컴퓨터교육학회 논문지 2024년 제27권 제8호
https://doi.org/10.32431/kace.2024.27.8.005

트랜스포머 기반 적외선 및 가시 이미지 융합 기술 연구*

Research on Transformer-based Infrared and Visible Image Fusion Technology

정헌석† · 허재혁†† · 양수미††† · 곽성범††††
Hunsuk Chung† · Jaehyeok Hur†† · Sumi Yang††† · Seongbeom Kwak††††

요약

적외선과 가시광선 이미지 융합의 목표는 대상을 강조하고 세부 질감 정보를 포함하는 융합 이미지를 생성하는 것이다. 하지만 기존 알고리즘은 종종 이미지의 시각적 품질에만 집중하고, 의미적 내용을 간과하는 경향이 있다. 이를 해결하기 위해, 본 연구에서는 트랜스포머 모델의 전역 특징 추출 기능과 대비 언어 이미지 사전 학습(CLIP: Contrastive Language Image Pre-training)을 통한 손실 함수를 활용하여 이미지 융합 과정을 최적화하는 방법을 제안한다. 먼저, 이미지에서 로컬 및 글로벌 정보를 추출하고 상호작용하기 위해 특징 안내 트랜스포머(FGT: Feature-Guided Transformer) 모듈을 개발한다. 이후, 두 가지 서로 다른 이미지를 적응적으로 융합하기 위해 특징 동적 융합(FDF: Feature Dynamic Fusion) 모듈을 설계한다. 또한, 수학적 손실 함수와 언어 기반 손실 함수를 결합하여 융합된 이미지의 시각적 품질과 의미적 정보를 동시에 향상시켰다. 공개 데이터 세트에 대한 종합적인 실험 결과, 제안된 방법이 기존의 융합 방법들에 비해 주관적 평가에서 우수한 성능을 보였음을 입증하였다.

주제어 적외선 영상, 가시광선 영상, 이미지 퓨전, 대비 언어 이미지 사전 학습 모델, 트랜스포머, 딥러닝

ABSTRACT

The goal of infrared and visible image fusion is to generate a fused image that emphasizes targets while retaining detailed texture information. However, conventional algorithms often focus solely on visual quality, neglecting the semantic content of the images. To address this issue, this study proposes a method to optimize the image fusion process by leveraging the global feature extraction capabilities of the transformer model and utilizing a loss function derived from Contrastive Language-Image Pre-training (CLIP). First, a Feature-Guided Transformer (FGT) module is developed to extract and interact with both local and global information from the images. Then, a Feature Dynamic Fusion (FDF) module is designed to adaptively fuse the two different types of images. Additionally, the method incorporates a combination of mathematical loss functions and language-based loss functions to simultaneously enhance the visual quality and semantic content of the fused images. Comprehensive experiments on public datasets demonstrate that the proposed method outperforms existing fusion methods in terms of subjective evaluations.

Keywords Infrared Image, Visible Light Image, Image Fusion, Contrastive Language Image Pre-training Model, Transformer, Deep Learning

†정회원 극동대학교 에너지IT공학과 교수
††정회원 극동대학교 친환경에너지공학과 석사과정
†††정회원 극동대학교 에너지IT공학과 조교수(교신저자)
††††정회원 주식회사 위즈윙

논문투고 2024년 08월 28일
심사완료 2024년 11월 05일
게재확정 2024년 11월 06일
발행일자 2024년 11월 20일

* 본 논문은 2024년도 행정안정부 및 산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(20025104).
본 논문은 2024년도 산업통상자원부 및 한국에너지기술평가원(KETEP) 연구비 지원에 의한 연구임(20224000000070).

1. 서론

영상 융합은 여러 소스에서 얻은 이미지나 비디오 데이터를 결합하여 하나의 통합된 표현을 생성하는 기술로, 다양한 응용 분야에서 중요한 역할을 한다 [1, 2]. 최근 딥러닝 기술의 발전은 영상 융합 방법론에 혁신을 가져왔으며, 이는 전통적인 방법에 비해 뛰어난 성능을 보이고 있다 [3, 4, 5]. 본 연구에서는 딥러닝 기반 이미지 융합 방법을 제안하고, 딥러닝 모델이 어떻게 다양한 데이터 소스를 효율적으로 융합하여 정보 손실을 최소화하고, 더 나은 화질의 결과물을 제공하는지에 대해 논의한다. 적외선 및 가시광선 이미지 융합은 두 이미지 유형에서 상호 보완적인 정보를 추출하여 보다 상세한 융합 이미지를 생성하는 중요한 개선 기법이다[6, 7, 8]. 적외선 이미지는 물체의 열 방출을 포착하여 야간이나 가시성이 낮은 조명 조건에서도 작동이 용이하다. 그러나 일반적으로 가시 이미지의 색상과 세부 정보가 부족하여 장면을 해석하고 인식하는 데 방해가 될 수 있다. 반대로 가시 이미지는 풍부한 질감과 구조 정보를 제공하여 콘텐츠 식별과 이해를 용이하게 한다. 그러나 가시 이미지는 조명 및 오클루전과 같은 요인에 민감하다. 영상 융합은 적외선 이미지의 열화상 데이터를 실화상의 디테일 및 색상과 통합하여 다양한 환경에서 시각적 인식을 개선하는 풍부한 융합 이미지를 생성하는 것을 목표로 한다. 영상 융합은 의료 분야, 보안 및 감시, 군사 및 국방 [9, 10], 자율 주행 [11], 원격 탐사와 같은 실용적인 애플리케이션에 적용이 가능하다.

최근 몇 년 동안 영상 융합을 위한 수많은 딥러닝 기반 알고리즘이 제안되었으며[12, 13], 이는 크게 컨볼루션 신경망(CNN) 기반 방법, 자동 인코더(AE) 기반 방법, 생성적 적대 신경망(GAN) 기반 방법의 세 가지 그룹으로 분류된다[14]. 이미지 융합 기술의 상당한 발전에도 불구하고 이러한 모델의 성능을 제약하는 지속적인 결함이 존재한다. 첫째, 대부분의 이미지 융합 방법은 주로 시각적 품질 향상에 초점을 맞추지만 이미지의 의미론적 일관성을 무시하는 경우가 많다. 이러한 경향은 융합된 이미지의 의미론적 불일치를 초래하여 상위 수준의 다운스트림 작업에서 잠재적으로 효율성을 저해할 수 있다. 둘째, 현재의 방법은 주로 특징 추출을 위해 컨볼루션 신경망(CNN)을 활용한다. 그러나 이러한 방법은 글로벌 정보를 효과적으로 처리하지 못하는 경우가 많아 이미지의 복잡하고 글로벌한 특징을 종합적으로 포착하고 활용하는 데 방해가 된다. 셋째, 기존 방법론에서 사용되는 융합 전략은 이미지의 다양한 콘텐츠와 특징에 대한 적응성이 부족한 기존의 수동 기법에 의존하는 경우가 많다. 따라서 이러한 전통적인 전략은 다양한 융합 상황에서 최적의 결과를 얻지 못할 수 있다.

이 연구에서는 적외선 및 실화상 이미지 융합을 위한 트랜스포머 기반 방법을 제안한다. 이 연구의 주요 기여는 다음과 같다:

- 우리는 소스 이미지의 글로벌 정보와 로컬 정보를 능숙하게 통합하는 영상 융합을 위한 새로운 트랜스포머 기반 방법을 제안한다. 또한 이미지 융합과 텍스트 프롬프트를 연결하는 브리지를 구축하여 융합된 이미지의 의미 정보를 풍부하게 한다.
- 멀티모달 특징의 상호 작용과 통합을 용이하게 하는 특징 가이드 트랜스포머 모듈을 제시하여 로컬 및 글로벌 보완 특징의 효율적인 통합을 가능하게 한다.
- 다양한 모달 특징의 융합 가중치를 적응적으로 조정하여 모달 간 동적 정보 통합을 용이하게 하도록 설계된 특징 동적 융합 모듈을 소개한다.
- 이미지 콘텐츠와 텍스트 프롬프트 간의 복잡한 관계를 쉽게 이해하고 캡슐화할 수 있는 언어 기반 손실 함수를 개발한다. 이 손실 함수는 융합된 이미지의 의미적 표현을 향상시킨다.

2. 기존 연구

딥러닝 기반 방법은 크게 자동 인코더 기반 방법, 심층 컨볼루션 신경망 기반 방법, 생성적 적대 신경망 기반 방법의 세 가지로 나눌 수 있다. 자동 인코더 기반 방법은 일반적으로 한 쌍의 인코더와 디코더를 사용한다. 인코더는 멀티모달 특징을 추출하고 디코더는 융합된 이미지를 재구성한다. 이 접근 방식은 멀티모달 정보를 효과적으로 통합하여 융합된 이미지의 전체적인 표현을 최적화한다. Li 등[15]은 처음으로 고밀도 블록을 활용하는 멀티모달 특징 추출을 위한 자동 인코더 기반 접근 방식인 DenseFuse를 소개하였다. 먼저 수동으로 설계된 융합 전략을 적용하여 특징을 통합하고 마지막으로 디코더를 사용하여 이미지 재구성을 수행하였다. Zhao는 인코더에서 이중 스케일 분해를 사용하여 소스 이미지의 배경과 세부 특징을 분리하는 이미지 융합 방법을 도입했다 [16]. Jian은 이미지 융합 작업에 주의 메커니즘을 통합하여 네트워크가 소스 이미지의 두드러진 특징과 질감 디테일에 집중할 수 있는 능력을 향상시켰다 [17].

심층 컨볼루션 신경망 기반 방법은 신중하게 설계된 손실 함수와 네트워크 아키텍처를 사용하여 효과적인 특징 추출, 융합 및 이미지 재구성을 용이하게 한다. Zhang은 소스 이미지의 강도와 그라데이션 특징을 개별적으로 학습하기 위해 이중 경로 고밀도 네트워크를 제안했다 [18]. 한편, 이들은 융합된 이미지의 강도와 그라데이션 정보 균형을 유지하기 위해 손실 함수를 설계했다. Li는 메타 학습과 컨볼루션 신경망을 결합하여 서로 다른 해상도에서 적외선과 실화상 이미지를 융합했다 [19]. Xu는 통합 이미지 융합 프레임워크를 제안했다 [20]. 이 프레임워크는 다양한 소스 이미지의 중요도를 적응적으로 평가할 수 있는 고밀도 네트워크와 정보 측정 네트워크를 통합한다.

생성적 적대 신경망 기반 방법은 소스 이미지에서 정보 전송의 균형을 맞추기 위해 확률적 분포 제약을 설정하는 것을 목표로 한다. FusionGAN [21]은 GAN을 이용한 이미지 융합 분야의 선구자이다. 적외선 이미지와 가시 이미지를 병합하기 위한 생성적 적대 프레임워크를 구축하여 융합된 이미지의 텍스처 구조를 크게 향상시켰다. Yang

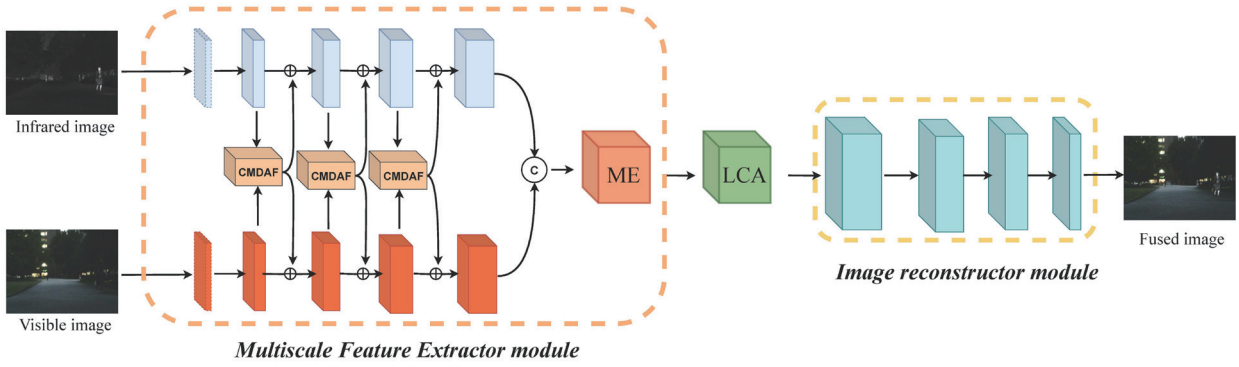


Figure 1. Architecture of the proposed method.

Table 1. 특징 안내 트랜스포머와 특징 동적 융합의 차이

항목	특징 안내 트랜스포머	특징 동적 융합
설계 철학	트랜스포머 기반의 전역적인 특징 추출	특징 분해와 융합에 초점, CNN 사용 가능
주요 목적	이미지 내 전역 패턴과 상관관계 학습	서로 다른 정보 요소를 분리하고 융합
동작 메커니즘	Self-attention을 통한 전역적 특징 학습	특징을 분리하여 고유 정보와 보완적 정보를 결합
주요 사용 단계	이미지의 고차원적 특징 추출 단계	적외선 및 가시광 이미지의 정보 융합 단계
융합 방식	각 이미지를 독립적으로 처리 후, 융합	각 이미지의 정보를 분해 후, 융합

은 IVF를 위한 텍스처 조건부 생성적 적대 네트워크를 도입하였다 [22]. 이 네트워크는 텍스처 맵을 사용하여 융합 프로세스 전반에 걸쳐 그라데이션 변화를 포착한다. Liu는 이미지 융합 및 물체 감지 작업을 위한 최적화 체계를 제안했다 [23]. 이 방식은 이미지 융합 기술을 통합하여 후속 고수준 비전 작업의 성능을 향상시킨다.

3. 제안 영상 융합 기술

3.1 개요

제안하는 방법은 (1) 특징 유도 트랜스포머 모듈, (2) 특징 동적 융합 모듈, (3) 이미지 재구성 모듈의 세 가지 모듈로 구성된다. 구체적으로 적외선 이미지 과 실화상 이미지를 특징 유도 트랜스포머 모듈에 개별적으로 입력하여 와이라는 두 가지 모달리티의 특징을 얻는다. 그 후 특징 유도 트랜스포머 모듈은 이러한 특징을 활용하여 융합된 특징을 생성한다. 마지막으로 이미지 재구성 모듈은 이러한 융합된 특징을 융합된 이미지 으로 재구성한다. 또한, 제안된 네트워크는 언어 기반 손실 과 수학적 손실 에 의해 제약을 받는다. 이러한 손실 함수는 소스 이미지의 구조적 특징과 특성을 보존하는 동시에 전역 및 로컬 디테일을 최적화한다. 표 1은 특징 안내 트랜스포머와 특징 동적 융합의 차이를 도시한다.

3.2 네트워크의 구조

[기능 안내 변환 모듈] 특징 유도 트랜스포머 모듈은 채널 주의 유닛과 두 개의 트랜스포머 유닛으로 구성된다. 처음에 소스 이미지는 컨볼루션 레이어에 의해 특징 표현으로 변환된 후 채널 주의 유닛에 입력된다.

채널 주의 유닛은 글로벌 평균 풀링을 사용해 특징에서 글로벌 컨텍스트 정보를 추출한다. 그런 다음 이 정보는 피처 내의 세밀한 디테일을 포착하도록 설계된 두 개의 추가 컨볼루션 레이어를 통해 정제된다. 마지막으로 시그모이드 함수는 입력 피처와 요소별 곱셈을 수행하여 중요한 피처는 강화하고 중복 정보는 억제하는 주의도 맵을 생성한다. 트랜스포머 유닛의 주요 기능은 복잡한 장면을 해석하는 데 필수적인 다중 헤드 자기 주의 메커니즘을 통해 특징 간의 장거리 의존성을 설정하는 것이다. 자기 주의 메커니즘을 통해 모델은 여러 헤드에 걸쳐 다양한 주의 분포를 학습할 수 있으므로 다양한 특징 표현을 포착하고 모델의 표현력을 강화할 수 있다. 자기 주의 메커니즘 이후에는 정규화 레이어와 다층 퍼셉트론 레이어가 구현되어 특징을 더욱 세분화하여 변별력을 향상시킨다.

[특징 동적 융합 모듈] 그림 1은 조밀하게 연결된 블록, 글로벌 평균 풀링, 시그모이드 활성화 기능으로 구성된 특징 밀도 융합 모듈의 아키텍처를 보여준다. 제안 아키텍처는 특징 추출 방식, 멀티스케일 정보 처리 능력, 적응형 융합 전략, 복잡한 패턴 학습 능력, 계산 효율성 측면에서 기존 모델과 차별성이 있다(표 2). 특징 밀도 융합 모듈의 핵심은 다양한 모달리티에 걸쳐 상호 보완적인 정보를 완벽

Table 2. 제안 아키텍처와 기존 모델과의 차이

항목	기존 모델	제안 트랜스포머 기반 모델
특징 추출 방식	계층적으로 지역적인 특징을 추출. 주로 작은 커널을 사용하여 이미지의 로컬 정보에 중점.	전역적인 컨텍스트 효율적으로 파악. 이미지 내 장거리 종속성을 모델링하는데 용이. 전체 이미지 정보를 균형 있게 고려하여 특징 추출.
멀티스케일 정보 처리	다중 스케일 피라미드나 특수한 CNN 아키텍처를 통해 멀티스케일 정보를 처리.	다중 스케일 정보를 자연스럽게 처리할 수 있는 메커니즘, self-attention 메커니즘을 통해 다른 픽셀과의 관계를 고려. 다양한 스케일에서의 복합 정보를 효과적으로 융합.
적응형 융합 전략	주로 고정된 융합 규칙이나 단순한 결합 방식을 사용하여 적외선 및 가시 이미지의 특징을 결합.	적응적으로 특징을 결합하는 능력을 통해 가시 및 적외선 이미지의 정보를 유연하게 융합. 픽셀마다 다른 융합 가중치를 부여하여 정교한 융합 결과 도출.
복잡한 패턴 학습 능력	로컬 패턴 학습에 강점이 있으나, 전역 패턴을 학습하는 데 한계가 있음.	전역적으로 복잡한 패턴을 학습하는 데 강점, 적외선 및 가시영상 간 복잡한 상호작용 포착.
계산 효율성	CNN은 비교적 계산량이 적지만, 이미지의 전역적인 정보를 처리하는 데 한계가 있음.	계산량이 많이 들 수 있으나, 최근 효율적인 트랜스포머 변형들이 제안되어 계산 비용을 줄이면서도 전역 정보를 잘 처리함.

하게 통합하는 기능이다. 적외선 및 가시광선 특징(ΦEO, ΦIR)은 특징 밀도 융합 모듈의 입력으로 사용된다. 처음에 이러한 특징은 별도의 밀도 블록을 통해 추가 추출을 거친다. 각 밀도 블록은 비선형 활성화 레이어가 있는 여러 컨볼루션 레이어로 구성되어 레이어 간 정보 교환을 향상시킨다. 그 후 글로벌 평균 풀링을 통해 공간 정보를 글로벌 특징 설명자로 통합한다. 그런 다음 시그모이드 활성화 함수를 적용하여 융합 가중치를 계산하고 융합된 특징에서 정보의 상대적 중요도를 결정한다. 마지막으로 특징 가중치 프로세스를 통해 융합된 특징이 생성된다.

특징 안내 트랜스포머는 주로 적외선과 가시광 이미지를 독립적으로 처리하면서, 각 이미지에서 전역적이고 풍부한 특징을 추출하는 데 중점을 둔다. 한편 특징 동적 융합은 추출된 특징을 기반으로 두 이미지의 중요한 정보를 효과적으로 융합하는 역할을 한다. 이는 정보를 분리한 후 재조합하는 특성이 있으며, 중복되거나 상충되는 정보를 최소화하여 최종 융합된 이미지를 생성한다. 따라서, 특징 안내 트랜스포머는 특징 생성에 집중하고, 특징 동적 융합은 그 생성된 특징을 최적의 방식으로 결합하는데 주력한다.

[이미지 재구성 모듈] IR 모듈은 융합된 특징 ΦFUSION을 입력으로 받아들인다. 그런 다음 컨볼루션 및 활성화 레이어를 통해 이미지의 공간 해상도와 특징 표현력을 점진적으로 증가시켜 융합 이미지를 생성한다. 첫 번째 컨볼루션 레이어는 커널과 256개의 입력 채널을 사용하여 융합된 특징의 공간 정보를 효과적으로 활용한다. 이후 컨볼루션 레이어는 이 구성을 따르지만 채널 수를 각각 128개, 64개, 32개로 점진적으로 줄이다. 이 설계는 계산 복잡성을 최소화하고 네트워크 구조를 최적화하는 동시에 심층 특징의 표현을 보존한다. 각 컨볼루션 레이어 다음에 LeakyReLU 활성화 함수가 적용된다. 최종 컨볼루션 레이어는 커널을 활용하고 쌍곡탄젠트 활성화 함수(Tanh)를 연결한다. 이 함수는 출력값을 범위로 정규화하여 융합 이미지를 생성한다.

3.3 손실 함수

시각적으로, 그리고 의미적으로 융합 이미지의 고품질을 보장하기 위해 언어 기반 손실 L_C와 수학적 손실 L_M을 포함하는 복합 손실 함수를 구축했다. 이 손실 함수의 목적은 융합 네트워크가 소스 이미지의 시각적 특성을 유지하면서 특정 의미적 설명을 준수하도록 유도하는 것이다. 우리는 대조 언어 이미지 사전 학습(CLIP)의 이미지-텍스트 인코딩 기능을 사용해 융합된 이미지와 해당 텍스트 프롬프트 간의 의미적 일관성을 평가하는 언어 기반 손실 함수를 개발했다. 이 함수는 고차원 특징 공간에서 최대한의 유사성을 보장하여 이미지와 텍스트 간의 의미적 연관성을 강화하고, 다음과 같이 공식화할 수 있다.

$$L_C = 1 - \frac{\epsilon_t \cdot \epsilon_I}{\|\epsilon_t\|\|\epsilon_I\|} \quad (1)$$

여기서 ϵ_t와 ϵ_I는 융합된 이미지와 텍스트 프롬프트 임베딩을 의미한다. 또한 수학적 손실 함수를 도입하여 네트워크를 더욱 제한함으로써 융합 이미지가 소스 이미지와 콘텐츠 일관성을 유지할 수 있도록 하였다. 수학적 손실 L_M는 다음과 같이 정의된다.

$$L_M = \alpha_1 \cdot L_{ssim} + \alpha_2 \cdot L_{grad} + \alpha_3 \cdot L_{per} \quad (2)$$

여기서 L_{ssim}, L_{grad}, 및 L_{per}은 구조적 손실, 그라데이션 손실 및 지각적 손실을 나타낸다. α_1, α_2, 및 α_3은 각 용어의 기여도를 제어하기 위한 가중치 인자이다. 구조적 손실은 구조, 휘도, 대비 측면에서 융합 이미지와 소스 이미지 간의 유사성을 평가한다. 이 손실 함수는 소스 이미지와 융합 이미지의 구조적 일관성을 유지하도록 보장하고, 다음과 같이 공식화할 수 있다.

$$L_{ssim} = 1 - SSIM(I_f, \max(I_{IR}, I_{EO})) \quad (3)$$

여기서 $SSIM(\bullet)$은 구조적 유사성 측정값을 나타낸다 [24]. 에지 정보와 이미지 디테일의 보존을 강화하기 위해, 융합 이미지와 소스 이미지 간의 그라데이션 불일치를 정량화하여 네트워크가 융합 프로세스 전반에 걸쳐 필수 구조적 디테일을 유지하도록 유도하기 위해 그라데이션 손실을 통합했다. 그러면 그라데이션 손실은 다음과 같이 공식화된다.

$$L_{grad} = \| \nabla I_f - \max(\nabla I_{IR}, \nabla I_{EO}) \|_2 \quad (4)$$

여기서 ∇은 그라디언트 연산이고 $\| \bullet \|_2$는 행렬 L_2노름을 나타낸다.

우리는 지각 손실을 활용하여 높은 수준의 특징과 텍스처 정보를 캡처하고, 사전 학습된 VGG-19 네트워크에서 추출한 특징을 활용하여 이미지 품질을 평가했다. 지각적 손실은 융합된 이미지의 스타일과 질감이 원본 이미지와 일관성을 유지하도록 보장하는데 다음과 같이 정의할 수 있다.

$$L_{per} = \| \phi(I_f) - \max(\phi(I_{IR}), \phi(I_{EO})) \|_2 \quad (5)$$

여기서 $\phi(\bullet)$는 이미지 특징 추출에 사용된 사전 훈련된 VGG-19 네트워크를 나타낸다. 따라서 총 손실은 다음과 같이 공식화할 수 있다.

$$L_{final} = L_M + \beta L_C \quad (6)$$

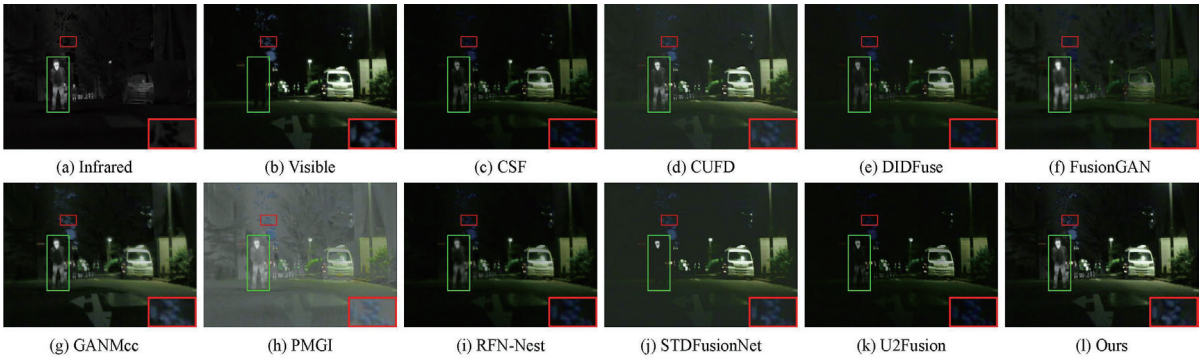


Figure 2. Qualitative evaluation results of the nine counterparts on typical image pairs from the MSRS image pairs.

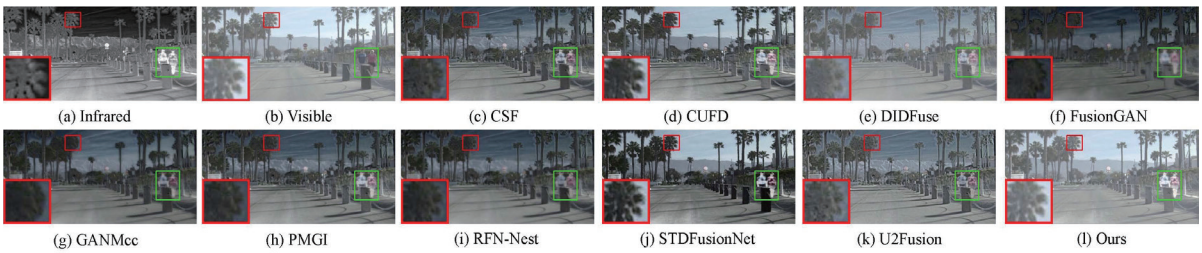


Figure 3. Qualitative evaluation results of the nine counterparts on typical image pairs from the ROAD image pairs.

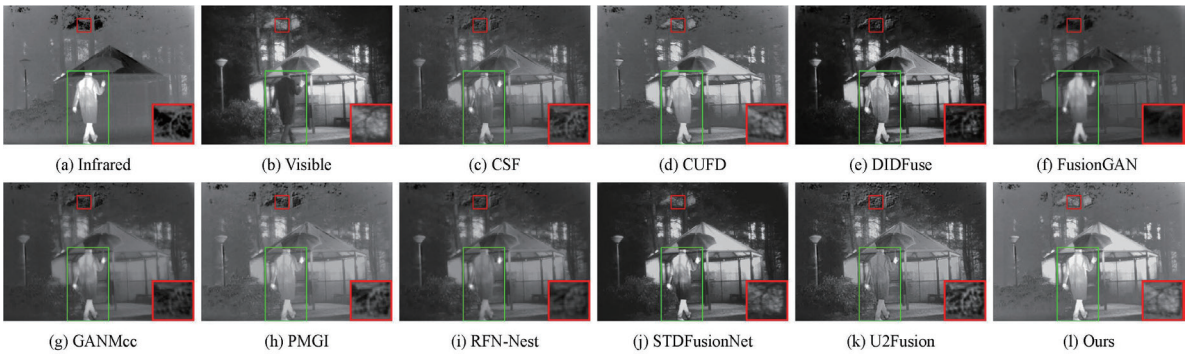


Figure 4. Qualitative evaluation results of the nine counterparts on typical image pairs from the TNO image pairs.

4. 실험 결과

4.1 데이터셋과 성능 측정 도구

우리는 제안된 모델을 훈련하기 위해 MSRS 데이터 세트의 이미지 1,000개를 훈련 세트로 사용했다. 데이터 증강을 위해 이러한 훈련 이미지를 크기의 이미지 패치로 무작위로 잘라냈다. 또한 361쌍의 MSRS 테스트 세트에서 이 방법의 효과를 검증했다. 그리고 221쌍의 로드씬 및 25쌍의 TNO 데이터 세트에서 제안한 방법의 일반화 능력을 평가했다.

4.2 비교 실험

제안하는 이미지 융합 방법의 적용 가능성과 효과를 검증하기 위해 MSRS(Multi-Sensor Remote Sensing) 데이터셋을 대상으로 성능 실험을 수행했다. 이는 주로 다양한 센서에서 수집된 원격 탐사 이미지 데이터를 포함하며, 이런 데이터셋은 다중 센서 융합 연구나 원격 탐사 응용 분야에서 중요한 역할을 한다. MSRS 데이터셋은 다양한 센서로부터 얻은 데이터를 통합하여 한 장소에 대한 다각적인 시각을 제공한다. MSRS의 특징으로는, (1) 다중 센서 데이터이고, (2) 융합 연구에 사용이 가능하고, (3) 다양한 응용 분야가 있으며, (4) 데이터가 다양성하다는 점을 들 수 있다. 따라서 공간적, 시간적, 스펙트럼적 다양한 데이터를 제공할 수 있고 이는 연구자들이 시간에 따른 변화나 다양한 스펙트럼 밴드에서의 정보를 분석할 수 있게 한다. 본 연구에서는 제안한 방법을 DenseFuse [15], FusionGAN [22], PMGI [18], UMF-CMGR [25], DATFuse [26], CrossFuse [27], YDTR [28], IRFS [29], ICAFusion [30] 등 9가지 최신방법과 비교하고, 그 결과를 그림 2-4에 도시했다.

4.3 정성적 비교

이미지 융합의 주관적 화질 평가 방법은 융합된 이미지의 품질을 평가하기 위해 사람의 주관적인 판단을 이용하는 방식이다. 이 방법은 수치적 분석이나 기계적인 측정 도구와는 달리, 인간의 시각적 및 심리적 반응을 바탕으로 이미지를 평가한다. 다중평가자 접근법 (Multiple Evaluator Approach), 절대 평가 (Absolute Rating), 상대 평가 (Comparative Rating), 주관적 평가 척도 (Subjective Evaluation Scales), 심리물리학적 평가 (Psychophysical Evaluation) 등을 들 수 있다. 본 연구에서는 MSRS 데이터 세트에 대한 정성적 비교를 진행하였다. 그 결과는 그림 2-4에 나와 있다. 결과를 볼 때, DenseFuse, DATFuse, CrossFuse, ICAFusion은 다양한 정도의 색상 왜곡을 보이는 것을 알 수 있다. 또한 PMGI, UMF-CMGR, YDTR은 에지 디테일 정보가 일부 손실되었다. FusionGAN과 IRFS의 배경 정보는 적외선 이미지의 배경에 의해 방해를 받는다. 반면, 제안하는 방법은 에지 디테일을 유지하며 시각적 인식, 목표 선명도, 배경 정보의 풍부함에서 더 나은 성능을 발휘한다.

연구를 진행하면서 확인한 MSRS(그림 2)과 TNO(그림 4) 데이터셋에서의 비교를 보면 SF (Spatial Frequency), AG (Average Gradient), Qcv 지표를 볼 때 제안하는 방식이 가장 우수한 성과를 보였다. 이는 제안된 기법이 소스 이미지의 중요한 내용과 그라디언트 정보를 완전히 유지하면서도 뛰어난 시각적 효과를 제공함을 입증한다. 한편, ROAD(그림 3) 데이터셋에서는 제안 방법이 SF, AG, Qcv에서 1위를 차지하였다. 이는 제안된 기법이 다른 경쟁 방법보다 명확한 우위를 보인다는 것을 증명한다. 종합해볼 때 제안 방법은 대부분의 성능 지표에서 우수한 결과를 나타내었으며, 특히 이미지의 중요한 정보와 그라디언트를 잘 유지하면서도 시각적 품질이 뛰어난 결과를 제공하는 데 강점을 가지고 있다.

5. 결론

이 연구에서는 트랜스포머 기반 이미지 융합 프레임워크를 제안한다. 제안한 방법은 트랜스포머 모델의 전역 특징 추출 기능과 CLIP 모델에 의해 유도된 손실 함수의 최적화 기능을 활용한다. 제안 방법은 융합된 이미지의 시각적 품질을 향상시키는 동시에 시맨틱 콘텐츠의 일관성을 보장한다. 특징 안내 트랜스포머와 특징 동적 융합 모듈을 개발함으로써 제안 방법은 소스 이미지의 로컬 및 글로벌 정보를 효과적으로 상호 작용하고 병합한다. 또한, 수학적 손실과 언어 기반 손실을 결합하여 풍부한 의미 정보를 보장하는 동시에 융합된 이미지의 시각적 품질을 향상시킨다. 세 가지 공개 데이터 세트에 대한 종합적인 실험을 통해 우리의 방법이 주관적, 객관적으로 SOTA 융합 방법보다 성능이 우수하다는 것을 알 수 있다.

참고문헌

[1] Wang, Y., Shao, Z., Lu, T., Wu, C., & Wang, J. (2023). Remote Sensing Image Super-Resolution via Multiscale Enhancement Network. *IEEE Geoscience and Remote Sensing Letters, 20,* 1-5. https://doi.org/10.1109/LGRS.2023.3248069

[2] Liu, N., Li, W., Sun, X., Tao, R., & Chanussot, J. (2023). Remote Sensing Image Fusion With Task-Inspired Multiscale Nonlocal-Attention Network. *IEEE Geoscience and Remote Sensing Letters, 20,* 1-5. https://doi.org/10.1109/LGRS.2023.3254049

[3] Wang, Y. et al. (2024). Remote Sensing Pan-Sharpening via Cross-Spectral-Spatial Fusion Network. *IEEE Geoscience and Remote Sensing Letters, 21,* 1-5. https://doi.org/10.1109/LGRS.2023.3337844

[4] Shi, J., Liu, W., Shan, H., Li, E., Li, X., & Zhang, L. (2023). Remote Sensing Scene Classification Based

on Multibranch Fusion Attention Network. *IEEE Geoscience and Remote Sensing Letters, 20*, 1-5. https://doi.org/10.1109/LGRS.2023.3262407

[5] Chen, G., Lu, H., Di, D., Li, L., Emam, M., & Jing, W. (2023). StfMLP: Spatiotemporal Fusion Multilayer Perceptron for Remote-Sensing Images. *IEEE Geoscience and Remote Sensing Letters, 20*, 1-5. https://doi.org/10.1109/LGRS.2022.3230720

[6] Mikaeili, M., & Bilge, H. Ş. (2023). Evaluating Deep Neural Network Models on Ultrasound Single Image Super Resolution. *2023 Medical Technologies Congress (TIPTEKNO)*, Famagusta, Cyprus, pp. 1-4. https://doi.org/10.1109/TIPTEKNO59875.2023.10359188

[7] Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., & Wu, W. (2019). Feedback Network for Image Super-Resolution. *arXiv:1903.09814 [cs.CV]*. https://doi.org/10.48550/arXiv.1903.09814

[8] Carreira, J., Agrawal, P., Fragkiadaki, K., & Malik, J. (2016). Human Pose Estimation with Iterative Error Feedback. *arXiv:1507.06550 [cs.CV]*. https://doi.org/10.48550/arXiv.1507.06550

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs.CV]*. https://doi.org/10.48550/arXiv.1512.03385

[10] Dong, C., Loy, C.C., He, K., & Tang, X. (2014). Learning a Deep Convolutional Network for Image Super-Resolution. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (Vol. 8692, Lecture Notes in Computer Science). Springer, Cham. https://doi.org/10.1007/978-3-319-10593-2_13

[11] Niu, A., Zhang, K., Pham, T. X., Sun, J., Zhu, Y., Kweon, I. S., & Zhang, Y. (2023). CDPMSR: Conditional Diffusion Probabilistic Models for Single Image Super-Resolution. *arXiv:2302.12831 [eess.IV]*. https://doi.org/10.48550/arXiv.2302.12831

[12] Jin, X., Chen, Y., Feng, J., Jie, Z., & Yan, S. (2016). Multi-Path Feedback Recurrent Neural Network for Scene Parsing. *arXiv:1608.07706 [cs.CV]*. https://doi.org/10.48550/arXiv.1608.07706

[13] Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Deep Back-Projection Networks For Super-Resolution. *arXiv:1803.02735 [cs.CV]*. https://doi.org/10.48550/arXiv.1803.02735

[14] Bevilacqua, M., Roumy, A., Guillemot, C.M., & Alberi-Morel, M. (2012). Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. *British Machine Vision Conference*.

[15] Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the Super-Resolution Convolutional Neural Network. *arXiv:1608.00367 [cs.CV]*. https://doi.org/10.48550/arXiv.1608.00367

[16] Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *In Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, Vancouver, BC, Canada. 2, 416-423. https://doi.org/10.1109/ICCV.2001.937655.

[17] Ding, Q., & Yang, J. (2023). Sparse-Aware Transformer for Single Image Super-Resolution. *2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, Chengdu, China, 145-149. doi: 10.1109/CBASE60015.2023.10439116.

[18] Zeyde, R., Elad, M., & Protter, M. (2012). On Single Image Scale-Up Using Sparse-Representations. In J.D. Boissonnat et al. (Eds.), *Curves and Surfaces 2010*, Lecture Notes in Computer Science, 6920, 711-730. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27413-8_47

[19] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. *arXiv:1807.02758 [cs.CV]*. To appear in ECCV 2018. https://doi.org/10.48550/arXiv.1807.02758

[20] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *arXiv:1809.00219 [cs.CV]*. To appear in ECCV 2018 workshop. https://doi.org/10.48550/arXiv.1809.00219

[21] Kim, J., Lee, J. K., & Lee, K. M. (2016). Deeply-Recursive Convolutional Network for Image Super-Resolution. *arXiv:1511.04491 [cs.CV*]. Oral presentation at CVPR 2016. https://doi.org/10.48550/arXiv.1511.04491

[22] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv*:1502.01852 [cs.CV]. https://doi.org/10.48550/arXiv.1502.01852

[23] Han, W., Chang, S., Liu, D., Yu, M., Witbrock, M., & Huang, T. S. (2018). Image Super-Resolution via Dual-State Recurrent Networks. *arXiv:1805.02704 [cs.CV]*. https://doi.org/10.48550/arXiv.1805.02704

[24] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual Dense Network for Image Super-Resolution. *arXiv:1802.08797 [cs.CV*]. To appear in CVPR 2018 as spotlight. https://doi.org/10.48550/arXiv.1802.08797

[25] Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *arXiv:1511.04587 [cs.CV]*. Oral presentation at CVPR 2016. https://doi.org/10.48550/arXiv.1511.04587

[26] Liang, M., Hu, X., & Zhang, B. (2015). Convolutional Neural Networks with Intra-Layer Recurrent Connections for Scene Labeling. *Neural Information Processing Systems*, Montreal, Canada, 1-9.

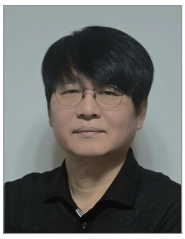
[27] Tong, T., Li, G., Liu, X., & Gao, Q. (2017). Image Super-Resolution Using Dense Skip Connections. *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 4809-4817. https://doi.org/10.1109/ICCV.2017.514.

[28] Zamir, A. R., Wu, T.-L., Sun, L., Shen, W., Malik, J., & Savarese, S. (2017). Feedback Networks. *arXiv:1612.09508 [cs.CV]*. https://doi.org/10.48550/arXiv.1612.09508

[29] Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. *arXiv:1707.02921 [cs.CV]*. To appear

in CVPR 2017 workshop. https://doi.org/10.48550/arXiv.1707.02921

[30] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs.CV]*. Presented at CVPR 2017. https://doi.org/10.48550/arXiv.1608.06993



정현석

· 1991년 서울과학기술대학교 전산학(공학사)
· 1993년 홍익대학교 데이터베이스 (이학석사)
· 1999년 홍익대학교 데이터베이스 (이학박사)
· 2003년 ~ 현재 극동대학교 에너지IT공학과 교수

관심분야 : 컴퓨터교육, 데이터베이스, 영상분석
hschung@kdu.ac.kr

허재혁

· 2023년 극동대학교 에너지IT공학과(이학사)
· 2023년~현재 극동대학교대학원 친환경에너지공학과 석사과정

관심분야 : 에너지IT, 드론영상분석, 신재생에너지
hgh9848@naver.com

양수미

· 1992년 공주대학교 물리학과(이학사)
· 1999년 충남대학교 물리학과 고체물리전공(이학석사)
· 2020년 ~ 현재 극동대학교 에너지IT공학과 조교수

관심분야 : 에너지IT, 드론영상분석, 신재생에너지
esther4853@kdu.ac.kr

곽성범

· 2017년 인천대학교 전자공학과(공학사)
· 2023년 인천대학교 임베디드시스템전공(석사과정)
· 2017년 ~ 현재 (주)위즈윙 부대표

관심분야 : 무인항공기, 임베디드시스템
rnd@wiziwng.co.kr