GEO DATA
eISSN 2713-5004 • https://geodata.kr

DATA ARTICLE
September 2024 • Vol. 6. No. 3, pp. 175-185
https://doi.org/10.22761/GD.2023.0049

딥러닝 기반 정지궤도 환경위성 황사 탐지 데이터셋

유진우[1,2], 박채원[3,4], 이원진[5], 이용미[6], 김유하[7], 정형섭[8,9,*]

[1]석박사통합과정생, 서울시립대학교 공간정보공학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국
[2]석박사통합과정생, 서울시립대학교 스마트시티학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국
[3]석사, 서울시립대학교 공간정보공학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국
[4]석사, 서울시립대학교 스마트시티학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국
[5]연구관, 국립환경과학원 환경위성센터, 인천광역시 서구 환경로 42, 22689, 대한민국
[6]연구사, 국립환경과학원 환경위성센터, 인천광역시 서구 환경로 42, 22689, 대한민국
[7]연구원, 국립환경과학원 환경위성센터, 인천광역시 서구 환경로 42, 22689, 대한민국
[8]교수, 서울시립대학교 공간정보학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국
[9]교수, 서울시립대학교 스마트시티학과, 서울특별시 동대문구 서울시립대로 163, 02504, 대한민국

Dataset for Deep Learning-based GEMS Asian Dust Detection

Jin-Woo Yu[1,2], Che-Won Park[3,4], Won-Jin Lee[5], Yong-Mi Lee[6], Yu-Ha Kim[7], and Hyung-Sup Jung[8,9,*]

[1]Integrated Master and PhD Student, Department of Geoinformatics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea
[2]Integrated Master and PhD Student, Department of Smart Cities, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea
[3]Master, Department of Geoinformatics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea
[4]Master, Department of Smart Cities, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea
[5]Senior researcher, Environmental Satellite Center, National Institute of Environmental Research, 42 Hwangyeong-ro, Seo-gu, 22689 Incheon, South Korea
[6]Researcher, Environmental Satellite Center, National Institute of Environmental Research, 42 Hwangyeong-ro, Seo-gu, 22689 Incheon, South Korea
[7]Research Official, Environmental Satellite Center, National Institute of Environmental Research, 42 Hwangyeong-ro, Seo-gu, 22689 Incheon, South Korea
[8]Professor, Department of Geoinformatics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea
[9]Professor, Department of Smart Cities, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, 02504 Seoul, South Korea

Received November 27, 2023
Revised June 14, 2024
Accepted July 15, 2024

Corresponding Author
Hyung-Sup Jung
Tel: +82-2-6490-2892
E-mail: hsjung@uos.ac.kr

In South Korea, Asian dust frequently occurs during the spring, causing various health issues, including respiratory diseases. Consequently, public awareness and concern about air pollutants have increased, leading to demands for improved air quality and accurate forecasting. To meet these demands, the Ministry of Environment has deployed the Geostationary Environment Monitoring Spectrometer (GEMS) on the GK2B satellite to monitor atmospheric pollutants and climate change-inducing substances in real-time. The current GEMS dust product, generated using thresholds of the UV-aerosol index and visible-aerosol index, has shown limitations in accurately detecting suspended particulate matter. This study aims to develop a comprehensive AI dataset for improving GEMS Asian dust detection. Data were collected from January to May 2021, focusing on dates with significant dust events. Label data were meticulously generated through annotations based on outputs from various satellites and ground-based observations. Subsequent data preprocessing and augmentation techniques, including normalization and cut-mix, were applied to enhance the dataset's robustness and generalizability. To evaluate the dataset, model training was conducted. The results predicted by the model showed improvements over the detection results of existing algorithms. Future datasets will be developed with improved labeling methods and accuracy verification techniques. These dataset improvements are expected to contribute to the development of deep learning models with superior predictive performance compared to current dust detection algorithms.

Keywords: GEMS; Artificial intelligence; Segmentation; Asian dust; Dataset

1. 서론

봄철 대한민국에서는 황사가 빈번하게 발생하며 이는 호흡기 질환을 비롯한 다양한 건강 문제의 주요 원인으로 작용하고 있다(Kim and Lee, 2009). 이러한 이유로 국민들의 대기 오염 물질에 대한 관심과 인식이 증가하고 있으며 이는 정확한 대기질 개선 및 예경보에 대한 요구로 이어지고 있다. 이러한 요구에 대응하기 위하여 환경부는 대기 환경 및 기후 변화 유발 물질을 실시간으로 모니터링하기 위해서 정지궤도 복합위성 2B호(GEO-KOMPSAT-2B, GK2B)에 환경탑재체(Geostationary Environment Monitoring Spectrometer, GEMS)를 탑재하였다(Kim et al., 2020).

GEMS는 대기 오염 물질의 발생과 이동을 지속적으로 모니터링하여 대기질 개선을 위한 중요한 산출물들을 제공하고 있다. 특히 GEMS 산출물 중 에어로졸 유형에는 황사 분류 산출물이 포함되어 있으며 이는 자외선 에어로졸 지수(ultraviolet-aerosol index, UVAI)와 가시광 에어로졸 지수(visible-aerosol index, VISAI)의 임계값을 활용하여 탐지한다(Cho et al., 2023). 기존 GEMS 황사 산출물(version 1)은 임계값이 UVAI 0, VISAI 0.2로 설정되어 있었다. 해당 임계값에서는 실제 황사 발생량보다 적게 예측되는 문제가 있다. 이를 개선하기 위해 2023년 이후의 version 2 산출물은 UVAI와 VISAI의 임계값을 각각 0.2, -0.3으로 조정하여 더 많은 양의 황사를 예측할 수 있게 되었다(Jung et al., 2023). 그러나 version 2의 산출물에도 특정 영역에서는 여전히 황사를 정확하게 산출하지 못하거나 황사 발원지 인근에서 과대 예측되는 문제가 남아 있어 이에 대한 개선이 필요하다.

GEMS의 황사 오탐지 문제를 해결하기 위해서는 기존 알고리즘 기반의 방식이 아닌 새로운 접근 방식이 필요하다. 최근 이미지 내 관심 영역을 탐지하는 최신 알고리즘들이 딥러닝을 기반으로 구현되고 있으며 기존 비딥러닝 기반 알고리즘의 성능을 능가하고 있다(Wang et al., 2017). 특히 딥러닝 기법을 해외 환경위성 영상에 적용한 연구에서는 기존 알고리즘 기반 산출물보다 우수한 성능을 보이고 있다(Bandara, 2022; Ghahremanloo et al., 2023). 이러한 성과는 딥러닝 기법이 GEMS의 황사 산출물 개선에 효과적일 수 있음을 시사한다.

따라서 GEMS의 황사 산출물을 개선하기 위해서는 딥러닝 기법의 적용이 필요하다. 딥러닝 모델 학습을 위해서는 대량의 규격화된 데이터가 필요하나 현재 딥러닝 기반 GEMS 황사 탐지를 위한 데이터셋은 구축되어 있지 않다. 이러한 배경을 바탕으로 본 연구에서는 딥러닝 기반 정지궤도 환경 위성의 황사 탐지 데이터셋을 제작하였다. 구체적으로 살펴보면 2021년 1월부터 2021년 5월까지 황사가 발생한 날짜의 영상 129장을 수집하였고 결측 지역 마스킹, 99% 최대-최소 정규화, 데이터 증강 기법을 적용하여 데이터셋을 제작하였다. 또한 라벨 데이터는 라벨링을 위한 참고 자료 선정 및 가이드라인 제작, 라벨링 수행 및 검수 과정을 거쳐 정밀하게 제작되었다. 이를 통해 딥러닝 기반 GEMS 황사 탐지를 위한 AI 학습 및 평가 데이터셋을 구축하였다.

2. 본론

2.1 입력 데이터 취득

황사 탐지를 위한 AI 학습 데이터셋을 구축하기 위해 GEMS level 2 자료를 활용하였다(Choi et al., 2023). GEMS 자료는 2021년 1월부터 5월까지의 모든 날짜의 영상 중 황사 일기도 및 기존 GEMS 황사 산출물에 황사가 탐지된 21일자의 영상 129장을 선정하였다. 황사가 발생한 일자는 Table 1과 같다.

Fig. 1은 연구 지역을 나타낸다. 연구 지역은 황사의 발생부터 한반도를 관통하여 동해안으로 나가는 전반적인 황사의 공간적 분포를 체계적으로 확인하기 위해 주요 발원지인 중국의 고비 사막과 몽골의 타클라마칸 사막을 포함하여 위도 30-45°, 경도 100-132°에 이르는 영역을 연구 지역으로 선정하였다. 발원지에서부터 최종 도달지에 이르는 경로를 포괄함으로써 황사의 발원, 이동 그리고 최종 영향을 파악할 수 있는 기반을 마련하고자 하였다.

입력 데이터는 GEMS의 level 2 영상 자료인 normalized radiance (NR) 자료와 UVAI 영상과 VISAI 영상을 선정하였다. NR은 354, 388, 412, 443, 477, 490 nm의 분광대역을 정규화한 영상으로 에어로졸 광학 두께(aerosol optical depth, AOD) 및 에어로졸 지수를 산출하는 데 활용된다(Yang et al., 2023). NR을 구하는 수식은 Eq. 1과 같다. 수식에서 I와 E, λ는 각각 radiance, irradiance, wavelength (354, 388, 412, 443, 477, 550 nm)를 나타낸다.

$$N_\lambda = \frac{I_\lambda}{E_\lambda} \quad \text{Eq. 1}$$

VISAI와 UVAI는 에어로졸의 흡수율과 입자의 크기를 평가하는 지표로 두 지수의 임계값을 활용하여 황사를 분류하고 있다. UVAI는 에어로졸의 흡수성을 보여주는 정보를 나타내는 지표로 에어로졸 입자의 광학적 흡수 특성이 강할수록 양의 방향으로 값이 커진다. VISAI는 에어로졸의 크기에 대한 정보를 나타내며 입자 크기가 커질수록 양의 방향으로 값이 증가한다(Go et al., 2020). 에어로졸 지수는 NR 값을 통해 계산되며 계산 수식은 Eq. 2와 같다. 수식에서 $N_{\lambda 1}$과 $N_{\lambda 2}$는 NR 값을 의미하며 VISAI에서는 477, 490 nm의 값을 사용하고 UVAI에는 354, 388 nm의 값을 사용한다. *means*와 *calc*는 각각 측정된 복사량과 계산된 복사량을 의미하며 *LER*은 Rayleigh 산란을 보정하여 추정된 지표 반사율을 나타낸다.

Table 1. List of dates selected for the dataset for Asian dust detection

Date of occurrence of Asian dust
January 12, 2021
January 13, 2021
January 14, 2021
January 15, 2021
March 27, 2021
March 28, 2021
March 29, 2021
March 30, 2021
April 15, 2021
April 16, 2021
April 17, 2021
April 26, 2021
April 27, 2021
April 28, 2021
April 29, 2021
May 6, 2021
May 7, 2021
May 8, 2021
May 22, 2021
May 23, 2021
May 24, 2021

$$AI = -100\left[log\left(\frac{N_{\lambda_1}}{N_{\lambda_2}}\right)_{means} - log\left(\frac{N_{\lambda_1}(LER_{\lambda_1})}{N_{\lambda_2}(LER_{\lambda_2})}\right)_{calc}\right] \quad \text{Eq. 2}$$

2.2 라벨 데이터 제작

라벨 데이터를 제작하기 위하여 기존 GEMS의 황사 산출자료, 타 위성의 황사 산출 자료, 지상 관측 자료를 활용하여 라벨링을 수행하였다. Fig. 2는 라벨 데이터 제작 과정을 나타낸다. 총 129장의 라벨 데이터를 제작하였으며 이 과정은 라벨링을 위한 기반 데이터 선정, 신뢰성 있는 자료 제작을 위한 라벨링 가이드라인 제작, GIS 프로그램(QGIS; ArcGIS; ESRI, Redlands, CA, USA)을 통한 라벨링 수행, 데이터의 일관성 유지를 위한 교차 검수 과정을 포함하였다. 황사 라벨은 위성 기반 자료로 라벨링된 데이터와 지상 관측 자료로 라벨링된 데이터를 통합하여 하나의 라벨로 제작되었다.

라벨링을 위한 참고 자료로는 GEMS의 황사 산출물 자료, advanced meteorological imager (AMI)의 황사 산출물 자료(aerosol detection products, ADPS), AMI의 Ash RGB Pink 자료를 선정하였으며, 지상 관측 자료로는 한국 및 중국의 지상 PM10 자료와 community multiscale air quality model (CMAQ) 자료를 선정하였다. 총 5가지의 자료를 활용하여 라벨링 기준을 설정하였다. 이렇게 라벨 데이터의 신뢰도를 확보하기 위해 다양한 자료를 사용하였지만 이들 자료는 각각 불확도를 포함할 수 있기 때문에 황사 영역에

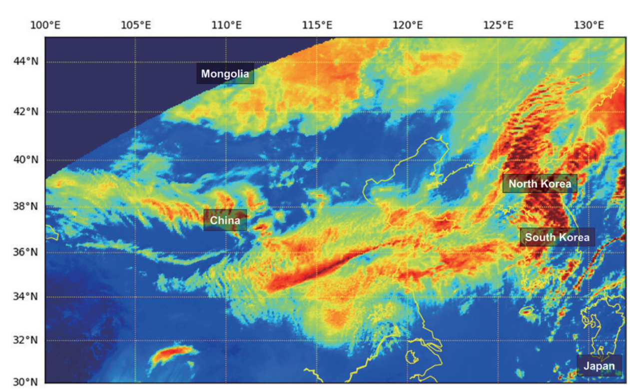

Fig. 1. Study area covering latitudes 30° to 45° and longitudes 100° to 132°, including the Gobi and Taklamakan Deserts. This region tracks dust storm paths from origin to the East Sea in Korea.

대한 신뢰도를 제공하기 어려운 측면이 있다. 실제로 황사에 대한 참값은 실측 지점에서 포인트 단위로 관측되므로 면 단위의 정답값을 얻기 위해서는 위성 알고리즘 및 모델 자료를 활용할 수밖에 없다. 이러한 자료의 부정확성을 최소화하기 위해 다양한 출처의 데이터를 종합적으로 참고하여 라벨링을 수행하였다.

위성 기반 자료를 활용한 라벨링은 GEMS AOD, AMI ADPS, AMI Dust RGB 자료 간 공통으로 탐지된 영역을 기준으로 수행되었다. 각 자료에 포함될 수 있는 불확도에 의한 영향을 줄이기 위해 세 개의 자료 중 두 개 이상에서 공통으로 황사가 탐지된 영역을 라벨링 기준으로 삼아 황사 라벨링의 정확도를 높이고자 하였다. 지상 관측 자료 기반 라벨링은 한국과 중국의 지상 관측 자료와 CMAQ PM10 예측 자료를 기반으로 수행되었다. 지상 관측 자료는 황사의 참값으로 볼 수 있는 자료이기 때문에 지상 관측 지점에 황사가 포함된 경우 이 포인트가 반드시 포함되게 하여 라벨링을 수행하였다.

GIS 소프트웨어를 사용하여 정해진 기준에 따라 정밀한 라벨링을 수행하였으며 이 과정에서 교차 검수를 통해 라벨링의 일관성을 높였다. 최종적으로 위성 기반 라벨링 자료와 지상 기반 라벨링 자료를 통합하여 라벨 데이터를 도출하였다. 이러한 접근은 다양한 출처의 자료를 종합적으로 활용함으로써 라벨 데이터의 신뢰성과 정확성을 최대한 확보하고자 한 것이다. 이와 같은 과정을 통해 제작된 라벨 데이터는 각 자료의 불확도를 고려하여 신뢰도를 높이기 위해 다각적인 검토와 교차 검수를 거쳤으며 실질적인 황사 탐지 지역에 대한 공간 라벨 자료로 활용될 수 있는 기준을 마련하였다.

2.3 데이터 전처리 및 증강

Fig. 3은 입력 데이터의 전처리 및 데이터 증강의 흐름도를 나타낸다. 해당 과정에는 null space 제거, 학습-평가 데이터 분할, 회전 및 반전 증강, cut-mix 증강이 포함되어 있다.

GEMS는 촬영 각도를 조금씩 틀어서 영상을 촬영하기 때문에 특정 시간대에 촬영된 영상은 좌측 혹은 우측 상단 부분에 null space가 포함될 수 있다. 이러한 null space는 모델 학습에 악영향을 미칠 수 있기 때문에 masking out을 수행하였다. Masking out은 VISAI와 UVAI 자료에 포함된 null value의 위치 정보를 기반으로 mask 자료를 생성하였으며 이를 8개의 입력 자료 모두에 적용하여 null space를

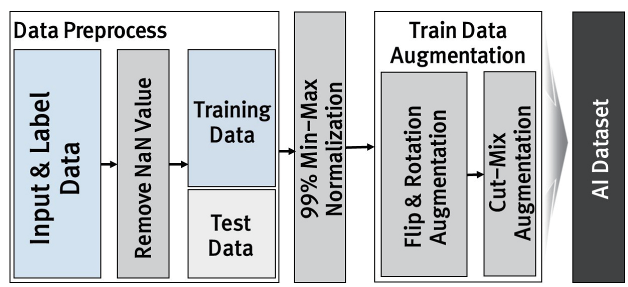

Fig. 3. Data pre-processing flow. The process includes null space elimination, training-test data split, rotation and inversion augmentation, and cut-mix augmentation.

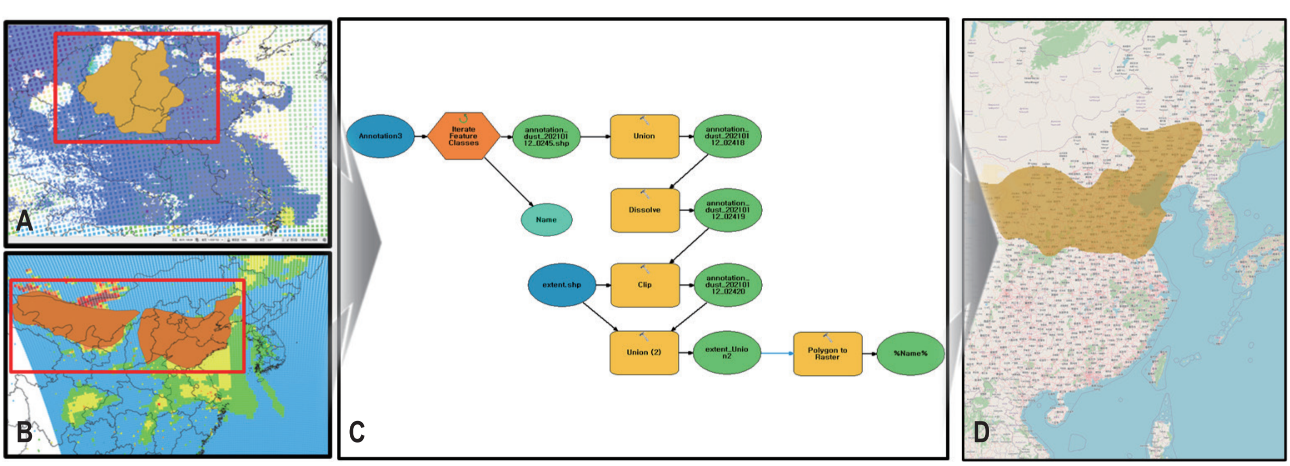

Fig. 2. Label data creation process. (A) Satellite-based annotation. (B) Ground-based annotation. (C) Annotation data integration. (D) Generated label data.

제거하였다. 이후 129장의 데이터를 모델 학습과 평가의 용도로 활용하기 위해 데이터를 약 8:2의 비율로 겹치지 않게 나누어 학습 데이터 100장, 평가 데이터 29장을 구축하였다. 8개의 입력 데이터는 각 데이터마다 값의 범위와 단위가 다르므로 딥러닝 기법 성능을 높이기 위해서는 데이터 값의 범위와 단위를 맞춰주는 과정이 필요하다. 또한 위성 영상 촬영 시 주변 밝기에 따른 센서 노이즈, 원본 이미지 압축 과정에서의 노이즈, 전송 오류로 인한 노이즈 등이 발생하게 되는데 이는 알고리즘 정확도 저하의 원인이 될 수 있다(Yu et al., 2022). 따라서 값의 단위와 범위를 일정하게 맞춰주고 영상 내에 포함되어 있는 이상치 값을 제거하기 위해 픽셀값을 크기별로 정렬한 후 0.5%와 99.5%에 해당하는 값을 제거하고 정규화를 수행하는 99% 최대-최소 정규화를 수행하였다. 99% 최대-정규화의 수식은 Eq. 3과 같다.

$$x' = \frac{x - min_{0.99}(x)}{max_{0.99}(x) - min_{0.99}(x)} \quad \text{Eq. 3}$$

딥러닝 학습 과정에서 훈련 데이터셋이 부족할 경우 학습된 모델이 검증 및 평가 데이터를 잘 설명하지 못하는 문제점이 발생한다. 과적합 문제를 방지하고 모델의 일반성 및 강건성을 확보하기 위해 학습 데이터에 데이터 증강 기법을 적용하였다(Shorten and Khoshgoftaar, 2019). 연구에 사용된 증강 기법은 데이터 회전 및 반전, random cut-mix 증강 기법을 적용하였고 이를 통해 학습 데이터의 양을 10배 증가시켜 1,000장의 학습 데이터를 구축하였다. 제작된 학습 데이터와 평가 데이터는 5×5 km로 격자화를 수행하여 640×300 pixel로 같은 크기로 변환하였으며 제작된 데이터는 일정한 규격으로 변환하여 데이터의 품질이 일관적으로 유지되게 하였다. Table 2는 설계된 학습 데이터의 규격을 나타낸다. 구축된 데이터셋은 Table 2의 규격에 맞춰 제공되며 입력 데이터 8종(NR 1-6, UVAI, VISAI)과 라벨 데이터로 이루어져 있다. 데이터 배포 및 사용의 효율성을 위해 파이썬(Python Software Foundation, Wilmington, DE, USA)의 바이너리 데이터 저장 형식인 pkl 데이터로 제공되며 데이터 증강 전후의 데이터셋이 모두 제공된다. 제작된 데이터셋에 대한 보다 자세한 내용은 아래에 기술하였다.

3. 결과

Fig. 4는 데이터 증강을 통해 생성된 학습 데이터의 예시를 보여준다. Fig. 4A-F는 NR 1-6을 나타내며 Fig. 4G는 UVAI, Fig. 4H는 VISAI를 나타내고 Fig. 4I는 라벨 데이터를 나타낸다. Fig. 4의 (1)-(4)는 각각 다른 학습 데이터의 예시를 보여준다. (1)은 회전 혹은 반전만 적용되어 있는 영상이며 이는 원본 데이터와 유사한 형태를 가진다. (2)-(4)는 회전 및 반전이 적용된 영상들의 일부를 랜덤하게 잘라내어 cut-

Table 2. Generated data specifications

Data	Data size	Data type	Resolution	Area
Input data				
Norm radiance 1	Width: 640 pixel	pkl	5 km	Latitude: 30°-45°
Norm radiance 2	Height: 300 pixel			Longitude: 100°-132°
Norm radiance 3	Channel: 8 channel			
Norm radiance 4				
Norm radiance 5				
Norm radiance 6				
GEMS UVAI				
GEMS VISAI				
Label data				
Labeled data about Asian dust	Width: 640 pixel			
	Height: 300 pixel			
	Channel: 1 channel			

GEMS, Geostationary Environment Monitoring Spectrometer; UVAI, ultraviolet-aerosol index; VISAI, visible-aerosol index.

mix 증강 기법을 적용한 결과이다. 이러한 데이터 증강 기법은 모델의 일반화 성능을 향상시키기 위해 사용되며 다양한 변형된 입력 데이터로 모델을 학습시켜 실제 환경에서 발생할 수 있는 다양한 형태의 황사 패턴을 보다 효과적으로 인식하고 예측할 수 있게 한다. 특히 cut-mix 증강 기법은 여러 이미지를 혼합하여 새로운 학습 샘플을 생성함으로써 모델이 특정 패턴에 과적합되지 않도록 하며 더 넓은 범위의 데이터 분포를 학습할 수 있도록 하였다.

Fig. 5는 평가 데이터셋의 입력 데이터의 예시로 2021년 1월 14일의 universal time coordinated (UTC) 4시 45분의 GEMS 데이터를 나타낸다. Fig. 5A-F는 전처리된 NR 영상들을 나타내며 Fig. 5G는 UVAI, Fig. 5H는 VISAI를 나타낸다. Fig. 5에서 각 영상들의 우측 상단은 전처리를 통해 null value 값이 masking되어 있는 것을 확인할 수 있다. 또

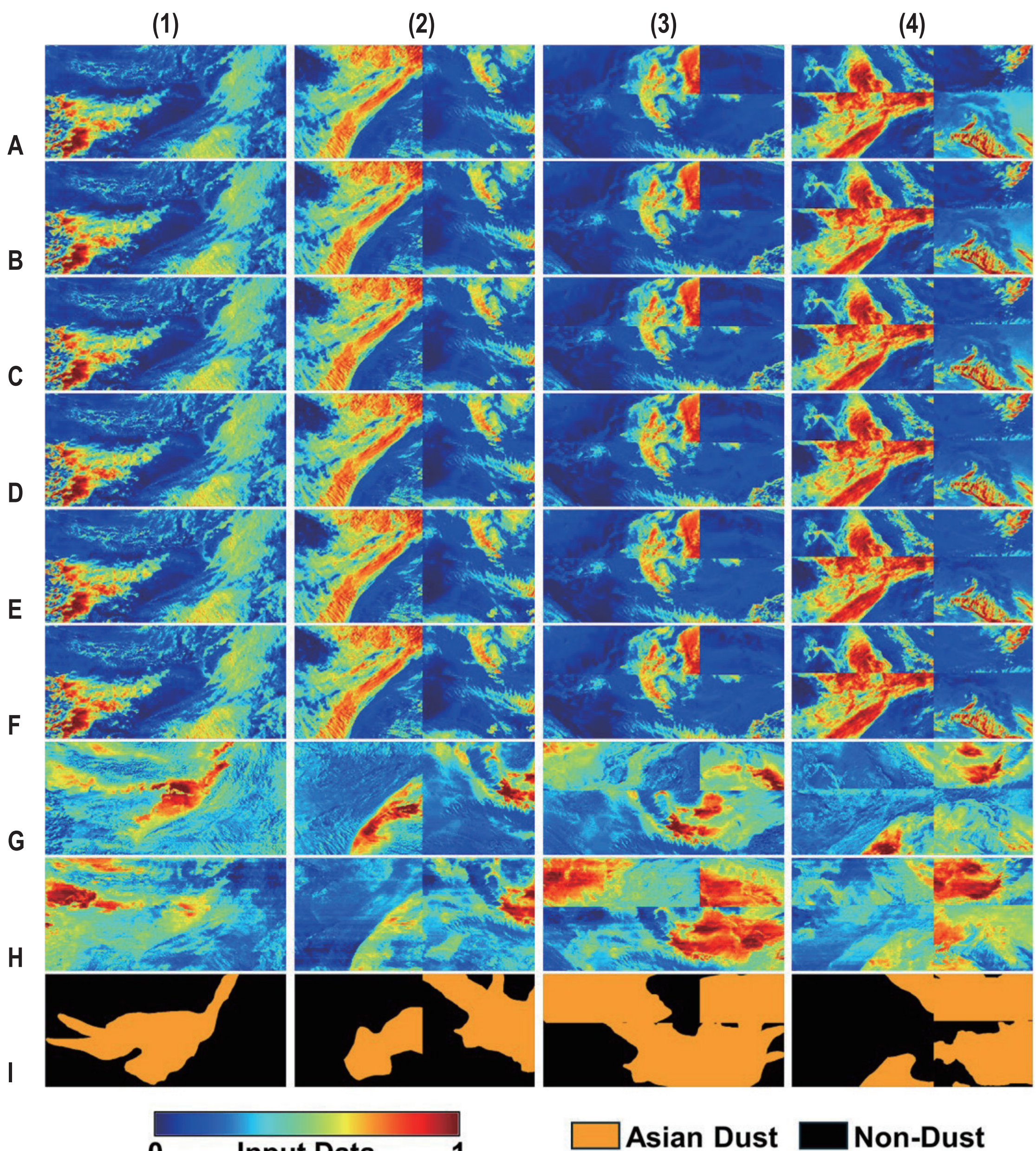

Fig. 4. Examples of augmented training data. (A-F) show NR 1 to 6, (G) the UV aerosol Index, (H) the visible aerosol index, and (I) the label data. (1)-(4) show different training data examples. (1) includes images with only rotation or flipping. (2)-(4) use cut-mix augmentation on randomly cut parts of rotated and flipped images. NR, normalized radiance; UV, ultraviolet.

한 NR 영상은 동일 시간에 다른 밴드 대역을 사용해 영상을 촬영하여 서로 유사한 형태를 띄고 있으나 약간의 값의 차이가 존재하였다. 해당 자료에서 구름 영역은 값이 높게 나타나고 있으나 황사 영역은 육안으로 확인하기 어려웠다. 이에 비해 VISAI와 UVAI에서는 NR 영상에 비해 황사를 좀 더 육안으로 확인하기 쉬웠다.

제작된 학습 데이터셋이 황사 탐지에 있어 유의미한지 확인하기 위해 이미지 객체 분할에서 일반적으로 활용되고 있는 U-NET에 적용하여 모델 학습을 수행하였다. U-NET 모델 학습에 있어 optimizer는 Adam, learning rate는 0.0001, loss function은 cross entropy, batch size는 10, epoch는 500으로 hyper parameter를 설정하였다. Fig. 6은 위성 산출물과 제작된 라벨 데이터, U-NET 모델을 통해 추론한 결과의 예시를 나타낸다. Fig. 6A는 2021년 1월 14일 UTC 4시 45분의 GEMS 황사 산출물을 나타내며, Fig. 6B는 이와 5분 차이나는 AMI 황사 산출물을 나타내고, Fig. 6C는 라벨링을 통해 제작된 라벨 데이터, Fig. 6D는 U-NET 모델이 예측한 황사 탐지 결과를 나타낸다. 기존 GEMS의 황사 산출물과 AMI의 황사 산출물을 비교해 보았을 때 GEMS의 황사 산출물이 level 2로 업데이트되면서 더 많은 양의 황사를 탐지하고 있는 것을 확인할 수 있다. 그러나 박스로 표시한 영역과 같이 해안 부근의 황사는 여전히 탐지되지 않는 문제를 가지

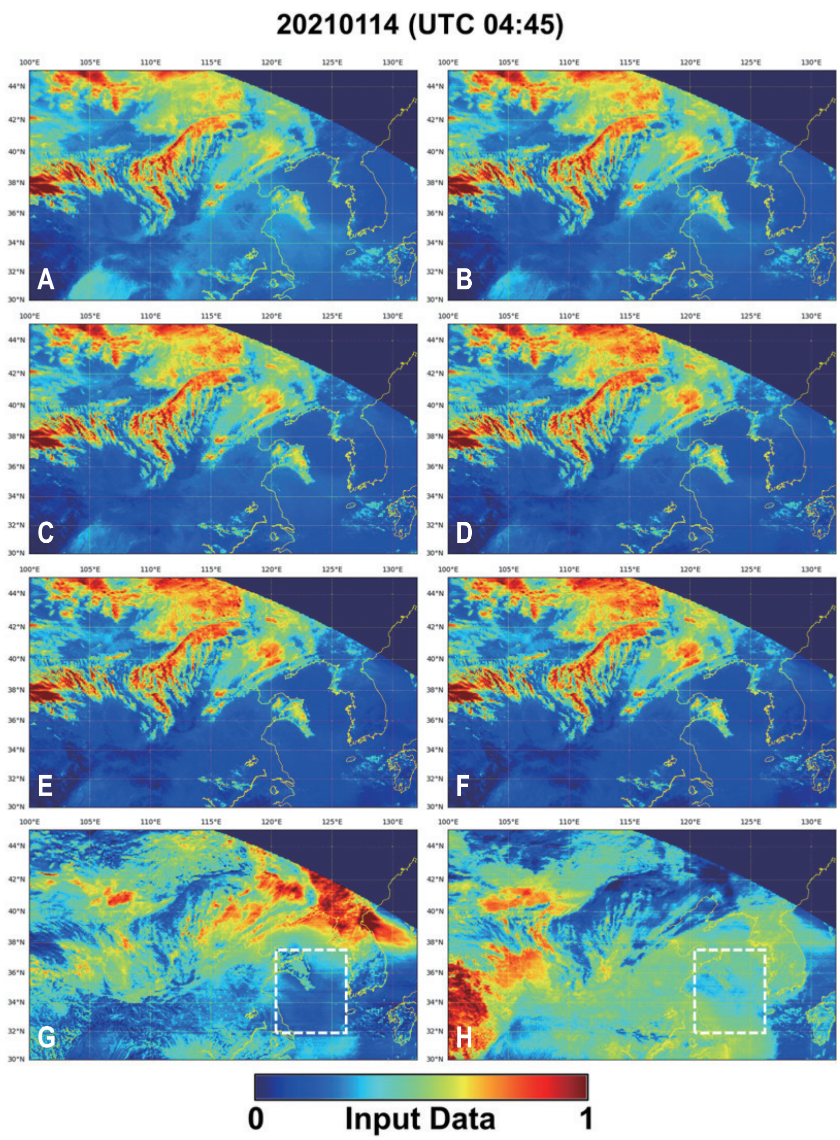

Fig. 5. Evaluation dataset showing input data, label data, and model inference results. (A-F) display preprocessed NR images, (G) shows the UV aerosol Index, and (H) shows the visible aerosol index. UTC, universal time coordinated; NR, normalized radiance; UV, ultraviolet.

고 있다. 이는 Fig. 5의 VISAI와 UVAI를 통해서도 확인할 수 있다. 일반적으로 자외선 대역에서는 해수면에 대한 반사도가 낮기 때문에 육지 부근에 비해 상대적으로 값이 낮게 도출된다. 그러나 VISAI에서는 높은 값을 보이는 것으로 보아 해당 부근에는 실제로 황사가 존재할 가능성이 높다. 이러한 결과는 자외선 대역에서의 반사도 특성으로 인해 기존의 임계값 기반의 모델이 해안가 부근의 황사를 정확하게 탐지하지 못하는 한계를 보여준다. 그러나 모델의 추론 결과는 이와 다르게 나타난다. U-NET 모델을 평가 데이터셋으로 성능을 분석한 결과 약 0.90의 높은 정확도를 보였으며 박스로 표시된 해수면 부분에서 황사가 잘 탐지되는 것을 확인할 수 있었다. 이는 본 연구에서 제작된 데이터셋이 황사 탐지 모델 학습에 있어 효과적임을 보여주었다. 해당 값이 정답값이라고 확언할 수는 없으나 기존의 GEMS 알고리즘이 해수면 영역에서 탐지하지 못하던 황사를 딥러닝 모델이 잘 탐지하고 있기 때문에 딥러닝 기존 알고리즘을 보완할 수 있을 것으로 판단된다.

또한 다른 시기의 황사 일기도와 기존의 GEMS 황사 탐지 결과와 딥러닝 예측 결과를 비교해 보면 기존 GEMS의 황사 산출물에서 황사로 포함되지 않던 부분이 딥러닝 예측 결과에서는 포함되어 있는 것을 확인할 수 있다(Fig. 7). 황사 일기도도 수치 모델을 통해 제작된 데이터이기 때문에 불확도를 포함하고 있을 수 있다. 그럼에도 불구하고 기존 GEMS 산출물에서 탐지하지 못하던 해수면 부근과 황사 일기도에 황사로 나타나고 있는 부분을 딥러닝 모델이 잘 탐지하고 있다. 이는 본 연구에서 제작된 데이터셋이 황사 탐지 딥러닝 모델을 학습하는 데 있어 효과적임을 보여준다. 또한 딥러닝 모델

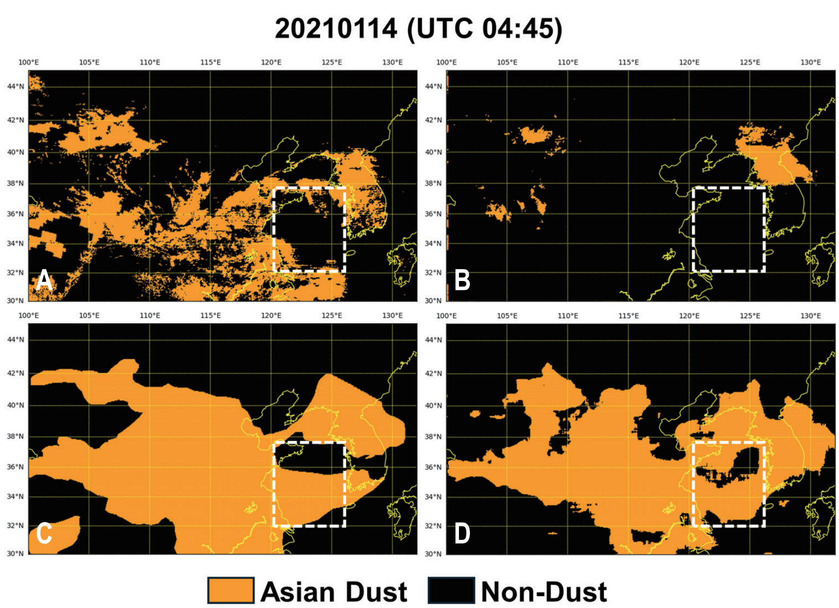

Fig. 6. Example of Asian dust detection results using the generated dataset and the U-NET model. (A) shows the GEMS dust product at 04:45 UTC on January 14, 2021. (B) displays the AMI dust product 5 minutes later. (C) presents the label data created through labeling, and (D) shows the dust detection results predicted by the U-NET model. UTC, universal time coordinated; GEMS, Geostationary Environment Monitoring Spectrometer.

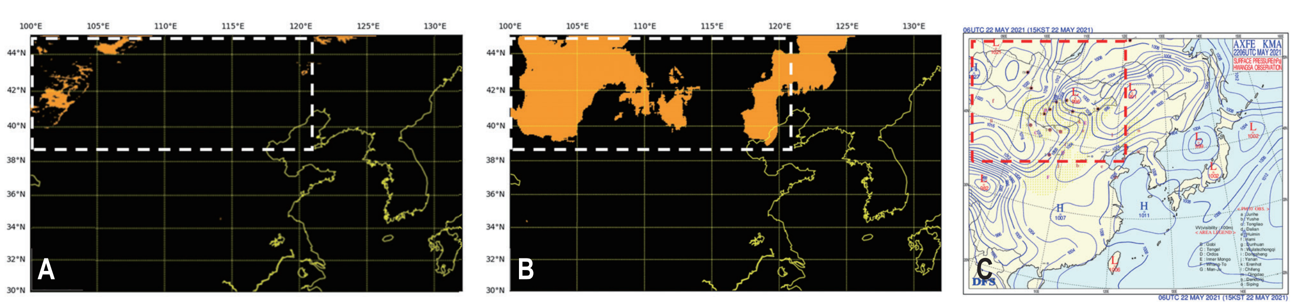

Fig. 7. Comparison of GEMS Asian dust detection results, deep learning predictions, and a dust weather chart for a different period. (A) shows the GEMS dust product, while (B) presents the deep learning prediction results for the same period. (C) is a dust weather chart. KST, Korea Standard Time; KMA, Korea Meteorological Administration; GEMS, Geostationary Environment Monitoring Spectrometer.

을 활용할 시 기존 황사 산출물을 개선할 있다는 시사점을 가진다. 그러나 기존 산출물을 개선하기 위해서는 보다 정확한 라벨 자료를 구축하는 것이 필요하며 라벨 자료의 정확도 검증 방안이 필요하다. 본 연구에서 제작된 데이터셋은 GEMS와 AMI의 황사 산출물, 지상 관측 자료를 참고하여 라벨링을 수행하였고 황사 일기도와의 비교를 통해 라벨 자료의 신뢰도를 분석하였으나 추후 연구에서는 lambertian equivalent reflectivity (LER) 및 구름 산출 자료와 같이 에어로졸 탐지와 큰 관련이 있는 자료들을 추가적으로 확보하여 라벨링을 수행하고 정밀한 라벨 데이터의 검증 방안을 마련하여 데이터 제작을 수행하여 라벨의 신뢰도를 확보할 예정이다.

4. 결론 및 토의

본 연구에서는 정지궤도 환경위성의 황사 산출물 개선을 위한 목적으로 AI 학습 데이터셋을 제작하였다. 대기 오염 발생이 증가함에 따라 이에 대한 국민적 관심과 욕구가 증가하고 있으며 이에 대응하기 위하여 환경부에서는 GK2B에 GEMS를 탑재하여 악화되고 있는 대기 환경 및 기후 변화 유발 물질의 공간적 분포를 파악하고 있다.

기존 GEMS의 황사 산출 알고리즘은 VISAI와 UVAI의 임계값을 통해 산출되고 있다. 그러나 GEMS는 자외선 대역을 활용하여 황사를 탐지하는 과정에서 대기 중에 부유하고 있는 황사를 비교적 잘 탐지하지 못하는 한계가 있었다. 이를 해결하기 위해서는 최근 이미지 처리 분야에서 기존 알고리즘의 성능을 능가하고 있는 딥러닝 기법을 적용이 필요하다. 딥러닝 모델을 학습하기 위해서는 대량의 규격화된 데이터가 필요하지만 현재 이에 대한 규격과 구축되어 있는 데이터셋은 존재하지 않는다. 이를 해결하기 위하여 본 연구에서는 GEMS 황사 탐지 딥러닝 모델을 위한 데이터셋 구축을 수행하였다. 데이터셋 구축을 위해 GEMS가 영상을 제공하기 시작한 2021년 1월부터 5월까지의 GEMS 데이터를 수집하였고 타 위성 산출물과 모델 및 지상 자료를 통해 황사 탐지 모델 학습을 위한 라벨 데이터를 제작하였다. 이후 전처리 및 데이터 증강 기법을 적용하여 최종적인 딥러닝 학습 및 평가 데이터셋을 구축하였다. 제작된 데이터셋을 통해 U-NET 모델을 학습시키고 평가한 결과 0.90의 정확도를 보였으며 기존 GEMS와 AMI가 탐지하지 못하던 해안 영역을 탐지하는 것을 확인할 수 있었고 황사 일기도와 비교하였을 때도 기존에 알고리즘 기반의 산출물들이 탐지하지 못하던 부분을 탐지하는 결과를 보였다. 해당 데이터셋을 통해 황사 탐지를 위한 딥러닝 모델을 학습시키고 황사를 예측하게 된다면 기존 황사 산출물보다 개선된 탐지 결과를 얻을 수 있을 것으로 보인다.

본 연구에서 구축한 데이터는 GEMS 및 타 환경위성 자료, 지상 관측 자료를 참고하여 라벨을 제작하였다. 그러나 해당 참고 데이터들이 불확도를 포함하고 있기 때문에 제작된 라벨 데이터 역시 완벽한 정답값으로 보기 어렵다. 따라서 어노테이션 가이드라인에 따른 체계적인 라벨링 및 검수 과정을 거쳐 정밀한 라벨링을 수행하였고 이러한 과정을 통해 라벨 데이터의 신뢰성과 정확성을 최대한 확보하고자 하였다. 추후에 연구에서는 GEMS L2 cloud 자료 및 LER 자료 등과 같이 라벨링에 사용될 참고 자료를 추가적으로 확보하고 라벨링 가이드라인 및 라벨의 신뢰도 평가 방안에 대한 보완을 수행하고자 한다. 이러한 개선을 통해 더욱 정밀하고 신뢰성 높은 라벨 데이터를 구축하고자 하며 이를 통해 딥러닝 기반의 황사 탐지 모델의 정확성과 효용성을 극대화할 수 있을 것이다.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding Information

This work was supported by National Institute of Environmental Research with the funding of the Ministry of Environment (NIER-2023-01-01-086).

Data Availability Statement

The data that support the findings of this study are openly available in DataON at https://doi.org/10.22711/idr/1031.

References

Bandara NS (2022) Ensemble deep learning for automated dust storm detection using satellite images. In: 2022 International Research Conference on Smart Computing and Systems

Engineering (SCSE), IEEE, Colombo, 1 Sep 2022

Cho Y, Kim J, Go S, et al (2023) First atmospheric aerosol monitoring results from Geostationary Environment Monitoring Spectrometer (GEMS) over Asia. Atmos Meas Tech Discussion:1-29

Choi WJ, Moon KJ, Kim G, Lee D (2023) Reliability analysis based on air quality characteristics in East Asia using primary data from the test operation of Geostationary Environment Monitoring Spectrometer (GEMS). Atmosphere 14(9):1458

Ghahremanloo M, Lops Y, Choi Y, Mousavinezhad S, Jung J (2023) A coupled deep learning model for estimating surface NO2 levels from remote sensing data: 15-year study over the contiguous United States. JGR: Atmospheres 128(2): e2022JD037010

Go S, Kim J, Park SS, et al (2020) Synergistic use of hyperspectral UV-visible OMI and broadband meteorological imager MODIS data for a merged aerosol product. Remote Sens 12(23):3987

Jung HS, Lee S, Yu JW, et al (2023) Studies on artificial intelligence model development for improving Geostationary Environmental Monitoring Spectrometer (GEMS) output. National Institute of Environmental Research, Incheon

Kim J, Jeong U, Ahn MH, et al (2020) New era of air quality monitoring from space: Geostationary Environment Monitoring Spectrometer (GEMS). Bull Amer Meteor Soc 101(1):E1-E22

Kim SY, Lee SH (2009) The study on occurrence of Asian dust and their controlling factors in Korea. J Geol Soc Korea 44(6):675-690

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60

Wang X, Yang W, Weinreb J, et al (2017) Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci Rep 7(1):15415

Yang Q, Kim J, Cho Y, et al (2023) A synchronized estimation of hourly surface concentrations of six criteria air pollutants with GEMS data. npj Clim Atmos Sci 6(1):94

Yu JW, Yoon YW, Lee ER, Baek WK, Jung HS (2022) Flood mapping using modified U-NET from TerraSAR-X images. Korean J Remote Sens 38(6):1709-1722

Meta Data for Dataset

Sort	Field	Subcategory#1	Subcategory#2
Essential	*Title of Dataset	Dataset for deep learning-based GEMS Asian dust detection	
	*DOI	https://doi.org/10.22711/idr/1031	
	*Category	utilitiesCommunication	
	*Temporal Coverage	Deep Learning Based Asian Dust Segmentation Dataset: Input (6 Nomralized Radiances, UVAI, VISAI), Label (Asian Dust Annotation Data). The dataset is organized in the form of a python pickle file.	
	*Spatial Coverage	WGS84 Latitude:30°-45° Longitude:100°-132°	
	*Personnel	Name	Jin-Woo Yu
		Affiliation	University of Seoul
		E-mail	jinwooy@uos.ac.kr
	*CC License	CC BY-NC	
	Summary of Dataset	AI training dataset for training a deep learning-based GEMS Asian dust detection model	
Optional	Project	Studies on artificial intelligence model development for improving Geostationary Environmental Monitoring Spectrometer (GEMS) output	
	Instrument		