일반논문 (Regular Paper)
방송공학회논문지 제28권 제4호, 2023년 7월 (JBE Vol.28, No.4, July 2023)
https://doi.org/10.5909/JBE.2023.28.4.470
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Vision Transformer를 활용한 단일 영상 카메라 캘리브레이션

이 승 용[a)], 한 종 기[a)‡]

Single Image Camera Calibration using Vision Transformer

Seung-Yong Lee[a)] and Jongki Han[a)‡]

요 약

카메라 캘리브레이션은 카메라의 초점거리, 주점, 왜곡 계수와 같은 내부 파라미터와 카메라의 위치 및 방향과 같은 외부 파라미터를 구하는 작업이다. 정확한 내부 파라미터를 구하는 작업은 컴퓨터비전 분야에서 매우 중요한 작업이지만 고전적인 방법들은 그 과정이 복잡하고 제약이 많다는 단점이 있다. 이를 극복하기 위해 딥러닝을 활용한 단일영상만을 이용한 카메라 캘리브레이션 연구들이 발표되고 있지만 정확도가 다소 떨어진다는 문제가 있다. 본 논문에서는 EfficientNetV2를 이용해 얻은 피처에 Token들을 추가하여 Vision Transformer를 이용해 카메라 내부 파라미터를 추정하는 모델을 제안하고 이전 단일 영상 캘리브레이션 연구들과 비교하여 더 높은 정확도를 얻을 수 있음을 확인했다.

Abstract

Camera calibration is the process of determining both the intrinsic parameters, such as focal length, principal point, and distortion coefficients, as well as the extrinsic parameters, such as the position and orientation of the camera. Accurately estimating the internal parameters is a crucial task in the field of computer vision, but traditional methods have limitations in terms of complexity and constraints. To overcome these limitations, research using deep learning for camera calibration with a single image has been proposed, but it suffers from reduced accuracy. In this paper, we propose a model that utilizes Vision Transformers by adding tokens to the features obtained using EfficientNetV2 to estimate the camera's internal parameters. We compare our approach with previous single-image calibration studies and confirm that our model achieves higher accuracy.

Keyword : Camera Calibration, Computer Vision, Deep Learning

a) 세종대학교(Sejong University)
‡ Corresponding Author : 한종기(Jongki Han)
E-mail: hjk@sejong.edu
Tel: +82-2-3408-3739
ORCID: https://orcid.org/0000-0002-5036-7199

※ This work was supported by the National Research Foundation of Korea (NRF) under Grant 2022R1F1A1071513 funded by the Korea government through the Ministry of Science and ICT (MSIT).
· Manuscript May 19, 2023; Revised July 11, 2023; Accepted July 11, 2023.

Ⅰ. 서 론

카메라를 통해 3차원 공간을 2차원 영상으로 만들 때 2차원 영상의 화소값들을 결정하는 것은 영상을 촬영하는 시점의 카메라의 방향과 위치이다. 하지만 실제로는 카메라의 방향과 위치뿐만 아니라 카메라 렌즈의 광학적 특성, 렌즈와 이미지 센서 사이의 거리, 이미지 센서의 위치 등 카메라 자체의 영향도 영상 데이터에 큰 영향을 끼친다. 컴퓨터 비전 분야에선 앞서 말한 카메라의 방향과 위치를 외부 파라미터, 카메라 렌즈에 의한 광학적 왜곡, 렌즈와 이미지 센서 사이의 거리(초점 거리), 카메라 광축과 이미지 센서의 교점 좌표(주점) 등을 내부 파라미터라고 한다.

최근 VR, AR, 메타버스와 같은 기술들이 발전하며 3차원 미디어의 수요가 증가하고 있고 이에 따라 3차원 미디어의 대표적인 예시인 6-DoF(Degrees of Freedom) VR영상 생성하는 3D 재구성 기술이 주목을 받고 있다. 다수의 2차원 영상을 통해 3차원 물체 또는 환경을 복원하는 3D 재구성 기술은 카메라의 외부 파라미터, 즉 3차원 위치 정보를 필요로 하고 이를 위해 2차원 영상들로부터 카메라의 위치와 방향을 추정하는 Structure from Motion(SfM)[1] 기술을 사용한다. SfM은 2차원 영상의 점들을 3차원으로 투영시키고 다시 이 점들을 다른 영상으로 재투영하는 과정을 거치며 카메라들 간의 상대적인 위치 정보를 파악하는데, 이 과정에서 카메라 내부 요인의 영향을 제거해야만 정확한 외부 파라미터를 얻을 수 있다. 때문에 내부 파라미터를 추정하는 기술이 필요하고 이 과정을 카메라 캘리브레이션이라고 한다.

고전적인 카메라 캘리브레이션 방법[2,3]은 그림 1과 같은 체커판(checker board) 또는 캘리브레이션 패턴을 인쇄한 종이를 여러 각도로 촬영한 영상을 활용하여 초점 거리, 렌즈 주점, 렌즈의 광학적인 왜곡을 추정한다. 이러한 카메라 캘리브레이션 방법은 비교적 정확하다는 장점이 있지만 그 과정이 번거롭고, 카메라의 자동 초점 기능(auto focus)에 따라 변화하는 초점 거리와 OIS(Optical Image Stabilization)에 의해 변화하는 주점은 추정할 수 없다는 단점이 있다.

[2,3]의 방법의 단점을 극복하기 위해 [4,5,6]와 같은 딥러닝을 이용한 단일 영상 카메라 캘리브레이션 방법들이 연구되었다. [4,5,6]은 CNN(Convolutional Neural Network)기반의 피처 추출기를 이용해 입력 영상의 피처를 구하고 FFN(Feed Forward Network)을 통과시켜 원하는 파라미터를 추정하는 방법을 사용한다. 이러한 방법은 [2,3]에서 특정 패턴을 촬영한 다수의 영상을 필요로 하던 것과 달리 단일 영상만 입력으로 사용한다는 장점이 있지만 추정한 내부 파라미터의 정확도가 다소 떨어진다는 단점이 있다. 이에 대한 원인 중 하나는 CNN으로 추출한 피처를 구조가 단순한 FFN을 통해 분석한다는 것이다. 따라서 본 논문에서는 위와 같은 문제를 해결하기 위해 단일 영상을 입력으로 EfficientNetV2-S[7]을 통해 구한 피처를 Vision Transformer[8]를 이용해 분석하여 카메라의 내부 파라미터를 추정하는 모델을 제안한다.

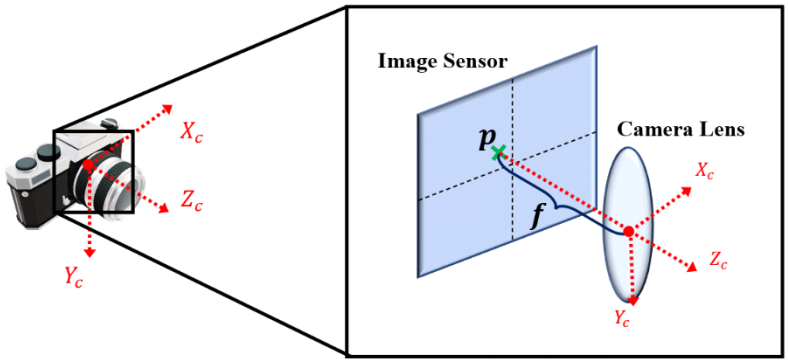

그림 1. 카메라의 내부 파라미터
Fig. 1. Camera's intrinsic parameter

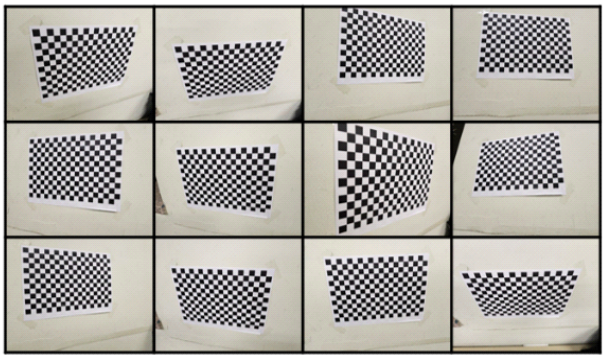

그림 2. 카메라 캘리브레이션을 위한 체커판 사진
Fig. 2. Checkboard image for camera calibration

본 논문의 구성은 다음과 같다. Ⅱ장에서 카메라 캘리브레이션 방법에 딥러닝을 활용한 다른 기존의 방법들과 문제들을 소개하고 Ⅲ장에서 제안하는 알고리즘을 설명한다. Ⅳ장에서 제안한 방법에 대한 실험결과와 기존의 방법들의 결과를 비교하고 Ⅴ장에서 결론짓는다.

Ⅱ. 기존 연구들 및 문제점

카메라 캘리브레이션 방법에 대한 기존 연구들은 크게 두 분류로 나눌 수 있다. 핸디 크래프트 방법들과 딥러닝을 적용한 방법들이다. 핸디 크래프트 방식들[2,3]은 여러 장의 영상을 분석하여 카메라의 내부 파라미터를 추정하는 방식을 사용하였다. 이러한 방법들의 가장 큰 문제점은 특정한 패턴을 촬영한 영상을 여러 장 필요로 하거나 영상에 다수의 직선이 존재하는, 특정 환경에 대해서만 적용할 수 있다는 것이다. 또는 사용자가 수동으로 영상의 정보를 추가적으로 제공해주어야 캘리브레이션이 가능하기 때문에 제약이 많고 번거롭다는 단점이 있다. 이러한 단점을 보완하기 위해 최근 연구들에서 딥러닝 모델을 적극적으로 활용하여 단일 영상만으로도 내부 파라미터를 추정하는 연구들이 등장하고 있다.

DEEPFOCAL[4]은 단일 영상을 입력으로 AlexNet[9]을 이용해 피처를 추출하고 3개의 fully connected layer를 통과시켜 영상의 초점 거리를 구하는 방법을 제안하였다. 단일 영상으로 내부 파라미터를 구하는 이전 연구들이 제한적이고 특정 환경에서만 가능했던 것에 반해 일반적인 환경에서도 적용이 가능하다. DeepCalib[5]는 Inception-V3[10]를 이용하여 단일 영상으로 초점 거리와 왜곡 계수를 구하는 딥러닝 모델을 제안하였다. 두 파라미터를 구하기 위해 하나의 백본 네트워크를 사용하는 SingleNet, 두개의 백본을 이용하여 두 파라미터를 각각 구하는 DualNet, 초점 거리를 먼저 구하고 그 값이 다른 백본에 영향을 주는 SeqNet을 발표하고, 이 중에서 SingleNet이 가장 효율적이고 좋은 결과를 보인다고 발표하였다. 또한 기존 연구들이 내부 파라미터 구하는 문제를 회귀(regression) 문제로 생각한 것과 달리 분류(classification) 문제처럼 다루는 방법을 제안하였다. M. Lopez et al.[6]는 densenet-161[11]을 백본 네트워크로 사용하여 피처를 추출하고 4개의 FFN을 이용하여 카메라의 초점 거리, 왜곡 계수, 팬, 틸트를 각각 추정하는 모델을 발표하였다.

이러한 딥러닝을 사용한 방법들은 기존의 다수의 영상을 입력으로 사용하는 방법에 비해 제약사항은 줄어들었으나 추정할 수 있는 카메라 내부 파라미터가 제한적이고 그 정확도가 다소 떨어진다는 문제가 있다. 이에 대한 이유는 딥러닝 모델이 너무 단순하여 영상을 정확히 분석하기 어렵기 때문이다. 이전 연구들[4,5,6]은 모두 ImageNet dataset[12]으로 pre-train 된 CNN기반의 백본 네트워크를 통해 얻은 피처맵을 FFN을 이용해 분석하여 각 파라미터 값을 추정한다. 하지만 FFN은 CNN을 통해 얻은 피처맵의 의미론적 정보(semantic information)만 사용하기 때문에 피처의 global context를 활용하지 못한다. Vision Transformer[8]는 이에 대한 해결책으로, 입력의 global context를 분석할 수 있다는 장점을 가지고 있고 최근 많은 연구들에서 이를 활용하고 있다[13]. 본 논문에서는 CNN 피처 추출기로 EfficientNetV2-S[7] 이용해 추출한 피처를 의미론적 정보와 피처의 global context를 모두 사용하여 분석하기 위해 focal token, principal point token, distortion token을 추가하고 Vision Transformer[8]를 사용하는 모델을 제안한다.

Ⅲ. 제안하는 모델

그림 3은 제안하는 딥러닝 모델을 도식화한 그림이다. 단

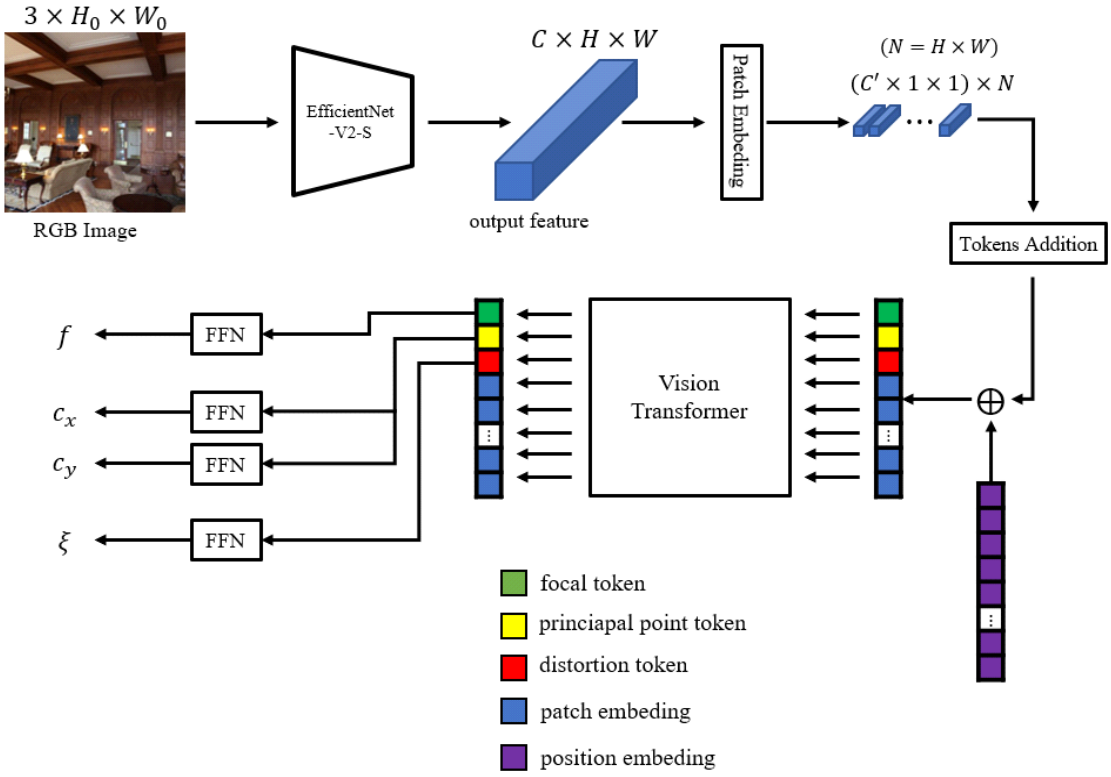

그림 3. 제안하는 모델 도식. 단일 RGB영상을 입력으로 받아 EfficientNetV2-S로 feature를 추출하고 Patch Embedding, Tokens Addition을 거치고 position embedding을 더하여 Vision Transformer의 입력으로 사용한다. Transformer Layer를 거친 결과에서 focal token, principal point token, distortion token을 FFN을 이용하여 내부 파라미터를 추정한다.
Fig. 3. Overview of the proposed model. It receives a single RGB image as an input, extracts feature with EfficientNetV2-S, goes through Patch Embedded and Tokens Addition, and adds position embedding to use it as an input to Vision Transformer. From the results of the Transformer Layer, the intrinsic parameters of local token, principal point token, and distortion token are estimated using FFN.

일 입력 영상으로부터 초점거리, 주점, 왜곡계수를 추정할 수 있다. CNN 백본 네트워크를 이용해 피처 추출을 수행하고 추출된 피처를 Vision Transformer의 입력으로 사용하여 각 파라미터를 추정한다.

1. Backbone Network

피처 추출을 위해 사용할 백본 네트워크는 ImageNet dataset[12]으로 pre-train된 EfficientNetV2-S[7]이다. EfficientNetV2는 progressive learning을 통해 빠른 학습속도와 좋은 성능을 낼 수 있다. 본 논문에서는 Image Classification을 위한 Network 최종 block의 average pooling layer와 classifier layer를 제거하고 $3 \times H_0 \times W_0$ 크기의 영상을 입력으로 받아서 $C \times H \times W (H = H_0/32, W = W_0/32)$ 크기의 피처맵을 출력하도록 하였다. 본 논문의 실험에서는 $H_0, W_0 = 512$, $C = 1280$, $H, W = 16$ 으로 그 값을 설정하였다.

2. Vision Transformer

EffcientNetV2-S[7]을 통해 얻은 피처를 Vision Transformer[8]의 입력으로 사용하기 위해 $C \times H \times W$ 크기의 피처를 $C \times 1 \times 1$ 크기의 patch로 나누어 N개의 patch로 변환하고 각 patch를 Patch Embedding을 통해 $C' \times 1 \times 1$ 크기로 변환한다. 본 논문에서는 $C' = 256$ 으로 설정하였다. 기존의 Vision Transformer는 Image Classification을 위해

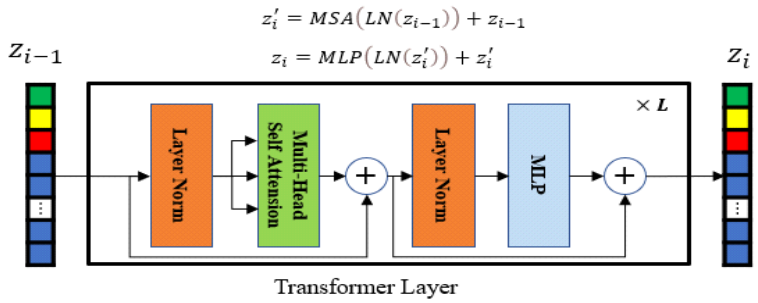

그림 4. Vision Transformer의 구조
Fig. 4. Architecture of Vision Transformer

Class Token 1개만 추가하지만 본 논문의 Tokens Addition에서는 크기를 가진 Focal Token, Principal Point Token, Distortion Token, 총 3개의 token들을 추가하여 각각 초점 거리, 주점, 왜곡 계수에 대한 representation vector를 얻을 수 있도록 한다. 이후 position embedding을 차원에 맞게 추가하여 Vision Transformer의 입력으로 사용한다. Transformer Layer의 구조는 그림 4와 같다. i번째 Transformer Layer의 입력의 집합을 z_i의 번째 원소를 이라고 할 때 아래와 같이 표현할 수 있다.

$$z_i = \left[z_i^1, z_i^2, z_i^3, \cdots, z_i^N, z_i^{N+1}, z_i^{N+2}, z_i^{N+3}\right] \quad (1)$$

Layer Normalization을 LN, Multi-Head Self Attention을 MSA 라고 할 때 그림 4의 구조는 아래와 같이 나타낼 수 있다.

$$z_i' = MSA(LN(z_{i-1})) + z_{i-1} \quad (2)$$

$$z_i = MLP(LN(z_i')) + z_i' \qquad i = 1,2, \cdots, L \quad (3)$$

각 layer의 Multi-Head Self Attention은 8개의 head를 가지고 있고 Layer의 개수 L은 6개로 정하였다. 8개의 head를 가지는 transformer layer 6개를 추가하며 증가하는 계산량은 약 11.5 GFLOPs이다[14].

3. Feed Forward Network

각 parameter를 추정하기 위해 4개의 FFN을 사용하였다. 초점 거리, 주점의 좌표, 왜곡 계수를 각각 f, c_x, c_y, ξ라고 할 때, f는 Focal Token, ξ는 Distortion Token을 사용하여 값을 추정하고 c_x, c_y는 Principal Point Token을 공유하여 추정한다. 각 FFN은 하나의 hidden layer를 가지고 있고 ReLU-activation function를 사용하였다.

4. 손실함수

제안한 모델을 outlier에 강건하게 학습시키기 위해 각 parameter에 $L1\ Loss$를 최소화하는 방향으로 훈련을 진행하였다.

$$L1\ \ Loss = \left\| y_{predicted} - y_{gt} \right\|_1 \quad (4)$$

학습에 사용된 최종적인 손실함수는 아래의 식(5)와 같다. 훈련에 사용된 dataset의 f, c_x, c_y, ξ의 ground truth 범위가 각각 200~600(pixel), 240~272(pixel), 0~1.2으로 각 parameter의 scale이 다르기 때문에 보다 학습이 잘 되도록 하기 위해 f에 0.1, c_x, c_y에 0.2, ξ에 20을 곱하여 scale을 보정한 값을 손실함수로 사용하였다.

$$l = 0.1l_f + 0.2l_{c_x} + 0.2l_{c_y} + 20l_\xi \quad (5)$$

Ⅳ. 실험 결과 및 분석

실험을 위해 SUN360 panorama dataset[15]을 이용해 생성한 영상을 사용하였다. ERP(Equirectangular Projection) 영상들로 이루어진 SUN360 panorama dataset의 영상을

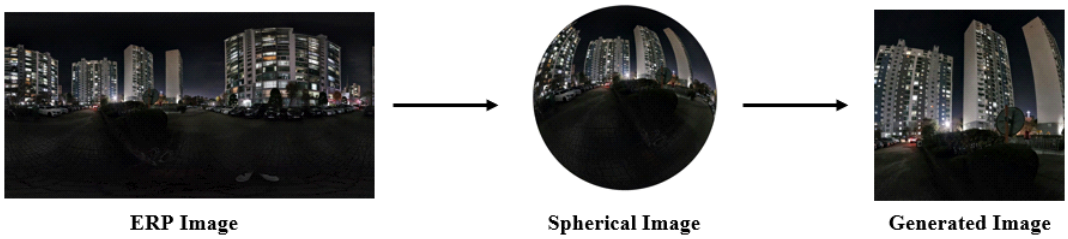

그림 5. Dataset 생성 방법
Fig. 5. How to generate dataset

sphere 형태로 projection하고 임의의 pitch, roll, yaw와 초점거리, 주점, 왜곡 계수를 이용해 크기의 RGB영상을 생성하고 내부 파라미터에 대한 ground truth를 얻었다. Training set, validation set, test set 생성을 위해 각각 446장, 36장, 36장을 사용하였고, pitch: -20~20(degree), roll: -30~30(degree), yaw: -180~180(degree)와 f: 200~600(pixel), c_x,c_y: 240~272(pixel), ξ: 0~1.2 범위에서 random하게 정하는 과정을 반복하여 29180개의 training set, 2918개의 validation set, 2918개의 test set을 생성하였다.

학습에 사용된 optimizer는 Adam을 사용하고 learning rate는 5×10^{-5}로 설정하여 10 epoch 마다 2로 나누어 보다 세밀한 학습을 할 수 있도록 하였다. 손실함수는 Ⅲ장에서 설명했듯이 각 파라미터의 L1 loss에 scale을 보정한 값을 사용하였다. 최대 Epoch는 100으로 설정하고 오버피팅을 방지하기 위해 validation dataset이 수렴하면 학습을 멈추도록 하였다.

그림 6은 epoch에 따른 validation dataset에서 각 parameter의 평균절대오차를 그래프로 나타낸 것이다. 학습이 진행됨에 따라 오차가 수렴하고 있음을 확인할 수 있고 epoch가 50을 초과한 후에는 학습 효과가 크게 나타나지 않는 것으로 보인다. 그림 7은 딥러닝 모델 구조에 따른 각 파라미터의 누적오차분포이다. Only FFN은 EfficientNetV2-S[7]

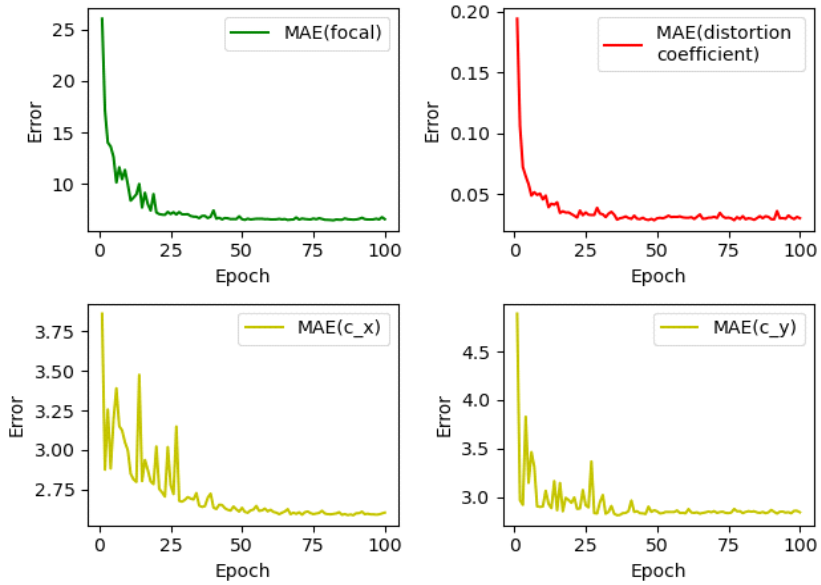

그림 6. Epoch에 따른 validation dataset의 평균절대오차
Fig. 6. Mean absolute error of validation data set according to Epoch

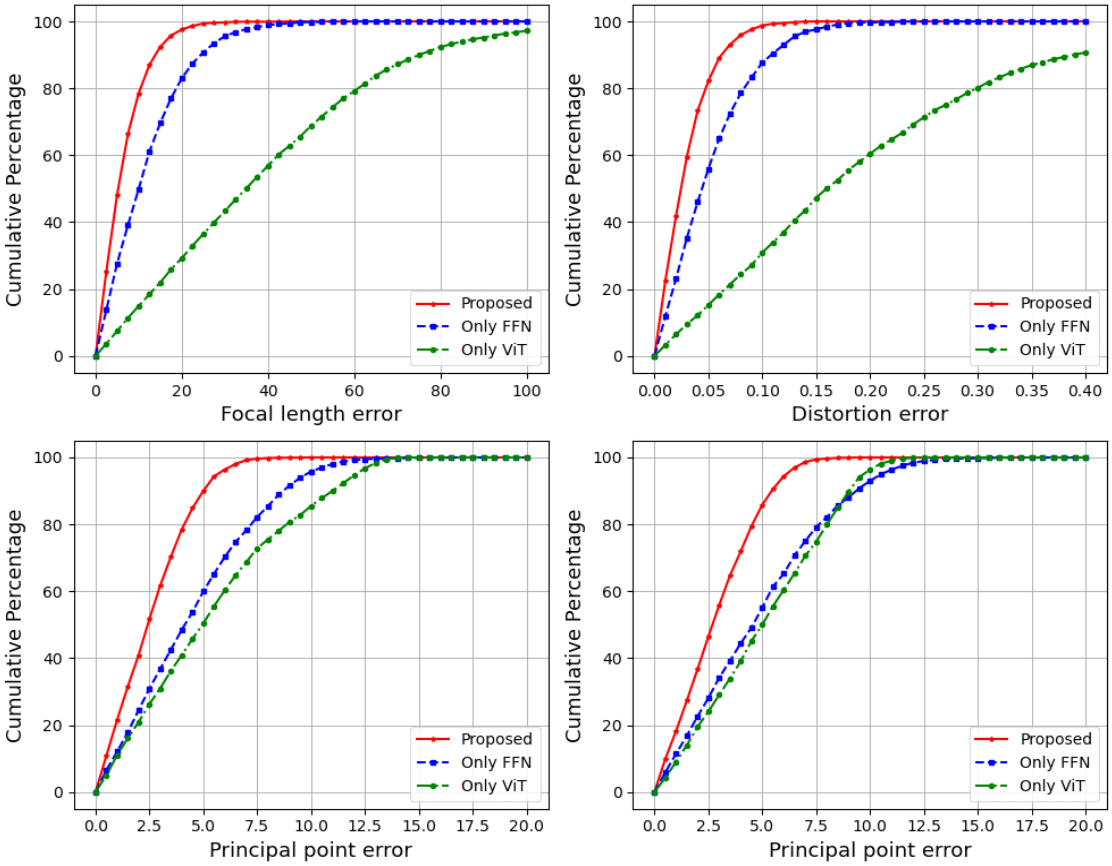

그림 7. 모델 구조에 따른 각 파라미터의 누적오차분포
Fig. 7. Cumulative error distribution of each parameter according to the model architecture

의 피처를 FFN만을 이용한 결과이고, Only ViT는 CNN 피처추출기 없이 Vision Transformer[8]만을 이용한 결과이다. 모든 파라미터에 대해서 본 논문에서 제안한 모델이 가장 우수한 성능을 보이고 Vision Transformer만 사용한 모델이 가장 성능이 떨어지는 것을 확인할 수 있다.

표 1은 다른 딥러닝 기반의 카메라 캘리브레이션 연구와 결과를 비교한 표이다. 표 안에 숫자는 각 모델에서 파라미터들의 절대오차를 표기한 것이다. 전통적인 카메라 캘리브레이션 방법들[2,3]은 체커보드를 촬영한 다수의 영상을 이용하여 카메라 파라미터를 구하는 기술이기 때문에, 본 논문에서는 비교 실험을 하지 않았다. DeepCalib[5]는 Ⅱ장에서 설명했듯이, 단일 영상을 이용하여 초점 거리와 왜곡 계수를 구하는 연구이다. DeepPTZ[16]는 같은 장면을 촬영한 다른 내부 파라미터를 가진 한 쌍의 영상을 입력으로 초점거리, 왜곡 계수와 pitch, roll, yaw를 추정하는 모델이다. [5]와 [16] 모두 본 논문과 같이 파노라마 영상을 이용해 생성한 영상으로 dataset을 구성하여 입력으로 사용한다.

표 1. 다양한 딥러닝 기술들 기반의 카메라 캘리브레이션 성능 비교
Table 1. Performance comparison between various algorithms based on deep learning

Model	MAE of	MAE of	MAE of	MAE of
DeepCalib[5]	41.734	0.179	*N/A*	*N/A*
DeepPTZ[16]	12.889	0.085	*N/A*	*N/A*
Ours (synthetic)	6.515	0.029	2.600	2.874
Ours (real world)	8.1287	0.032	4.113	4.267

Ours는 훈련 영상과 같은 방식으로 생성한 테스트 영상을 이용한 결과와 실제 카메라로 촬영한 250장의 영상을 이용한 결과를 따로 표기하였다. 표 1의 결과를 보면 본 논문에서 제안한 모델은 단일 영상을 사용하는 DeepCalib[5]에 비해 월등히 좋은 성능을 보이고 두 장의 영상을 사용하는 DeepPTZ[16]에 비해서도 우수한 성능을 보인다. 또한 두 모델은 모두 영상의 주점을 이미지의 중심으로 가정하고 실험을 진행하기 때문에 주점에 대한 추정을 하지 않지만, 본 논문의 모델은 주점을 추정한 결과가 포함 되어있다. 실제 카메라로 촬영한 real world 영상을 입력으로 사용했을 때, synthetic 영상에 비해 성능이 다소 떨어지는 결과를 보이지만, 여전히 [5], [16]의 결과보다 우수한 성능을 보인다. 이러한 결과들을 통해 본 논문에서 제안한 EfficientNetV2[7]을 백본 네트워크로 사용하고 Vision Transformer[8]를 이용해 피처를 분석하여 카메라의 내부 파라미터를 추정하는 모델이 효과적이고 이전 연구들에 비해 우수한 성능을 보일 수 있음을 알 수 있다.

그림 8은 본 논문에서 제안한 모델로 예측한 파라미터를 이용하여 왜곡된 영상을 왜곡 보정한 결과들이다. 왼쪽 영상들은 왜곡에 의해 영상 속 물체가 휘어지는 현상이 발생

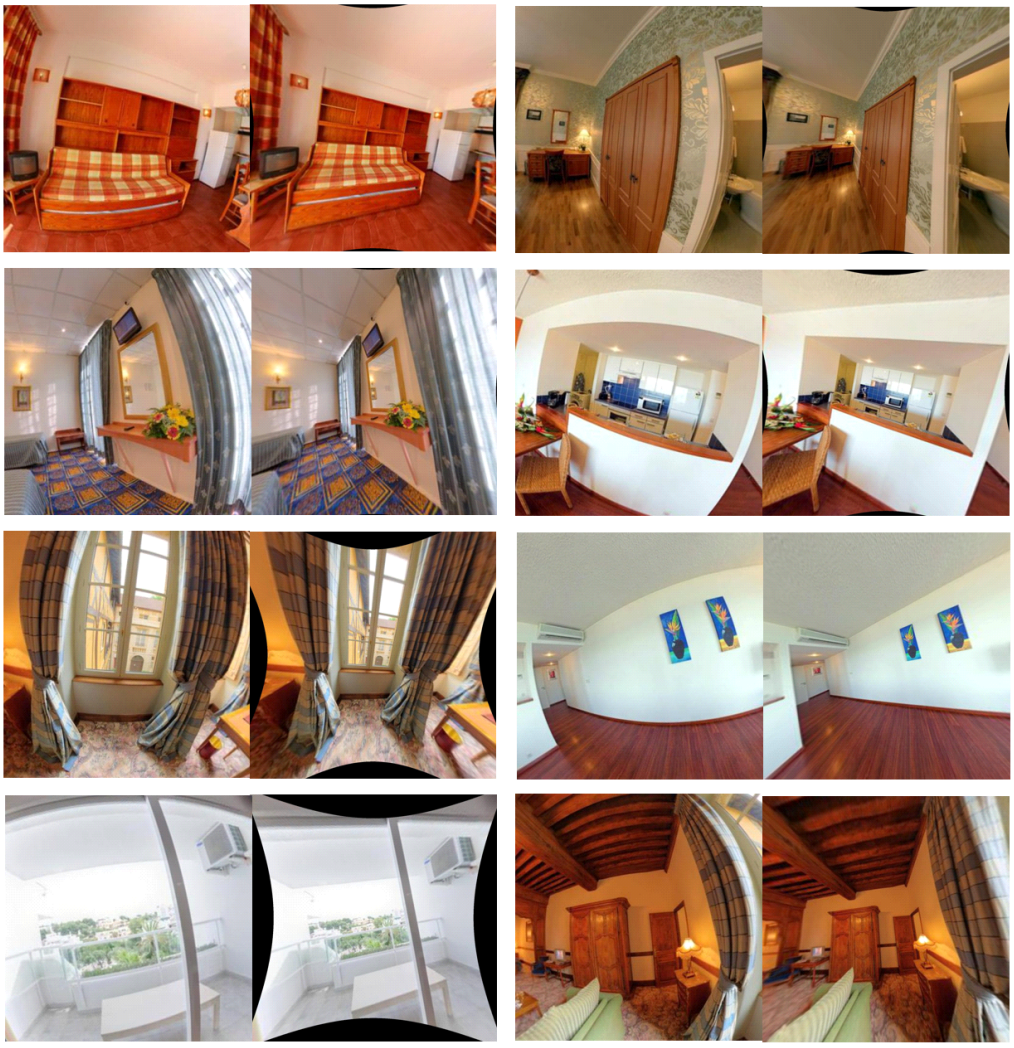

그림 8. 예측한 카메라 파라미터를 이용한 왜곡 보정 결과. 좌: 원본영상, 우: 왜곡 보정된 영상
Fig. 8. Undistortion results using predicted camera parameters. Left: Original image, Right: Undistorted image

한다. 이를 보정한 오른쪽 영상들은 휘어진 물체들이 바르게 표현되는 것을 확인할 수 있다.

Ⅴ. 결 론

기존의 딥러닝 기반 단일 영상 카메라 캘리브레이션 방법은 ImageNet dataset[12]으로 pre-train된 CNN 피처 추출기로 얻은 피처를 FFN으로 분석하여 카메라 내부 파라미터를 추정하는 방식을 사용한다. 이러한 방법은 피처 분석에 한계가 있기 때문에 정밀한 카메라 내부 파라미터를 추정하기에 적합하지 않다. 본 논문에서는 이를 해결하기 위해 피처를 patch embedding하고 Focal Token, Principal Point Token, Distortion Token을 추가하여 Vision Transformer[8]의 입력으로 활용하여 카메라의 내부 파라미터를 추정하는 모델을 제안하였고 실험을 통해 제안한 모델이 기존 연구들에 비해 정밀한 카메라 내부 파라미터 추정을 할 수 있음을 확인했다.

참 고 문 헌 (References)

[1] P. Moulon, P. Monasse, and R. Marlet, “Adaptive structure from motion with a contrario model estimation,” In Asian Conf. Comput.Vision, pp.257-270, Springer, 2013. Available: github.com/openMVG/openMVG/.
doi: https://doi.org/10.1007/978-3-642-37447-0_20

[2] Z. Zhang, “A Flexible New Technique for Camera Calibration,” IEEE Transactions on PatternRecognition and Machine Intelligence (PAMI), vol. 22, no. 11, 2000
doi: https://doi. org/10.1109/34.888718

[3] Bouguet, J. Y. “Camera Calibration Toolbox for Matlab.” Computational Vision at the California Institute of Technology

[4] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, “DEEPFOCAL: A method for direct focal length estimation,” in 2015 IEEE International Conference on Image Processing (ICIP), sep 2015, pp. 1369 – 1373. 2, 3, 7
doi: https://doi.org/10.1109/ICIP.2015.7351024

[5] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. DeepCalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In Proc. European Conf. Vision Media Production, pages 1 – 10, 2018.
doi: https://doi.org/10.1145/3278471.3278479

[6] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, “Deep single image camera calibration with radial distortion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11817 – 11825.
doi: https://doi.org/10.1109/CVPR.2019.01209

[7] M. Tan and Q. Le, “Efficientnetv2: smaller models and faster training,” in Proceedings of the International Conference on Machine Learning, pp. 10096 – 10106, New York, USA, July, 2021.
doi: https://doi.org/10.48550/arXiv.2104.00298

[8] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021.
doi: https://doi.org/10.48550/arXiv.2010.11929

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2012.
doi: https://doi.org/10.1145/3065386

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, et al., “Rethinking the inception architecture for computer vision,” in CVPR, 2016, pp. 2818 – 2826.
doi: https://doi.org/10.48550/arXiv.1512.00567J.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks.” In CVPR, vol. 1, no. 2, 2017, p. 3.
doi: https://doi.org/10.48550/arXiv.1608.06993

[12] Deng and Dothers, “ImageNet: A Large-Scale Hierarchical Image Database,” in CVPR, 2009.10]
doi: https://doi.org/10.1109/CVPR.2009.5206848

[13] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A Survey on Visual Transformer. arXiv, 2020.
doi: https://doi.org/10.1109/TPAMI.2022.3152247

[14] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. arXiv:2106.04560, 2021.
doi: https://doi.org/10.48550/arXiv.2106.04560

[15] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, “Recognizing scene viewpoint using panoramic place representation,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2695 – 2702
doi: https://doi.org/10.1109/CVPR.2012.6247991

[16] C. Zhang, F. Rameau, J. Kim, D. M. Argaw, J.-C. Bazin, and I. S. Kweon, ‘‘DeepPTZ: Deep self-calibration for PTZ cameras,’’ in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 1030 – 1038.
doi: https://doi.org/10.1109/WACV45572.2020.9093629

저 자 소 개

이 승 용

- 2018년 3월 ~ 현재 : 세종대학교 전자정보통신공학과
- ORCID : https://orcid.org/0009-0005-7843-1083
- 주관심분야 : 영상 신호처리, VR

한 종 기

- 1992년 : KAIST 전기및전자공학과 공학사
- 1994년 : KAIST 전기및전자공학과 공학석사
- 1999년 : KAIST 전기및전자공학과 공학박사
- 1999년 3월 ~ 2001년 8월 : 삼성전자 DM연구소 책임연구원
- 2001년 9월 ~ 현재 : 세종대학교 전자정보통신공학과 교수
- 2008년 9월 ~ 2009년 8월 : University California San Diego (UCSD) Visiting Scholar
- ORCID : https://orcid.org/0000-0002-5036-7199
- 주관심분야 : 비디오 코덱, 영상 신호처리, 정보 압축, 방송 시스템