중치들을 최적화하고, 동시에 정확한 관심영역 설정이 필수임을 알수있다.

손실 함수	목적	기반 원리 및 제약 조건
포토메트릭 손실	예측된 변위 필드의 정확한 재구성 유도	포토메트릭 일관성(Photometric Consistency): 변형 전후 픽셀 강도 불변 가정
변형률 일관성 손실	예측된 변위와 변형률 간의 역학적 일관성 보장	변형률-변위 관계(Strain-Displacement Relations): 연속체 역학의 기본 법칙
부드러움 손실	예측된 변형 필드의 노이즈 억제 및 물리적 현실성 보장	부드러움(Smoothness) 제약: 물리적 변형의 국부적 평활성

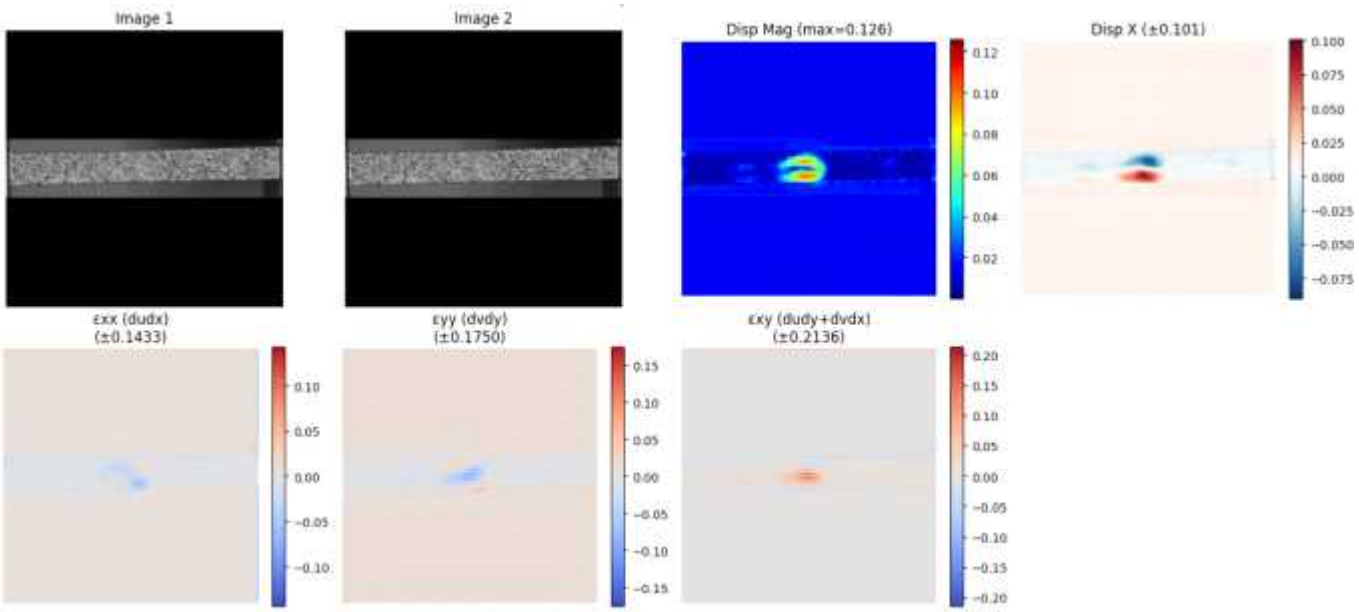
(그림 2) (본 연구에서 사용된 손실함수 구성)



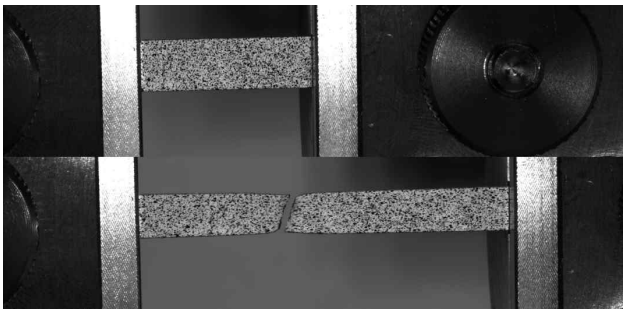
(그림 3) (실제실험에서 손실값을 0으로 수렴시키려는 경향을 보였음)

3.3 실제 실험 데이터의 결정적 역할

실제 인장 실험에서 획득한 데이터는 모델의 신뢰성 검증에 핵심적인 요소이다.[그림5] 본 연구에서는 라벨링 되지않은 실제 실험데이터를 활용하여 자기지도 학습 모델을 훈련 하였으며, 이를 통해 모델은 현실적인 노이즈와 이미지 결함 속에서도 일반적인 인장 변형을 예측하려는 능력을 보였다. 그러나 큰 변형을 포착하는데에는 한계를 드러냈다. 이러한 결과는 합성데이터만으로 충분히 모사하기 어려운 실제 데이터의 복잡성 속에서 모델이 학습하는 방식의 중요성을 시사한다. 인장 실험 중 시편 표면에 도포된 스페클 패턴과 고해상도 카메라로 촬영된 이미지에는 조명 변화, 패턴 불균형, 미세한 흔들림 등 모델이 반드시 학습해야 할 현실적인 요소들이 포함되어 있다.[그림5] 따라서 실제 실험 데이터를 활용하는것은 단순히 데이터 소스를 변경하는 수준을 넘어, 시뮬레이션 기반 학습이 가지는 한계를 극복하고, 실제 재료거동을 반영할 수 있는 견고한 DIC 모델 개발을 위한 필수적인 접근방식임을 확인하였다. 그러나 모델의 예측 성능을 극대화하기 위해서는 데이터 전처리 과정과 학습 목표설정에 대한 추가적인 최적화가 반드시 필요하다.



(그림 4) (모델이 예측한 인장 시편의 변형 전후 이미지와 함께 변위크기, x방향변위, x및 y방향 수직변형률(εxx, εyy), 그리고 전단변형률(εxy)분포를 보여준다. 특히 시편 중앙부에서 발생하는 국부적인 변위와 변형률 집중현상을 시각화한다.)



(그림 5) (조명, 시편보다 큰 배경등 전처리가 필요한 실제 촬영된 원본 인장실험 데이터)

4. 결론 및 향후연구

본 연구는 라벨링 된 데이터 없이도 자기지도 학습을 활용하여 인장 실험의 초기 변형 구간을 예측할 수 있음을 보여주었다.[그림 4] 그러나 파단 예측에 있어서는 모델이 실제 변형을 포착하지 못하고 변위 0에 가까운 수렴을 보여주며 한계를 드러냈다.[그림3] 이는 손실함수의가중치 설계와 최적화 설계상의 제약을 반영하며, Photometric Loss 와 Strain Loss의 특성때문에 급격한 변형을 과도하게 페널티로 간주함으로써 모델이 해당 영역을 정확히 예측하는 데 어려움을 겪은것이 원인으로 파악된다. 또한 데이터 전처리 과정에서 시편의 동적 변형과 배경 노이즈로 인해 관심영역(ROI)의 추적이 완벽하지 않았으며, 크롭 이미지 사이즈의 일관성 문제 역시 예측성능에 영향을 미쳤다. 향후 연구에서는 다음과 같은 개선방향을 제시한다. 첫째, 딥러닝 기반 객체 감지모델을 활용하여 이미지 시퀀스 전체의 변형을 추적하고, 파단면을 포함하는등 동적ROI를 자동으로 설정하는 방법에 대한 연구 진행. 둘째, 자동화된 하이퍼파라미터 최적화 기법(예 : 베이지안 최적화)등을 도입하여 모델 성능 극대화 방법 연구. 마지막으로 다른 데이터셋으로 사전 학습 된 모델 활용을 통해 데이터 다양성 확보. 이를통해 제안된 모델은보다 견고하게 발전 할 수 있으며, 재료의 미세변형 분석 및 예측 유지보수 시스템에 적용 가능한 실용적 성능을 확보할수 있을것으로 기대된다.

감사의 글

본 과제(결과물)는 2025년도 교육부 및 광주광역시의 재원으로 광주RISE센터의 지원을 받아 수행된 지역혁신중심대학지원체계(RISE)의 결과입니다.(2025-RISE-05-013),

본 연구성과물은 2024년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. RS-2024-00463238)

참 고 문 헌

[1] Khoo, S-W., Saravanan Karuppanan, and C-S. Tan. "A review of surface deformation and strain measurement using two-dimensional digital image correlation." *Metrology and Measurement Systems* 23.3 (2016): 461-480.

[2][Yang, Ru, et al. "Deep DIC: Deep learning-based digital image correlation for end-to-end displacement and strain measurement." Journal of Materials Processing Technology 302 (2022): 117474.]

[3] Niu, Bangyan, and Jingjing Ji. "Convolutional neural network based denoising for digital image correlation reconstructing high-fidelity deformation field." 2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2023.

[4] Huang, Jiasheng, Simone Paganoni, and Emanuele Zappa. "3D-DIC for large displacement with laser-based speckle patterns." *IEEE Transactions on Instrumentation and Measurement* 72 (2023): 1-8.

[5] Wigger, Tim, Colin Lupton, and Jie Tong. "A p

arametric study of DIC measurement uncertainties o n cracked metals." *Strain* 54.6 (2018): e12291.
[6] Jing, Longlong, and Yingli Tian. "Self-supervise d visual feature learning with deep neural network s: A survey." *IEEE transactions on pattern analysi s and machine intelligence* 43.11 (2020): 4037-4058.
[7] Florence, Peter, Lucas Manuelli, and Russ Tedra ke. "Self-supervised correspondence in visuomotor p olicy learning." *IEEE Robotics and Automation Let ters* 5.2 (2019): 492-499.
[8] Liu, Pengpeng, et al. "Selflow: Self-supervised l earning of optical flow." *Proceedings of the IEEE/ CVF conference on computer vision and pattern re cognition*. 2019.
[9] Effect of DIC Spatial Resolution, Noise and Int erpolation Error on Identification Results with the VFM

Attention-Enhanced Optimized Deep Ensemble Network For Effective Facial Emotion Recognition

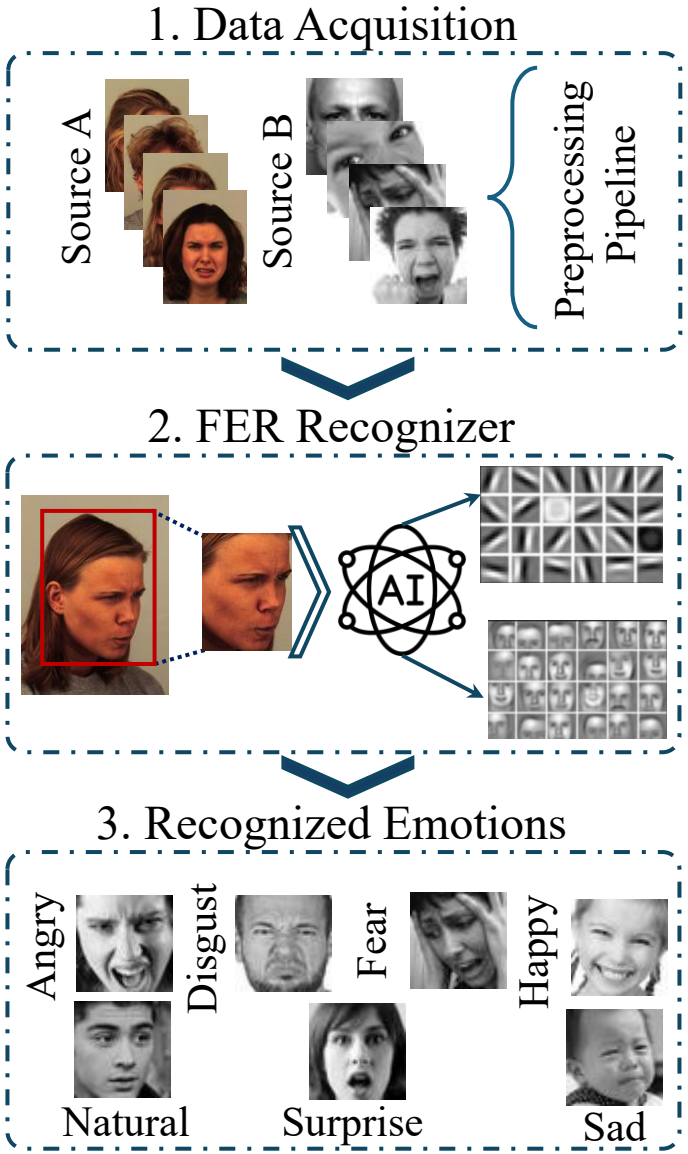
Taimoor Khan, Chang Choi
Department of Computer Engineering, IT Convergence, Gachon University
E-mail : taimooricp@gmail.com, changchoi@gachon.ac.kr

가천대학교
Gachon University

Table of Content

1. Overview

- Facial Expression Recognition (FER) aims to automatically recognize human emotions by analyzing facial expressions using AI models.
- FER has diverse applications, including enhancing human-computer interaction, enabling emotion-aware healthcare systems, improving intelligent surveillance, personalizing entertainment experiences, and supporting affective learning in educational settings.
- **Problem in Existing Methods:** Existing FER methods are suffered due to:
  - *High Computational Cost:* Complex models require significant processing power, making real-time FER challenging on devices with limited resources.
  - *Difficulty Recognizing Small Facial Changes:* Existing methods often struggle to detect emotions when there are slight facial changes or when the input data distribution varies from the training data.
  - *Sensitivity to Noisy Inputs:* Performance drops with blurry, unclear, or low-quality images. Models that perform well in controlled settings often struggle in unpredictable, real-world environments.
- **Proposed Solution:** We propose EA-Net, an attention-based ensemble framework for accurate FER comprising two phases: preprocessing and model training. Preprocessing applies data augmentation and super-resolution to boost data quantity and quality. The model uses parallel EfficientNetB0 and InceptionV3 for feature extraction, followed by C_{AM} and S_{AM} for key feature selection, and FC layers for final emotion classification.



2. Applications

Real-Time Applications of FER Across different Domains

01

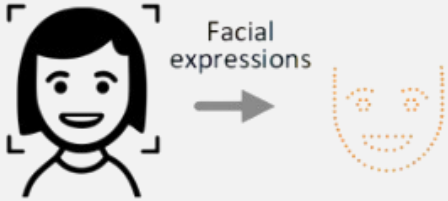
Healthcare and Mental Health

Monitoring patient emotions in therapy or pain detection.

02

Security and Surveillance

Identifying suspicious behavior through facial cues.



03

Driver Monitoring Systems

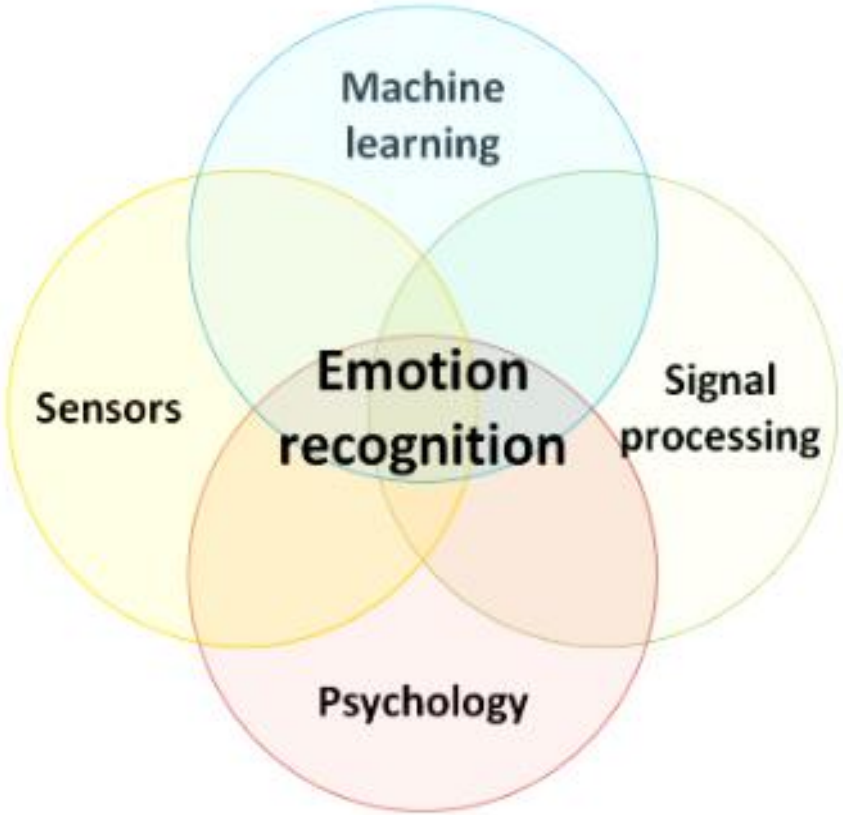
Detect drowsiness, distraction, or emotional distress to trigger alerts or safety measures.

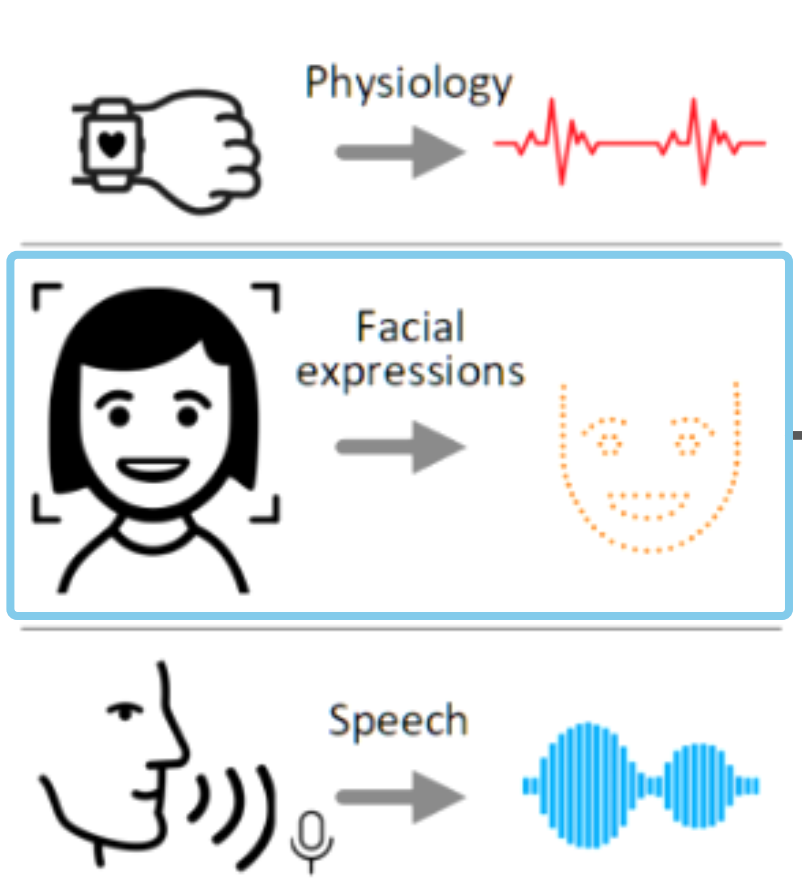
04

Education

Gauging student engagement and emotional response during learning.

3. Literature Review





Landmark-based Methods

Machine Learning

Deep Neural Learning

Problem Statement & Research Objectives

- **Problem Statement:** Current FER models struggle with subtle expressions and noisy inputs, leading to reduced performance. There is a need for a more accurate and robust FER framework focused solely on maximizing recognition performance in real-world conditions.

- **Research Objectives:** The objectives of this research work is outlined as:
  - **Preprocessing** techniques to improve data quality and quantity. Super-resolution enhances low-resolution images by restoring finer details, while data augmentation (rotation, flipping, resizing, de-colorization) increases dataset diversity, boosting the model's generalization and FER performance.
  - **Improve** the accuracy of facial emotion recognition (FER) by designing a dual-backbone architecture that captures both fine-grained textures and high-level semantic features.
  - **Enhance** the feature representation through the integration of modified dual attention modules, allowing the model to focus dynamically on the most expressive facial regions.
  - **Achieve** robust performance across real-world FER benchmarks (FER and KDEF datasets) despite challenges like subtle expressions and noisy inputs.

4. Proposed Method

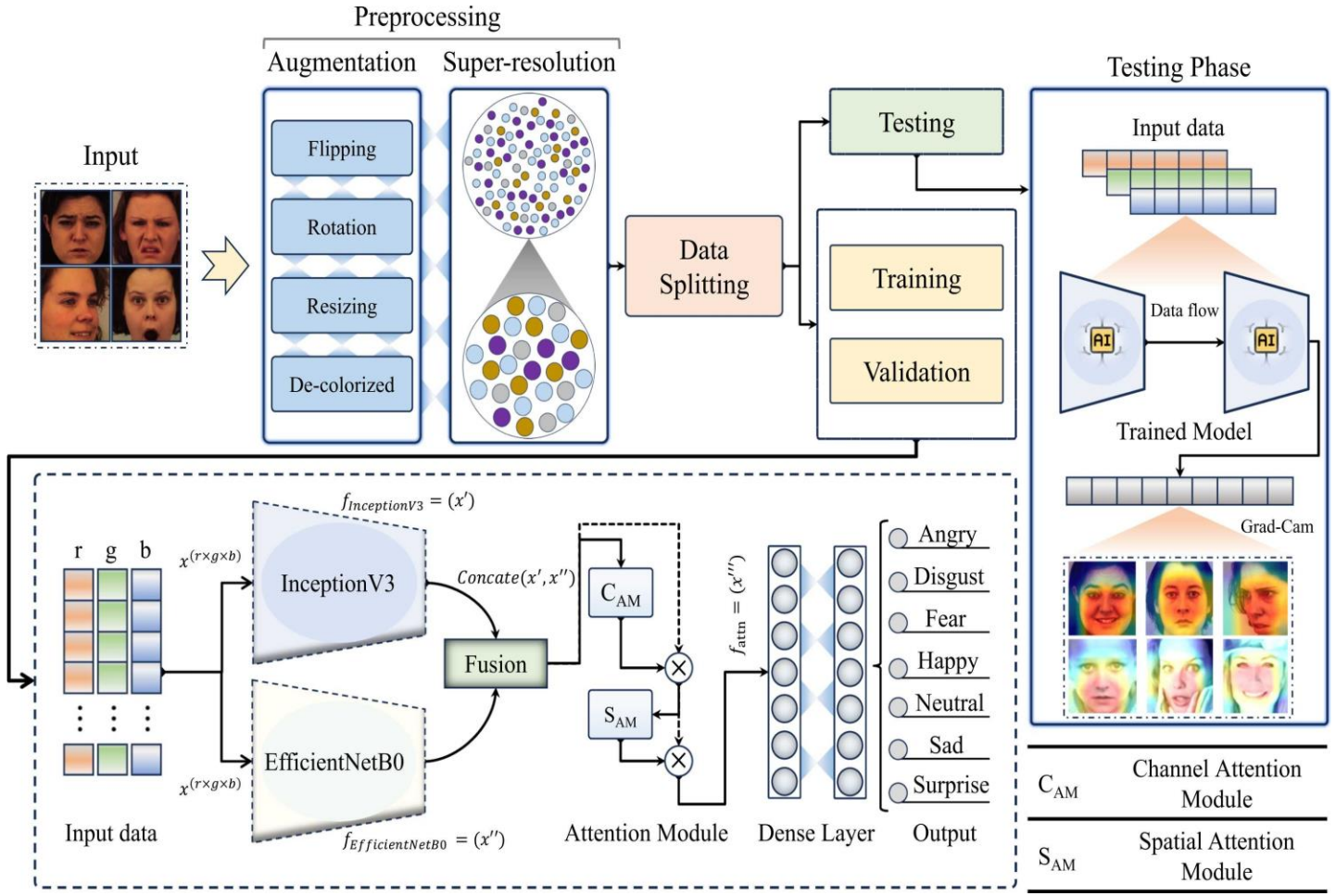


Fig. 1. High-level framework for effective FER: Input data undergo preprocessing (data augmentation and super-resolution) before being fed into the proposed EA-Net for FER.

- **Data Enhancement:** To address data scarcity and low resolution in public FER datasets, we apply data augmentation (rotation, flipping, resizing, de-colorizing) and super-resolution to improve input quality and quantity.
- **EA-Net Framework:** We propose EA-Net, an ensemble of EfficientNetB0 and InceptionV3 for robust feature extraction, combining outputs via addition and refining them through attention modules.
- **Attention Mechanism:** Sequential channel and spatial attention modules (CAM & SAM) enhance relevant features, followed by FC layers with ReLU and SoftMax for classification.
- **Evaluation:** Extensive experiments on FER and KDEF benchmarks, including ablation studies, show EA-Net outperforms SOTA methods across precision, recall, F1-score, and accuracy.

5. Results and Discussion

Table 1. Detailed information about the KDEF and FER datasets.

KDEF dataset				FER				
Classes	Number of Samples	Number of images before augmentation	Number of images after augmentation	Classes	Number of Samples	Image resolution before upscaling	Image resolution after upscaling	Total number of images
Angry	840	4900	5868	Angry	4953	48x48	196x196	35,887
Disgust	918			Disgust	547			
Fear	762			Fear	5031			
Happy	858			Happy	8989			
Neutral	912			Neutral	6198			
Sad	975			Sad	6078			
Surprise	603			Surprise	4002			

Fig. 5. Image samples from the KDEF and FER datasets.

$$Precision = \left(\frac{TP}{TP + FP}\right),$$

$$Recall = \left(\frac{TP}{TP + FN}\right),$$

$$F1 - score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right),$$

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right).$$

5. Results and Discussion

Table 2. Comparative analysis of proposed network and SOTA techniques on the FER and KDEF datasets. Here, bold text in dicates the best performance; red text represents the second-best performance.

Method	FER dataset	KDEF
	Accuracy	Accuracy
VGG [23]	65.80%	86.75%
Mollahosseini [40]	66.40%	--
DenseNet201 [16]	68.52%	92.52%
FaceLiveNet [20]	68.60%	-
InceptionV3 [17]	68.86%	90.25%
Inception-ResNetV2 [18]	69.72%	94.70%
Dense_FaceLiveNet [21]	69.99%	95.89%
Deep Fusion [28]	--	98.30%
PDREP [39]	73.5%	76.33%
transfer learning DCNN [41]	62.30%	--
FMA + MLP [42]	59.77%	92.275%
FMA + LD [42]	66.60%	93.665%
FMA + SVM [42]	61.11%	92.045%
DCNN [43]	63.80%	89.54%
DBN [44]	--	90.22%
GA-Dense-FaceLiveNet [45]	--	99.17%
CBiLSTM [26]	58.09%	94.23%
iVABL [10]	69.60%	95.63%
VGG [47]	69.65%	95.92%
GA [46]	77.4%	--
Proposed Model	**78.60%**	**99.30%**

5. Results and Discussion

Table 3. Comparative analysis of the proposed network and SOTA techniques in terms of precision, recall, and F1-score on t he FER and KDEF datasets. Here, bold text indicates the best performance; red text indicates the second-best performance.

Method	KDEF			FER		
	Precision	F1-score	Recall	Precision	F1-score	Recall
transfer learning DCNN [41]	0.86	0.86	0.86	0.597	0.61	0.63
FMA + MLP [42]	0.73	0.72	0.73	0.59	0.60	0.60
FMA + LD [42]	0.78	0.78	0.79	0.62	0.64	0.67
FMA + SVM [42]	0.72	0.71	0.72	0.59	0.60	0.61
DCNN [43]	0.87	0.86	0.85	0.60	0.62	0.62
DBN [44]	0.89	0.88	0.87	--	--	--
CBiLSTM [26]	0.93	0.92	0.92	0.55	0.56	0.58
iVABL [10]	0.95	0.94	0.94	0.66	0.68	0.70
VGG [47]	0.96	0.96	0.96	--	--	--
GA [46]	--	--	--	**0.77**	0.77	0.77
Proposed EA-Net	**0.99**	**0.99**	**0.98**	0.76	**0.77**	**0.79**

5. Results and Discussion

Table 4. Detailed ablation study of different pretrained models before preprocessing on the FER and KDEF datasets.

S: No	Technique	C_{AM}	S_{AM}	FER				KDEF			
				Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
1	InceptionV3	×	×	0.410	0.375	0.392	39.50%	0.737	0.783	0.760	74.47%
2		✓	×	0.400	0.419	0.382	41.50%	0.757	0.804	0.780	76.60%
3		×	✓	0.402	0.414	0.408	42.29%	0.774	0.822	0.798	78.72%
4		✓	✓	0.392	0.430	0.410	42.50%	0.815	0.831	0.823	81.05%
5	EfficientNetB0	×	×	0.504	0.500	0.502	40.50%	0.767	0.800	0.783	77.66%
6		✓	×	0.376	0.417	0.395	42.00%	0.798	0.814	0.806	79.79%
7		×	✓	0.516	0.512	0.514	42.50%	0.800	0.824	0.812	80.32%
8		✓	✓	0.411	0.442	0.426	43.50%	0.820	0.845	0.832	82.45%
9	Ensemble	×	×	0.392	0.425	0.408	42.00%	0.849	0.865	0.857	84.21%
10		✓	×	0.402	0.436	0.418	43.00%	0.886	0.851	0.868	86.10%
11		×	✓	0.415	0.446	0.430	44.50%	0.917	0.899	0.908	90.27%
12		✓	✓	**0.446**	**0.479**	**0.462**	**46.50%**	**0.938**	**0.920**	**0.929**	**92.43%**

5. Results and Discussion

Table 4. Detailed ablation study of different pretrained models after preprocessing on the FER and KDEF datasets.

S: No	Model	C_{AM}	S_{AM}	FER				KDEF			
				Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
1	InceptionV3	×	×	0.4519	0.4796	0.4653	46.00%	0.7615	0.7981	0.7793	76.50%
2		✓	×	0.4811	0.5100	0.4951	48.00%	0.7727	0.8095	0.7907	77.50%
3		×	✓	0.5370	0.5472	0.5421	51.00%	0.8381	0.8381	0.8148	80.00%
4		✓	✓	0.6087	0.6195	0.6140	56.00%	0.8393	0.8624	0.8507	83.50%
5	EfficientNetB0	×	×	0.5455	0.5660	0.5556	52.00%	0.8581	0.8316	0.8446	83.96%
6		✓	×	0.5714	0.5872	0.5792	53.50%	0.8466	0.8817	0.8638	85.73%
7		×	✓	0.5913	0.6126	0.6018	55.00%	0.8580	0.9025	0.8797	87.33%
8		✓	✓	0.6460	0.6293	0.6376	59.71%	0.8964	0.9240	0.9100	90.42%
9	Ensemble	×	×	0.6387	0.6496	0.6441	58.00%	0.9703	0.9333	0.9515	94.71%
10		✓	×	0.6441	0.6667	0.6552	61.17%	0.9806	0.9712	0.9758	97.34%
11		×	✓	0.7155	0.7281	0.7217	68.93%	0.9961	0.9751	0.9855	98.40%
12		✓	✓	**0.7610**	**0.7996**	**0.7798**	**78.60%**	**0.9961**	**0.9865**	**0.9913**	**99.30%**

5. Results and Discussion

Figure 6. Image localization performance of the proposed network on KDEF and FER datasets.

6. Conclusion

- This study proposes the advanced EA-Net framework to significantly improve facial emotion recognition (FER) performance.
- Preprocessing techniques, including data augmentation and super-resolution, are applied to increase the input dataset size and enhance image quality, boosting model accuracy.
- The model uses an ensemble approach combining EfficientNetB0 and InceptionV3 backbones in parallel for rich feature extraction.
- Channel Attention Module (C_{AM}) and Spatial Attention Module (S_{AM}) are sequentially integrated to focus on the most relevant facial features.
- Extensive experiments on FER and KDEF datasets show EA-Net achieves 78.60% and 99.30% accuracy respectively, outperforming state-of-the-art methods.
- Ablation studies demonstrate the impact of preprocessing and attention modules, confirming the model's superior precision, recall, F1-score, and accuracy..

6. Study Limitation and Future Direction

- Although the proposed EA-Net exhibits promising capabilities, it has limitations, notably in terms of performance and c omputational complexity.
- In this study, the proposed EA-Net is trained on only two publicly available datasets, along with the application of differ ent data preprocessing strategies.
- However, the proposed model should be explored on other well-known datasets, including CK+, 4DFAB, MMI, JAFFE, Oulu–CASIA, EmotioNet, and AffectNet, and compared with SOTA techniques to further evaluate its generalizability.
- In addition, the proposed EA-Net is larger than the comparison techniques; thus, it requires optimization in terms of mo del size, parameters, Mega floating-point operations, and frame per second for real-time decision making on edge devic es.
- In the future, we plan to further optimize the proposed network using different techniques, including GAs, pruning, and quantization, to reduce the number of network parameters with optimal performance and enable real-time implementati on on resource-constrained platforms, e.g., Raspberry Pi and Jetson Nano.
- In addition, we plan to explore other challenging datasets for FER and further compare the proposed EA-Net with existi ng SOTA networks to evaluate its efficiency and effectiveness on diverse data.

References

1. Mannepalli, K., P.N. Sastry, and M. Suman, *A novel adaptive fractional deep belief networks for speaker emotion recognition.* Alexandria Engineering J ournal, 2017. **56**(4): p. 485-497.
2. Nan, Y., et al., *A-MobileNet: An approach of facial expression recognition.* Alexandria Engineering Journal, 2022. **61**(6): p. 4435-4444.
3. Jeong, M. and B.C. Ko, *Driver's facial expression recognition in real-time for safe driving.* Sensors, 2018. **18**(12): p. 4270.
4. Shen, F., et al., *EEG-based emotion recognition using 4D convolutional recurrent neural network.* Cognitive Neurodynamics, 2020. **14**: p. 815-828.
5. Yun, S.S., et al., *Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system.* Autism Research, 201 7. **10**(7): p. 1306-1323.
6. Kaulard, K., et al., *The MPI facial expression database—a validated database of emotional and conversational facial expressions.* PloS one, 2012. **7**(3): p. e32321.
7. Canal, F.Z., et al., *A survey on facial emotion recognition techniques: A state-of-the-art literature review.* Information Sciences, 2022. **582**: p. 593-617.
8. Mellouk, W. and W. Handouzi, *Facial emotion recognition using deep learning: review and insights.* Procedia Computer Science, 2020. **175**: p. 689-69 4.
9. Ali, M., et al., *Facial expressions can detect Parkinson's disease: Preliminary evidence from videos collected online. NPJ digital medicine. 2021; 4 (1): 1–4*. DOI.
10. Rakshith, M. and H.H. Kenchannavar, *Hybrid Deep Optimal Network for Recognizing Emotions Using Facial Expressions at Real Time.*

Carcagnì, P., et al., *Facial expression recognition and histograms of oriented gradients: a comprehensive study.* SpringerPlus, 2015. **4**(1): p. 1-25.

12. Chen, L., C. Zhou, and L. Shen, *Facial expression recognition based on SVM in E-learning.* Ieri Procedia, 2012. **2**: p. 781-787.

Acknowledgement

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2023-00259004)

Thank You!

멀티모달 기반 패션 아이템 판매량 예측 시스템 연구

이새봄[1], 최창[1]*

가천대학교 컴퓨터공학과[1]

e-mail : dltoqha@gachon.ac.kr, changchoi@gachon.ac.kr

목차

요약

- 네 가지 모달리티(Temporal, Text, Vision, Trend)를 통합하는 Transformer 기반 M4FT(Multimodal Quad Fusion Transformer)을 제안함
- M4FT는 계층적 2단계 Fusion Network를 통해 모달리티 고유의 특성을 보존하면서도 세밀한 모달리티 간 상호작용을 학습함
- 실제 전자상거래 데이터에서 최첨단 베이스라인 모델 대비 약 19% 낮은 성능을 보여, 멀티모달 데이터를 효과적으로 융합할 수 있음을 보여줌

1. 연구 목표

- 본 연구는 신제품 수요 예측의 불확실성을 줄이고 정확성을 높이기 위해 멀티모달 기반 패션 아이템 판매량 예측 시스템, M4FT(Multimodal Quad Fusion Transformer)를 제안함

- M4FT는 시계열 판매 데이터(Temporal), 제품 설명(Text), 이미지 정보(Vision), 인터넷에서 수집한 외부 검색도(Google Trend) 간 네 가지 이질적 모달리티를 동시에 활용함

- M4FT는 단순히 데이터를 병합하는 방식이 아니라, 모달리티 별 특성을 보존하면서도 모달리티 간의 정교한 상호작용을 학습할 수 있도록 2단계 융합 전략을 도입함

- 1단계에서는 Temporal-Text Feature Fusion Network를 통해 시계열과 텍스트 정보를 결합하여 시간적·언어적 특성을 반영하고, 2단계에서는 Text-Vision Feature Fusion Network를 통해 텍스트와 이미지 정보를 융합한 뒤 외부 트렌드를 최종적으로 통합함

- 본 모델은 최첨단 베이스라인 모델 대비 약 34% 낮은 WAPE와 약 19% 낮은 MAE를 기록함

2. 연구의 필요성

- 디지털 경제의 성장으로 전자상거래 소비자들의 구매 행동이 복잡하고 다양해지고 있음
- 특히, 패션 산업은 시장 반응이 빠르고 제품 수명이 짧아 수요 예측에 어려움이 큼
- 신제품 출시 직후 수요를 정확히 예측하는 일은 재고 운영, 공급망 관리, 마케팅 전략 등 전반적인 경영 의사 결정에 직접적으로 연결되기 때문에 기업에 매우 중요한 과제임
- 그러나 실제 신제품 수요 예측은 유행 주기 단기화, 계절성, 소셜 미디어 실시간 트렌드 확산 등으로 인해 높은 불확실성과 변동성 동반함
- 또한, 기존 시계열 기반 예측 모델은 과거 데이터 충분할 때는 안정적 성능 보이지만, 판매 이력 부족하거나 전혀 없는 신제품에는 성능 급격히 저하되는 한계가 존재함
- 따라서, 따라서 신제품 수요를 안정적이고 정확하게 예측하기 위해서는 복잡한 시장 변동성과 다양한 요인을 반영할 수 있는 새로운 접근 방법이 필요함

3. 연구 내용

- **Dataset : VISUELLE**
  - E-commerce Sales Data from Italian Fast-Fashion Company(Nunalie) (2016.10 ~ 2019.12)
  - **Product Image** : 배경이 제거된 정면 패션 아이템 이미지 (Total : 5,577)
  - **Product Attribute Text** : 색상, 원단, 카테고리 등 텍스트
  - **Product Attribute Temporal** : 출시 이후 최대 26주간 판매량, 할인율, 판매 가격 등
  - **External Google Trend** : Google Trends 기반 제품 속성 관련 키워드 주간 검색 인기도

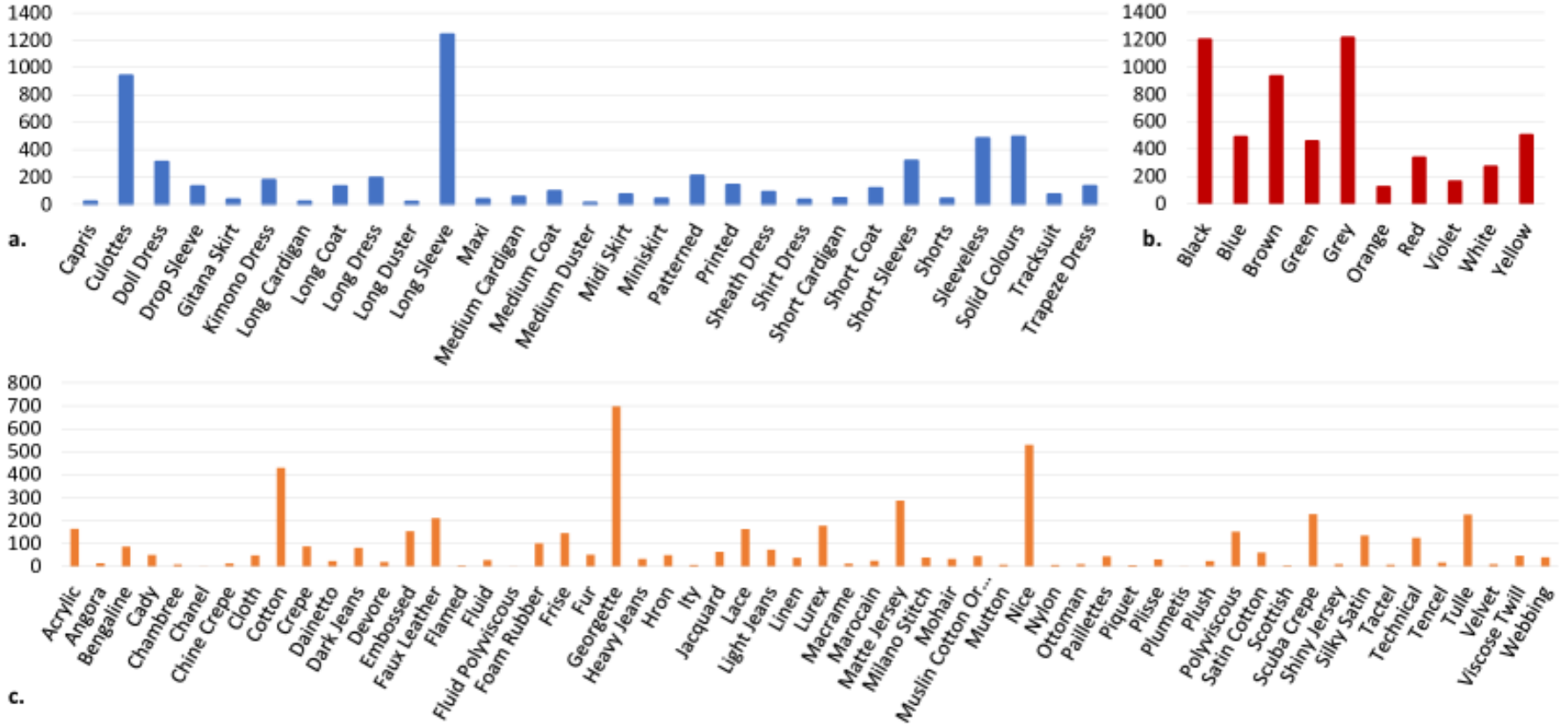
3. 연구 내용

- **Dataset : VISUELLE**
  - **Product Image** : 배경이 제거된 정면 패션 아이템 이미지 (Total : 5,577)

(그림 1) VISUELLE 데이터 내 다양한 제품 카테고리 이미지 샘플

3. 연구 내용

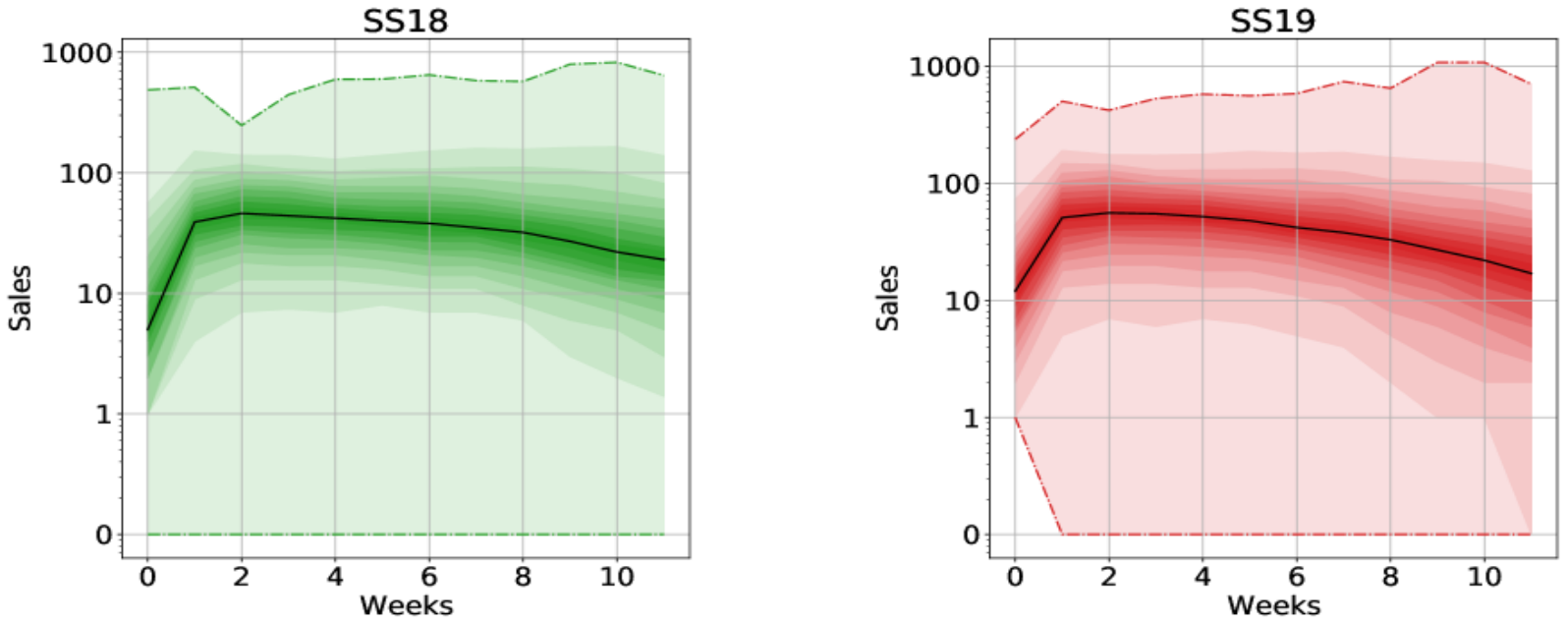
- **Dataset : VISUELLE**
  - **Product Attribute Text** : 색상, 원단, 카테고리 등 텍스트 정보



(그림 2) (a) 의상의 종류 (b) 색상 (c) 원단에 따른 데이터 분포도

3. 연구 내용

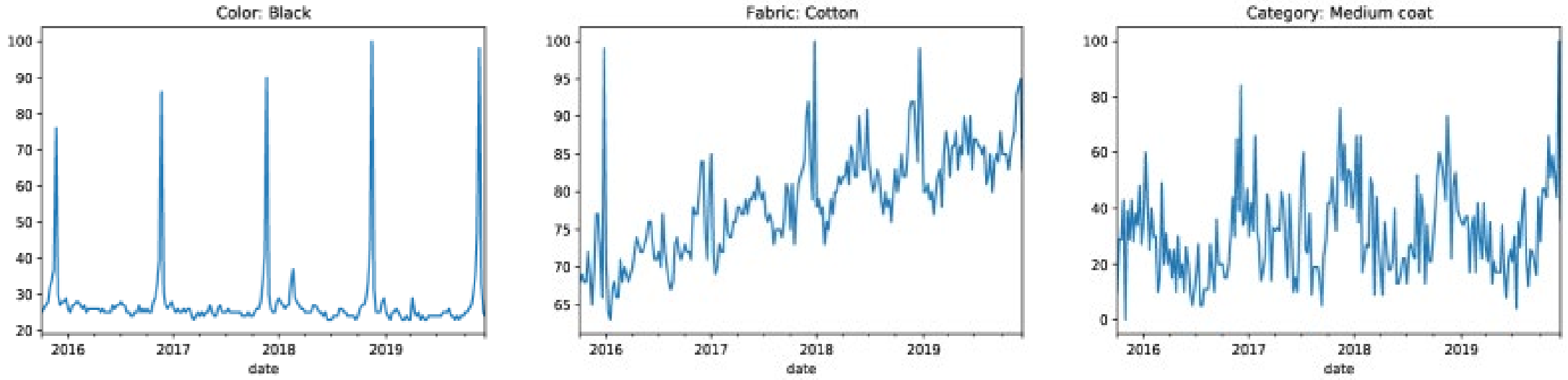
- **Dataset : VISUELLE**
  - **Product Attribute Temporal** : 출시 이후 최대 26주간 판매량, 할인율, 판매 가격 등
  - 판매가 첫 주에 가장 많고 시간에 따라 하락하는 형태를 보임



(그림 3) SS16~SS19 기간 동안 가장 많은 특징이 두드러진 SS18, SS19 Density plots

3. 연구 내용

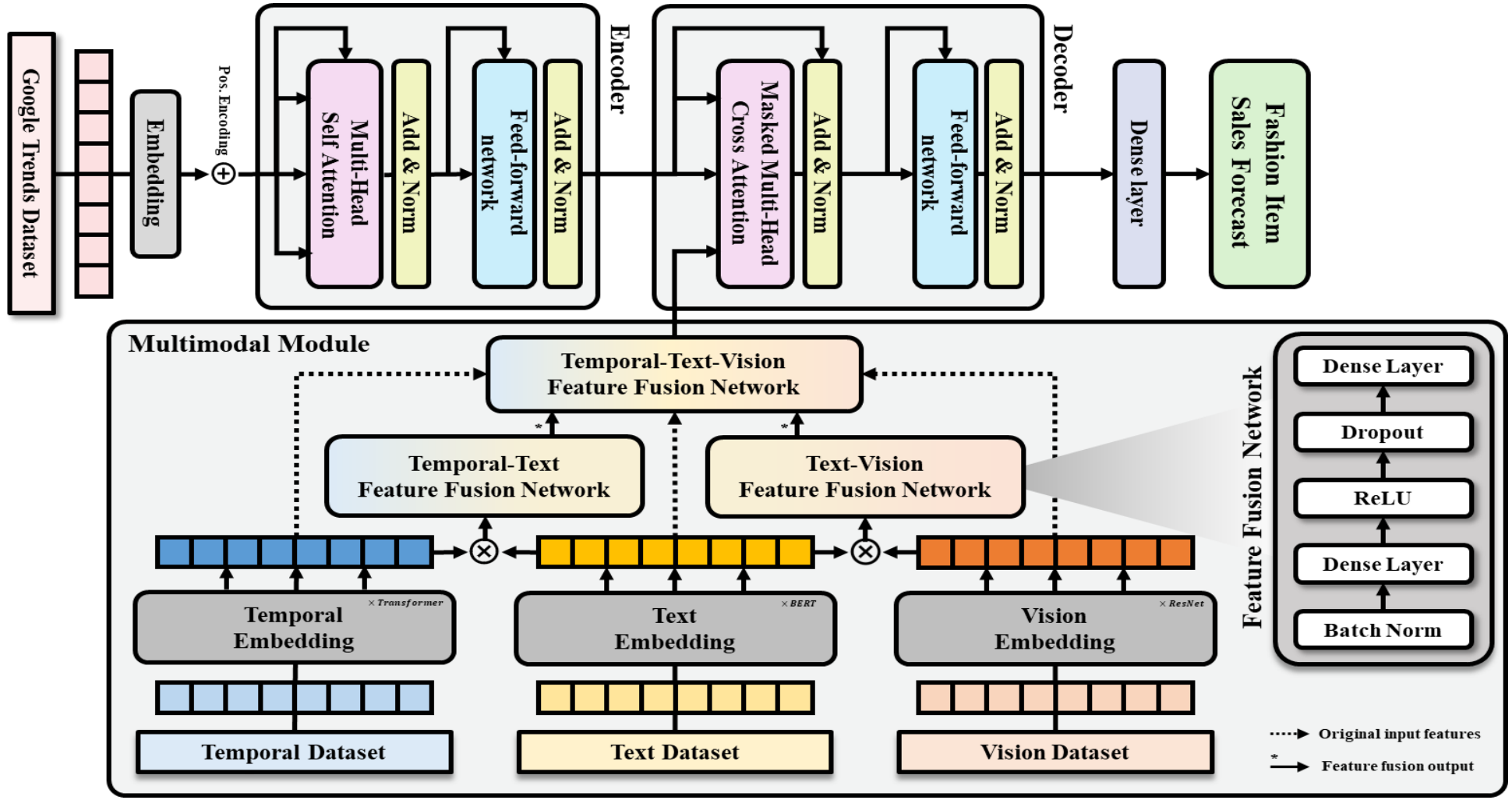
- **Dataset : VISUELLE**
  - **External Google Trend** : Google Trends 기반 제품 속성 관련 키워드 주간 검색 인기도
  - 키워드 : 색상, 원단, 카테고리 / 0부터 100까지의 인기도를 나타냄



(그림 4) 제품 속성 키워드(Color: Black, Fabric: Cotton, Category: Medium cost) 시각화 예시

3. 연구 내용

- **Proposed Method - M4FT: Multimodal Quad Fusion Transformer**



3. 연구 내용

- **Experiment Result**
  - **Mean Absolute Error (MAE)**: 예측 값과 실제 값 간의 절대 오차 평균 (적대적 차이)
  - **Weighted Absolute Percentage Error(WAPE)**: 전체 절대 오차를 실제 판매량 총합으로 정규화 (상대적 차이)
  - Input/Out Sequence 길이에 따라 다음과 같이 설정
    - **Input Sequence: 28주 / 52주**
      - In 28: 제품 출시 약 2개월 전, 초기 발주 시점
      - In 52: 시즌 별 제품 출시 직전 시점
    - **Output Sequence: 12주**
      - Out 12: 제품 출시 후 첫 12주간 순수 수요 구간에서의 예측

3. 연구 내용

- **Experiment Result**

  - **Mean Absolute Error (MAE)**: 예측 값과 실제 값 간의 절대 오차 평균 (적대적 차이)
  - **Weighted Absolute Percentage Error(WAPE)**: 전체 절대 오차를 실제 판매량 총합으로 정규화 (상대적 차이)

(표 1) Experiment Result

Methods	Iuput	**In:52, Out:12** (epoch=100)		**In:28, Out:12** (epoch=100)	
		WAPE	**MAE**	**WAPE**	**MAE**
GTM-Transformer AR	[G+A+T+V]	59.6	32.5	59.4	32.1
GTM-Transformer		55.2	32.1	58.7	31.0
Cross-Attention RNN+A		59.0	30.2	56.8	31.0
MuQAR		53.61	29.28	54.51	30.1
M2TFM(SOTA)		**52.61**	**29.28**	**54.13**	**29.75**
Proposed method*		**55.30**	**26.61**	**19.38**	**10.47**
Proposed method* AR		**17.85**	**10.07**	**16.53**	**10.03**

- Trend Data (Google Trends) [G]
- Product Attribute Temporal (Season, Release data, Timeframe/day, week, month, year) [A]
- Product Attribute Text (Category, Color, Fabric) [T]
- Product Texture Image (FW /SS 17, FW/SS 18, FW/SS 19) [V]

4. 결론 및 향후연구

- 본 연구는 전자상거래 환경에서 신제품 판매량을 보다 정밀하게 예측하기 위해, 네 가지 모달리티(Temporal, Text, Vision, Trend)를 통합적으로 활용하는 M4FT(Multimodal Quad Fusion Transformer) 모델을 제안함

- 제안한 M4FT는 모달리티 간 의미론적 상호작용을 계층적으로 학습하였으면, 최첨단 베이스라인 모델 대비 약 34% 낮은 WAPE와 약 19% 낮은 MAE를 기록함

- M4FT는 신제품 판매 예측이라는 고차원적 과제를 해결하는데 있어 강력하고 효율적인 AI 기반 접근법으로, 콜드 스타트 문제 해결에 도움을 줄 수 있을 것으로 기대됨

(표 2) 예측 판매량을 활용하여 추천 후보 제공 예시

제품명	속성 정보	판매 이력	추천 가능 여부
신규 블랙 코트	Color=Black, Category=Coat	x	x
판매 예측 모델(M4FT)	예측 판매량 75점	**추천 후보 가능**	-

참고문헌

1. Skenderi, Geri, et al. "Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends." *Journal of Forecasting* 43.6 (2024): 1982-1997.
2. Chen, Gang, et al. "Attending to customer attention: A novel deep learning method for leveraging multimodal online reviews to enhance sales prediction." *Information Systems Research* 35.2 (2024): 829-849.
3. Yan, Xiamin, and Haihua Hu. "New product demand forecasting and production capacity adjustment strategies: Within-product and cross-product word-of-mouth." *Computers & Industrial Engineering* 182 (2023): 109394.
4. Shilong, Zhang. "Machine learning model for sales forecasting by using XGBoost." *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 2021.

감사의 글

This work was supported by the IITP(Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2023-00259004)

MAGL-YOLO: 다종 경로 적응적 전역 특징 융합 기반 조류 객체 탐지 알고리즘

왕흠요[1], 김판구[2*]
조선대학교 컴퓨터공학과[1]
조선대학교 AI소프터웨어학부(컴퓨터공학전공)[2*]
e-mail : yourkerwxy@pm.me, pkkim@chosun.ac.kr

MAGL-YOLO: A Bird Object Detection Algorithm Based on Multi-path Adaptive Global Feature Fusion

Wang Xin Yao[1], Pan-Koo Kim[2*]
Dept of Computer Engineering, Chosun University[1]
Department of AI Software, Computer Engineering
Chosun University[2*]

요 약

Bird detection is vital for ecology, aviation safety, and agriculture. However, existing YOLO models, especially YOLOv8, face difficulties in avian scenes with large scale variation and clutter. We propose MAGL-YOLO, a lightweight framework with three key designs: (i) MAFRBlock for adaptive feature extraction, (ii) GFPNEF for efficient multi-scale fusion, and (iii) LSCSBND for parameter-efficient detection. Experiments on the Only Bird dataset show that MAGL-YOLO achieves 85.23% mAP50 and 46.67% mAP50-95, surpassing YOLOv8 while using only 2.8M parameters and 7.2 GFLOPs. These results confirm its strong accuracy - efficiency balance and practical potential for real-time edge applications.

1. 서 론

Bird detection plays a vital role in ecological conserv ation, aviation safety, and agricultural production. Tradit ional manual observation methods are inefficient and un suitable for large-scale, long-term monitoring. With the rapid development of computer vision and deep learning, automated bird detection has emerged as an effective s olution. For instance, improved YOLOv8-based methods have been applied to wetland surveillance videos, signifi cantly enhancing detection accuracy under complex bac kgrounds[1]. In the aviation domain, real-time detection frameworks have been introduced to mitigate bird strike risks and improve flight safety[2]. In natural scenes, the YOLO-Bird model demonstrated that feature enhanceme nt and lightweight designs are effective for detecting s mall avian objects[3]. Furthermore, researchers have co nstructed specialized datasets and introduced feature fus ion mechanisms to strengthen the generalization ability of detection models[4]. Nevertheless, existing YOLOv8 models remain limited when addressing large scale vari ations, diverse morphologies, and cluttered backgrounds in avian-specific scenarios. To address these challenges, this paper proposes MAGL-YOLO, a lightweight bird d etection algorithm designed to achieve a superior balanc 2e between accuracy and computational efficiency.

2. 관련 연구

2.1 Two-Stage Detectors

Two-stage algorithms, such as Faster R-CNN, Mask R-CNN, and Cascade R-CNN, generate region proposals and then perform classification and refinement. They achieve strong accuracy but i nvolve high computational costs, limiting real-time deployment [5][6][7].

2.2 One-Stage Detectors

One-stage methods directly predict categories and bounding bo xes on dense feature maps. YOLO pioneered this paradigm, SSD improved multi-scale adaptability with default boxes, and Lin et introduced Focal Loss to address class imbalance, enabling fast a nd accurate detection[8][9].

2.3 Yolov8

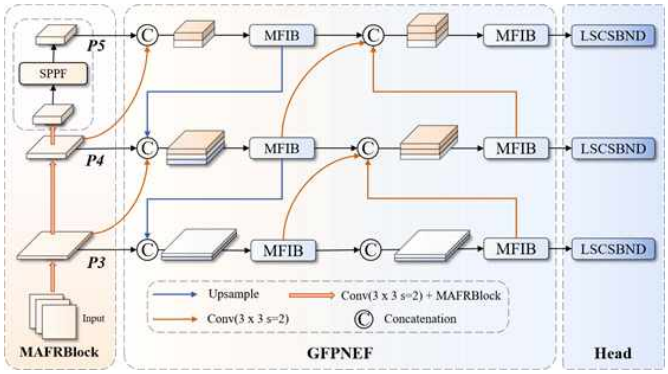


Figure 1: MAGL-YOLO Algorithm Architecture

The latest YOLOv8 integrates C2f modules to enhance gradient flow, a bidirectional FPN+PAN neck for multi-scale fusion, and an anchor-free decoupled head that separates classification and regression. These innovations achieve an effective balance between accuracy and speed, making YOLOv8 a strong baseline for real-time bird detection[10].

3. 본 론

Built on YOLOv8, MAGL YOLO addresses the characteristic challenges of avian scenes including severe scale variation, morphological diversity, and background clutter through a coordinated design that first strengthens nonlinear representation, then optimizes cross scale fusion, and finally reduces redundancy while preserving stable optimization.The detailed architecture is shown in Figure 1.

In the backbone, MAFRBlock uses parallel transformation paths together with a selective emphasis mechanism to preserve fine grained textures and suppress distractors. This enhances cross scale context modeling and boundary sensitivity at modest computational cost and improves the separability of small targets and targets near image boundaries.

In the neck, GFPNEF performs staged multi branch integration following a split then refine then reintegrate paradigm to align semantics across scales while maintaining local details. Reparameterizable paths are folded at inference into efficient equivalents, combining expressive training structures with fast execution, which promotes semantic coherence and detail flow and increases recall and localization in dense scenes with many small objects.

In the detection head, LSCSBND shares convolutional kernels across scales to remove parameter redundancy while assigning each scale its own normalization statistics to respect distributional differences. Lightweight channel interactions finalize classification and regression in a decoupled manner.

The result is a favorable balance of parameter efficiency, stable convergence, and cross scale generalization. Overall, MAGL YOLO forms a complementary triad of nonlinear discrimination, cross scale coherence, and lightweight efficiency that supports real time deployment in bird strike warning, agricultural protection, and wetland monitoring with reliable accuracy.

4. 시 험 결 과

Table 1: Comparative Experimental Analysis of Different Mainstream SOTA Models

Model	GFLOPS	Params(M)	P(%)	R(%)	mAP50(%)	mAP50-95(%)
Yolov5n	4.5	1.9	79.32	70.23	72.8	38.45
Yolov10n	6.5	2.2	83.17	76.06	82.03	44.27
Yolov11n	6.3	2.5	85.08	77.88	83.15	44.97
hyper-yolot	8.9	3	84.45	78.16	82.79	44.72
RT-DETR-r18	56.9	19.8	87.46	74.9	82.64	42.49
Yolov5s	16.5	7.2	79.23	72.45	75.55	40.65
Yolov10s	21.4	7.2	84.97	78.26	83.89	46.39
Yolov11s	21.3	9.4	85.46	77.87	83.19	46.51
hyper-yolo	10.8	3.9	83.94	78.07	82.7	45.21
RT-DETR-r32	88.8	31	88.01	75.86	82.24	43.32
yolov8(base)	8.1	3	83.94	75.15	81.97	43.32
MAGL-yolo(ours)	7.2	2.8	87	79.28	85.23	46.67

The comparative experimental results demonstrate that the MAGL-YOLO model achieves significant superiority in both accuracy and efficiency over existing mainstream approaches. Compared with the YOLOv8 baseline, MAGL-YOLO improves the mAP50 by 3.26 percentage points, reaching 85.23%, and enhances the mAP50-95 by 3.35 percentage points. Remarkably, these gains are attained with a 7% reduction in parameter count and an 11% decrease in computational cost. With a lightweight architecture of only 2.8M parameters and 7.2 GFLOPs, MAGL-YOLO surpasses YOLOv5n, YOLOv10n, and YOLOv11n by 12.4%, 3.2%, and 2.1% in detection accuracy, respectively, and even outperforms the RT-DETR-r18 model, which contains 19M parameters.

In summary, MAGL-YOLO delivers state-of-the-art (SOTA) accuracy while maintaining the smallest model size, demonstrating strong generalization ability and deployment efficiency, thereby providing an effective and practical solution for bird object detection.

5. 결 론

The proposed MAGL-YOLO model achieves an outstanding balance between accuracy and efficiency in bird detection tasks. By integrating three core modules—MAFRBlock, GFPNEF, and LSCSBND—the model effectively addresses challenges such as scale variation, morphological diversity, and background clutter. With only 2.8M parameters and 7.2 GFLOPs, MAGL-YOLO not only surpasses the YOLOv8 baseline but also outperforms other mainstream lightweight models, confirming its superior precision, generalization, and real-time applicability. Consequently, MAGL-YOLO offers a practical and efficient solution for ecological monitoring, aviation safety, and agricultural protection.

감사의 글

This research was supported by the Regional Innovation System & Education(RISE) program through the (Gwangju RISE Center), funded by the Ministry of Education(MOE) and the (Gwangju Metropolitan City), Republic of Korea.(2025-RISE-05-013)

참 고 문 헌

[1] Ma, Jianchao, et al. "An improved bird detection method using surveillance videos from Poyang Lake based on YOLOv8." Animals 14.23 (2024): 3353.

[2] Qu, Yi, et al. "Dynamically Optimized Object Detection Algorithms for Aviation Safety." Electronics 14.17 (2025): 3536.

[3] Felix-Jimenez, Axel Frederick, et al. "Integration of YOLOv8 Small and MobileNet V3 Large for Efficient Bird Detection and Classification on Mobile Devices." AI 6.3 (2025): 57.

[4] Chen, Xian, et al. "An efficient method for monitoring birds based on object detection and multi-object tracking networks." Animals 13.10 (2023): 1713.

[5] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[6] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[7]Cai, Zhaowei, and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[8] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Cham: Springer International Publishing, 2016.

[9] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[10] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Local Feature Enhancement via Feature Fusion for Spoof Fingerprint Detection with Receptive Field–Wise Learning

Md Al Amin[1], Naim Reza[1], Ho Yub Jung*

Department of Computer Engineering, Chosun University, Gwangju 61452, Republic of Korea[1]

ahmedalamin@chosun.ac.kr, naim@chosun.ac.kr, hoyub@chosun.ac.kr

*Corresponding author: Ho Yub Jung (e-mail: hoyub@chosun.ac.kr)

Abstract

With the increasing reliance on fingerprint-based biometric systems, ensuring resilience against spoofing has become crucial. A major obstacle in developing effective convolutional neural networks (CNNs) models for this task is the limited number of fingerprint images available in existing datasets. To address this challenge, we propose a receptive field–wise feature learning framework. In our approach, a feature integration block enables information from multiple branches to be jointly leveraged, resulting in richer feature representations. Feature maps from two branches are then fused and compacted by averaging spatial activations into a single value, effectively increasing the number of effective labels during training while reducing overfitting. By enhancing the discriminative power of the learned features, this strategy achieves an average accuracy of 96.71% on the LivDet-2015 dataset, surpassing the performance of several prior methods.

1 Introduction

Fingerprint-based biometric authentication systems are widely adopted in applications such as mobile devices, border control, and financial transactions because of their convenience and reliability [1]. However, they are highly vulnerable to presentation attacks where fabricated fingerprints created from materials such as silicone, gelatin, or printed images are used to deceive the sensor [2]. Ensuring reliable detection of such attacks is therefore essential to maintaining the security and trustworthiness of biometric systems.

In recent years, CNNs have achieved remarkable success in fingerprint presentation attack detection by automatically learning discriminative features directly from the images [3]. While CNN-based approaches have demonstrated promising results, their effectiveness is often hindered by the limited size of available fingerprint spoof datasets, and many existing methods rely on external strategies such as transfer learning [4] or patch-based decomposition [5] to compensate for the limited size of fingerprint spoof datasets. Other studies have used generative models for data augmentation [6].

In this work, receptive-field–wise feature learning is employed to reduce overfitting by compacting activations, while the fusion of CNN-extracted features with those from a custom DenseNet enriches local feature representation. The fused maps are then compacted into a single representation, which enhances the effective labels. This design focuses on enriching local features through multi-branch integration, which strengthens the network's ability to discriminate between live and spoof fingerprints across different sensors and materials. The main contributions of this work are as follows:

- We introduce a feature integration block that fuses multi-branch information, yielding richer representations and more robust spoof detection.
- We compact feature maps by averaging spatial acti-

vations, reducing overfitting and improving accuracy to 96.71% on LivDet-2015 [7], outperforming several previous approaches.

2 Related Work

Fingerprint liveness detection has drawn significant attention due to the vulnerability of biometric systems to spoofing. Early works largely relied on handcrafted local descriptors. Ghiani et al. [8] introduced Local Phase Quantization, which is robust to blurring and captures spectral differences between live and fake fingerprints. Gragnaniello et al. [9] extended this with the Local Contrast Phase Descriptor, combining spatial contrast and frequency-phase features for improved discrimination. On the other hand, hybrid approaches, such as HyFiPAD, combined Local Adaptive Binary Patterns with other descriptors and ensemble classifiers to improve generalization across datasets [10].

While effective, these methods often required careful feature engineering and still struggled with unseen spoof materials. With the success of deep learning, CNN-based solutions have become dominant. Nogueira et al. [11] demonstrated that transfer learning from pretrained networks such as VGG and AlexNet significantly reduced classification errors, establishing a state of the art and winning the LivDet 2015 competition. More recent studies have explored deeper architectures and feature-combining strategies to further improve performance.

Based on the concept of feature-merging, we propose a receptive-field–wise feature learning framework that fuses CNN features with those from a custom DenseNet. Our approach compactly integrates multi-branch feature maps to reduce overfitting while enriching local detail. This design enhances discriminative power across sensors and spoof materials, achieving higher accuracy and improving generalization ability than previous many approaches.

3 Proposed Method

The proposed methodology, illustrated in Figure 1, is built upon a two-branch architecture. This design integrates the CNN branch with a custom DenseNet stream through a feature fusion block, enabling rich feature representation to improve accuracy while minimizing overfitting on small fingerprint datasets. The detailed methodology is outlined in the following subsections.

3.1 Multi-Branch Feature Extraction

Given an input fingerprint image $I \in \mathbb{R}^{H\times W}$, discriminative features are extracted through two parallel streams: a CNN branch with five convolutional layers and a DenseNet branch with three blocks and a growth rate of 6. Their outputs are expressed as

$$F_{\text{cnn}} = f_{\text{cnn}}(I), \quad F_{\text{dense}} = f_{\text{dense}}(I), \tag{1}$$

where $f_{\text{cnn}}(\cdot)$ and $f_{\text{dense}}(\cdot)$ denote convolutional transformations. The feature maps obtained from the two branches are concatenated and subsequently refined through a convolutional layer, producing a single-channel representation:

$$F_{\text{fusion}} = f_{\text{conv}}\big(\text{Concat}(F_{\text{cnn}}, F_{\text{dense}})\big), \tag{2}$$

where $f_{\text{conv}}(\cdot)$ represents the convolutional operation applied to the concatenated features. This refinement step enables to represetn richer features.

3.2 Receptive-Field–Wise Compaction

To mitigate overfitting on limited datasets, we adopt receptive-field–wise learning. After feature fusion and refinement, the output is a single-channel feature map $F_{\text{fusion}} \in \mathbb{R}^{H\times W}$. The spatial activations are compacted using global average pooling, expressed as

$$y' = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{\text{fusion}}(i, j), \tag{3}$$

where y' denotes the compacted representation that serves as the liveness score. This operation effectively treats each receptive field as a training signal, enhancing generalization while reducing model complexity. The model is then optimized using binary cross-entropy loss.

3.3 Dataset and Training Configuration

We evaluated our approach on the LivDet-2015 dataset, which includes four sensors, CrossMatch, DigitalPersona,

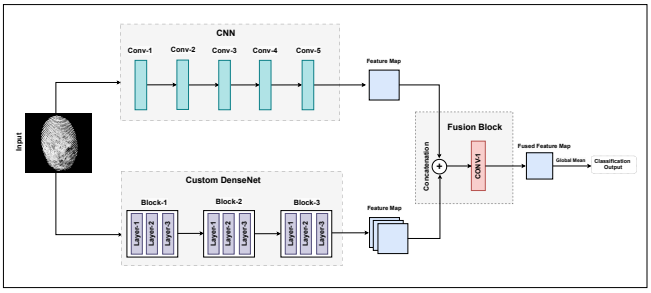


Figure 1: Proposed dual-branch architecture combining CNN and DenseNet streams with receptive-field–wise feature fusion for spoof fingerprint detection.

GreenBit, and HiScan, each providing 1000 to 1500 live and spoof samples, along with additional unknown spoofs. Images were converted to grayscale, inverted, and the fingerprint region extracted and centered on a 512×512 black background. Data augmentation with random flips and rotations in the range of $[-30°, +30°]$ was applied.

The model was trained using SGD with a learning rate of 0.01 and a batch size of 2 for 1500 epochs. Despite the limited dataset, the receptive-field–wise learning framework reduced overfitting, allowing the model with only 120,568 parameters to achieve high accuracy across sensors.

4 Results

Model	DigitalPersona	GreenBit	CrossMatch	HiScan	Avg. Acc.
CNN-VGG [11]	93.72	95.40	98.10	94.36	95.39
ALDRN [12]	93.20	95.23	96.54	93.76	94.68
Jomaa et al. [13]	91.96	94.68	97.29	95.12	94.87
Ulian et al. [14]	–	–	95.00	–	95.00
LFLDNet [15]	93.52	98.56	98.18	96.00	96.56
OPG-CNN [5]	93.50	97.63	97.51	96.18	96.20
Proposed	**94.35**	97.55	**99.11**	95.84	**96.71**

Table 1: Performance comparison of fingerprint liveness detection models on LivDet-2015.

We evaluated the proposed model using classification accuracy and Average Classification Error (ACE), which is the mean error rate of live and spoof classes, providing a balanced measure of performance. On the LivDet-2015 dataset, the model achieved an average accuracy of 96.71% shown in Table 1, using a dual-branch architecture that combines DenseNet and CNN features to capture fine-grained texture details.

Dataset	Sensor	Training Materials	Testing Materials	LivDet15 Winner[16]	SA-R-CNN[17]	Proposed
LivDet-2015	Digital Persona	Ecoflex, Gelatin, Latex, Wood Glue	Liquid Ecoflex, RTV	6.00%	5.46%	3.30%
	GreenBit	Ecoflex, Gelatin, Latex, Wood Glue	Liquid Ecoflex, RTV	7.40%	4.84%	2.50%
	HiScan	Ecoflex, Gelatin, Latex, Wood Glue	Liquid Ecoflex, RTV	5.80%	6.03%	2.90%
Average	–	–	–	6.4	5.44	**2.9**

Table 2: Comparison of cross-material robustness measured by ACE for the proposed method on the LivDet-2015 dataset.

Moreover, tour model improves generalization, which explains the lower error rates like 2.9% on average against unknown spoof materials compared to the LivDet-2015 Winner and SA-R-CNN shown in Table 2. Overall, the results demonstrate that the proposed model achieves better accuracy compared to many previous studies while maintaining robustness against spoofing materials.

5 Conclusion

This work presented a dual-branch framework for fingerprint spoof detection that combines a custom DenseNet branch with a CNN-based branch through feature fusion. The proposed model leverages receptive-field–wise learning to compact spatial activations, which helps to mitigate overfitting in small-scale fingerprint datasets. Experimental evaluation on the LivDet-2015 benchmark confirmed the effectiveness of the approach, achieving an average accuracy of 96.71% across multiple sensors and demonstrating robustness against unknown spoof materials. These results validate the strength of integrating complementary feature representations for enhanced discriminative power.

Funding

This study was supported by research fund from Chosun University, 2025.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Professor Ho Yub Jung, for his invaluable guidance, support, and encouragement throughout this work.

References

[1] A. Kalra, "Emerging technologies in biometrics: Artificial intelligence and machine learning," *Biometrics*, pp. 3–26.

[2] P. Rehan, A. Bhandari, and A. Bathla, "Feature-based fingerprint presentation attack detection: System integration, taxonomy, cross-material analysis, and open challenges," *Taxonomy, Cross-Material Analysis, and Open Challenges*.

[3] T. Riaz, A. Anjum, M. H. Syed, and S. Rehman, "Improving presentation attack detection classification accuracy: Novel approaches incorporating facial expressions, backdrops, and data augmentation," *Sensors*, vol. 25, no. 7, p. 2166, 2025.

[4] M. Cheniti, Z. Akhtar, and P. K. Chandaliya, "Dual-model synergy for fingerprint spoof detection using vgg16 and resnet50," *Journal of Imaging*, vol. 11, no. 2, p. 42, 2025.

[5] A. Rai, A. Anshul, A. Jha, P. Jain, R. P. Sharma, and S. Dey, "An open patch generator based fingerprint presentation attack detection using generative adversarial network," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 27723–27746, 2024.

[6] M.-H. Hsu, Y.-C. Hsu, and C.-T. Chiu, "Inpainting diffusion synthetic and data augment with feature keypoints for tiny partial fingerprints," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[7] "Livdet - fingerprint liveness detection competitions." https://www.livdet.org/competitions.php. Accessed: 2025-09-11.

[8] L. Ghiani, G. L. Marcialis, and F. Roli, "Fingerprint liveness detection by local phase quantization," in *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pp. 537–540, IEEE, 2012.

[9] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "Local contrast phase descriptor for fingerprint liveness detection," *Pattern Recognition*, vol. 48, no. 4, pp. 1050–1058, 2015.

[10] D. Sharma and A. Selwal, "Hyfipad: a hybrid approach for fingerprint presentation attack detection using local and adaptive image features," *The Visual Computer*, vol. 38, no. 8, pp. 2999–3025, 2022.

[11] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint liveness detection using convolutional neural networks," *IEEE transactions on information forensics and security*, vol. 11, no. 6, pp. 1206–1213, 2016.

[12] C. Yuan, Z. Xia, X. Sun, and Q. J. Wu, "Deep residual network with adaptive learning framework for fingerprint liveness detection," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 3, pp. 461–473, 2019.

[13] R. M. Jomaa, H. Mathkour, Y. Bazi, and M. S. Islam, "End-to-end deep learning fusion of fingerprint and electrocardiogram signals for presentation attack detection," *Sensors*, vol. 20, no. 7, p. 2085, 2020.

[14] D. M. Uliyan, S. Sadeghi, and H. A. Jalab, "Anti-spoofing method for fingerprint recognition using patch based deep learning machine," *Engineering Science and Technology, an International Journal*, vol. 23, no. 2, pp. 264–273, 2020.

[15] K. Zhang, S. Huang, E. Liu, and H. Zhao, "Lfld-net: lightweight fingerprint liveness detection based on resnet and transformer," *Sensors*, vol. 23, no. 15, p. 6854, 2023.

[16] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. A. Schuckers, "Livdet 2015 fingerprint liveness detection competition 2015," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, 2015.

[17] C. Yuan, Z. Xu, X. Li, Z. Zhou, J. Huang, and P. Guo, "An interpretable siamese attention res-cnn for fingerprint spoofing detection," *IET Biometrics*, vol. 2024, no. 1, p. 6630173, 2024.

음향 장면 생성을 위한 Flow 기반 모델의 비교 분석

김어진, 박유정, 이건우, 전찬준*
조선대학교 AI소프트웨어학부
e-mail : {jjj333, pak01, geonwoo, cjchun}@chosun.ac.kr

Comparative Analysis of Flow-based Models for Acoustic Scene Generation

Eojin Kim, Yujeong Pak, Geon Woo Lee, Chanjun Chun*
School of AI Software, Chosun University

요 약

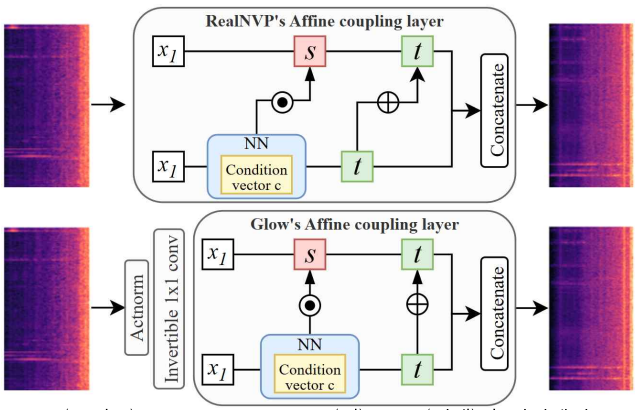
본 연구에서는 Normalizing flow 계열의 대표적 생성 모델인 조건부 RealNVP, Glow를 대상으로 Acoustic scene generation(음향 장면 생성) 성능을 비교·분석하였다. 특히, 다양한 환경 조건에 따라 소리를 생성하는 조건부 환경 소음 생성 문제에 초점을 맞추어 각 모델의 구조적 특성과 결과를 검증하였다. 본 연구는 Normalizing flow 기반 모델의 특성을 비교함으로써, 향후 조건부 환경 소음 생성 및 Acoustic scene generation 연구에 활용 가능한 결과를 제시한다.

1. 서 론

음향 장면 생성이란 공원, 사무실, 기차 등 특정 환경을 반영하는 오디오를 합성하는 과제를 의미한다. 이는 소리 시뮬레이션, 데이터 증강, 제어 가능한 오디오 합성과 같은 응용을 지원한다. 장면 조건에서 사실적인 합성을 달성하려면 그럴듯한 신호를 생성하는 것뿐만 아니라 의도된 환경적 맥락을 반영할 수 있는 모델이 필요하다. Normalizing flow는 단순한 기저 분포를 복잡한 데이터 공간으로 변환하여 정확한 우도(likelihood) 추정을 제공하는 가역적 생성 모델이다. RealNVP[1], Glow[2]와 같은 모델은 이미지와 음성 등 다양한 도메인에서 효과적임이 입증되었다. 그러나 입력을 단순히 연결(concatenation)하는 방식은 주어진 조건과 생성된 출력 간의 정렬(alignment)을 강제하지 못해 구조화된 조건을 다루는 데 한계가 있다. 이를 해결하기 위해, 본 논문에서는 장면 조건부 멜 스펙트로그램 생성을 위한 Conditional normalizing flow(CNF) 모델을 제안한다. CNF는 RealNVP, Glow 기반 구조에 원-핫 장면 벡터를 통합하여, 완전히 가역적인 프레임워크 내에서 조건 특화 변환을 학습할 수 있게 한다. 이 설계는 음향 장면과 일치하는 일관된 생성을 가능하게 하며, 계산 가능성을 해치지 않고 종단 간 제어 가능한 합성을 지원한다.

2. 조건부 정규화 흐름을 통한 음향 장면 생성

본 연구는 환경적 특성을 반영한 멜 스펙트로그램 생성을 목표로 하는 조건부 생성 모델을 제안한다. 모델은 RealNVP, Glow 등의 normalizing flow 계열의 설계 원리를 바탕으로 구현되며, 장면을 나타내는 조건 벡터가 모든 변환 단계에 구조적으로 통합된다. 그림 1은 제안한 Real NVP와 Glow 기반의 CNF 모델의 구조를 보여준다. flow 블록의 핵심은 조건부 affine coupling layer이다. coupling



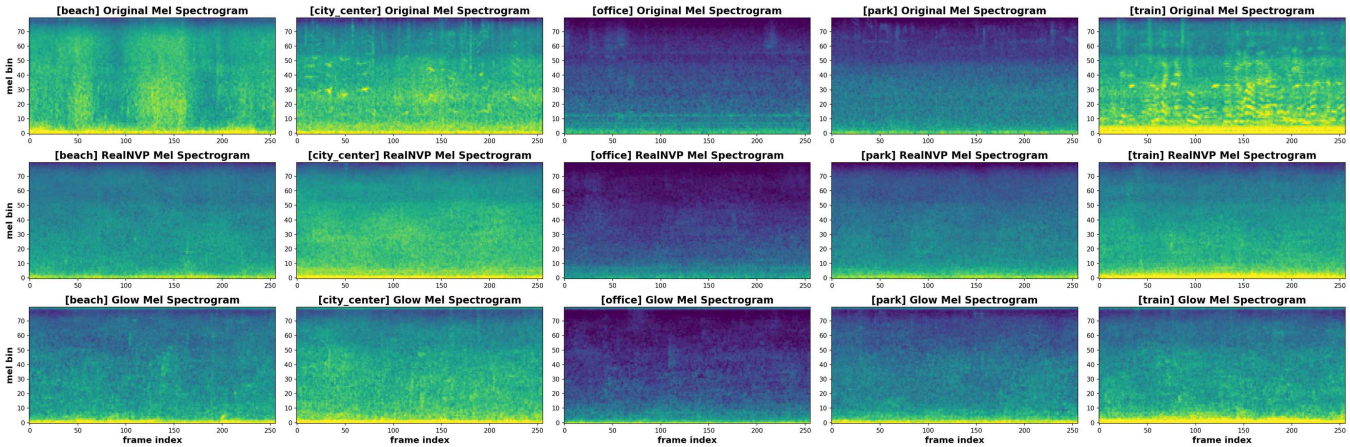
(그림 1) Conditional RealNVP(위), Glow(아래)의 아키텍처

layer는 입력을 두 부분으로 분할하여 한쪽 부분을 고정하고 다른 쪽 부분을 고정된 쪽의 정보와 조건으로 변환하는 방식으로 동작한다. 본 논문에서 사용된 변환은 수식 (1),(2)과 같다.

$$y_a = x_a, \tag{1}$$

$$y_b = x_b \odot \exp(s(x_a, c)) + t(x_a, c) \tag{2}$$

여기서 입력 x은 두 부분 x_a, x_b으로 나뉘며, x_a는 그대로 유지된다. 반면, x_b는 x_a와 장면 조건 c에 의해 계산된 스케일 함수 s와 평행이동 함수 t에 따라 변환된다. 조건 c는 one-hot 벡터로 주어지며, 선형 투영을 통해 특정 공간으로 임베딩된 후 x_a와 결합되어 네트워크에 입력된다. 이 구조의 장점은 가역성이다. 즉, y_b에서 변환을 역으로 수행하면 원래의 x_b를 쉽게 복원할 수 있어, 데이터의 잠재 표현과 재생성이 모두 가능하다. CNF 모델은 추가적으로 다중 스케일 구조를 채택하여 초기에는 전역적인 원본의 멜 스펙트럼 형태를, 이후 단계에서는 음성의 하모닉이나 잡음 같은 세밀하고 구조적인 요소를 학습하게 된다.



(그림 2) 5개의 scene에 대한 원본 멜 스펙트로그램(왼쪽)과 RealNVP로 생성한 멜 스펙트로그램(중앙), Glow로 생성한 멜 스펙트로그램(오른쪽)

(표 1) CNF로 생성한 audio의 FAD와 분류 정확도 평가

Metric	CNF	beach	city center	office	park	train
FAD(Frechet audio distance)	RealNVP	10.6	11.8	6.6	11.5	7.3
	Glow	2.7	3..1	1.2	2.3	2.5
Classification Accuracy	RealNVP	0.59	0.65	0.98	0.42	0.58
	Glow	0.85	0.87	0.99	0.74	0.87

3. 실험 및 결과

실험에는 TUT Acoustic Scenes 2017 데이터셋[5]을 사용하였다. 데이터셋은 15개 실제 환경에서 녹음된 4,680개의 10초 오디오 클립으로 구성되며, 샘플링 레이트는 44.1kHz이다. 각 클립은 80개의 주파수 bin과 640개의 시간 프레임을 갖는 멜 스펙트로그램으로 변환되었으며, 파형 재구성을 위해 HiFi-GAN 보코더와 정렬되었다. 비교 실험은 RealNVP, Glow 두 가지 flow 기반 모델과 제안한 CNF를 대상으로 수행하였다. 평가에서는 다섯 개의 대표 장면 클래스(해변, 도심, 사무실, 공원, 기차)를 선택하였다. 훈련 시 장면 라벨은 15차원 원-핫 벡터로 인코딩되어 모든 coupling layer에 주입되었다. 성능 평가는 Fréchet audio distance (FAD)[5]와 분류 정확도를 활용하였다. FAD는 사전 학습된 임베딩 공간에서 실제 오디오와 생성 오디오의 분포 차이를 측정하며, 값이 낮을수록 두 분포가 유사함을 의미한다. 분류 정확도는 PANNs로 pretrained 시킨 가중치를 활용하여 각 scene별 생성된 audio가 환경에 맞게 분류가 잘 되는지의 척도를 의미한다. FAD의 결과, 조건부 RealNVP는 조건 반영이 개선되어 FAD가 낮았으나, 복잡한 장면에서는 여전히 분포 차이가 컸다. 조건부 Glow는 평균적으로 가장 낮은 FAD를 달성하여, 구조적 환경에서 실제 음향 장면과의 높은 유사성을 입증하였다. 분류 정확도 역시 조건부 Glow가 높은 정확도로 환경을 예측하였으며, RealNVP는 조건부 Glow에 비해 성능이 떨어지는 모습을 보였다. 그림 2는 멜 스펙트로그램 도메인에서의 원본 멜과 모델의 출력으로 나온 생성된 멜 스펙트로그램을 비교한 그림이다. 생성된 멜 스펙트로그램은 최종적으로 HiFi-GAN 보코더를 통해 파형이 재구성된다.

4. 결론

본 논문에서는 특정 환경 조건을 반영한 멜 스펙트로그램 생성을 위해 설계된 Conditional normalizing flow(CNF) 모델을 제안하였다. CNF는 모든 장면에서 안정적이고 높은 조건 충실도를 나타냈으며, 특히 FAD 평가에서 낮은 값을 기록하여 실제 데이터와의 분포 차이가 가장 적음을 확인하였다. 이는 단순 분류 정확도 기반의 평가를 넘어, 생성된 오디오가 품질과 조건 일관성 모두에서 우수함을 보여준다. 따라서 CNF는 조건 인식 오디오 생성을 위한 확장 가능하고 해석 가능한 프레임워크임을 입증하였다. 향후 연구에서는 연속적인 조건 표현 학습, 다운스트림 분류기와의 통합, 다중 모달 생성 및 scene-to-audio 합성과 같은 과제를 통해 본 연구의 범위를 더욱 확장할 수 있을 것이다.

감사의 글

본 연구는 2024년도 연구개발특구진흥재단의 '지역의 미래를 여는 과학기술 프로젝트' 사업으로 수행되었음. (과제고유번호: 2022-DD-UP-0312)

참 고 문 헌

[1] Laurent Dinh, Jascha Sohl-Dickstein, Samy Bengio, "Density estimation using Real NVP," in *Proc. The International Conference on Learning Representations (ICLR)*, 2017.

[2] Diederik P. Kingma, Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. The Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[3] Laurent Dinh, David Krueger, Yoshua Bengio, "NICE: nonlinear independent components estimation," in *Proc. The International Conference on Learning Representations (ICLR)*, 2015.

[4] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, "Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions," in *Proc. DCASE workshop*, 2020.

[5] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, Matthew Sharifi, "Fréchet Audio Distance: A metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019.

검색 증강 생성과 소형 언어 모델을 활용한 산업안전 조치 의사결정 지원 시스템에 관한 연구

김수아[1], 정연비[2], 최기도[3], 김원열[1]*
조선대학교 인공지능공학과[1], 조선대학교 정보통신공학과[2], ㈜머제스[3]
e-mail : ksa6352@chosun.ac.kr, jeongyeonbi@chosun.ac.kr, merzes@merzes.com, kwy00@chosun.ac.kr

A Study on Industrial Safety Decision Support System Using Retrieval-Augmented Generation and Small Language Models

Su-A Kim[1], Yeon-Bi Jeong[2], Ki-Do Choi[3], Won-Yeol Kim[1]*
Dept of Artificial Intelligence Engineering, Chosun University[1],
Dept of Information and Communication Engineering, Chosun University[2],
Merzes Inc.[3]

요 약

최근 대규모 언어 모델(LLM)은 다양한 전문 분야에 적용되고 있으나, 산업 안전보건 분야에서는 환각(Hallucination) 현상과 데이터 보안 문제로 인해 신뢰성 높은 실시간 조치 제공에 한계가 있어 도입이 제한적이었다. 본 연구에서는 산업 현장에서 발생하는 안전사고에 대해 신속하고 정확한 안전보건 조치 수립을 지원하기 위해 RAG 기반 소형 대규모 언어 모델 의사결정 지원 시스템을 제안한다. 본 시스템에서는 기존 대규모 언어 모델이 지닌 법규에 대한 최신성 부족과 환각 현상 문제를 극복하기 위해 검색 증강 생성 기술과 소형 언어 모델을 효과적으로 결합한 접근법을 적용하였다. 제안 시스템은 ko-sroberta-multitask 기반의 사례 검색, 프롬프트 템플릿을 통한 맥락 통합, 그리고 Llama3.1-8B 전문가 모델의 안전조치 생성을 핵심 구성 요소로 하며, 19,210개의 실제 산업재해 사례로 구성된 데이터셋을 활용해 평가한 결과, 답변 유효성 65.67%, 조회 정확도 85.59%를 기록하여 기존 모델들에 비해 우수한 성능을 보였다.

1. 서 론

1.1 소개

산업 현장의 안전사고는 신속하고 정확한 대응이 이루어지지 않을 때 대형 재해로 이어질 수 있으며, 따라서 즉각적이고 법규에 근거한 의사결정 지원 도구의 필요성이 무엇보다 크다. 특히 안전관리자가 긴급 상황에서 적합한 법적 조치를 빠르게 도출하지 못하면 초기 대응에 실패할 위험이 높아지고, 이는 곧 인명 피해와 더불어 사회적·경제적 손실로 이어질 수 있다.

최근 산업안전 분야에서 최근 자연어 처리 기술, 특히 대규모 언어 모델(LLM)의 활용 가능성이 주목받고 있다. 그러나 기존의 범용 LLM은 해당 분야로의 활용에 있어 다음과 같은 한계를 가진다. 첫째, 법규 최신성 부족으로 인해 학습 시점 이후 개정된 법규를 반영하지 못한다. 둘째, 환각 발생으로 인해 실제 존재하지 않는 조항이나 잘못된 조치를 생성할 위험이 있다.

본 연구는 위와 같은 한계를 극복하기 위해, 로컬 환경에서 구동되는 소형 언어 모델(sLLM)과 검색 증강 생성(RAG) 기술을 결합한 산안법 질의응답 시스템을 제안한다.

2. 관련 연구

2.1 산업안전 분야에서의 대규모 언어 모델

산업안전 분야에서 인공지능과 대규모 언어 모델(LLM)의 적용 가능성은 최근 몇 년간 빠르게 주목받고 있다. 다양한 산업환경에서 안전 데이터의 자동 분석, 규정 준수, 사고 예방 지원 등 LLM 기반 기술의 잠재적 가치가 실증적으로 검토되고 있다. LLM을 활용한 실제 안전 관리 사례에서, LLM이 건설사고 보고서의 핵심 정보를 자동 추출·분류함으로써 안전관리 효율화에 기여할 수 있음이 확인됐다[1]. 건설사고 사례 보고서의 주요 속성(사고 원인, 신체 부위, 심각도 등)을 LLM 기반 분석 시스템이 자동으로 분류하고, 다양한 모델 특성을 비교·분석한 연구는 관련 데이터의 신속한 처리와 현장 안전 수준 향상 가능성을 제시한다[2]. 또한, LLM은 복잡한 산업 현장에서 안전 전문가의 판단을 보조하는 역할로도 적용되고 있다. 안전 현장의 위험성 평가, 규정 준수 검토 등 실제 업무에 LLM을 활용해본 결과, LLM 기반 시스템이 복잡한 컨텍스트를 해석하여 전문가의 의사결정 과정을 지원할 수 있음이 실증적으로 입증됐다[3]. 이러한 LLM의 산업 현장으로의 적용 가능성을 토대로, 안전보건 조치 수립을 지원하

* 교신저자

는 시스템을 연구하고자 한다.

2.2 검색 증강 생성 (RAG, Retrieval-Augmented Generation)

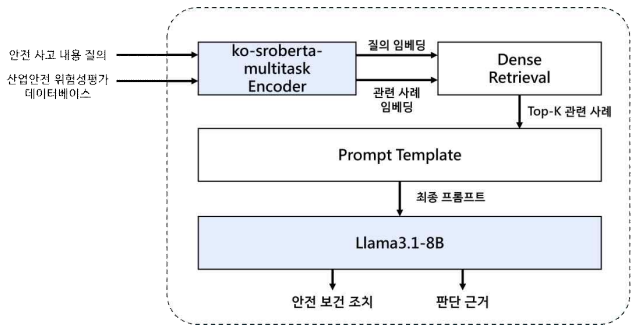
이러한 LLM의 산업 현장으로의 적용 가능성을 토대로, 안전보건 조치 수립을 지원하는 시스템을 연구하고자 한다. RAG는 LLM이 답변을 생성할 때 모델 내부 지식에만 의존하지 않고, 외부의 신뢰할 수 있는 데이터베이스에서 관련 정보를 검색하여 이를 참고하도록 하는 기법이다[4]. 이 방식은 두 가지 중요한 장점을 가진다. 첫째, 환각 현상을 억제할 수 있는데, 이는 모델이 사실과 다른 내용을 생성하는 문제를 크게 완화한다. 둘째, 최신성 확보가 가능하여, LLM 학습 이후 추가되거나 개정된 정보를 실시간으로 반영할 수 있다[5]. 산업안전보건법과 같이 법령이 지속적으로 개정·보완되는 영역에서는, 단순히 사전 학습된 언어 모델만으로는 최신 법규를 반영하기 어렵다. 그러나 RAG 기법을 적용하면, 법규 원문과 시행령, 시행규칙, 고시 등 외부 데이터베이스에서 관련 조항을 즉시 검색하여 답변에 반영할 수 있으므로, 신뢰도 높은 법규 기반 응답이 가능하다.

3. 본 론

3.1 제안하는 방법

제안하는 시스템은 그림 1과 같이 입력 안전사고 상황과 관련된 사례를 DB에서 검색하는 언어 모델과, 검색된 사례의 내용을 구조화하는 프롬프트 템플릿, 그리고 안전보건 조치를 수립하는 전문가 LLM으로 구성된다.

우선, BERT 기반의 언어 모델인 ko-sroberta-multitask를 검색 언어 모델로 사용하여 데이터베이스 내의 사례들과의 벡터 유사도를 각각 계산한다. 계산된 벡터 유사도 중 상위 K개의 사례는 프롬프트 템플릿을 통해 전문가 언어 모델이 참고할 맥락으로 통합된다. 프롬프트 템플릿에서는 계산된 벡터 유사도 중 상위 K개의 사례와 전문가 언어 모델의 출력 규칙을 통합한다. 통합된 입력 프롬프트는 Llama3.1-8B 모델에 입력되어 최종적으로 안전사고 상황에 대해 안전관리자가 수행할 안전보건 조치를 출력하며, 판단 근거 또한 함께 제시한다.



(그림 1) 제안하는 시스템 구조

3.2 실험 환경

RAG의 유효성 및 시스템의 성능을 검증하기 위해 자체 데이터셋과 비교 모델을 활용하였다.

산업안전 위험성평가 데이터베이스: 실제 산업재해 사례 19,210개로 구성

또한 아래의 세가지 조건을 비교 실험하였다.

Baseline 1: RAG 없이 로컬 sLLM(Llama-3.1-8B)만 사용

Baseline 2: OpenAI GPT-4o API (외부 LLM 활용)

Proposed: RAG + 로컬 sLLM

평가 지표로는 두 가지를 사용하였다.

답변 유효성: 생성된 안전보건 조치가 해당 상황에 적합한 핵심 법규와 조치 사항을 올바르게 포함하는 정도를 측정하는 지표이다. 실제 안전보건 조치와 시스템이 생성한 안전조치를 비교하여, 필수 조치 항목의 일치율을 계산한다. 각 안전조치 항목에 대해 정확히 포함된 경우 1점, 부분적으로 포함된 경우 0.5점, 포함되지 않은 경우 0점을 부여하여 전체 점수를 산출한다.

조회 정확도: 시스템이 관련 문서나 정보를 정확하게 검색하여 활용하는 정도를 측정하는 지표이다. 본 연구에서 제안한 API 기반 LLM 시스템은 사전 학습된 모델을 직접 활용하여 안전조치를 생성하는 방식으로, 별도의 외부 문서 검색 과정이 없다. 따라서 해당 지표는 측정 대상에서 제외하고 0.0으로 처리하였다.

3.6 실험 결과

실험 결과는 표 1에 요약되어 있다. Baseline 1, 2는 법규 지식 부족과 환각 현상으로 인해 낮은 정확도를 보였다. 반면, 제안 모델은 조회 정확도 85.59%로 실제 안전보건 조치를 잘 참고할 수 있도록 하였으며, 이를 통해 가장 우수한 답변 유효성을 보였다. 이는 제안하는 시스템이 산업 현장에서 요구되는 신뢰성·정확성을 만족할 수 있음을 입증한다.

(표 1) 실험 결과

모델	답변 유효성	조회 정확도
Baseline 1	52.44%	-
Baseline 2	45.24%	-
Proposed	65.67%	85.59%

4. 결 론

본 연구는 산업안전 분야에 특화된 RAG 기반 안전보건 조치 의사결정 지원 시스템을 설계·구현함으로써, 산업 현장에서 필요한 신뢰성 높은 안전보건 조치 생성의 가능성을 입증하였다. 제안 시스템은 특히 환각 현상의 억제와 최신 법령 정보의 실시간 반영을 통해 기존 범용 LLM의 한계를 극복하였다. 실험 결과, RAG 기법 적용 이후 실제 안전보건 조치를 85% 이상의 조회 정확도로 검색하였으며, 이를 통해 답변 유효성이 13%p 이상 향상되었다. 또한 GPT-4o 모델과 비교하였을 때는 20%p 이상 향상되었다. 향후 다양한 산업 분야로의 확장, 데이터 전처리, 실시간 법규 업데이트 기능 구현을 통해 산업 현장 적용성을 강화할 계획이다.

감사의 글

본 연구는 과학기술정보통신부와 정보통신기획평가원(IIT

P)이 주관하는 대학-기업 협력 소프트웨어 아카데미 국가 프로그램(과제번호: 2022-0-01112)의 지원을 받아 수행되었음.

참 고 문 헌

[1] E. Ahmadi, et al. "Automatic construction accident report analysis using large language models (LLMs)." Journal of Intelligent Construction 3.1 (2025): 1-10.

[2] S. Charalampidou, et al. "Hazard analysis in the era of AI: Assessing the usefulness of ChatGPT4 in STPA hazard analysis." Safety Science 178 (2024): 106608.

[3] Y. Wang, et al. "Causation analysis of crane-related accident reports by utilizing ChatGPT and complex networks." Journal of Building Design and Environment 3.2 (2025): 202535-202535.

[4] P. Lewis, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, vol. 33, 2020.

[5] Y. Jeong, et al. "Retrieval-augmented visual parcel invoice understanding transformer for address correction." Engineering Applications of Artificial Intelligence 158 (2025): 111542.

Mask-and-Reconstruct (MAR) on Noisy WiFi CSI for Human Pose Estimation

Iftikhar Ahmad
Computer Engineering Department
Chosun University
Republic of Korea
iftikhar@chosun.ac.kr

Wooyeol Choi
School of Computer Science and Engineering
Chung-Ang University
Republic of Korea
wchoi@cau.ac.kr

ABSTRACT

This study presents a long short-term memory (LSTM)-based deep learning (DL) model structured in a U-Net shape with an auxiliary self-supervised training strategy, Mask-and-Reconstruct (MAR), for WiFi-based human pose estimation. During training, MAR randomly masks small time × subcarrier patches of the input CSI, and the network is optimised to reconstruct only the masked entries (auxiliary task) while simultaneously predicting human pose (primary task). The proposed method is evaluated using WiFi CSI data collected from four volunteers with a single 3×3 multiple-input multiple-output (MIMO) setup. The robustness of the proposed method is assessed by adding Additive white Gaussian noise (AWGN) at Signal-to-noise ratio (SNR) levels ranging from 0 to 15 dB. The results show that MAR yields small but consistent gains under noisy CSI data.

Keywords: LSTM, CSI, MAR, PCK, Human pose, WiFi

INTRODUCTION

Recent research has concentrated on WiFi-based human pose estimation due to its privacy-preserving qualities, effectiveness in occluded and dark environments, and cost efficiency [1]. In WiFi sensing, channel state information (CSI) is a physical-layer measurement whose amplitude and phase fluctuate over time and subcarriers because of human movement, enabling the estimation of human pose [2]. Moreover, deep learning (DL) models have shown significant potential in capturing the complex, nonlinear relationships between WiFi CSI and human poses. However, CSI is often affected by noise and environmental factors, which require robust techniques to enhance the stability and accuracy of pose estimation [3]. Inspired by prior studies [4] and [5], we propose a lightweight Mask-and-Reconstruct (MAR) auxiliary task for a long short-term memory (LSTM) model, which reduces noise-induced degradation in the CSI. During training, it masks small time × subcarrier CSI patches and requires the network to reconstruct only the hidden entries while simultaneously learning the pose.

PROPOSED METHOD

In this work, we propose a teacher–student framework for human pose estimation. The teacher network extracts human poses from camera images by employing YOLOv3 [6] for person detection and RMPE [7] for pose regression. The teacher network's output serves as ground truth to supervise the student network during training. The student network then learns to estimate human poses from WiFi CSI, which contains the same pose information as the camera data, since both modalities are synchronised through timestamps. After training, the student network operates independently, enabling reliable human pose estimation solely from WiFi CSI signals. Furthermore, the student network is implemented as a U-Net architecture built with LSTM units. The encoder comprises five layers with 128, 64, 32, 16, and 8 LSTM units, while the decoder mirrors this configuration in reverse order across five layers. Attention-based skip connections are included between corresponding encoder and decoder layers, and a final fully connected layer outputs 17 human body keypoints. Moreover, our key contribution is a MAR auxiliary objective for LSTM-based WiFi-CSI pose estimation. During training, we mask small time × subcarrier patches in the input CSI and train the model to both predict poses (primary task) and reconstruct only the masked CSI entries (auxiliary task). We apply MAR with 2 × 6 patches and a 0.30 masking ratio—i.e., about 30% of the input CSI grid is hidden at the block level. The auxiliary loss is computed only on the hidden cells, encouraging the LSTM to exploit local temporal–spectral context and thereby enhance tolerance to noisy or partially missing CSI. During testing, no masking is applied, allowing the model to predict poses directly without additional costs. This lightweight regularisation yields small, consistent gains, particularly under noisy conditions.

EXPERIMENTAL SETUP

The experimental setup features a single WiFi transmitter and receiver, each equipped with three antennas, forming a 3×3 MIMO system. The transmitter operates at 2.4 GHz using orthogonal

frequency division multiplexing (OFDM) technology, with 30 subcarriers over a bandwidth of 20 MHz. The sampling frequency of the WiFi setup is set to 100 Hz and synchronised with a camera running at 20 frames per second, such that each video frame corresponds to five consecutive CSI samples. Accordingly, the model input has the shape (5, 30, 3, 3), where the first dimension represents five consecutive CSI packets, 30 is the number of subcarriers, and the last two dimensions indicate the number of transmit and receive antennas. Data collection was carried out in the Smart Networking Laboratory, Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea. The dataset includes 3,500 CSI samples recorded from four volunteers performing various activities, including walking, standing, sitting, squatting, and raising hands. Finally, to evaluate robustness, additive white Gaussian noise (AWGN) was added to the original CSI at signal-to-noise ratios (SNRs) of 0 dB, 5 dB, 10 dB, and 15 dB to assess the performance of the proposed method. With 2×6 patches (time × subcarriers) and a 0.30 masking ratio, MAR conceals 30% of the input CSI grid (5 × 30 × 3 × 3) at the block level.

RESULTS AND DISCUSSION

Table 1 shows the performance evaluation of the proposed method under different noise conditions, measured in terms of percentage of correct keypoints (PCK) @5 and PCK@10, both with and without MAR. The values in the table represent the average PCK scores, calculated by taking the mean performance across all 17 body keypoints for the entire dataset. As expected, increasing the SNR from 0 dB to 15 dB consistently enhances performance across all metrics, since higher SNR indicates reduced noise interference and more reliable CSI-based feature extraction. At low SNR levels (0–10 dB), including MAR yields significant improvements, with gains observed in both PCK@5 and PCK@10. For instance, at 0 dB SNR, PCK@5 increases from 2.41% to 2.69%, and PCK@10 from 9.04% to 9.87% when MAR is applied. These findings demonstrate the effectiveness of MAR in mitigating noise-induced degradation, particularly in challenging conditions.

At higher SNR levels, the performance gap between the two configurations diminishes, indicating that MAR provides diminishing returns as noise levels drop. For example, at 15 dB SNR, PCK@5 increases only slightly from 9.40% without MAR to 9.54% with MAR, while PCK@10 shows a slight rise from 29.62% to 29.81%. These findings suggest that the proposed MAR mechanism is especially effective in low-SNR conditions, where signal corruption is severe, but its impact plateaus under moderate-to-high SNR scenarios. Overall, these results confirm that MAR primarily enhances robustness in challenging environments and ensures stable performance when WiFi CSI measurements are heavily influenced by noise.

Table 1. Performance comparison of the proposed method with and without MAR under AWGN.

Noise Level	PCK@5		PCK@10	
	Without MAR	With MAR	Without MAR	With MAR
SNR 0	2.41%	2.69%	9.04%	9.87%
SNR 5	4.86%	5.06%	16.27%	17.31%
SNR 10	6.73%	6.94%	22.27%	22.70%
SNR 15	9.40%	9.54%	29.62%	29.81%

CONCLUSION

This paper presented an LSTM-based U-Net for WiFi CSI–based human pose estimation, enhanced with a lightweight Mask-and-Reconstruct (MAR) auxiliary objective. By randomly masking small patches in the time and subcarrier dimensions of the input CSI during training and reconstructing only the hidden entries while simultaneously learning the pose, the model is encouraged to leverage local temporal–spectral context. Experiments conducted on data collected with a single 3×3 MIMO setup from four participants, including evaluations under AWGN at 0–15 dB SNR, demonstrate minor but consistent improvements in PCK@5 and PCK@10 with MAR, with the most notable gains observed at lower SNRs. Overall, the results suggest that MAR is a practical, low-overhead regularizer for noisy CSI.

REFERENCES

[1] He, Ying, et al. "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi." *IEEE Internet of Things Journal* 7.9 (2020): 8296-8317.

[2] Ahmad, Iftikhar, Arif Ullah, and Wooyeol Choi. "WiFi-based human sensing with deep learning: Recent advances, challenges, and opportunities." *IEEE Open Journal of the Communications Society* 5 (2024): 3595-3623.

[3] Ratnam, Vishnu V., et al. "Optimal preprocessing of WiFi CSI for sensing applications." *IEEE Transactions on Wireless Communications* 23.9 (2024): 10820-10833.

[4] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[5] Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[6] Redmon, Joseph. "Yolov3: An Incremental Improvement." arXiv preprint arXiv:1804.02767 (2018).

[7] Fang, Hao-Shu, et al. "RMPE: Regional Multi-Person Pose Estimation." *Proceedings of the IEEE international conference on computer vision, Venice Italy,* 22-29 October, 2017, pp. 2334-2343.

Crack Segmentation Using U-Net and Transformer Combined Model

노주현, 조정운, 양희덕

조선대학교 컴퓨터공학과

e-mail : narak@chosun.ac.kr, jwnstj8032@naver.com, heedeok_yang@chosun.ac.kr

1. 연구 목표

균열 감지의 중요성

- 인프라의 안정성과 수명 유지 필수적
- 시간 및 자원 측면에서 비효율적

기존 방법의 한계

- 육안 검사: 주관성으로 인한 결과 불일관성
- 이미지 처리 기술: 복잡한 환경에서의 감지 어려움
- 대규모 인프라 적용 시 시간, 인력, 자원 소모

딥러닝 기반 접근법

- **이미지 분류**: AlexNet, VGG, GoogLeNet, ResNet
- **객체 감지**: R-CNN, Fast R-CNN, Faster R-CNN, YOLO, SSD
- **이미지 분할**: FCN, U-Net, DeepLab, PSPNet, Mask R-CNN

CNN & 트랜스포머의 한계 극복

- CNN: 지역 특징 추출에 강하지만 전역적 의존성을 포착 못함
- 트랜스포머: 전역적 의존성을 잘 포착하지만 미세 세부 정보 감지에 어려움

연구 목표

- U-Net 인코더-디코더 구조에 비전 트랜스포머 통합
- 복잡한 균열 패턴 효과적 학습
- 자동화되고 정밀한 균열 분할 가능하게 하기
- 수동 검사 시간 및 비용 획기적 감소

기대 효과

- 시간 및 자원 효율성 획기적 향상
- 미세 균열 정확한 감지 및 측정
- 사후 대응에서 사전 예방으로의 전환

2. 관련 연구 동향

CNN 기반 균열 감지 및 분할

연구 발전

- **초기:**YOLO 모델을 스마트폰 이미지에 적용하여 균열 및 포트홀 감지
- **확장:**배치 정규화 및 드롭아웃 기술을 컨볼루션 모델에 통합하여 콘크리트 구조물의 균열 감지 정확도 향상
- **최신:**Inception 및 ResNet 아키텍처의 강점을 결합한 모델 개발

대표 모델

FCN
완전 연결 계층을 1x1 컨볼루션으로 대체

U-Net
대칭적인 U자형 인코더-디코더 구조

DeepCrack
FCN의 확장, 계층적 특징 학습

CrackNet
다중 스케일 특징 추출을 위한 풀링 계층 대체

비전 트랜스포머 기반 균열 분할

문제 해결

CNN 기반 분할 모델의 본질적 한계인 고정된 커널 크기와 제한된 수용 필드 해결

주요 모델

SETR (Segmentation Transformer)
CNN 기반 인코더를 ViT 인코더로 대체
↑ 다단계 특징 집계(multilevel feature aggregation) 방식이 가장 좋은 성능

DiNAT (Dilated Neighborhood Attention Transformer)
NAT의 개선된 버전, 확장된 창을 도입하여 수용 필드 확장
✓ 지역 및 전역 특징을 동시에 포착

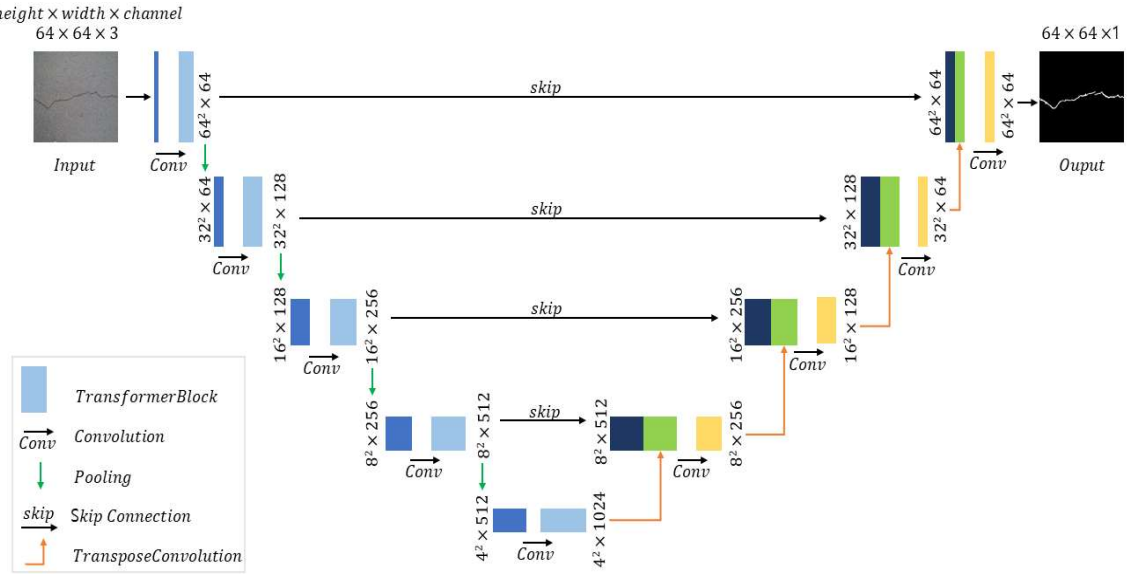
트랜스포머의 주요 특성

- 전역 특징 포착 능력
- 장거리 의존성 학습
- NAT: 창 기반 주의 메커니즘으로 계산 오버헤드 감소
- DiNAT: 다중 스케일 특징 포착

3. 연구 내용

U-Net 비전 트랜스포머 아키텍처

- U-Net의 대칭적인 인코더-디코더 구조를 사용
- 인코더: 입력된 이미지로부터 균열을 식별하는 데 필요한 핵심 특징을 추출하는 역할
- 인코더는 64x64x3 → 4x4x1024 특징 표현 생성
- 디코더: 인코더가 압축한 고차원 특징을 다시 원본 이미지 크기로 점진적으로 확대, 픽셀 단위의 최종 분할 맵을 생성하는 역할
- 디코더는 4x4x1024 → 64x64x64 → 64x64x1 출력 이미지 생성

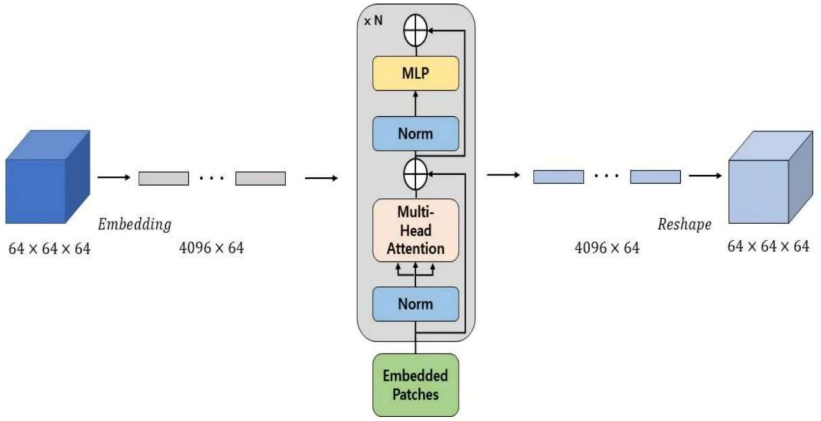


(그림 1) 제안하는 모델 아키텍처

3. 연구 내용

ViT 인코더의 핵심 구성 요소

- ViT 인코더는 기존 CNN이 이미지를 픽셀의 격자(Grid)로 보는 것과 달리, 자연어 처리에서 문장을 단어의 나열로 처리하는 방식을 차용
- 임베딩된 벡터(단어)들은 트랜스포머의 핵심인 '다중 헤드 셀프 어텐션(Multi-Head Self-Attention)' 계층으로 전달
- 본 모델은 단어가 아닌 이미지 패치를 전달함으로써 , 자연어 처리 모델이 문장 속 단어들의 관계를 파악하듯, 이미지 전체를 구성하는 각 영역(패치)들의 상호 관계와 중요도를 분석할 수 있음
- '다중 헤드'는 이러한 어텐션 과정을 여러 개의 '헤드'가 동시에, 독립적으로 수행하는 것을 의미
- 이러한 과정을 통해 ViT 인코더는 CNN의 제한된 시야를 뛰어넘어, 이미지 전체를 아우르는 거시적인 관점에서 특징을 추출하고 이해할 수 있게 됩니다.



(그림 2) ViT 인코더

4. 실험 설정 및 데이터셋

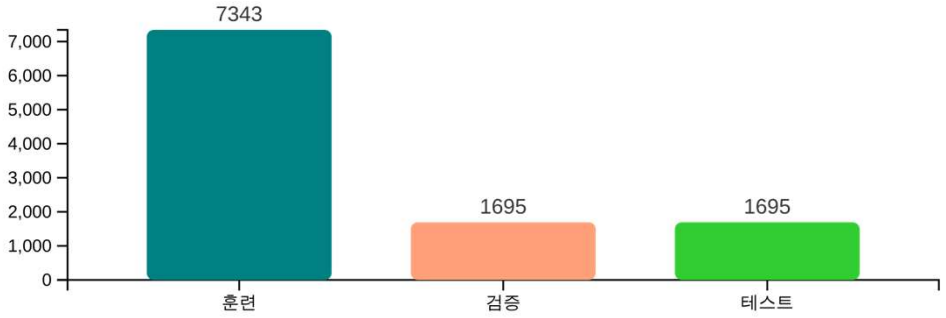
CrackSeg9k 데이터셋

데이터셋 구성
9,255개 → 11,298개의 이미지로 확장

데이터 분할
훈련: 7,343개, 검증: 1,695개, 테스트: 1,695개

데이터셋 원본
Crack500, DeepCrack, SDNet, CrackTree, Gaps, Volker, Rissbilder, NonCrack, Masonry, Ceramic 등

전처리
원본 400x400 이미지를 U-Net ViT 입력 크기에 맞추기 위해 128x128로 이중 선형 보간법으로 크기 조정



(그림 3) CrackSeg9k 데이터셋 구성

실험 설정

훈련 에포크
50 에포크 동안 훈련

학습률
1e-4에서 1e-6으로 점진적 감소

배치 크기
1

옵티마이저
Adam

손실 함수
이진 교차 엔트로피

(표 1) 하이퍼파라미터

Table 2. Hyper parameter.

Epoch	Learning rate	Batch size
50	1e-4 ~ 1e-6	1
Optimizer	Loss function	
Adam	Binary Cross Entropy	

4. 실험 결과 및 비교 분석

제안 모델의 정량적 결과

정밀도 (Precision)
0.6739

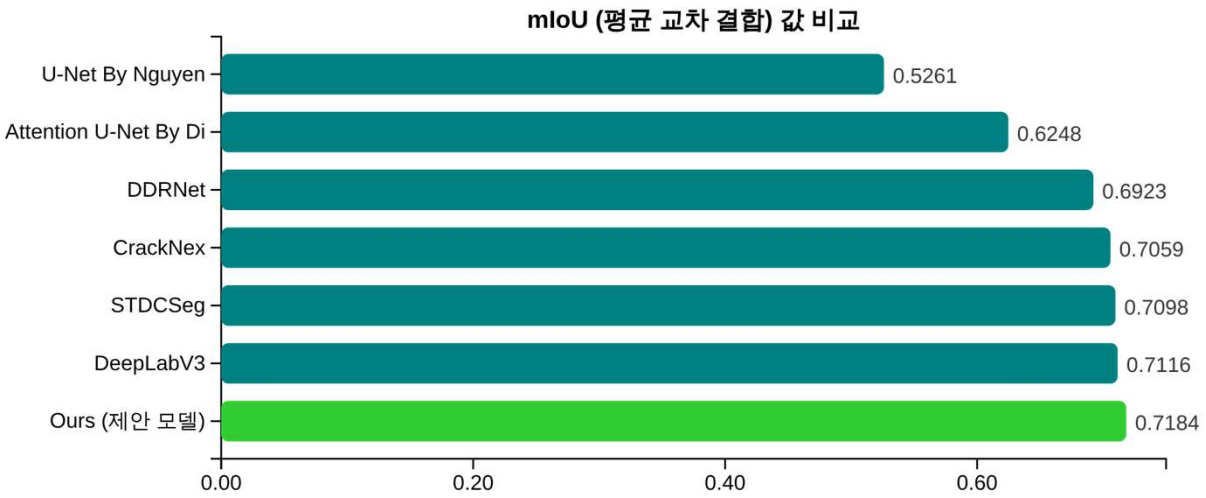
재현율 (Recall)
0.6003

F-1 점수
0.6350

mIoU
0.7184

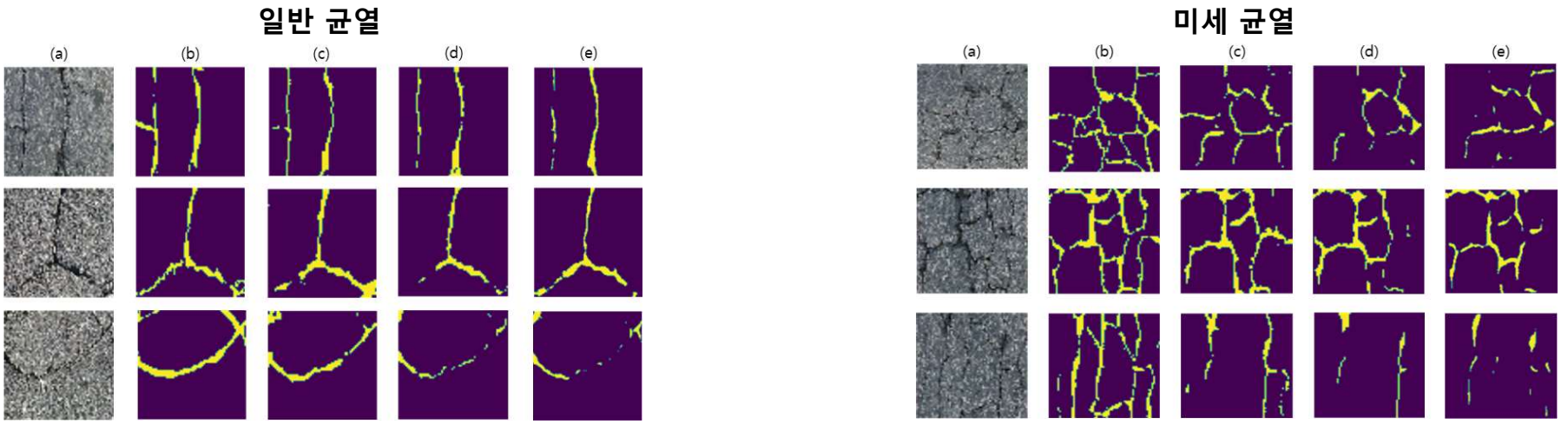
- 제안 모델은 0.7184의 가장 높은 mIoU를 달성
- 모든 평가 지표에서 개선된 성능
- 통계적 유의미한 개선 결과
- 성능과 계산 효율성의 균형 제공
- 드론 등 자원 제약 환경에 적합한 모델

모델 성능 비교 (mIoU)



(그림 4) 다른 모델과의 비교

4. 실험 결과 및 비교 분석



(그림 5) 크랙 검출 결과 Input image, (b) Label image, (c) Proposed model, (d) Attention U-Net [31], (e) U-Net.

- 다른 모델과 비교하여 미세 균열을 더 잘 검출하는 모습을 보여줌
- 연속적인 균열 검출에 있어 성공적으로 분할하는 모습
- 전체 구조를 더 잘 파악하고 있음
- 더 일관된 분할 마스크를 생성
- 이러한 시각적 결과는, 제안된 모델이 단순히 픽셀 단위의 색상이나 질감을 넘어 균열의 구조적, 문맥적 정보를 이해하는 높은 수준의 능력을 갖추었음을 증명

4. 결론 및 향후연구

연구 결론

U-Net 인코더-디코더 구조에 비전 트랜스포머를 성공적으로 통합

CrackSeg9k 데이터셋에서 0.7184의 mIoU 달성
→ 기존 U-Net 및 Attention U-Net 모델보다 우수한 성능

균열 세그먼트 연결성에서 명확한 개선
→ 다른 모델이 놓친 균열도 감지

복잡하고 불규칙한 균열 패턴 분할에 효과적

연구의 한계

정보 손실로 인한 미세 균열 분할 성능 저하

그림자나 텍스처에서 오탐 발생

DeepLabV3에 비해 mIoU 개선 미미(0.7%)

계산 비용 분석 부족

향후 연구 방향

고해상도 이미지 처리

패치 기반 처리 전략이나 더 효율적인 트랜스포머 아키텍처 탐색

오탐지 감소

다양한 비균열 이미지로 훈련 데이터셋 보강

일반화 능력 검증

다른 공개 균열 데이터셋에 대해 테스트

계산 비용 분석

CNN 전용 모델과의 런타임 및 자원 사용량 비교

최신 기반 모델 활용

VLM(Vision-Language Models)과 같은 최신 기반 모델 탐색

참고문헌

1. Ronneberger, O., et al. (2015). U-Net: Convolutional networks for biomedical image segmentation.
2. Vaswani, A., et al. (2017). Attention is all you need.
3. Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.
4. Long, J., et al. (2015). Fully convolutional networks for semantic segmentation.
5. Chen, L.C., et al. (2017). Rethinking atrous convolution for semantic image segmentation.
6. He, K., et al. (2016). Deep residual learning for image recognition.
7. Zheng, S., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers.
8. Kulkarni, S., et al. (2022). CrackSeg9k: A collection and benchmark for crack segmentation datasets and frameworks.
9. Di Benedetto, A., et al. (2023). U-Net-based CNN architecture for road crack segmentation.
10. Cha, Y.J., et al. (2017). Deep learning-based crack damage detection using convolutional neural networks.

감사의 글

본 연구는 2025년도 과학기술정보통신부(MSIT) 및 정보통신기획평가원(IITP)의 SW중심대학지원사업의 지원을 받아 수행된 연구임 (과제번호: 2024-0-00062).