특집논문 (Sepecial Paper)
방송공학회논문지 제27권 제1호, 2022년 1월 (JBE Vol.27, No.1, January 2022)
https://doi.org/10.5909/JBE.2022.27.1.69
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Normal map 생성을 이용한 물질 이미지 분류

남 현 길[a)], 김 태 현[a)], 박 종 일[a)‡]

Material Image Classification using Normal Map Generation

Hyeongil Nam[a)], Tae Hyun Kim[a)], and Jong-Il Park[a)‡]

요 약

본 연구에서는 이미지 물질의 표면의 특성을 나타내는데 사용되는 노말 맵(normal map) 이미지를 생성하고, 이를 활용하여 원본 물질 이미지의 분류 정확도를 향상시키는 방법을 제안한다. 우선, (1) 이미지 내에서 물질의 표면 특성을 반영하고 있는 노말 맵을 생성하기 위해서 Generator로 Attention-R2 Gate를 적용한 U-Net을 사용하고, 생성된 노말 맵과 원본 노말 맵의 유사도를 Reconstruction loss로 활용한 Pix2Pix 기반의 방법을 사용하였다. 그 다음으로 (2) 앞서 만들어진 노말 맵 이미지를 분류 네트워크의 Attention Gate에 적용하여 원본 물질 이미지를 분류의 정확도를 개선할 수 있는 네트워크를 제안한다. 그리고 Pixar Dataset을 이용하여 생성된 노말 맵에 대해서, Ground Truth에 해당하는 노말 맵 사이의 유사도를 평가한다. 이 때, 유사도 측정 방식에 따라 다르게 적용된 reconstruction loss function의 결과를 비교한다. 또한 물질 이미지 분류에 대한 평가를 위해서 MINC-2500과 FMD 데이터셋을 기준으로 제안된 방법과 선행연구의 비교 실험을 통해 보다 정확하게 구분할 수 있음을 확인하였다. 본 논문에서 제안된 방법은 이미지 내에서 물질을 파악하는 할 수 있는 다양한 이미지 처리 및 네트워크 구축에 기반이 될 수 있을 것으로 기대된다.

Abstract

In this study, a method of generating and utilizing a normal map image used to represent the characteristics of the surface of an image material to improve the classification accuracy of the original material image is proposed. First of all, (1) to generate a normal map that reflects the surface properties of a material in an image, a U-Net with attention-R2 gate as a generator was used, and a Pix2Pix-based method using the generated normal map and the similarity with the original normal map as a reconstruction loss was used. Next, (2) we propose a network that can improve the accuracy of classification of the original material image by applying the previously created normal map image to the attention gate of the classification network. For normal maps generated using Pixar Dataset, the similarity between normal maps corresponding to ground truth is evaluated. In this case, the results of reconstruction loss function applied differently according to the similarity metrics are compared. In addition, for evaluation of material image classification, it was confirmed that the proposed method based on MINC-2500 and FMD datasets and comparative experiments in previous studies could be more accurately distinguished. The method proposed in this paper is expected to be the basis for various image processing and network construction that can identify substances within an image.

Keyword : Normal map, Image classification, Material property

Ⅰ. 서 론

물질 이미지의 색상, 모양 등 다양한 정보를 이용하여 이미지 내에서 물질을 분류하는 이슈는 그동안 많은 시도가 있었다[1,2,3]. 이미지 분류 이슈는 이미지 만으로 물질 정보를 인식하거나 현실 정보에 기반하여 산업으로의 적용가능성이 높은 이슈이기도 하다. 이미지 내에서 물체를 포착한 정보를 이용하여 장면 내에 물질이 포함되어 있는 지의 여부에 따라서 다양한 응용 프로그램들을 만들 수 있다.

이미지 내에서의 물질 분류 이슈는 이미지 내에 객체의 분류 및 추적 등과 같은 맥락에서 함께 다루어 져왔다. 기존의 선행연구에서는 딥 러닝을 기반으로 이러한 분류 문제를 해결하기 위해서, 원본 이미지를 무작위로 이동, 회전, 리라이팅 등의 데이터 첨가(Data augmentation) 방법들을 적용하여 분류 네트워크에 반영하였다. 본 연구에서는 선행연구와 달리 이미지 내의 물질 특성을 반영하여 물질 분류 문제를 해결해보고자 한다. 그래서 거칠기 등 물질의 표면 특성을 나타내는데 효과적인 노말 맵(normal map)을 이용하여 원본 영상과 함께 물질 분류를 위한 입력으로 반영하였다. 그리고 생성된 노말 맵의 특징을 가이드로 하여 원본 물질 이미지의 분류 정확도를 개선하는 딥러닝 기반의 분류 네트워크를 제안한다.

보다 구체적으로, 물체의 표면을 표현하는데 특화되어 있는 노말 맵을 원본 영상으로부터 생성하여 물질 분류에 반영하였다. 원본 이미지를 Conditional GAN(Generative Adversarial Network)에 해당하는 Pix2Pix 방법을 이용하여, 노말 맵을 생성하였다. 이 때, 선행연구와 달리 Pix2Pix에 적용되는 생성자에 attention-R2 gate를 적용한 U-Net과 학습 과정에 필요한 reconstruction loss function에는 영상의 구조와 텍스처 유사성을 나타내는 DISTS(Deep Image Structure and Texture Similarity)을 loss function으로 활용한다. 이 과정에서 Pixar 데이터셋을 이용하여, DISTS 이외의 다른 유사도 metric들을 loss function으로 적용하여 비교해본다.

그 다음으로, 원본 물질 이미지의 분류 정확도를 향상시키기 위해서, 네트워크를 더욱 깊게 넓게 만들면서도 빠르게 학습할 수 있는 Efficient Net 방법을 기반으로 생성된 노말 맵과 원본 물질 이미지의 특징들을 추출한다. 그리고 생성된 노말 맵의 특징들을 attention gate에 적용하여 원본 이미지가 더욱 잘 학습 될 수 있도록 반영하여 분류기 성능을 개선하고자 한다. 제안된 방법의 효과를 파악하기 위해서, 선행연구의 네트워크 방법들과 분류의 정확도를 이용하여 비교 평가를 진행한다. 그리고 제안된 방법을 대표적인 물질 분류 데이터 셋인 MINC-2500 데이터 셋과 FMD (Flickr Material Database) 데이터 셋을 이용하여 선행연구의 방법들과 비교 평가를 수행한다.

본 연구에서는 이미지 내에서 물질을 분류할 때 원본 영상으로부터 물질과 관련된 더 많은 정보를 확보하고 활용할 수 있는 방법론이 되도록 하였다. 이를 통해서 이미지 내에서 분류 하고자 하는 물질과 관련된 특성들을 분류 문제를 해결하는데 활용할 수 있는 기반 연구가 될 수 있을 것으로 판단된다.

본 연구의 Contribution은 다음과 같다.

- 원본 물질 이미지를 이용해 노말 맵으로 생성하기 위해서, attention-R2 gate를 포함한 U-Net을 생성자로 하고, DISTS(Deep Image Structure and Texture Similarity) metric을 Reconstruction loss function에 적용하여 Pix2Pix 기반의 노말 맵을 생성 방법을 개선함
- 생성된 노말 맵의 특징을 가이드로 하여 원본 물질 이미지 분류 정확도를 개선할 수 있는 학습 모델을 제안함
- 비교 실험을 통해서 해당 방법이 노말 맵을 유사도 높게 노말 맵을 생성할 수 있으며, 이를 이용해서 물질 분류 방법에 효과적임을 실험으로 확인함

본 연구의 구성은 다음과 같다. 2장에서는 관련 연구들을 살펴보고, 3장에서는 본 연구의 구체적인 방법론에 대해서

a) 한양대학교 컴퓨터소프트웨어학과(Department of Computer Science, Hanyang University)
‡ Corresponding Author : 박종일(Jong-Il Park)
E-mail: jipark@hanyang.ac.kr
Tel: +82-2-2220-4368
ORCID: https://orcid.org/0000-0003-1000-4067
※이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019R1A4A1029800).
※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(NRF-2019R1A4A1029800).
· Manuscript received December 2, 2021; Revised January 12, 2022; Accepted January 19, 2022.

설명한다. 그 다음으로 4장에서는 제안한 방법을 비교 실험을 통해서 그 효과를 확인해본다. 마지막으로 5장은 결론에 해당한다.

Ⅱ. 관련 연구

1. 노말 맵(Normal map) 생성

이미지 상에 물질의 표면 특성을 나타내는 경우에 특히나 텍스처의 노말 맵을 함께 렌더링하여 빛에 반사되었을 때 물질의 표면 재질을 입체감 있게 표현해 낼 수 있다. 노말 맵을 생성하기 위해서는 실측 기반의 방법과 딥 러닝 기반의 방법으로 크게 구분이 가능하다[4,5,6]. 우선 실측 기반의 선행연구에서는 특정한 fabric의 종류를 구분하기 위해서 옷감의 색을 나타내는 앰비언트 맵(ambient map)과 노말 맵을 함께 결합하여 옷감의 표면 특성을 분류 하고자 하였다[6]. 해당 연구에서는 각 옷감에 대한 노말 맵을 획득하기 위해서 4장 이상의 이미지들을 일정한 각도에서 촬영하여 계산하였다[6]. 그러나 옷감을 제외한 다양한 상황에서 노말 맵을 실측하여 매번 획득하기에는 현실적으로 어려움이 존재한다.

딥 러닝을 기반으로 노말 맵을 생성하는 방법은 분류 대상이 되는 단일의 이미지만으로 normal map을 획득할 수 있다. 딥 러닝을 기반으로 한 normal map 생성을 위한 방법은 Auto encoding방식의 U-Net을 형태의 생성 방법이 있었다[4,5]. 해당 방법은 Encoding 과정에서 발생하는 정보 손실을 보완하기 위해서 Up-sampling 시에 Encoding에서 추출된 정보들을 반영하여 정보의 손실을 줄이고자 하였다[4,5]. 더 나아가서는 GAN(Generative Adversarial Networks)를 활용한 연구 또한 있었다[7]. 여러 GAN model들 중에서도 선행연구는 Conditional GAN 기반으로 이미지 도메인 간의 변환에 용이한 Pix2Pix를 사용하였다[7]. Pix2Pix의 모델 구조는 기본적으로 생성자(Generator)는 U-Net을 구조를 사용한다[8]. 또한 선행연구에서 Pix2Pix와 일반적인 U-Net의 방법을 이용하여 노말 맵을 생성하였을 때를 비교하였을 때, Pix2Pix로 생성된 노말 맵이 Ground truth에 해당하는 노말 맵과 더 유사함을 알 수 있다[7]. 본 연구에서도 선행연구와 마찬가지로 U-Net을 생성자로 하여 Pix2Pix를 사용하여 normal map을 생성하고자 한다[7]. 그러나 U-Net구조를 사용하는 generator의 성능 향상을 위해서 U-Net구조에 attention block과 R2 Gate를 반영하여 선행연구 보다 성능을 향상시키고자 하였다. Attention gate를 반영한 U-Net은 Decoding될 때, skip-connection 부분에서 Attention gate를 적용하여 학습 되어야 할 위치를 더욱 명확하게 되도록 하는 결과를 확인할 수 있다[9]. 그리고 R2(Recurrent Residual) gate를 Encoding과 Decoding 시에 반영하여 깊이 있는 학습이 가능하도록 하였다[10,11]. 또한, 이미지를 생성하거나 변환하는 딥 러닝 모델에서 고려되는 Reconstruction loss를 어떻게 활용하느냐에 따라 다르게 성능에 차이가 발생한다[12,13]. 따라서 본 논문에서는 기존의 U-Net에서 사용되는 L1 loss function이 아닌, 노말 맵의 구조와 텍스쳐의 유사도를 나타내는 DISTS(Deep Image Structure and Texture Similarity)을 Reconstruction loss function으로 사용하여 Ground Truth와 더욱 유사한 노말 맵을 생성하고자 한다[14]. 그리고 이렇게 생성된 normal map을 원본 이미지와 함께 분류 시에 반영함으로써 분류의 정확도를 향상시키고자 한다.

2. 물질 이미지 분류 네트워크

이미지를 분류하기 위한 딥 러닝 기반의 여러 가지 네트워크들이 존재한다[1]. CNN(Convolutional Neural Network)를 중심으로 이미지를 처리할 수 있는 방대한 방법들이 제시되었는데, AlexNet에서부터 ResNet을 거쳐 이미지 분류의 정확도가 점차 높아져왔다[2,3]. 다수의 CNN을 이용한 이미지 분류 네트워크는 여러 단계의 Convolutional layer와 Pooling layer를 거쳐 이미지 내에서 특징들을 뽑아내고 이를 최종적으로 Fully connected layer를 거쳐서 어떠한 이미지 인지 판별을 하게 된다[1].

딥 러닝 기반의 학습 모델들은 깊고 넓게 하여 학습의 결과를 향상시키기 위해서 시키고자 한다[10]. ResNet은 지금도 널리 사용되는 네트워크로 Residual Block을 중심으로 네트워크의 아키텍처가 구성되어 있다[3]. Residual block은 3x3 Convolution이 반복되는 과정에서, Convolution 연산 사이에 출력 결과에 대해 identity shortcut에 해당하는 입력 특징 맵을 그대로 더해 주거나 linear projection을 한 후에 더해주는 방식을 사용하게 된다[3]. 이러한 방식을 포

함한 Residual block을 사용함으로써, 모델의 Depth scaling을 확대시키는 것과 동시에 vanishing gradient에 강인 해지고, 학습의 결과도 향상되었다. 또한 Gpipe 라이브러리는 학습 시에 더욱 방대하게 하기 위해, 여러 레이어를 병렬적으로 처리하여 거대한 크기의 학습 모델을 효율적으로 학습이 가능하도록 하였다[15]. 나아가서 최근 학습 모델의 깊이, 너비, 입력 이미지의 해상도를 효율적으로 조절할 수 있는 Compound scaling 방법을 제안한 Efficient Net 기반의 연구들이 이미지 분류를 비롯한 다양한 분야에 적용되고 있다[11]. Compounding scaling 방식은 baseline에 해당하는 모델의 깊이, 너비, 입력 이미지의 해상도를 조정하기 위해서 각 각을 α, β, ɣ 의 비율로 균등하게 증가시키는 방법으로 효과적인 CNN의 학습이 가능해진다[11]. 이러한 Efficient Net 방법은 ResNet이나 MobileNet 등의 CNN 기반 구조에서도 효과적인 결과를 보여주고 있다[11]. 이렇게 본 연구에서도 Mobile-size baseline모델을 사용하는 Efficient Net 방법을 적용하여 원본 이미지와 노말 맵의 특징들을 각각 추출하지만, 선행연구와 달리 노말 맵의 특징을 가이드로 하여 원본 물질 이미지의 분류 정확도를 향상시키고자 attention gate를 활용한다.

Ⅲ. 분류 방법

1. 전체 아키텍처

본 연구에서 제안하는 이미지 내의 물질을 분류하는 전체 시스템의 구조는 [그림 1]과 같다. 먼저 (1) 원본 물질 이미지를 이용하여 물질의 표면 특성을 표현할 수 있는 노말 맵을 생성한다. 그 다음으로, (2) 원본 물질 이미지와 생성된 노말 맵에서 각 특징들을 추출하고, 이를 활용하여 분류 결과를 도출하는 형태로 구성되어 있다.

2. 노말 맵(Normal map) 생성

RGB 이미지에서 Pix2Pix를 기반으로 노말 맵을 생성할 수 있도록 하였다. Conditional GAN의 구조를 가지고 있는 Pix2Pix 모델은 일반적으로 생성자의 loss function에는 L1(reconstruction loss)와 cGAN(adversarial loss)을 이용한다[8]. 이 때 discriminator에서는 PatchGAN을 사용하여 patch 단위로 생성된 이미지인지, ground-truth 이미지인지를 판단하고 이 차이를 평균으로 계산하여 고주파 영역에 집중하여 학습되도록 한다. 그리고 생성자에서 생성된 이미지와 ground-truth 이미지 사이의 차이인 Reconstruction loss를 L1 loss를 사용하여 저주파 영역이 학습에 용이하도록 적용된다[13]. 그러나 본 연구에서는 생성된 이미지 형태의 map을 비교할 수 있는 구조와 텍스쳐의 유사도를 나타내는 DISTS(Deep Image Structure and Texture Similarity)을 Reconstruction loss function으로 적용하였다[그림 2 (a)][14]. 이를 통해서 자연물이나 옷감 등의 물질 이미지 내에서 나타나는 패턴들에 대해 보다 효과적으로 학습 될 것이라 판단된다. 기본적으로 이미지 품질 평가를 위해서 개발된 알고리즘은 DISTS은, 사람이 이미지 인지 시에 이미지의 밝기, 콘트라스트 등의 구조 변화에 민감하다는 것에 착안한 기존 방법에 나아가서, 텍스쳐 리샘플링(texture re-sampling)을 허용하여 더 사람이 지각하는 방식과 유사한 metric 방법이다[14]. DISTS 알고리즘은 먼저 원본 이미지($\tilde{x}_j^{(i)}$)와 비교 할 이미지($\tilde{y}_j^{(i)}$)를 사전 학습된 VGG16을

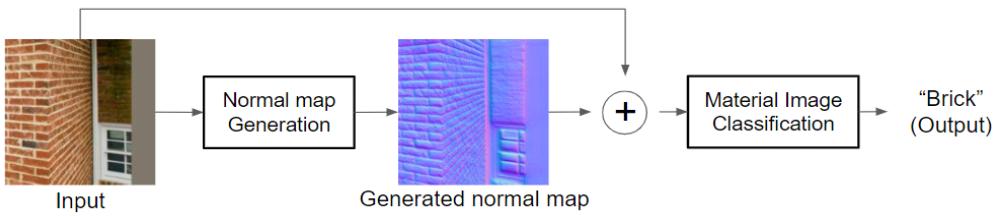


그림 1. 생성된 노말 맵을 이용한 물질 이미지 분류를 위한 전체 시스템
Fig. 1. The overall system for classifying material images using the generated normal map.

거쳐서 얻은 각 각의 특징 맵들을 이용한다. 여기서 i는 총 m개의 convolution layer 중에 i번 째, j는 i번째 convolution layer의 특징 맵 개수(최대 n_i개)를 나타낸다. 그리고 이를 이용해 크게 2가지의 metric으로 유사도를 측정한다. 하나는 수식 (1)과 같이, 원본 이미지의 특징 맵과 비교할 이미지의 특징 맵의 평균을 비교하여 텍스쳐의 유사도(l)를 측정하고, 다른 하나는 수식 (2)와 같이 각 특징 맵들의 상관관계(correlation)을 비교하여 구조의 유사도(s)를 측정한다. 구체적인 수식은 다음과 같다.

$$l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\mu_{x_j}^{(i)}\mu_{y_j}^{(i)} + c_1}{(\mu_{x_j}^{(i)})^2 + (\mu_{y_j}^{(i)})^2 + c_1} \quad (1)$$

$$s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\sigma_{x_j y_j}^{(i)} + c_2}{(\sigma_{x_j}^{(i)})^2 + (\sigma_{y_j}^{(i)})^2 + c_2} \quad (2)$$

여기에서 $\mu_{x_j}^{(i)}$, $\mu_{y_j}^{(i)}$는 각 각 x와 y 이미지에 대한 I 번째 convolution layer의 j 개의 평균을 나타내고, $\sigma_{x_j}^{(i)}$, $\sigma_{y_j}^{(i)}$는 각 각 x와 y 이미지에 대한 I 번째 convolution layer의 j 개의 공분산을 이용한다. 그리고 c_1, c_2는 0으로 나눠지는 것을 방지하기 위한, 0에 가까운 작은 constant 값에 해당한다. 결과적으로 입력 이미지를 비교하여 VGG를 통해서 나오게 되는 각 특징 맵에 대한 l와 s의 점수를 전체를 이용하여 metric을 구한다. 이를 이용해 수식 (3)과 같이 가중치를 부여해서 모두 합 값을 1에서 빼서 최종 점수를 얻게 된다. 전체 점수 D를 얻기 위한 수식은 다음과 같다.

$$D(x,y;\alpha,\beta) = 1 - \sum_{i=0}^{m}\sum_{j=1}^{n_i}(\alpha_{ij}l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) + \beta_{ij}s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)})) \quad (3)$$

여기서 가중치는 α_{ij}과 β_{ij}은 학습을 통해서 얻게 된다. 이 때의 α_{ij}와 β_{ij}는 $\sum_{i=0}^{m}\sum_{j=1}^{n_i}(\alpha_{ij}+\beta_{ij}) = 1$을 만족한다. 또한 α_{ij}와 β_{ij}를 학습시킬 때 두 개의 손실 함수를 합한 것을 사용하는데 이는 각 각 이미지 품질 평가, 텍스쳐 분류 데이터셋과 연관된 것이다. 이렇게 되면, ground-truth 이미지 품질 라벨과 예측된 D 사이의 절대 차가 작아지도록 α_{ij}와 β_{ij}가 조정되는 동시에, 텍스쳐가 같은 것에 대해서는 D가 작아지

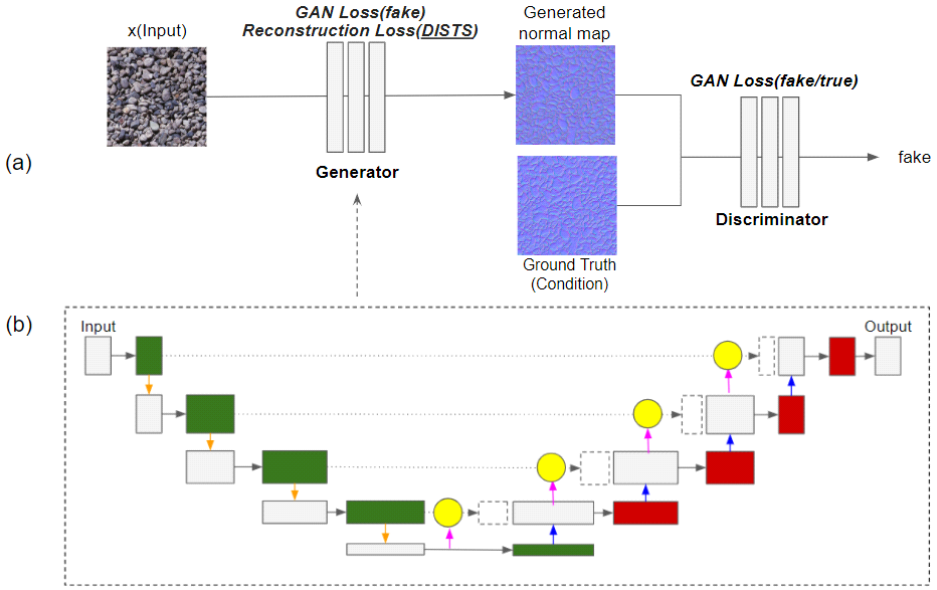


그림 2. (a) 노말 맵 생성을 위한 모델의 구조, (b) 생성자에 해당하는 Attention-R2(Residual-Recurrent) gate를 포함한 U-Net(초록색 박스: ReLU를 포함한 Recurrent Conv, 빨간색 박스: ReLU를 포함한 Recurrent Up sampling(Conv), 주황색 선: Max-pooling, 파란색 선: Up-Conv, 분홍색 선: 결합, 노란색 동그라미: Attention gate, 점선: 인코딩 과정의 입력 결합)

Fig. 2. (a) Structure of model for generating normal maps, (b) U-Net model including Attention-R2(Residual-Recurrent) gate as a generaoter(Green Box: recurrent conv including ReLU, red box: recurrent up sampling(conv) including ReLU, orange line: max-pooling, blue line: up-conv, pink line: concatenation, yellow circle: attention gate, dot line: input combination of encoding process).

도록 α_{ij}와 β_{ij}가 조정된다. 이렇게 되어서 구조적 변화에 민감함을 유지하면서 텍스쳐 리샘플링을 허용할 수 있게 된다.

나아가서 학습 과정에서 본 연구에서는 선행연구와 달리 generator에 해당하는 U-Net 네트워크에 Attention block과 Repeated Recurrent(R2)를 넣어서 정보 손실이 줄이게 된다[그림 2 (b)]. 보다 자세하게는 Attention U-Net은 Up-sampling될 입력 특징들과 동일한 단계에서 이전에 Encoding과정의 특징들 결합하고 ReLU, (1x1x1 conv), Sigmoid 함수를 거쳐 입력 특징과 동일한 특징 맵 크기의 리샘플러를 만든다. 이렇게 만들어진 리샘플러를 Encoding 과정의 특징들에 곱해져서, Up-sampling될 입력 특징과 결합되어 단계별로 학습이 진행된다[9,11]. 그리고 R2는 Residual과 Recurrent를 지칭하는데, Encoding과 Decoding시에 Recurrent Convolution Network를 적용하여 보다 깊이 있는 학습이 가능하도록 하였다[10,11]. 이러한 점들을 모두 적용한 Attention Recurrent Residual U-Net을 생성자로 하여 원본 이미지의 고주파, 저주파 영역의 특성 및 정보 손실을 최소화하여 영상을 노말맵으로 변환할 수 있었다. 노말 맵 생성을 위한 학습의 데이터 셋은 Pixar Texutres를 사용한다[16]. 총 231개의 color RGB image(256x256)와 노말 맵(256x256)의 각 각 입력과 출력 조건에 해당하는 데이터 쌍을 학습 시에는 80%의 데이터를 무작위로 뽑아 훈련 데이터로 사용한다.

3. 물질 이미지 분류를 위한 네트워크

원본 이미지와 생성된 노말 맵을 이용하여 효과적인 물질 분류 네트워크를 구축하기 위해서는 원본 이미지와 노말 맵을 결합하여 단일의 합성 데이터를 입력으로 두는 것이 아닌, 물질의 표면 특성과 원본 이미지의 추출된 특징들이 누락되지 않도록 하는 것이 중요하다. 따라서 제안하는 네트워크는 원본 이미지의 특징과 물질의 표면 특성을 나타내는 노말 맵 이미지에서 각 각 특징들을 추출한다. 그리고 생성된 노말맵을 분류의 정확도를 향상시킬 수 있는 가이드 역할을 할 수 있도록 네트워크를 구성한다.

먼저, 이미지에서 특징을 추출하기 위해서는 EfficientNet을 활용한다. 선행 연구에서 제시된 여러 형태의 Efficient Network들 중에서 ImageNet 데이터셋에서 높은 성능

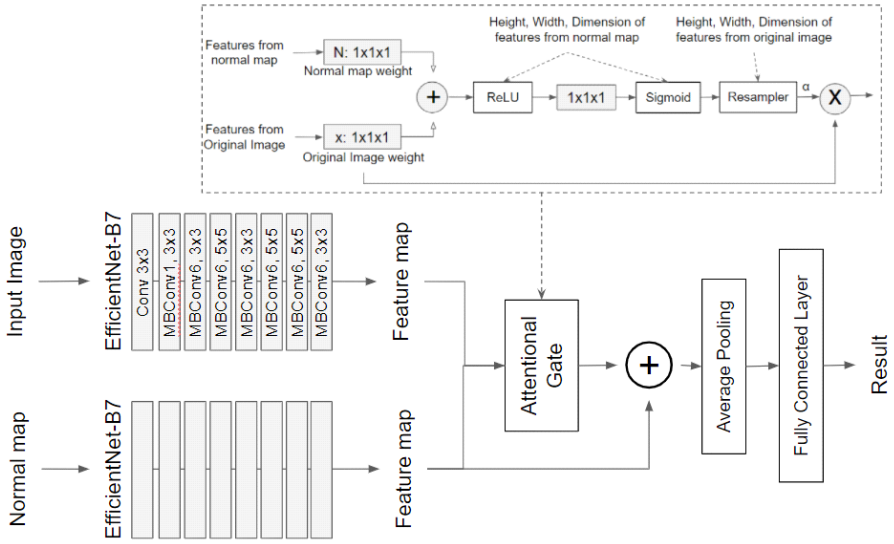


그림 3. 생성된 노말맵을 이용한 물질 이미지 분류 네트워크. 먼저 EfficientNet- B7을 이용하여 원본 이미지와 노말 맵에서 각 각 추출하고, 노말 맵에서 추출된 특징을 이용해서 학습할 원본 미지를 더 정확하게 초점을 맞출 수 있도록 리샘플러를 활용하고, 특징 맵을 원본 이미지와 결합하여 분류 작업을 수행함.
Fig. 3. Material image classification network using the generated normal map. First, the original image and the normal map are extracted using EfficientNet-B7, the resampler is used to focus more accurately on the original image to be learned using the features extracted from the normal map, and the classification is performed by combining the feature map with the original image.

을 보여주었던, EfficientNet-B7을 기준으로 본 연구에서도 각 이미지의 특징들을 추출한다[11]. 먼저, Efficient Net은 compound scaling 방법으로 특정한 비율로 전체 네트워크의 깊이, 너비, 이미지의 해상도를 균등하게 증가시킬 수 있는 compound coefficient ϕ를 활용한다[11]. 기본적으로 Efficient Net의 아키텍처는 Mobile inverted Bottleneck Convolution(MBConv)를 기본 구성요소 단위로 가지는데, MBConv는 입력을 확대시키고 줄이는 과정에서 Depth-wise Conv를 거친 데이터를 다시 확장시킨 결과에 곱해주고 최적화(excite)를 거치게 된다. 이러한 MBConv block들을 EfficientNet의 baseline network들은 각기 다른 형태로 구성되는데, 본 연구에서 적용된 EfficientNet-B7은 [그림 3]과 같이 filter의 크기, stride, channel수에 따라 7개의 서로 다른 MBConv block으로 나누어진다.

그 다음으로 원본 이미지와 생성된 노말 맵에서 추출된 각 2560 채널의 데이터들을 attentional gate와 average pooling layer를 각각 거친 후에 결합을 진행하고 최종적으로 fully connected layer를 통해서 각 이미지 특징들이 가지고 있는 정보를 바탕으로 물질 이미지가 분류되도록 한다. 보다 구체적으로 [그림 3]에서의 attentional gate를 보면 노말 맵과 원본 이미지 특징들을 개별 1x1x1 변환을 하고, ReLU함수에 넘어가게 되고, 다시 1x1x1를 수행하여 Sigmoid 함수를 거치게 된다. 그리고 복셀 방식의 마스크 타입이 생성되어, 리샘플러를 통해 원본 이미지 특징들의 크기와 동일한 피쳐 맵 크기를 만들어서, 원본 이미지 특징들과 마지막으로 결합되어 다음 단계로 넘어가게 된다. 이를 통해서 원본 이미지만이 아닌 물질의 특성을 나타내는 맵의 형태로 변환된 이미지를 함께 분류 작업에 포함시켜 학습함으로써, 분류에 대한 더 많은 정보들을 활용할 수 있게 된다.

Ⅳ. 비교 평가를 위한 실험

1. 각기 다른 Reconstruction loss function에 따라 생성된 노말 맵의 유사도 평가

선행연구에서 Pix2Pix를 이용하여 이미지를 생성할 때, Reconstruction loss function으로 L1을 사용하였으나, 본 연구에서는 구조적 유사도를 나타내는 DISTS (Deep Image Structure and Texture Similarity)을 Reconstruction loss function으로 적용한다. 이 때, Reconstruction loss function에 따라서, 생성된 노말 맵이 얼마나 ground-truth의 이미지와 유사한 지를 Ablation Study를 통해서 평가해보고자 한다. 각기 다른 Reconstruction loss function으로는 (1) L1 loss, (2) DISTS, (3) SSIM(Structural Similarity Index measure), (4) DISTS + SSIM 을 비교 평가한다[8,17,14]. 각 방법은 (1)은 L1 distance를 이용한 loss function, (2)는 본 연구에 적용된 metric으로 이미지의 구조 변화에 민감하면서 텍스쳐 리샘플링을 감안하고, (3)은 이미지의 구조적 유사도를 밝기, 콘트라스트, 구조를 고려한 metric, (4)는 (2)와 (3)을 모두 적용하여 loss function을 구성하였다. 학습에 적용되었던 Pixar 데이터 셋의 테스트 케이스들(46개)을 이용하여 비교 평가가 가능했다.

각 loss function들로 생성된 노말 맵들과 ground-truth에 해당하는 노말 맵 간의 결과를 비교 평가하기 위한 방법으로는 크게 3가지를 이용하였다. 앞서 이미지의 유사도를 나타내는 척도인 SSIM과 DISTS의 점수 뿐만 아니라, LPIPS (Learned Perceptual Image Patch Similarity)로 비교한다[18]. LPIPS는 딥러닝을 통해서 나온 결과를 이용하여, perceptual similarity를 비교한 것으로, 본 연구에서는 AlexNet을 통해 나온 결과를 이용하여 distance 점수를 평가하였다. 이 때, 점수 범위는 모두 0에서 1사이의 점수로 계산하였다. 그리고 DISTS와 LPIPS의 점수는 0에 가까울수록(점수가 낮을 수록), SSIM의 점수는 1에 가까울수록(점수가 높을 수록) ground-truth의 노말 맵과 유사도가 높다.

[그림 4]에서는 loss function이 달리 적용되었을 때, 각각 테스트 케이스들에 대해 (위-오른쪽)은 DISTS 점수를 나타내고, (위-왼쪽)은 LPIPS 점수, (아래)는 SSIM 점수를 나타낸다. 그 결과, 초록색을 나타내고 있는 DISTS을 Reconstruction loss function으로 적용했을 때, 전체적으로 노말 맵 결과 유사도 평가 metric으로써 적용된 DISTS, LPIPS 점수가 낮았으며 SSIM 점수가 높게 나타났다. 다시 말해 ground-truth 노말 맵과 유사도가 높은 것을 확인할 수 있다. 또한 [표1]에서는 해당 테스트 케이스에 대해서 각 metric의 점수의 평균 점수를 나타낸 결과이다. 여기서도 역시, DISTS를 Reconstruction loss function으로 적용했을 때, ground-truth와 가장 높은 유사도를 확인 할 수 있다.

표 1. Reconstruction loss에 각기 다른 metric이 적용되어 생성된 노말 맵과 ground-truth 노말 맵 간의 평균 유사도 비교(소수점 넷째 자리에서 반올림)
Table 1. Comparing the average similarity between the ground-truth normal map and the normal map generated by applying different metric to the reconstruction loss(Rounded up from the 4th decimal place).

Reconstruction loss function	Similarity Index (Higher similarity condition)		
	DISTS (closer to 0)	LPIPS (closer to 0)	SSIM (closer to 1)
L1	0.246	0.172	0.547
DISTS(Applied in this)	0.160	0.110	0.750
SSIM	0.275	0.190	0.471
DISTS + SSIM	0.173	0.122	0.737

2. 물질 이미지 분류 정확도 평가

본 연구에서 제안한 물질 이미지 분류의 효과를 입증하기 위해서 선행연구에서 적용된 이미지 분류 방법들을 동일한 데이터셋에 적용하여, 그 효과를 확인하고자 한다. 실험에 적용된 데이터 셋은 MINC-2500과 FMD(Flickr Material Database)에 나타난 물질의 형태를 표현하기로 하였다[19,20]. MINC-2500 데이터셋은 다양한 표면 텍스처, 기하학적 조건, 조명 조건 등이 결합한 일상 상황의 장면들에서

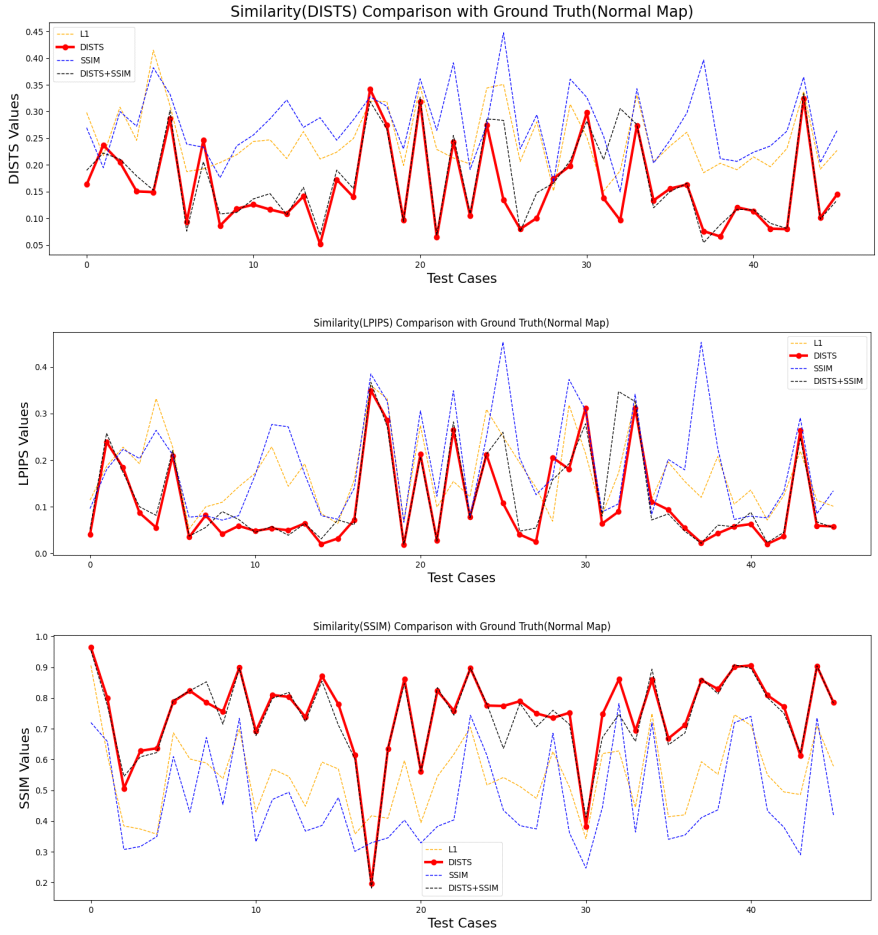


그림 4. Reconstruction loss에 각기 다른 metric이 적용되어 생성된 노말 맵과 ground-truth에 해당하는 노말 맵 간의 유사도 비교. (위): DISTS 점수(0에 가까울수록 유사), (중간): LPIPS 점수(0에 가까울수록 유사), (아래): SSIM 점수(1에 가까울수록 유사). 그래프 공통적으로 주황색 점선: L1 loss function, 빨간색 실선: DISTS loss function(본 연구에 적용된 metric), 파란색 점선: SSIM loss function, 검정색 점선: DISTS + SSIM loss function
Fig. 4. Comparison of similarity between normal maps generated by applying different metric to reconstruction loss and normal maps corresponding to ground-truth. (above): DISTS score (0 closer to ground-truth), (intermediate): LPIPS score (0 closer to ground-truth), (below): SSIM score (1 closer to ground-truth). Orange dotted line: L1 loss function, red solid line: DISTS loss function, blue dotted line: SSIM loss function, black dotted line: DISTS + SSIM loss function

물질을 분류해야 하기 때문에, 비교 실험에 적합한 데이터 셋이라고 판단하였다[19]. 데이터셋은 총 57,500장의 이미지 데이터 중 23개의 클래스로 구분되어 있다. 그 중에서 40%(23,000장)의 이미지에 대해 테스트 데이터로 사용하고 나머지 데이터를 트레이닝에 반영하였다. 또한 FMD 데이터 셋의 경우에는 총 10개의 클래스에 대해서 100장의 이미지가 있어, 총 10,000장 이미지 중에 30%(300장)의 이미지에 대해 테스트 데이터로 이용하고, 나머지를 트레이닝 데이터로 적용하였다.

비교 실험을 위해서 선행연구들에서 제안된 분류 모델들을 기반으로 동일한 실험 조건에서 학습을 진행하였다. 먼저 Attention Module을 이용하여 이미지에서 중요한 특징을 포착하여 출력에 해당하는 값을 정제하고자 하였던, Residual Attention Network를 사용하였다[21]. 또한 Texture Encoding Network의 발전된 형태로 나무 줄기와 같이 복잡한 순서의 디테일 등에 불리한 점들을 보완하기 위해서 등장한 Deep Encoding Pooling Network[22]를 비교 대상으로 잡았다. 그리고 이미지를 patch 단위로 나누어서 각각을 Encoding 하는 방식을 취하여 이미지를 인식 할 수 있는 Visual Transformer 방법과 각기 다른 patch 크기로 Visual Transformer에 대해 적용하였던 Cross visual transformer 방법을 이용한다[23,24].

또한 원본 이미지만을 이용하여 학습한 EfficientNet의 결과 비교 대상으로 선정하였다. 이렇게 EfficientNet에서 노말 맵을 사용하지 않고 추출된 특징만으로 분류를 진행한 것과 생성된 노말 맵의 가이드를 활용한 방법이 적용된 방법(제안된 방법)을 추가로 비교하도록 하였다. 실험 환경은 Nvidia의 3080 GPU를 이용하여 모두 동일하게 학습을 진행하였으며, epoch, batch size는 각각 300, 4로 설정하고 학습을 진행하였다. [표 2]에서는 해당 실험에 대한 결과는 다음과 같다. 각 네트워크의 최종적인 데이터의 선행연구에 비해서 제안된 연구가 MINC-2500 dataset에 대해서 79.204%, FMD 데이터 셋에 대해서는 76.860%로 높은 분류 정확도를 보여주었다. 해당 결과를 바탕으로 제안된 연구의 효과를 확인할 수 있었다.

표 2. MINC-2500과 FMD 데이터 셋에 대한 분류 학습의 정확도 비교(%)
Table 2. Comparison of the accuracy of classification for MINC-2500 and FMD datasets(%)

Models	Datasets	
	MINC-2500	FMD
Residual Attention Network	72.012	64.582
Deep Encoding Pooling Network	74.032	71.004
Visual Transformer	70.191	68.303
Cross Visual Transformer	72.094	71.728
Efficient Net-B7(with only original image)	78.383	77.000
Our Method(with Normal map)	79.204	76.860

V. 결 론

물질 이미지를 분류하기 위해서 본 연구에서는 원본 이미지로부터 GAN(Generative Adversarial Network)을 기반으로 한 노말 맵을 생성하여, 생성된 노말 맵을 가이드로 활용하여 원본 이미지에 대한 분류 성능을 향상시키고자 하였다. 구체적으로 먼저 원본 이미지에서 노말 맵을 생성하기 위해서는 Attention과 R2(Recurrent Residual)를 포함하고 있는 U-Net을 생성자로 하고 DISTS(Deep Image Structure and Texture Similarity)를 loss function metric으로 사용하여 Pix2Pix 성능을 향상 시켰다. 그리고 이렇게 만들어진 노말맵 이미지와 원본 이미지를 딥러닝 기반으로 분류하기 위한 물질 분류 네트워크를 제안하였다. 각 이미지를 EfficientNet을 통해서 특징들을 추출하고 생성된 노말 맵에서 추출된 특징들을 가이드로 하여 attention gate를 통과하여 원본 이미지의 분류 성능이 향상되도록 하였다.

노말 맵 생성의 성능의 파악하기 위해서 본 연구에서 적용한 DISTS 이외의 Reconstruction loss metric들로 생성된 노말 맵과 ground-truth에 해당하는 노말 맵의 유사도를 비교하였다. 그리고 물질 이미지 분류 정확도를 파악하기 위해서, MINC-2500 데이터 셋과 FMD 데이터 셋을 이용하여서 선행연구들과 정확도를 비교하였다. 그 결과 각각 79.204%, 76.860%로 높은 결과를 보여주었으며, 이는 원본 이미지와 물질의 표면 특성 정보를 함께 학습시킴으로써 물질 이미지 분류의 정확도를 향상시킬 수 있었음을 확인하였다. 본 연구로 이미지에서 물질의 다양한 정보들을 이용하여 분류를 수행할 수 있는 연구의 초석이 될 수 있을 것으로 판단된다.

참 고 문 헌 (References)

[1] He, Zhengyu. "Deep Learning in Image Classification: A Survey Report." 2020 2nd International Conference on Information Technology and Computer Application (ITCA). IEEE, pp. 174-177. 2020

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25. pp. 1097-1105. 2012.

[3] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778. 2016.

[4] GAO, DUAN & Li, Xiao & Dong, Yue & Peers, Pieter & Xu, Kun & Tong, Xin. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. ACM Transactions on Graphics. 38. pp.1-15. 2019. 10.1145/3306346.3323042

[5] Deschaintre, Valentin et al. “Single-image SVBRDF capture with a rendering-aware deep network.” ACM Transactions on Graphics (TOG) 37. pp 1 - 15. 2018.

[6] Kampouris, Christos, et al. "Fine-grained material classification using micro-geometry and reflectance." European Conference on Computer Vision. Springer, Cham, pp.778-792. 2016.

[7] Hyeongil Nam, and Jong-Il Park. " Normal map generation based on Pix2Pix for rendering fabric image", Proceedings of the Korean Society of Broadcast Engineers Conference. The Korean Society of Broadcast and Media Engineers. pp. 166-169. 2020.

[8] Isola, Phillip et al. “Image-to-Image Translation with Conditional Adversarial Networks.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967-5976. 2016.

[9] Oktay, Ozan, et al. "Attention u-net: Learning where to look for the pancreas." arXiv preprint arXiv:1804.03999 2018.

[10] So-hyun Lim and Jun-chul Chun " Image-to-Image Translation Based on U-Net with R2 and Attention" J. Internet Comput. Serv. 21.4. pp 9-16.2020

[11] Zuo, Qiang, Songyu Chen, and Zhifang Wang. "R2AU-Net: Attention Recurrent Residual Convolutional Neural Network for Multimodal Medical Image Segmentation." Security and Communication Networks 2021. 2021.

[12] Li, Yanchun, Nanfeng Xiao, and Wanli Ouyang. "Improved generative adversarial networks with reconstruction loss." Neurocomputing. 323. pp. 363-372. 2019.

[13] Shi, Haoyue, et al. "Loss Functions for Person Image Generation." BMVC. 2020.

[14] Ding, Keyan, et al. "Image quality assessment: Unifying structure and texture similarity." arXiv preprint arXiv:2004.07728. 2020.

[15] Huang, Yanping, et al. "Gpipe: Efficient training of giant neural networks using pipeline parallelism." Advances in neural information processing systems 32. pp.103-112. 2019.

[16] Pixar One Twenty Eight by Pixar Animation Studios, https://renderman.pixar.com/

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. IEEE transactions on image processing, 13(4):600, 2004.

[18] ZHANG, Richard, et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586-595. 2018.

[19] Bell, Sean, et al. "Material recognition in the wild with the materials in context database." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479-3487. 2015.

[20] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Accuracy and speed of material categorization in real-world images", Journal of Vision, vol. 14, no. 9, article 12, 2014

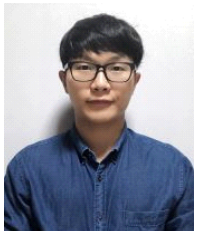
[21] Wang, Fei, et al. "Residual attention network for image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156-3164. 2017.

[22] Xue, Jia, Hang Zhang, and Kristin Dana. "Deep texture manifold for ground terrain recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 558-567. 2018.

[23] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929. 2020.

[24] Chen, Chun-Fu, Quanfu Fan, and Rameswar Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification." arXiv preprint arXiv:2103.14899. 2021.

저 자 소 개



남 현 길

- 2018년 : 한양대학교 경영학부 졸업(학사)
- 2018년 ~ 현재 : 한양대학교 컴퓨터소프트웨어학과 석박통합과정
- ORCID : https://orcid.org/0000-0002-6017-8869
- 주관심분야 : 컴퓨터 비전, 증강현실, 가상현길, HCI

김 태 현

- 2011년 ~ 2016년 : 서울대학교 전기정보공학부 박사
- 2016년 ~ 2018년 : 막스플랑크 연구소 연구원
- 2018년 ~ 현재 : 한양대학교 컴퓨터소프트웨어학부 조교수
- ORCID : https://orcid.org/0000-0002-7995-3984
- 주관심분야 : 컴퓨터비전, 머신러닝

박 종 일

- 1987년 : 서울대학교 전자공학과 졸업(학사)
- 1989년 : 서울대학교 전자공학과 졸업(석사)
- 1995년 : 서울대학교 전자공학과 졸업(공학 박사)
- 1992년 ~ 1994년 : 일본 NHK방송기술연구소 객원 연구원
- 1995년 ~ 1996년 : 한국방송개발원(현 한국콘텐츠진흥원) 선임연구원
- 1996년 ~ 1999년 : 일본 ATR지능영상통신연구소 연구원
- 1999년 ~ 현재 : 한양대학교 공과대학 교수
- ORCID : https://orcid.org/0000-0003-1000-4067
- 주관심분야 : 컴퓨터 비전/그래픽스, 증강현실 가상현실, HCI