일반논문 (Regular Paper)
방송공학회논문지 제28권 제2호, 2023년 3월 (JBE Vol.28, No.2, March 2023)
https://doi.org/10.5909/JBE.2023.28.2.230
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

관심 영역 추출과 영상 분할 지도를 이용한 딥러닝 기반의 이미지 검색 기술

유 민 정[a)], 조 은 혜[a)], 김 병 준[a)], 김 선 옥[a)‡]

Deep Image Retrieval using Attention and Semantic Segmentation Map

Minjung Yoo[a)], Eunhye Jo[a)], Byoungjun Kim[a)], and Sunok Kim[a)‡]

요 약

자율주행은 4차 산업의 핵심 기술로 차, 드론, 자동차, 로봇 등 다양한 곳에 응용 가능하다. 그 중 위치 추정 기술은 GPS, 센서, 지도 등을 활용하여, 객체나 사용자의 위치를 파악하는 기술로 자율주행을 구현하기 위한 핵심적인 기술 중 하나이다. GPS나 LIDAR 등의 센서를 이용하여 위치 추정이 가능하지만, 이는 매우 고가이고 무거운 장비를 탑재해야 하며 지하 혹은 터널 등 전파 방해가 있는 곳의 경우 정밀한 위치 추정이 어렵다는 단점이 있다. 본 논문에서는 이를 보완하기 위해 저가의 비전 카메라로 획득한 컬러 영상을 입력으로 하여 관심 영역 추출 네트워크와 영상 분할 지도를 이용한 영상 검색 기술을 제안한다.

Abstract

Self-driving is a key technology of the fourth industry and can be applied to various places such as cars, drones, cars, and robots. Among them, localiztion is one of the key technologies for implementing autonomous driving as a technology that identifies the location of objects or users using GPS, sensors, and maps. Locilization can be made using GPS or LIDAR, but it is very expensive and heavy equipment must be mounted, and precise location estimation is difficult for places with radio interference such as underground or tunnels. In this paper, to compensate for this, we proposes an image retrieval using attention module and image segmentation maps using color images acquired with low-cost vision cameras as an input.

Keyword : Localization, Image Retrieval, Semantic Segmentation, Attention, Deep Learning

a) 한국항공대학교(Korea Aerospace University)
‡ Corresponding Author : 김선옥(Sunok Kim)
E-mail: sunok.kim@kau.ac.kr
Tel: +82-2-300-0262
ORCID: https://doi.org/0000-0002-9665-4214
※ This was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP)(NRF-2021R1C1C2005202).
· Manuscript January 20, 2023; Revised March 30, 2023; Accepted March 30, 2023.

Ⅰ. 서 론

정확한 자율주행을 위해서 다양한 센서로부터 획득된 정보를 이용하여 현재 위치를 나타내는 위치 추정 기술이 매우 중요하다. 현재 많은 기술들은 위치 추정을 하기 위해 GPS 혹은 LIDAR를 이용하고 있다. 이러한 장비들은 정밀한 측정이 가능한 고성능 장비이기 때문에 매우 고가이며 모든 차량에 탑재하기엔 무겁다거나 크기가 매우 크다는 단점과 지하 혹은 터널과 같이 전파방해가 발생하는 지역에서는 GPS를 신뢰하기 어렵다는 단점이 있다. 이러한 비용적인 문제를 해결하기 위해 비교적 저가인 비전 카메라를 이용하여 위치 추정을 수행하는 연구가 활발히 진행되고 있다.

이미지 검색을 이용한 위치 추정 기술의 최신 연구인 Patch-NetVLAD[1]는 딥러닝을 기반으로 같은 객체의 공통된 픽셀값을 랜드마크로 특정하여 유사성 점수를 평가하는 이미지 검색 기술이다. 해당 방법은 특징이 명확히 보이는 낮 이미지에서 좋은 성능을 보이지만, 특징이 흐릿한 밤 이미지에 대해서는 성능이 현저히 낮아지는 문제점이 있다. 이는 환경 변화에도 강인해야하는 자율주행 기술에서 치명적인 단점이 된다. 이에 따라 본 논문에서는 밤에도 성능이 보장될 수 있도록 밤 이미지를 학습에 사용할 수 있는 네트워크를 제안한다. 흐릿한 이미지 정보를 보완해 줄 영상 분할 정보를 추가하고, 정보의 중요도를 판단하여 가중치를 부여하는 관심 영역 모듈을 추가하였다. 이러한 연구를 통해서 우리는 자율주행 기술의 비용 절감 효과와 더불어 보다 정확한 위치 추정 기술 개발을 기대한다.

본 논문은 이전에 진행되었던 관련 연구들을 Ⅱ절에서 설명하며, 제안한 네트워크의 각 모듈에 대한 자세한 설명을 Ⅲ절에서 다룬다. Ⅳ절에서 기존 연구와 성능 비교 후, 결론으로 마무리 하는 순서로 진행한다.

Ⅱ. 관련 연구

장소 인식을 위한 최신의 이미지 검색 기술인 Patch-NetVLAD[1]는 다양한 크기(multi-scale)의 패치 사이즈를 이용하여 지역적 특징자, 전역적 특징자를 추출하고 이를 융합(fusion)하여 지역적 특징자(Local Feature)와 전역적 특징자(Global Feature)를 결합함으로써 크기 변화, 각도 변화 등의 여러 가지 척도에 대응한 랜드마크를 특정한다. 랜드마크는 공간 점수화(Spatial Scoring) 및 인접 영역(Nearest Neighbor을) 통해 장소 인식을 위한 유사성 점수가 가장 높은 이미지를 추정한다. 해당 기술[1]에서는 컬러 이미지 하나만을 입력으로 하는데, 이는 객체의 구분이 뚜렷하지 않은 밤 이미지를 검색 하는 경우 검색의 정확도가 현저히 감소하게 된다. 이에 본 논문에서는 Patch-NetVLAD[1]를 이미지 검색 기술의 기반으로 사용하며 영상 분할 정보 이미지와 관심영역을 추가하여, 각 픽셀의 의미 정보를 통해 검색의 정확도를 높이고 밤에 대해서도 잘 적응할 수 있는 네트워크를 제안하였다.

다른 저명한 이미지 검색 기술로서 MAC[2], R-MAC[2]이 있다. MAC[2]은 이미지 정보에 대한 max-pooling 연산을 통해 이미지 표현 방식을 K크기의 벡터로 변환하며, 이 벡터 간의 코사인 유사도를 계산하는 방식이다. 가장 활성화가 높은 위치의 정보만을 취득하여 유사도를 비교하는 것

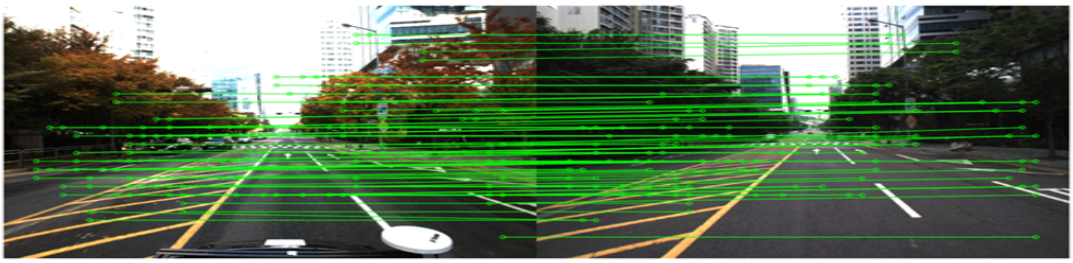


그림 1. Patch-NetVLAD[1]에 Mobiltech 데이터를 적용한 결과. (좌) 검색 이미지, (우) 결과 이미지
Fig. 1. Results of applying Mobiltech data to Patch-NetVLAD[1]. (Left) query image, (Right) result image

으로, 이때 이미지의 지역 정보를 반영하지 못하는 단점이 생긴다. R-MAC[2]은 여러 사이즈의 R영역으로 나누어 다양한 영역에 따른 유사도를 계산함으로써 지역적인 특성 손실을 보완하였다. 즉, 이미지 전체의 유사도를 구하는 것이 아니라 유사한 영역(R)을 먼저 선택한 후 해당 영역에 대한 유사도를 계산하는 리랭킹(re-ranking) 과정이 추가된 것이다. 이를 발전 시켜서 삼중 손실(triplet loss)을 제안한 논문[3] 또한 이미지 검색에서 많은 영향을 주었다.

Ⅲ. 본 론

1. 제안하는 프레임워크

그림 2는 본 논문에서 제안한 프레임워크의 전체적인 모습이다. 기존 Patch-NetVLAD[1]의 경우 낮 이미지에 대해서만 학습하여 밤 이미지에는 취약한 결과를 보였다. 이에 본 논문에서는 환경변화에 대응하기 위해 낮과 밤 이미지를 모두 학습에 사용하여 기상이나 시간 변화에도 강인하도록 하였다. 밤 이미지를 얻기 위해서 AU-GAN[4]을 통해 낮 이미지를 밤 이미지로 증강하였으며 그림 3에서와 같이 결과 이미지를 확인하였다. 또한, 네트워크의 성능을 높이기 위해 영상 분할 정보 이미지와 관심 영역 추출 모듈을 추가하였다. 입력으로 들어간 낮 시간대의 데이터 셋은 밤 시간대로 증강한 이미지와 영상 정보 분할 이미지로 변환되며, 각각의 이미지는 특징자로 추출된다. 이후 본 논문에서 제안한 관심 영역 추출 모듈을 거쳐 이미지 검색 모듈에 입력되며 최종적으로 입력 이미지와 같은 위치로 추정되는 이미지를 도출함으로써 위치를 추정한다.

2. 영상 분할 기술

기존 방법에서는 동일한 환경인 낮 이미지로 학습 및 테스트를 진행하여 환경 변화가 생겼을 때 성능이 감소하는

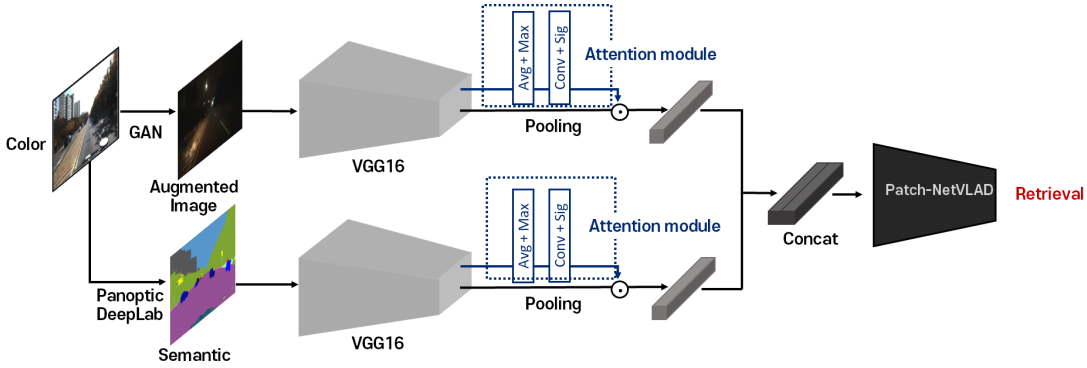
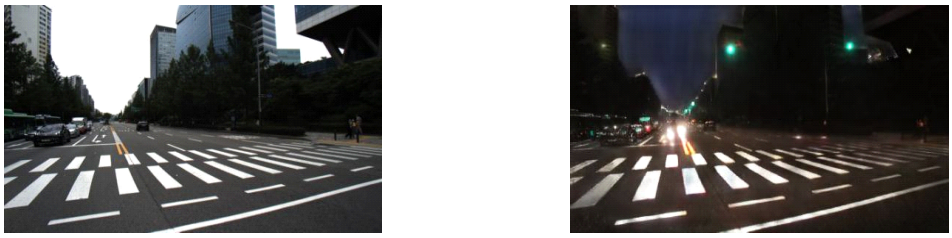


그림 2. 본 논문에서 제안하는 프레임워크
Fig. 2. The network architecture of the proposed framework



(좌)
(Left)

(우)
(Right)

그림 3. AU-GAN[4]을 이용하여 얻은 이미지. (좌) 원본 이미지, (우) 변환된 밤 이미지
Fig. 3. Images obtained with AU-GAN[4]. (Left) original image, (Right) converted night image

문제점이 있었다. 이를 해결하기 위해 밤 영상을 학습에 이용하였으나, 추가하는 것 만으로는 밤 이미지에 대한 학습이 잘 이루어지지 않았다. 따라서 성능 개선을 위해 영상 분할 정보 이미지를 추가적으로 사용하였다. 영상 분할 기술이 메인 네트워크와 동시에 학습이 이루어질 경우, 학습해야하는 파라미터가 방대해져 훈련 시간이 증가하게 된다. 따라서 본 논문에서는 영상 분할 정보 이미지를 얻기 위해 이미 학습되어있는 Cityscapes datasets[6]을 이용하여 ResNext-101[7]을 베이스 라인으로 사전 학습한 Panoptic-DeepLab[5]을 사용하였다. 해당 기술을 적용함으로써 위치 추정을 할 때 불필요한 자동차 사람 등의 객체가 아닌 도로, 건물, 나무 등의 중요한 객체를 중점적으로 파악하여 정확도를 높였다. 순간적으로 변하는 자동차나 인물 등에 집중하여 학습하게 되면 비슷한 인물이 있는 다른 위치로 추정할 가능성이 높아진다고 판단되어 위치를 추정할 때 중요한 고정 객체들을 판단하고자 영상 분할 정보 모듈을 추가하였다. 해당 모듈의 입력으로 밤 이미지가 아닌 낮 이미지에 대한 영상 분할 정보를 사용하였는데, 이는 밤 이미지의 경우 객체가 뚜렷하게 보이지 않아 정보가 대부분 왜곡되었기 때문이다. 표 2와 표3에서 확인할 수 있듯이 영상 분할 정보를 추가함으로써 평균 오차가 크게 감소하였다.

3. 관심 영역 검출

관심 영역은 입력된 각각의 이미지에서 중요하게 보아야 할 위치에 대한 가중치 값을 의미한다. 앞서 언급한바와 같이 위치를 추정할 때에는 일시적인 현상이 아닌 계속해서 유지되는 특징을 잡아서 매칭을 해야 하기 때문에 이동하는 객체의 중요도는 낮게, 고정 객체의 중요도는 보다 높고 명확하게 판단하여 학습에 영향을 줄 수 있는 관심 영역 검출을 도입하였다.

관심 영역은 그림 5와 같이 총 4개의 레이어로 이루어져 있으며, 평균 풀링(Average pooling)과 2x2 최대값 풀링(Max pooling) 레이어를 거친 후에 3x3 Convolution layer를 통과한다. 이후 VGG16에서 추출된 특징자들과 곱하기 위해 Sigmoid 레이어를 거쳐 최종적으로 관심 영역 검출이

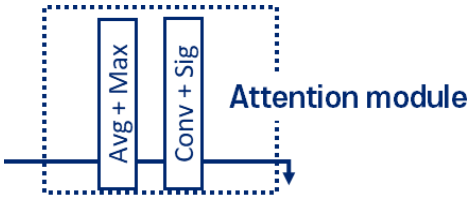
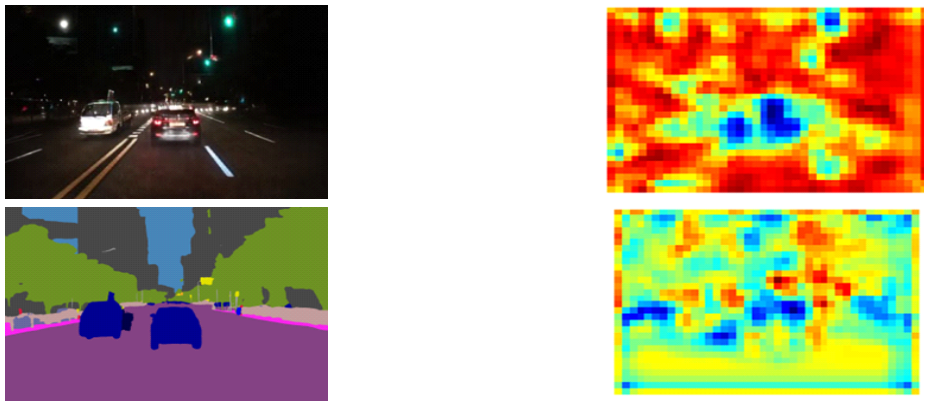


그림 5. 관심 영역 모듈
Fig 5. Attention module

(좌)
(Left)

(우)
(Right)

그림 4. (좌) Panoptic-DeepLab[5]에 입력된 원본 이미지, (우) 최종적으로 얻은 영상 분할 정보 이미지
Fig. 4. (Left) Original image for Panoptic-DeepLab[5], (Right) finally obtained segmentation image



(좌)
(Left)

(우)
(Right)

그림 6. (좌) 입력 이미지로 밤 이미지(위), 영상 분할 이미지(아래), (우) 각 이미지에 대한 관심 영역 이미지
Fig. 6. (Left) Night image(top), segmentation image(bottom) as input image, (Right) Attention image for each image

된다. 해당 모듈을 구성하기 위해 기존 연구[8][9]의 관심 영역 검출 네트워크를 참고하였으며, 본 논문의 취지인 자율주행에서 시간(real-time)은 중요한 도전 과제이기 때문에 최대한 단순화하여 모듈을 구성하였다.

그림 6에서 확인할 수 있듯 관심 영역은 중요한 부분일수록 붉은색으로 나타나게 된다. 밤 이미지에서는 명확한 고정 객체인 주변 나무나 도로를 중요하게 보아 붉게 나타나는 것을 볼 수 있다. 영상 분할 이미지에서는 밤 이미지에서는 볼 수 없었던 또 다른 고정 객체인 건물이나 표지판 등 세부적인 부분을 중요하게 판단하고 이동 객체인 차와 사람에 있어서는 중요도가 낮은 파란색으로 나타나는 것 또한 확인하였다. 관심 영역 이미지를 이용함으로써 환경이 달라져도 낮과 밤 이미지에서 각각 중요도를 판단해 학습하게끔 하여 환경 변화에도 강인한 아키텍처를 만들었다.

Ⅳ. 실험 및 결과

1. Dataset 및 평가 방법

학습 데이터셋은 Mobiltech 기업에서 자체 제작한 Mobiltech Dataset을 사용하였다. 해당 데이터셋은 GPS의 정확도를 저하시키는 전파가 다수 분포하는 상암동 일대를 약 30분간 주행한 것이며, 프레임 단위로 분할 하여 총 15,787장의 이미지로 구성되어 있다. 테스트 데이터는 같은 위치의 다른 계절의 데이터셋에서 110장을 선택하여 구성하였다.

표 1. 학습용 데이터셋과 테스트 데이터셋에 대한 정보
Table 1. Training/testing dataset configuration

	학습용 데이터셋 training dataset	테스트 데이터셋 test dataset
촬영 위치 location	상암동 일대의 도심 the urban area of Sangam-dong	상암동 일대의 도심 the urban area of Sangam-dong
촬영 시간 time	2021년 11월 2시경, 30분 2021, Nov. PM 2:00	2021년 8월 2시경, 30분 2021, Aug. PM 2:00
이미지 개수 # of images	15,787	110
해상도 resolution	1,920 x 1,200	1,920 x 1,200
특징	- 프레임 단위로 분할하여 유사한 장면이 많음 - 테스트 데이터의 경우 다양한 장면들로 구성이 되게끔 선별하여 구성함 - 모든 이미지는 GPS 상의 위치 정보 및 위도와 경도 정보를 가지고 있음 - There are many similar scenes since this dataset is divided video into frames - Test dataset is composed of various scenes. - All imges have location, latitude, and longitude information on GPS	

성능 확인을 위해 테스트 결과와 실제 정답의 위도-y, 경도-x를 이용한 L2 Normalization으로 이미지 검색 기술의 오차를 계산하였다. 또한 테스트 이미지를 검색할 때 정

답과 오답으로 분류를 하여 정답률을 계산 하였는데, 정답으로 인정하는 오차 범위는 현재 알려진 GPS의 평균 오차(15m~33m)와 실제 자율주행 차에 사용 가능한 오차 범위를 고려하여 3m를 기준으로 정오를 판단하였다.

2. 실험 결과

본 실험은 python 3.7, CUDA 11.2.0, NVIDIA GeForce RTX 3070 기반의 컴퓨터에서 실험하였으며, pytorch를 바탕으로 네트워크를 구성하였다. 비교군을 실험하기 위해 Patch-NetVLAD[1]의 저자가 제공하는 공식 코드를 그대로 사용하였으며, 이미 학습되어있는 파라미터를 사용하여 테스트를 진행하였다. 테스트 이미지로는 낮 이미지 110장을 이용하였으며, 검색 결과는 밤 데이터로만 구성된 데이터베이스를 사용하였다.

그림 7은 본 논문이 제안한 네트워크를 학습하여 테스트한 결과이며 좌측 이미지가 검색 이미지, 우측 이미지가 Top 1 검색 결과 이미지다. 시간 변화에도 강인한지 확인하기 위해 밤으로 구성된 데이터 베이스에서 낮 이미지를 검색하였다. 이를 통해 해당 네트워크가 위치 추정에 있어서 환경 변화에 영향을 받지 않음을 확인할 수 있었다.

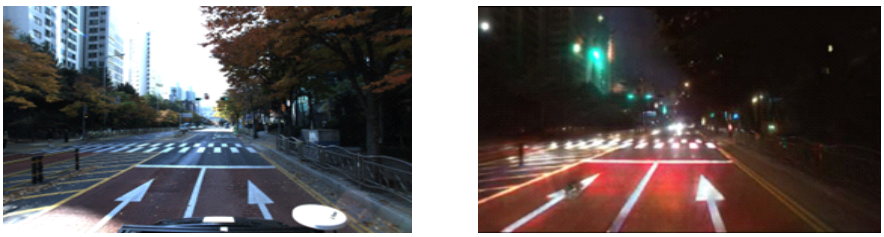
표 2는 임의의 테스트 이미지 5장에 대한 실제 GPS 거리 오차를 보여주고 있다. 해당 표는 이미지를 검색했을 때 나온 추정 위치와 실제 위치에 대한 오차를 나타낸 표로, 단위는 미터(m)이다. 표 3은 110개의 테스트 데이터셋에 대한 결과를 나타내고 있다. 기존의 Patch-NetVLAD[1]로 낮 이미지를 밤 데이터 베이스에서 검색했을 때 평균 225.7130m 오차를 보였으며, 전체 정답률 15%로 대부분의 이미지에서 위치 검색에 실패하였다. 영상 분할 이미지를 추가했을 때는 평균 오차 45.7036m, 정답률 67%로 오차가 약 5배 감소하였고, 정답률은 약 4배 증가하였다. 제안한 네트워크를 사용하였을 때는 평균 오차가 0.0198m 감소하였고 정답률이 1% 향상되었다. 해당 결과로 비추어보아 본 논문에서 제안한 네트워크가 기존 연구에 비해 보다 나은 성능을 보임을 확인할 수 있다.

표 2. 검색 이미지에 대한 검색 오차 (단위:m)
Table 2. Search error for query image (in m)

이미지 번호 No.	Patch-NetVLAD[1]	영상 분할 정보 이미지만 추가 Result with segmentation map only	제안한 방법 Result of the proposed method
1	613.7344	0.1081	0.1081
2	716.6734	0.0683	0.0683
3	1.3785	0.6307	0.3762
4	300.1484	0.0768	0.0768
5	29.3167	4.1942	2.2660

표 3. 테스트 이미지 110개의 결과
Table 3. Results of 110 test images

	Patch-NetVLAD[1]	영상 분할 정보 이미지만 추가 Result with segmentation map only	제안한 방법 Result of the proposed method
평균 오차 (m) average error(m)	224.7130	45.7036	45.6838
정답 개수 (개) # of correct answers	17	74	75
정답률 accuracy	15%	67%	68%



(좌) (우)

그림 7. 제안한 네트워크를 통해 얻은 결과. (좌) 검색 이미지, (우) 결과 이미지
Fig. 7. Results from the proposed network. (Left) query image, (Right) result image

Ⅴ. 결 론

기존의 이미지 검색 기술에 영상 분할 이미지와 관심 영역 검출을 적용함으로써 시간 등의 환경 변화에서도 강인한 이미지 검색의 성능을 높일 수 있음을 확인하였다. 이와 같이 다른 네트워크에도 간단하게 적용할 수 있는 추가적인 모듈을 제안함으로써 앞으로 비전 분야의 다양한 기술에 영향을 줄 수 있을 것으로 기대된다. 또한, 해당 연구는 GPS의 대체 또는 보완 기술로서 비전 카메라를 이용한 위치 추정 기술의 성능 향상을 기대할 수 있다.

참 고 문 헌 (Reference)

[1] Hausler, S., Garg, S., Xu. M., Milford, M., & Fischer, T, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.01392

[2] G. Tolias, S. Ronan, and J. Hervé, "Particular object retrieval with integral max-pooling of CNN activations," arXiv preprint, 2015.
doi: https://doi.org/10.48550/arXiv.1511.05879

[3] A. Gordo, J. Almazan, J. Revaud, & D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, Vol. 124, No. 2, pp. 237-254, 2017.
doi: https://doi.org/10.1007/s11263-017-1016-8

[4] J. G. Kwak, Y. Jin, Y. Li, D. Yoon, D. Kim, and H. Ko, "Adverse Weather Image Translation with Asymmetric and Uncertainty-aware GAN," arXiv preprint, 2021.
doi: https://doi.org/10.48550/arXiv.2112.04283

[5] B. Cheng, D. C. Maxwell, Z. Yukun, L. Ting, S. H. Thomas, A. Hartwig, C. Liang, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020.
https://doi.org/10.1109/CVPR42600.2020.01249

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, "The cityscapes dataset for semantic urban scene understanding," Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR), 2016.
https://doi.org/10.1109/CVPR.2016.350

[7] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2017.
doi: https://doi.org/10.1109/CVPR.2017.634

[8] S. Kim, S. Kim, D. Min, K. Sohn, "Laf-net: Locally adaptive fusion networks for stereo confidence estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00029

[9] S. Woo, J. Park, JY. Lee, I.S. Kwoen, "Cbam: Convolutional block attention module," *Proceedings of the European conference on computer vision (ECCV)*, 2018.
doi: https://doi.org/10.1007/978-3-030-01234-2_1

저 자 소 개

유 민 정

- 2019년 ~ 2023년 : 한국항공대학교 소프트웨어학과 학사
- 2023년 ~ 현재 : 한국항공대학교 인공지능학과 석사과정
- ORCID : https://orcid.org/0009-0003-4644-9671
- 주관심분야 : 컴퓨터비전, 영상분할

조 은 혜

- 2019년 ~ 2023년 : 한국항공대학교 소프트웨어학과 학사
- ORCID : https://orcid.org/0009-0008-5517-9550
- 주관심분야 : 딥러닝, 컴퓨터비전

저 자 소 개

김 병 준

- 2016년 ~ 2022년 : 한국항공대학교 소프트웨어학과 학사
- 2022년 ~ 현재 : 한국항공대학교 인공지능학과 석사과정
- ORCID : https://orcid.org/0009-0002-8906-1004
- 주관심분야 : 컴퓨터비전, 3차원 영상처리, 객체 탐지

김 선 옥

- 2009년 ~ 2014년 : 연세대학교 전기전자공학과 학사
- 2014년 ~ 2019년 : 연세대학교 전기전자공학과 박사
- 2019년 ~ 2021년 : 연세대학교 박사후연구원
- 2021년 ~ 현재 : 한국항공대학교 소프트웨어학과 조교수
- ORCID : https://orcid.org/0000-0002-9665-4214
- 주관심분야 : 컴퓨터비전, 인공지능, 3차원 영상처리