특집논문 (Special Paper)
방송공학회논문지 제29권 제3호, 2024년 5월 (JBE Vol.29, No.3, May 2024)
https://doi.org/10.5909/JBE.2024.29.3.252
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

멀티모달 특징 분해 트랜스포머 기반 적외선 및 가시광선 영상 합성 기법

김 가 현[a], 원 해 앙[b] ,이 철[a]‡

Infrared and Visible Image Fusion via Multi-modal Feature Decomposition Transformer

Gahyeon Kim[a], Duong Hai Nguyen[b], and Chul Lee[a]‡

요 약

본 논문은 멀티모달 영상 간 공통 및 보완 특징을 분해하는 트랜스포머 기반 적외선 및 가시광선 영상 합성 알고리즘을 제안한다. 먼저, 제안하는 알고리즘은 적외선 및 가시광선 영상에 대한 다중 스케일 특징맵을 추출한다. 다음으로 특징맵을 공통 및 보완 구성 요소로 분해하는 공통 및 세부 특징맵 분해 트랜스포머를 개발한다. 분해 성능을 향상시키기 위해서 공통 특징맵은 서로 상관되지만 보완 특징은 서로 상관되지 않도록 유도하는 분해 손실 함수를 개발한다. 마지막으로, fusion block은 공통 특징맵과 보완 특징맵을 결합하여 합성 영상을 생성한다. 모의 실험을 통해 제안하는 기법이 기존의 알고리즘보다 더욱 효과적으로 적외선 및 가시광선 영상을 합성할 수 있음을 확인한다.

Abstract

We propose an infrared and visible image fusion algorithm using common and complementary multi-modal feature decomposition transformers. First, the proposed algorithm extracts multiscale shallow features from infrared and visible images. Then, we develop common and complementary feature decomposition transformers, which decompose the features into common and complementary components for each modality. For better decomposition, we develop a decomposition loss by constraining the common features to be correlated, while the complementary features are uncorrelated. Finally, the fusion block generates the fused image by combining the common and complementary features. Experimental results on a public dataset demonstrate that the proposed algorithm outperforms conventional algorithms in quantitative and qualitative comparison.

Keyword : Visible and infrared image fusion, Feature decomposition, Contrastive learning, Transformer

a) 동국대학교 컴퓨터·AI학과(Department of Computer Science and Artificial Intelligence, Dongguk University)
b) 동국대학교 멀티미디어공학과(Department of Multimedia Engineering, Dongguk University)
‡ Corresponding Author : 이철(Chul Lee)
E-mail: chullee@dongguk.edu
Tel: +82-2-2260-3339
ORCID: https://orcid.org/0000-0001-9329-7365
· Manuscript April 2, 2024; Revised May 3, 2024; Accepted May 7, 2024.

Ⅰ. 서 론

영상 합성은 서로 다른 정보를 가진 여러 개의 영상을 결합하여 더욱 유의미한 영상을 생성하는 기술이다[1,2]. 특히, 적외선과 가시광선 영상 합성은 영상의 정보가 풍부해지고 후속 처리에 용이하기 때문에 활발하게 연구되어 왔다[1]. 가시광선 영상에는 풍부한 질감 세부 정보가 포함되어 있으나 조명 변화와 같은 환경 조건에 영향을 많이 받는다. 반면, 적외선 영상은 날씨와 같은 환경 조건 변화에는 강인하지만, 세부 정보를 제공하지 못한다. 적외선과 가시광선 영상의 합성은 각 영상의 상호 보완적인 특성으로 장면과 객체를 더 잘 표현하여 객체 감지 및 추적, 장면 분할, 군중 집계 등의 컴퓨터 비전 성능을 향상시킨다[1,3-5]. 그림 1은 적외선 및 가시광선 영상 합성으로 객체 감지 성능을 향상시키는 예를 도시한다[6]. 따라서 적외선과 가시광선 영상의 효과적인 합성을 위해 다양한 알고리즘이 개발되었다.

예를 들어, 적외선 및 가시광선 영상 합성 초기 연구에는 다중 스케일 변환, 희소 표현, 부분 공간 기반, 돌출 기반 및 하이브리드 모델[1]과 같은 수학적 모델을 사용하여 입력 영상에서 서로 보완적인 정보를 추출한 후, 합성 영상을 생성하기 위한 휴리스틱 합성 규칙을 개발하였다[7]. 그러나 이러한 기존의 모델링은 적외선 및 가시광선 특징을 표현하는 능력이 떨어지고, 입력 영상의 상호 보완적인 정보를 보존하기에 어려워 합성 성능이 저하될 수 있다.

최근에는 CNN(convolutional neural network)[8-11] 또는 GAN(generative adversarial network)[12-15]을 사용하는 딥러닝 기반 적외선 및 가시광선 합성 알고리즘이 개발되었다. CNN 기반 알고리즘은 수용 필드를 이용해 지역적 특징을 추출하여 각 입력 영상에서 의미론적 특징을 형성한다. 그러나 CNN 기반 알고리즘은 큰 수용 필드를 달성하기 위해 깊은 레이어를 쌓아야 하기 때문에[16] 에지 같은 낮은 수준의 정보가 손실되어 각 양식의 정보 손실 및 합성 성능 저하를 야기한다. GAN 기반 알고리즘은 적대적 학습을 사용하여 화소값 분포를 보존하여 적외선 및 가시광선 영상을 합성한다. 그러나 입력 영상 간의 전역 컨텍스트를 캡처하지 못할 수도 있을 뿐만 아니라, 적대적 학습에 의존하기 때문에 학습에 어려움이 있다. 또한, CNN 및 GAN 기반 알고리즘은 convolution layer에서 고정된 크기의 윈도우를 사용하여 특징을 추출하기 때문에 입력 영상 간의 전역 상호 작용을 간과하는 경우가 있다.

최근 트랜스포머의 우수한 장거리 종속성 모델링 성능을 기반으로[17], self-attention 메커니즘을 통해 입력 영상 간의 전역 상호 작용을 캡처할 수 있는 다양한 트랜스포머 기반 영상 합성 알고리즘[18,19]이 개발되었다. 그러나 전역적인 상호 작용을 고려하지 않고 각 입력 영상에 대해 self-atten-

(a) (b) (c)

그림 1. 적외선 및 가시광선 영상 합성을 통한 객체 탐지 성능 향상의 예시[6] (a) 적외선 영상 (b) 가시광선 영상 (c) 합성된 영상
Fig. 1. An example of object detection performance improvement using infrared and visible image fusion[6] (a) Infrared images (b) visible images (c) fused images

tion을 사용하면 적외선 및 가시광선 영상의 상호 보완적인 특성을 완전히 활용하지 못할 수도 있다. 이 문제를 해결하기 위해 Park 등은 입력 영상의 중복성을 제거하여 상호 보완 영역을 결정하는 교차 모달 트랜스포머[20]를 제안하였다. 그러나, 객체가 두 입력 영상 모두에서 구별되는 경우, 보완 정보를 제거하여 합성 영상에 아티팩트를 생성할 수도 있다.

본 논문은 입력 영상을 공통 특징맵과 보완 특징맵으로 분해하여 효과적인 적외선 및 가시광선 영상 합성 알고리즘을 제안한다. 제안하는 기법은 입력 특징맵이 배경이나 대규모 환경 특성과 같은 공유 정보와 각 영상의 고유한 특성인 세부 정보로 구성된다고 가정한다. 먼저, 각 입력 영상의 다중 스케일 특징맵을 추출하고, 제안하는 멀티모달 특징 분해 기반 트랜스포머를 이용하여 전역 상호 작용을 캡처하여 관련성 및 비관련성 맵을 추정한다. 다음으로, 공통 특징맵 간의 유사성과 보완 특징맵 간의 차이점을 정량화하는 분해 손실 함수를 개발한다. 공통 특징맵과 보완 특징맵을 결합하여 최종 합성 영상을 생성한다. 모의 실험을 통하여 제안하는 기법이 다양한 데이터셋에서 기존 기법[10-11,20]에 비해서 효과적으로 합성 영상을 생성함을 확인한다.

Ⅱ. 제안하는 기법

1. 네트워크 구조

본 논문은 멀티모달 특징 분해 기반 트랜스포머를 이용한 적외선 및 가시광선 영상 합성 기법을 제안한다. 제안하는 네트워크는 그림 2가 도시하는 것과 같이 입력 영상에서 공통 및 보완 특징맵을 추출하도록 구성된다.

입력 영상인 한 쌍의 적외선 및 가시광선 영상 $\{\mathrm{I}_{\mathrm{ir}},\mathrm{I}_{\mathrm{vis}}\}$은 3×3 convolution layer를 이용하여 세 단계의 얕은 특징 피라미드인 $\{\mathrm{f}^{l}_{\mathrm{ir}}\}^{3}_{l=1}$와 $\{\mathrm{f}^{l}_{\mathrm{vis}}\}^{3}_{l=1}$를 추출한다. 각 레벨 l에서, 가장 상위 레벨을 제외하고 한 쌍의 특징맵은 1×1 convolution layer와 업샘플링을 통해 이전 레벨의 특징과 합쳐진다. 다음 convolution layer에서는 residual feature distillation block (RFDB)[21]을 통해 더 강력한 특징

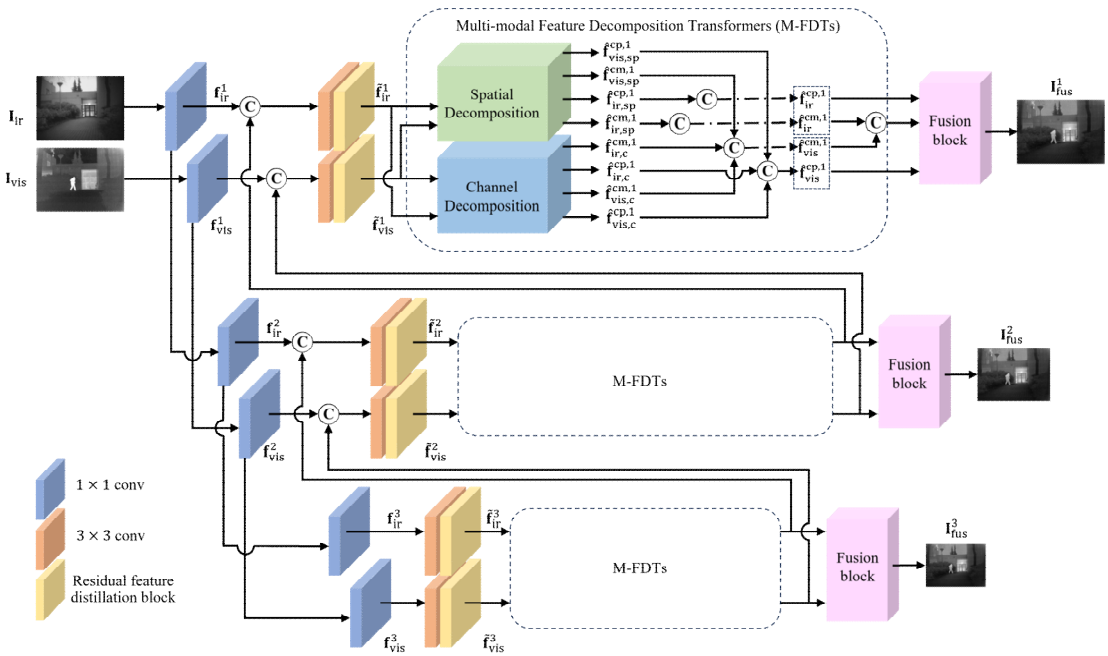


그림 2. 제안하는 전체 네트워크 구조
Fig. 2. Overview of the proposed fusion algorithm

표현을 학습한다. 제안하는 멀티모달 특징 분해 기반 트랜스포머 (M-FDTs)는 공간 및 채널 영역에서 각 입력에 대해 공통 및 고유 특징으로 분해한다. 마지막으로 fusion block은 각 특징들을 합쳐 최종 합성 영상인 I^{l}_{fus}를 생성한다.

2. Multi-modal Feature Decomposition Transformers (M-FDTs)

영상 합성을 위해 유의미한 정보를 추출하기 위해서는 지역적 정보뿐만 아니라 전역적 정보를 포착하는 것이 중요하다[20]. 이에 각 입력 영상의 화소 간 전역적인 상호 작용을 잘 포착하기 위해 트랜스포머 기반 알고리즘이 개발되었다[18,20,22]. 특히, 공간적인 특성을 고려하기 위해 공간 영역에서 self-attention을 사용하여 전역 상호 작용을 캡처하거나[23] 색상, 질감 등과 같은 시각적 특성을 포함하는 채널 영역에서 전역 정보를 얻고자 하는 연구가 진행되어 왔다[24,25]. 다양한 시각적 특성을 고려하고, 영상의 구조적 전역 정보를 효과적으로 고려할 수 있도록 공간 및 채널 영역을 함께 사용하여 전역 컨텍스트 정보의 상호 작용이 가능하도록 한다. 그중, 교차 모달 트랜스포머 (CMTs)[20]는 공간 영역과 채널 영역 모두에서 전역 상호 작용을 캡처하여 우수한 성능을 달성했으나, 보완 특징맵을 추출하기 위해 비관련성 맵만 추정하므로 두 입력 영상에서 공유되는 공통 정보가 손실되어 합성 영상에서 중요한 정보가 손실될 수 있다. 이러한 한계를 극복하기 위해, 그림 2와 3이 도시하는 것과 같이 제안하는 멀티모달 특징 분해 기반 트랜스포머는 각 입력 영상의 공간과 채널 영역에서 보완 특징 및 함께 공유되는 공통 정보가 손실되지 않도록 공통 특징을 분해하여 추출하는 spatial decomposition 및 channel decomposition을 제안한다.

3. Spatial Decomposition Transformer

공간 영역에서의 관련성과 비관련성을 활용하여 입력 영상 사이의 공통 및 고유한 정보를 추출하기 위한 공간 분해를 개발한다. 그림 3은 제안하는 공간 분해 트랜스포머의 구조를 도시한다. 적외선과 가시광선 특징을 추출하기 위해 두 개의 트랜스포머로 구성된다. 각 공간 분해 트랜스포머는 특징맵 $\{\tilde{f}_{ir}, \tilde{f}_{vis}\}$을 입력받아 공통 특징맵

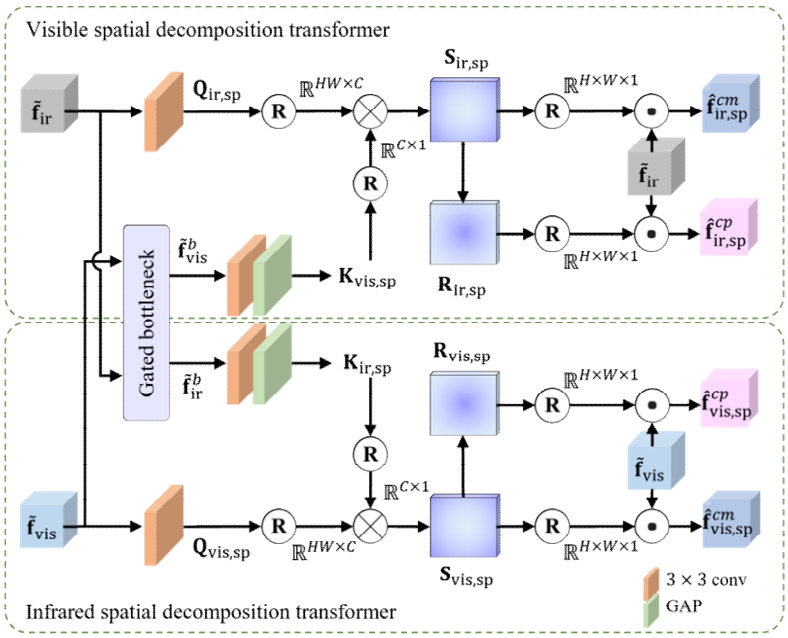


그림 3. 가시광선 및 적외선 기반 공간 영역 분해 트랜스포머의 구조
Fig. 3. Architectures of the visible and infrared spatial decomposition transformer

$\{\hat{f}_{ir}^{cm}, \hat{f}_{vis}^{cm}\}$과 고유 특징맵 $\{\hat{f}_{ir}^{cp}, \hat{f}_{vis}^{cp}\}$를 출력한다.

공간 분해는 트랜스포머는 1×1 convolution layer를 사용하여 쿼리 특징맵 $Q_{vis,sp}$를 생성한다. 입력 영상에는 서로 다른 양식의 공통된 정보가 포함되어 있으므로 양식 간의 교차로 정보를 교환하는 gated bottleneck[20]을 통해 키 특징맵 $K_{vis,sp}$를 생성한다. 다음, $Q_{vis,sp}$와 $K_{vis,sp}$를 각각 $HW \times C$ 및 $C \times 1$ 행렬로 재구성하여 관련성 맵 $S_{vis,sp}$와 비관련성 맵 $R_{vis,sp}$를 아래와 같이 추정한다.

$$S_{vis,sp} = g(Q_{vis,sp} \otimes K_{ir,sp}) \tag{1}$$

$$R_{vis,sp} = 1 - S_{vis,sp} \tag{2}$$

여기에서 $g(\cdot)$은 시그모이드 함수를 H, W, C는 각각 특징맵의 높이, 너비, 그리고 채널을 나타내고, $\otimes$는 행렬 곱셈을 의미한다.

최종적으로 공통 및 보완 특징은 아래와 같이 얻는다.

$$\hat{f}_{vis,sp}^{cm} = S_{vis,sp} \odot \tilde{f}_{vis} \tag{3}$$

$$\hat{f}_{vis,sp}^{cp} = R_{vis,sp} \odot \tilde{f}_{vis} \tag{4}$$

여기에서 $\odot$은 요소별 곱셈을 의미한다.

4. Channel Decomposition Transformer

적외선 및 가시광선 영상에는 전역 대조 및 화소값 분포와 같은 공간 및 전체 상황 정보가 포함되어 있다. 따라서 공간 분해 외에도 전역 상호 작용을 활용하는 채널 분해 트랜스포머를 개발한다. 그림 4는 개발한 채널 분해 트랜스포머의 구조를 나타낸다.

가시광선 영상에 대한 쿼리 특징 $Q_{vis,c}$는 1×1 convolution layer를 사용하여 구한다. 그런 다음, gated bottleneck을 통해 키 특징맵 $K_{vis,c}$를 생성한다. $Q_{vis,c}$와 $K_{vis,c}$을 각각 $C \times HW$ 및 $HW \times 1$로 재구성하여 관련성 맵 $S_{vis,c}$와 비관련성 맵 $R_{vis,c}$를 아래와 같이 추정한다.

$$S_{vis,c} = g(Q_{vis,c} \otimes K_{ir,c}) \tag{5}$$

$$R_{vis,c} = 1 - S_{vis,c} \tag{6}$$

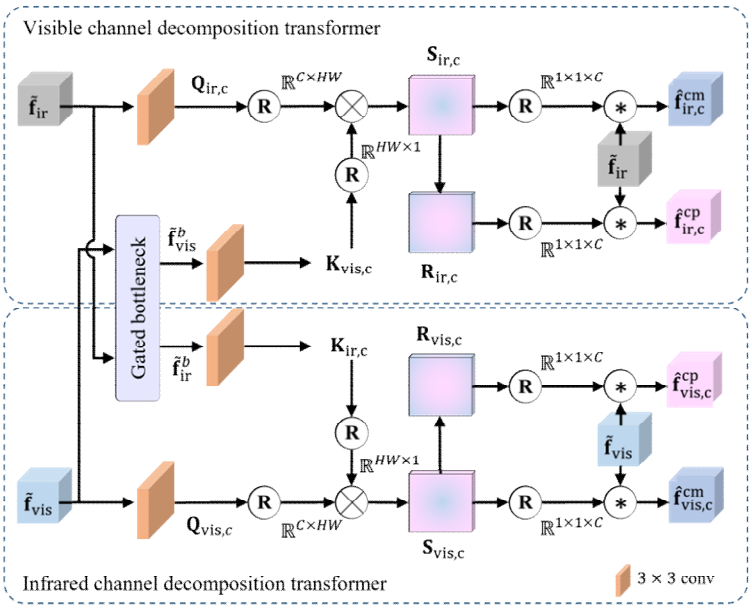


그림 4. 가시광선 및 적외선 기반 채널 영역 분해 트랜스포머의 구조
Fig. 4. Architectures of the visible and infrared channel decomposition transformer

최종적으로 채널 영역에서 공통 및 보완 특징맵은 아래와 같이 얻는다.

$$\hat{f}_{vis,c}^{cm} = S_{vis,c} * \tilde{f}_{vis} \tag{7}$$

$$\hat{f}_{vis,c}^{cp} = R_{vis,c} * \tilde{f}_{vis} \tag{8}$$

여기에서 *은 채널별 곱셈을 의미한다.

5. Feature Fusion

제안하는 M-FDTs는 그림 3와 4에서 공간 영역과 채널 영역에서 공통 및 보완 특징을 독립적으로 추출한다. 구체적으로, 두 개의 공통 특징맵 $\{\hat{f}_{vis,sp}^{cm}, \hat{f}_{vis,c}^{cm}\}$와 보완 특징맵 $\{\hat{f}_{vis,sp}^{cp}, \hat{f}_{vis,c}^{cp}\}$을 가시광선 영상에 대해 얻는다. 같은 과정을 통해 적외선 영상에서도 공통 및 보완 특징맵을 생성한다.

각 영역에서 고유한 정보가 포함될 수 있기에 공간 및 채널 분해의 특징맵 내에서 필수 정보를 전달하기 위해 1×1 convolution layer를 사용하여 각 분해 트랜스포머의 공통 및 보완 특징맵을 결합한다. 즉, 각 양식 $m \in \{vis, ir\}$에 대해 융합된 공통 및 보완 특징맵을 아래 수식과 같이 획득한다.

$$\hat{f}_{m}^{cm} = Conv^{1\times1}(\hat{f}_{m,sp}^{cm} \oplus \hat{f}_{m,c}^{cm}) \tag{9}$$

$$\hat{f}_{m}^{cp} = Conv^{1\times1}(\hat{f}_{m,sp}^{cp} \oplus \hat{f}_{m,c}^{cp}) \tag{10}$$

여기에서 $\oplus$는 concatenation을 의미한다.

6. Fusion Block

Fusion block은 분해된 공통 및 보완 특징맵을 효과적으로 합성하기 위해 self-fusion convolution (SFC)[26]를 기반으로 구성된다[20]. 먼저, 단일 1×1 convolution을 통해 공통 특징맵 $\hat{f}_{ir}^{cm}$과 $\hat{f}_{vis}^{cm}$를 합성한다. 그런 다음, 합성된 공통 특징맵과 두 개의 보완 특징맵 $\{\hat{f}_{ir}^{cp}, \hat{f}_{vis}^{cp}\}$을 SFC block의 입력으로 사용하여 최종 합성 영상인 I_{fus}^{l}을 생성한다.

7. Training

제안하는 네트워크를 훈련하기 위해 fusion loss L_{fus}[19]와 decomposition loss L_{decomp}의 합을 total loss L_{total}를 정의한다.

$$L_{total} = L_{fus} + \lambda_{decomp} L_{decomp} \tag{11}$$

입력 영상 사이의 공통 정보를 포함한 공통 특징 $\{\hat{f}_{ir}^{cm}, \hat{f}_{vis}^{cm}\}$을 정의하고, 각 공통 특징은 배경 및 대규모 환경과 같은 공유 정보를 더 많이 포함하므로 상관성이 높다고 가정한다. 대조적으로, 가시광선 영상의 질감이나 세부 정보나 적외선 영상의 열복사 정보 및 에지 정보 등 상관성이 낮은 보완 특징 $\{\hat{f}_{ir}^{cp}, \hat{f}_{vis}^{cp}\}$은 상호 보완적인 정보를 나타낸다. 학습 시 경사 하강법에서 L_{decomp}으로 인해 $CC(\hat{f}_{ir}^{cp}, \hat{f}_{vis}^{cp})$는 0에 가까워지고, $CC(\hat{f}_{ir}^{cm}, \hat{f}_{vis}^{cm})$은 더 커진다. 이러한 특성을 활용하여 아래와 같은 특징 분해를 위한 decomposition loss L_{decomp} 를 정의한다.

$$L_{decomp} = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{(CC(\hat{f}_{ir}^{cp}, \hat{f}_{vis}^{cp}))^2}{CC(\hat{f}_{ir}^{cm}, \hat{f}_{vis}^{cm}) + \epsilon}\right) \tag{12}$$

여기서 L은 레벨의 개수, $CC(\bullet, \bullet)$은 피어슨 상관 계수를 의미하며, ε는 1.01로 설정하여 학습한다.

본 논문은 KAIST 데이터셋[27]을 이용하여 학습한다. KAIST 데이터셋은 640×512 크기의 적외선 및 가시광선 영상 쌍으로 구성되어 있으며, 입력 영상을 무작위로 256×256 크기로 crop하고 grayscale로 변환 후 학습한다.

수식 (11)의 λ_{decomp}는 2로 고정한다. L_{fus}는 [20]과 동일한 세팅을 통해 학습한다.

Ⅲ. 실험 결과

본 논문에서는 학습에 사용되지 않은 무작위로 선택하여 구성된 640×512 크기를 가진 200쌍의 KAIST 데이

표 1. U2Fusion[10], ReCoNet[11], CMTFusion[20], 및 제안하는 기법의 KAIST dataset에서의 정량적 비교
Table 1. Quantitative comparison of U2Fusion[10], ReCoNet[11], CMTFusion[20], and the proposed algorithm on the KAIST dataset

	KAIST dataset					
	$Q^{AB/F}$(↑)	MS-SSIM(↑)	SCD(↑)	BRISQUE(↓)	NIQE(↓)	Avg. rank(↓)
U2Fusion[10]	0.5928	0.9584	1.5590	43.79	3.707	3.40
ReCoNet[11]	0.4260	0.8813	1.2105	40.05	3.226	3.40
CMTFusion[20]	0.6647	0.9662	1.6994	41.37	3.174	2.00
Proposed	0.6539	0.9717	1.7643	39.73	3.139	1.20

표 2. U2Fusion[10], ReCoNet[11], CMTFusion[20], 및 제안하는 기법의 TNO dataset에서의 정량적 비교
Table 2. Quantitative comparison of U2Fusion[10], ReCoNet[11], CMTFusion[20], and the proposed algorithm on the TNO dataset

	TNO dataset					
	$Q^{AB/F}$(↑)	MS-SSIM(↑)	SCD(↑)	BRISQUE(↓)	NIQE(↓)	Avg. rank(↓)
U2Fusion[10]	0.3923	0.9217	1.7026	30.61	3.929	2.60
ReCoNet[11]	0.3947	0.7153	1.5349	31.63	4.220	3.40
CMTFusion[20]	0.4810	0.9296	1.8421	26.73	4.327	1.40
Proposed	0.4566	0.8957	1.7617	28.96	5.116	2.60

터셋[27] 영상과 280×280에서 768×576 사이의 크기를 가진 20쌍의 TNO 데이터셋[28]을 이용하여 U2Fusion[10], ReCoNet[11] 및 CMTFusion[20]과 비교하여 영상 합성 성능을 평가한다.

표 1은 KAIST 데이터셋, 표 2는 TNO 데이터셋의 검증 영상에 대한 각 알고리즘의 정량적 결과를 비교한다. $Q^{AB/F}$[29], MS-SSIM[30] 및 SCD[31]는 점수가 높을수록, BRISQUE[32] 및 NIQE[33], average rank는 점수가 낮을수록 합성 성능이 우수함을 나타낸다. 제안하는 알고리즘은 KAIST, TNO 데이터셋에서 가장 높거나 두 번째로 높은 $Q^{AB/F}$, MS-SSIM 및 SCD 점수를 보여주며, 이는 제안하는 알고리즘이 입력 영상 내의 구조 및 질감 세부 사항을 잘 보존함을 의미한다. 또한, 제안하는 알고리즘의 BRISQUE 및 NIQE 점수와 알고리즘의 전반적인 성능을 평가하는 average rank는 가장 높거나 기존 알고리즘과 유사하다. 이는 제안하는 알고리즘이 적외선 및 가시광선 영상 간의 유의미한 공통 및 상호 보완 정보를 충실히 추출하고 이를 효과적으로 합성함을 의미한다. 또한, 제안하는 알고리즘은 모든 지표에 대해 유사한 경향성을 보이므로 일반화 능력이 우수함을 확인한다.

그림 5는 KAIST 데이터셋(1, 3행) 및 TNO 데이터셋(5, 7행)을 이용하여 각 알고리즘을 통해 얻은 합성 결과를 정성적으로 비교한다. 그림 5(c)-(d)의 U2Fusion 및 ReCoNet은 아티팩트와 노이즈를 생성하여 합성 성능이 저하되고, 세부 정보가 손실되어 흐릿한 결과 영상을 생성한다. 그림 5(e)의 CMTFusion은 입력 영상의 보완 특징을 보존하여 더 높은 품질의 영상을 생성하지만, 입력 영상 간 공통 정보를 삭제하기 때문에 정보 손실이 발생한다. 반면, 그림 5(f)의 제안하는 알고리즘은 입력 영상에서 공통 및 보완 정보를 모두 효과적으로 보존하여 가장 우수한 영상 합성 성능을 보임을 확인할 수 있다.

마지막으로, 본 연구에서 제안하는 특징 분해 트랜스포머 내 공간 및 채널 분해 모듈 기여도를 분석하기 위한 ablation study를 진행하였다. 표 3은 본 연구에서 진행한 각 모듈별 효과를 비교한다. 공간 분해 트랜스포머 또는 채널 분해 트랜스포머를 사용하면 해당 공간에서 전역 상호작용을 캡처할 수 있어 합성 성능을 향상시킨다. 공간 분해 트랜스포머와 채널 분해 트랜스포머를 모두 사용하면 각

그림 5. 합성 결과 영상 2, 4, 6, 8행은 KAIST 데이터셋(1, 3행) 및 TNO 데이터셋(5, 7행)의 빨간색 상자를 확대한 영상을 나타낸다
(a) 적외선 영상 (b) 가시광선 영상 (c) U2Fusion[10] (d) ReCoNet[11] (e) CMTFusion[20] (f) 제안하는 기법
Fig. 5. Comparison of fusion results and their magnified parts (a) Infrared images (b) visible images (c) U2Fusion[10] (d) ReCoNet[11] (e) CMTFusion[20] (f) the proposed algorithm

공간에서 공통 및 보완 정보를 보다 효과적으로 추출하여 합성 성능이 더욱 향상됨을 확인할 수 있다.

표 3. 서로 다른 모듈의 효과 분석
Table 3. Impacts of different combinations of networks

Spatial	Channel	Avg. rank
√		2.10
	√	2.31
√	√	1.82

Ⅳ. 결 론

본 논문은 적외선 및 가시광선 영상 합성을 위한 멀티모달 특징 분해 기반 트랜스포머를 제안하였다. 제안하는 네트워크는 멀티모달 영상의 공통 및 세부 정보 특징 분해 기반 트랜스포머를 통해 각 영상에 대한 공통 및 상호 보완적인 특징맵을 추출한다. 또한, 공통 특징맵은 서로 상관관계가 크고 보완 특징맵은 서로 상관관계가 적다는 가정을 기반으로, 분해 손실 함수를 개발하였다. 마지막으로, 공통

특징맵과 보완 특징맵을 결합하여 합성 영상을 생성하였다. KAIST 데이터셋 및 TNO 데이터셋을 이용한 모의실험을 통해서 제안하는 기법이 기존의 기법들에 비해서 영상 합성을 효과적으로 할 수 있음을 확인한다.

참 고 문 헌 (References)

[1] X. Zhang and Y. Demiris, “Visible and infrared image fusion using deep learning,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. 45, pp. 10535 – 10554, Mar. 2023.
doi: https://doi.org/10.1109/TPAMI.2023.3261282

[2] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, “CDDFuse: Correlationdriven dual-branch feature decomposition for multimodality image fusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5906 – 5916, June 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00572

[3] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, “CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers,” IEEE Trans. Intell. Transp. Syst., Vol. 24, No. 12, pp. 14679 – 14694, Dec. 2023.
doi: https://doi.org/10.1109/TITS.2023.3300537

[4] Z. Xie, F. Shao, G. Chen, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, “Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection,” IEEE Trans. Circuit Syst. Video Technol., Vol. 33, No. 8, pp. 4149 – 4163, Aug. 2023.
doi: https://doi.org/10.1109/TCSVT.2023.3241196

[5] B. Tang, Z. Liu, Y. Tan, and Q. He, “HRTransNet: HRFormer-driven two-modality salient object detection,” IEEE Trans. Circuit Syst. Video Technol., Vol. 33, No. 2, pp. 728 – 742, Feb. 2023.
doi: https://doi.org/10.1109/TCSVT.2022.3202563

[6] S. Park and C. Lee, “Multiscale progressive fusion of infrared and visible images,” IEEE Access, vol. 10, pp. 126 117 – 126 132, 2022.

[7] H. Yan, J.-X. Zhang, and X. Zhang, “Injected infrared and visible image fusion via L1 decomposition model and guided filtering,” IEEE Trans. Comput. Imag., Vol. 8, pp. 162 – 173, 2022.
doi: https://doi.org/10.1109/TCI.2022.3151472

[8] D. Wang, J. Liu, X. Fan, and R. Liu, “Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration,” in Proc. Int. Joint Conf. Artif. Intell., pp. 3508 – 3515, July 2022.
doi: https://doi.org/10.24963/ijcai.2022/487

[9] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, “UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion,” IEEE Trans. Circuit Syst. Video Technol., Vol. 32, No. 6, pp. 3360 – 3374, June 2022.
doi: https://doi.org/10.1109/tcsvt.2021.3109895

[10] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, “U2Fusion: A unified unsupervised image fusion network,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. 44, No. 1, pp. 502 – 518, Jan. 2022.
doi: https://doi.org/10.1109/TPAMI.2020.3012548

[11] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, “ReCoNet: Recurrent correction network for fast and efficient multi-modality image fusion,” in Proc. European Conf. Comput. Vis., pp. 539 – 555, Oct. 2022.
doi: https://doi.org/10.1007/978-3-031-19797-0_31

[12] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, “Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5792 – 5801, June 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00571

[13] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, “Infrared and visible image fusion via texture conditional generative adversarial network,” IEEE Trans. Circuit Syst. Video Technol., Vol. 31, No. 12, pp. 4771 – 4783, Dec. 2021.
doi: https://doi.org/10.1109/TCSVT.2021.3054584

[14] H. Zhang, J. Yuan, X. Tian, and J. Ma, “GAN-FM: Infrared and visible image fusion using GAN with fullscale skip connection and dual Markovian discriminators,” IEEE Trans. Comput. Imag., Vol. 7, pp. 1134 – 1147, 2021.
doi: https://doi.org/10.1109/TCI.2021.3119954

[15] Y. Gao, S. Ma, and J. Liu, “DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion,” IEEE Trans. Circuit Syst. Video Technol., Vol. 33, No. 2, pp. 549 – 561, Feb. 2023.
doi: https://doi.org/10.1109/TCSVT.2022.3206807

[16] Q. Dong, C. Cao, and Y. Fu, “Incremental transformer structure enhanced image inpainting with masking positional encoding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 11348 – 11358, June 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01107

[17] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent., May 2021.

[18] W. Tang, F. He, and Y. Liu, “YDTR: Infrared and visible image fusion via Y-shape dynamic transformer,” IEEE Trans. Multimedia, Vol. 25, pp. 5413 – 5428, 2023.
doi: https://doi.org/10.1109/TMM.2022.3192661

[19] Z. Chang, Z. Feng, S. Yang, and Q. Gao, “AFT: Adaptive fusion transformer for visible and infrared images,” IEEE Trans. Image Process., Vol. 32, pp. 2077 – 2092, 2023.
doi: https://doi.org/10.1109/TIP.2023.3263113

[20] S. Park, A. G. Vien, and C. Lee, “Cross-modal transformers for infrared and visible image fusion,” IEEE Trans. Circuit Syst. Video Technol., Vol. 34, No. 2, pp. 770 – 785, Feb. 2024.
doi: https://doi.org/10.1109/TCSVT.2023.3289170

[21] J. Liu, J. Tang, and G. Wu, “Residual feature distillation network for lightweight image super-resolution,” in Proc. European Conf. Comput. Vis. Workshops, pp. 41 – 55, Aug. 2020.

[22] V. Vs, J. M. J. Valanarasu, P. Oza, and V. M. Patel, “Image fusion

transformer," in Proc. IEEE Int. Conf. Image Process., pp. 3566–3570, Oct. 2022.
doi: https://doi.org/10.1109/ICIP46576.2022.9897280

[23] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," IEEE Trans. Intell. Transp. Syst., vol. 24, no. 12, pp. 14679–14694, Dec. 2023.
doi: https://doi.org/10.1109/tits.2023.3300537.

[24] Z. Xie, F. Shao, G. Chen, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, "Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection," IEEE Trans. Circuits Syst. Video Technol., 2023.
doi: https://doi.org/10.1109/tcsvt.2023.3241196

[25] B. Tang, Z. Liu, Y. Tan, and Q. He, "HRTransNet: HRFormer-driven two-modality salient object detection," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 2, pp. 728–742, Feb. 2023.
doi: https://doi.org/10.1109/tcsvt.2022.3202563

[26] S. Gong, S. Zhang, J. Yang, and P. C. Yuen, "Self-fusion convolutional neural networks," Pattern Recognit. Lett., Vol. 152, pp. 50–55, Dec. 2021.
doi: https://doi.org/10.1016/j.patrec.2021.08.022

[27] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1037–1045, June 2015.
doi: https://doi.org/10.1109/cvpr.2015.7298706

[28] A. Toet, "TNO image fusion dataset," 2014, https: //figshare.com/articles/dataset/TNO_ Image_Fusion_Dataset/1008029.

[29] C. Xydeas and V. Petrovic, "Objective image fusion performance measure," Electron. Lett., Vol. 36, No. 4, pp. 308–309, 2000.
doi: https://doi.org/10.1049/el:20000267

[30] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," IEEE Trans. Image Process., Vol. 24, No. 11, pp. 3345–3356, Nov. 2015.
doi: https://doi.org/10.1109/TIP.2015.2442920

[31] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," AEU-Int. J. Electron. Commun., Vol. 69, no. 12, pp. 1890–1896, Dec. 2015.
doi: https://doi.org/10.1016/j.aeue.2015.09.004

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process., Vol. 21, No. 12, pp. 4695–4708, Dec. 2012.
doi: https://doi.org/10.1109/TIP.2012.2214050

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," IEEE Signal Process. Lett., Vol. 20, No. 3, pp. 209–212, Mar. 2013.
doi: https://doi.org/10.1109/LSP.2012.2227726

저 자 소 개

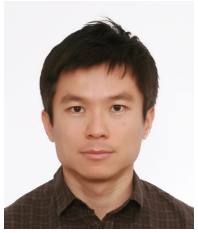
김 가 현

- 2023년 : 동국대학교 멀티미디어공학과 공학사
- 2023년 ~ 현재 : 동국대학교 컴퓨터·AI학과 석사과정
- ORCID : https://orcid.org/0009-0001-6527-1413
- 주관심분야 : RGB/IR 영상 합성

원 해 양

- 2015년 : B.S. in Math and Computer Science, University of Science, Vietnam
- 2018년 : 전남대학교 컴퓨터공학과 공학석사
- 2022년 ~ 현재 : 동국대학교 멀티미디어공학과 박사과정
- ORCID : https://https://orcid.org/0000-0002-3068-4728
- 주관심분야 : 영상처리

저 자 소 개



이 철

- 2003년 : 고려대학교 전기전자전파공학부 공학사
- 2008년 : 고려대학교 전자전기공학과 공학석사
- 2013년 : 고려대학교 전자전기공학과 공학박사
- 2002년 ~ 2006년 : ㈜바이오스페이스 (현 ㈜인바디)
- 2013년 ~ 2014년 : Postdoctoral Scholar, Pennsylvania State University
- 2014년 ~ 2015년 : Research Scientist, The University of Hong Kong
- 2015년 ~ 2019년 : 부경대학교 컴퓨터공학과 조교수
- 2019년 ~ 현재 : 동국대학교 AI소프트웨어융합학부 부교수
- ORCID : https://orcid.org/0000-0001-9329-7365
- 주관심분야 : 영상처리, 계산사진학, 컴퓨터비전