https://doi.org/10.7840/kics.2023.48.11.1457

객체 분할 정보를 활용한 선화 생성 모델의 성능 개선

최 재 웅◆, 이 재 구°

Improving Line Drawing Generation with Instance Segmentation Information

Jaewoong Choi◆, Jaekoo Lee°

요 약

선화 생성 모델은 원본 사진을 선화 사진으로 스타일 전이(style transfer) 된 사진을 생성하는 모델이다. 기존 선화 모델은 원본 사진에 대한 선화 쌍이 없더라도 선화를 추출하는데, 이를 위해 사진에 대한 의미론적(semantic) 정보와 기하학적(geometric) 정보를 학습한다. 특히 의미론적(semantic) 정보는 CLIP(contrastive language-image pretraining) 모델을 통해 학습하며 기하학적 정보를 위해 깊이 추정(depth estimation) 방법만을 사용하였다. 하지만, 사용된 깊이 추정의 경우, 깊이에 대한 정확한 값 정보가 부족하여 깊이 추정 방법을 통해 추정된 값을 생성하여 사용하기 때문에 선화 모델의 성능이 떨어질 수 있다. 따라서 본 논문은 정확한 기하학적 정보를 사용할 수 있는 객체 분할(segmentation) 방법을 추가하여 선화 성능을 올리고자 한다. 실제 객체 분할 정보를 추가함으로써, 깊이 추정에 대한 정보가 상대적으로 부정확하다는 단점을 보완하였으며, 최종적으로 배경과 음영 등의 정보도 기하학적 정보를 통해 추가하였다. 결과적으로 제안 방법을 다양한 영역의 공개 사진 데이터 집합에 정성적, 정량적 향상된 결과를 확인할 수 있었다.

키워드 : 선화, 스타일 전이, 멀티모달 학습, 사진 분할, 깊이 추정, 생성형 인공지능
Key Words : Line Drawing, Style Transfer, Multi-modal Learning, Image Segmentation, Depth Estimation, Generative AI

ABSTRACT

The line drawing generation model aims to generate stylized line drawing images from original photographs using style transfer techniques. Prior work trained semantic and geometric information from photographs to generate line drawing without paired line drawing. It utilized CLIP (contrastive language-image pretraining) for semantic information and employed depth estimation for geometric information. Due to lack of ground-truth depth information, baseline used pseudo-labels instead of ground-truth, showing lower performance than using ground-truth. Therefore, our approach aims to generate high-quality line drawing images by incorporating an object segmentation method that utilizes ground-truth of depth information. By adding object segmentation information, it compensates insufficient ground-truth information for depth estimation and added geometric information such as background and shading. As a result, our approach achieved better performance on the publicly available various image datasets for line drawing.

※ 본 연구는 과학기술이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(No.RS-2022-00167194)의 지원과 한국연구재단의 지원(No.RS-2023 00212484)을 받아 수행된 연구임
◆ First Author : College of Computer Science, Kookmin University, justinday123@naver.com, 학생회원
° Corresponding Author : College of Computer Science, Kookmin University, jaekoo@kookmin.ac.kr, 정회원
논문번호 : 202306-116-0-SE, Received May 31, 2023; Revised August 6, 2023; Accepted August 15, 2023

Ⅰ. 서 론

1.1 개요

원본 사진(source image)으로부터 다양한 대상 사진(target image)을 적용하는 과업인 스타일 전이(style transfer)는 콘텐츠로써 많이 활용되고 있다.

원본 사진에 대해 선화로 그리는 방법으로는, 선화 쌍이 있는 사진으로 스타일 전이를 하는 모델[3]과 쌍이 없는 사진으로 스타일 전이를 하는 모델[1]이 있다. 선화 쌍이 있는 사진으로 스타일 전이를 하는 모델[3]의 경우, 선화 사진을 잘 생성하나, 선화 쌍 데이터가 부족하다는 제한이 있다. 이러한 제한을 보완하여 선화 쌍이 없는 사진으로 스타일 전이를 하는 모델[1]이 연구되었다. 선화 쌍이 없는 사진으로 스타일 전이를 하는 모델의 경우, 생성한 선화의 CLIP(contrastive language-image pretraining)[2] 특징(feature)을 원본 사진의 CLIP 특징과 일치시킴으로써 의미론적 정보를 학습하였다. 또한 생성한 선화의 깊이 정보(depth)를 원본사진의 깊이 정보와 일치시킴으로써 기하학적 정보를 학습하였다. 하지만, 여전히 배경, 세부 정보, 음영과 같은 정보들을 잘 표현하지 못하는 문제가 확인되었다.

제안 모델은 이를 보완하기 위해 객체 분할(instance segmentation)[8] 정보를 추가하였다. 이를 통해 다양한 원본 사진 데이터 집합[9,10,12]에서 배경과 음영 등의 정보를 잘 보여주었다. 또한 다양한 선화 스타일[14,15]에서도 좋은 성능을 보이는 것을 CLIP 점수와 사용자 선호도 조사(user study) 평가지표를 통해 확인하였다.

1.2 관련 연구

1.2.1 스타일 전이 (Style Transfer)

사진과 사진 변형(image-to-image translation)[6,11]은 원본 사진을 대상 사진의 스타일의 (style)로 변환시키는 방법이다. 이 방법은 하나의 사진에 다른 사진의 스타일을 조합하여 새로운 사진을 생성하는 방법이다. 그러나 해당 방법은 고품질의 선화를 잘 생성해 내기 어렵다는 문제가 있다. 반면, 나아진 성능의 선화를 생성하는 최근의 방법[3]은 원본 사진과 일치하는 쌍의 선화 사진이 필요하다. 해당 모델은 좋은 성능을 보이기는 하나 원본 사진과 일치하는 쌍의 선화 사진을 구하기 힘들다는 단점이 있다.

1.2.2 밀집 예측 (Dense Prediction)

사진의 특징을 이해하고, 이 특징을 기반으로 객체 분할, 깊이 추정, 객체 탐지 등과 같은 과업을 밀집 예측(dense prediction)이라고 한다. 밀집 예측을 기반으로 선화 스타일 전이를 하는 과업[5]은 이전부터 연구되었으나, 밀집 예측 중 깊이 추정만 사용하였을 때, 세밀한 정보, 외곽선의 유실, 음영에 대한 표현력의 부족함이

그림 1. KITTI 데이터 집합[12]에 대한 제안 모델과 기존 선화 모델[1]의 결과 시각화
Fig. 1. Visualization of ours and baseline[1]. Data: KITTI[12]

확인되었다.

1.2.3 언어-사진 변환

최근의 연구에는 언어-사진 변환 기법인 CLIP (contrastive language-image pretraining)을 활용하여 생성 혹은 스타일 전이 모델에 좋은 성과를 이루었다. 생성 모델에 CLIP을 적용하는 방법은 CLIP에 대한 추가적인 훈련이 필요하지 않으며, 사전 학습 된 (pre-trained) CLIP 모델만을 필요로 한다. CLIP을 사용하는 많은 모델[13]은 입력값인 언어와 사진을 각각 CLIP 모델을 거쳐서 나온 임베딩(embedding)간의 코사인 거리(cosine distance)를 줄이는 방향으로 학습하였다. 이러한 과정을 통해 사진, 혹은 언어에 대한 의미론적(semantic) 정보를 학습했다.

Ⅱ. 본 론

본 논문에서는 의미론적 정보와 기하학적 정보를 잘 담고 있는 선화를 생성하는 것을 목표로 한다. 기존 모델[1]의 경우 기하학적 정보를 위해 실측값에서 깊이 추정 모델이 생성한 추정값으로 대체하였다. 하지만 실측값이 아닌 깊이추정 모델이 생성한 추정값으로 학습하였기 때문에, 기하학적 정보가 정확하지 않다는 단점이 존재하였다.

따라서 본 논문은, 정확한 정보인 실측값이 존재하는 객체 분할 정보를 학습에 사용하여 추정값을 사용한다는 단점을 보완하였다. 또한, 객체 분할 정보를 추가하였기에, 음영이나 배경 등을 잘 고려하고, 추정값을 사용하는 모델에 비해 성능의 일관성을 보장해 주었으며, 더욱 세밀하고 높은 성능의 결과를 얻을 수 있었다. 사용된 손실함수는 총 5가지이다. 첫째, 적대적(adversarial) 손실함수[7] L_{GAN}은 입력 사진[9,10,12] a를 통해 선화 스타일 데이터집합[14,15]을 따르는 선화 b를 생성하도록 한다. 식(1)에서 G_A는 원본 사진 a를 통해 선화 $b(=G_A(a))$를 생성하는 생성자이며, G_B는 생성한 선화 b를 통해 원본사진의 복원본인 $\hat{a}(=G_B(b))$ 를 생성하는 생성자이다. 또한 판별자 D_B는 생성된 선화가 선화 스타일 데이터집합[14,15]의 도메인과 일치하도록 하며, 판별자 D_A는 복원본이 원본사진의 도메인과 일치하도록 한다.

$$L_{GAN} = E_{a\sim A}[D_A(a)^2] + E_{b\sim B}[(1-D_A(G_B(b)))^2] + E_{b\sim B}[D_B(b)^2] + E_{a\sim A}[(1-D_B(G_A(a)))^2] \quad (1)$$

둘째, 깊이 손실함수[4] L_{Geom}은 사진과 선화의 깊이가 유사하도록 학습한다. 식(2)에서 I는 특징 추출기이고 F는 깊이 추정모델[4]이며, F(a)는 원본 사진에 대한 깊이추정 사진이다.

$$L_{Geom} = \| G_{Geom}(I(G_A(a))) - F(a) \| \quad (2)$$

셋째, CLIP[2] 손실함수 L_{CLIP}은 원본 사진과 선화 간 CLIP 임베딩 거리의 차이를 줄이는 방향으로 학습된다. CLIP은 언어-사진(text-image)간 임베딩(embedding) 거리를 가깝도록 학습한 모델이다. 해당 모델은 임베딩 거리를 가깝게 함으로써 특정 사진에 대한 의미론적(semantic)인 정보를 언어로 학습한다. 식(3)은 이러한 사전 학습된 CLIP 모델을 사용하였다. 사전 학습된 CLIP을 통해 생성한 선화의 임베딩 결과와 원본사진의 임베딩 결과의 거리가 가까워지도록 한다.

$$L_{CLIP} = \| CLIP(G_A(a)) - CLIP(a) \| \quad (3)$$

넷째, Cycle 손실함수[6]는 선화에 입력 사진의 모습을 보존하기 위한 손실함수이다. 식(4)의 좌항은 원본사진 a를 통해 생성한 선화 b로 다시 원본사진을 복원한 $G_B(G_A(a))$와 a와의 거리를 줄인다. 마찬가지로 우항은 선화 b를 통해 복원한 원본사진으로 다시 생성한 선화 $G_A(G_B(b))$와 b와의 거리를 줄인다.

$$L_{Cycle} = \| G_B(G_A(a)) - a \| + \| G_A(G_B(b)) - b \| \quad (4)$$

마지막으로, 그림 2에서 점선으로 표시된 새로운 손실함수인 분할[8] 손실함수 L_{seg}이다. 이는 선화의 분할 결과와, 원본 사진의 분할 결과가 유사하도록 학습하는 손실함수이다. 원본 사진의 정보를 잘 담고 있는 선화를 생성하였다면, 원본 사진의 분할 결과와 선화의 분할 결과 또한 유사할 것이라 가정한다.

$$L_{Seg} = CE(G_A(a),\ gt(a)) \quad (5)$$

L_{seg} 는 U-Net[8]을 사용하여 선화의 분할 결과를 얻었으며, gt(a)는 원본 사진의 분할 실측값이다. 식(5)의 CE는 크로스 엔트로피(cross-entropy)이며, 선화의 분할과 실측값의 분할 차이를 크로스-엔트로피(cross-entropy)를 통해 줄였다. 이를 통해 선화에 대한 분할 정보를 추가하였다.

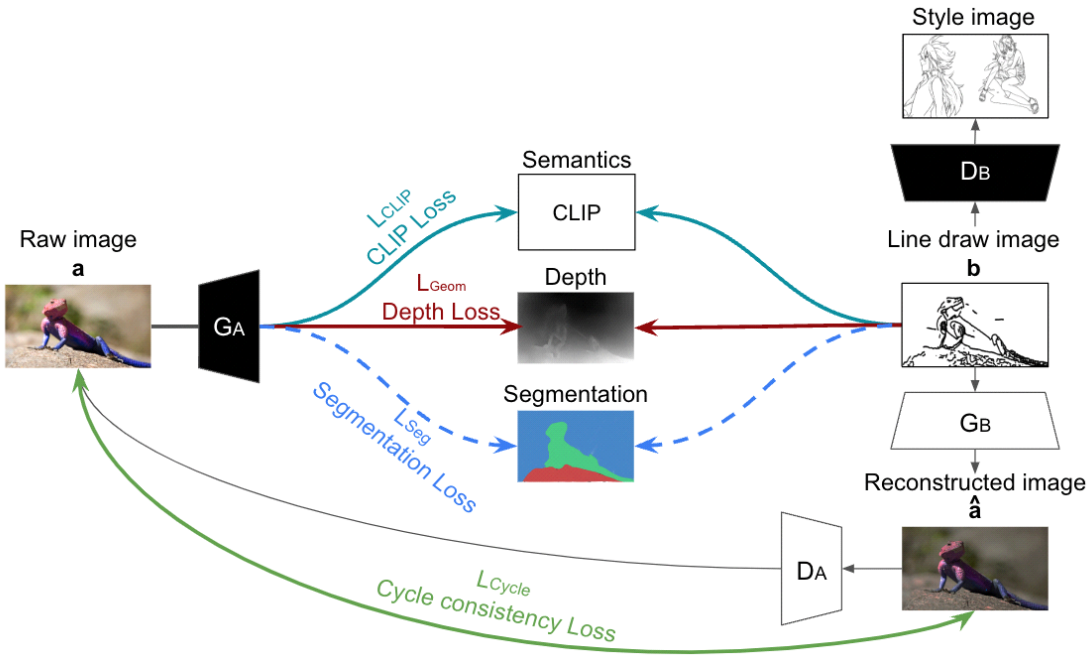

그림 2. 제안한 방법의 개요
Fig. 2. Overview of the proposed method

Ⅲ. 실 험

제안된 모델은 자연에 존재하는 다양한 사진을 담고 있는 COCO-Stuff[9] 데이터 집합으로 학습 되었다. 평가를 위해 원본사진으로 MIT-Adobe FiveK[10], COCO-Stuff[9], KITTI[12] 데이터 집합이 사용되었다. MIT-Adobe FiveK는 고화질의 분류(classification)가 가능한 사진을 제공하였으며 이를 통해 CLIP 점수를 평가하였다. 또한, COCO-Stuff의 경우, 객체들에 대한 표현력을 확인할 수 있었으며 KITTI의 경우, 풍경에 대한 정보가 많이 담겨있어 이를 통해 음영이나 배경에 대한 표현을 평가하였다. 그림 3는 선화의 생성과정을 훈련 시점(epoch)별로 출력하였으며, 이를 통해 생성 과정 중에도 더 나은 결과를 생성하는 것을 확인하였다. 그림 4는 원본 사진에 대한 제안 모델과 기존 선화 모델[1]을 비교하여 수정된 선을 시각화하였다. (빨간 선: 추가, 파란 선: 제거). 그림 1은 풍경에 대한 정보가 많이 담긴 KITTI에 대한 결과이다. 또한 우리는 선화의 스타일을 위해 선화 데이터집합인 Contour 데이터 집합[15]과 스케치 데이터집합[14]을 사용하였으며, Contour 스

그림 3. MIT-Adobe FiveK[10] 데이터에 기존 선화 모델[1]과 제안모델의 훈련 시점별 변화 시각화
Fig. 3. Visualization of lined rawing differences by epochs for baseline[1] and ours. Data: MIT-Adobe FiveK[10]

그림 4. MIT-Adobe FiveK[10] 데이터에서 기존 선화 모델[1]과 제안모델 간의 정성적 차이 비교 시각화
Fig. 4. Visualization of comparison with baseline[1] and ours. Data: MIT-Adobe FiveK[10]

그림 5. COCO-Stuff[9] 데이터에서 스케치 스타일[14]을 적용한 제안 모델과 기존 선화 모델[1] 시각화
Fig. 5. Visualizations of baseline[1] and ours adapting sketch style[14]. Data: COCO-Stuff[9]

타일은 그림 1, 그림 3, 그림 4, 에서 시각화하였으며, 스케치 스타일은 그림 5를 통해 시각화하였다. 그림 5의 경우, 원본사진을 위해 COCO-Stuff를 사용하였다. 최종적으로 스케치 스타일에서도 제안모델이 기존 선화 모델보다 음영에 대한 처리가 선명한 것을 확인하였다. 표 1은 CLIP 점수 지표를 사용하여 정량적으로 평가하였으며, 사용자 선호도(User study) 조사를 통해 정성적으로 평가한 결과이다. CLIP 점수 지표는 CLIP 모델에 선화 사진을 넣어 해당 사진이 어떤 문맥(context)을 담고 있는 사진인지 분류 결과를 확률적으로 보여주는 평가지표이다. 해당 지표는 MIT-Adobe Fivek를 사용하였으며, 제안 모델이 기존 선화 모델보

표 1. MIT-Adobe FiveK[10]와 COCO-Stuff[9], KITTI[12] 데이터에서 제안 모델과 기존 모델에 대한 CLIP 점수와 사용자 선호도 지표 정량적 결과 비교
Table 1. Perceptual evaluation using CLIP Score and User Study for baseline[1] and ours in MIT-Adobe FiveK[10], COCO-Stuff[9], KITTI[12] datasets

	Contour Style				Sketch Style
	CLIP 점수(↑) MIT-Adobe FiveK	User Study(↑) MIT-Adobe FiveK	User Study(↑) KITTI	User Study(↑) COCO-Stuff	User Study(↑) COCO-Stuff
기존 모델	0.4147	0.332	0.098	0.185	0.062
제안 모델	0.4394	0.668	0.901	0.814	0.937

다 CLIP 점수가 2.47% 높았다. 사용자 선호도 지표의 경우, MIT-Adobe FiveK, COCO-Stuff, KITTI 데이터 집합에 대해 그림 6의 방법으로 조사하였다. 사용자 선호도 지표 또한 본 모델이 세 가지 데이터 집합에 대해 기존 선화 모델보다 높은 것이 확인되었으며, 스케치 스타일에서도 더 좋은 지표가 확인되었다. 이를 통해 제안 모델이 기존 선화 모델보다 음영과 배경, 그리고 객체에 대한 표현력이 더 뛰어나다는 것이 확인되었다.

기존 선화 모델의 학습 시간은 약 198초가 소요되었으며, 제안 모델의 학습 시간은 약 276초가 소요되었다. 또한 기존 선화 모델의 추론 시간은 100장을 추론하는데 약 2.9초가 소요되었으며, 제안 모델도 동일하게 약 2.9초가 소요되었다.

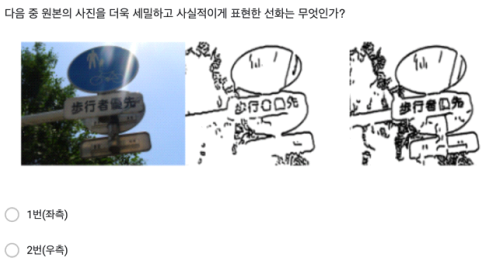

그림 6. 사용자 선호도 지표 평가 예시
Fig. 6. Example for evaluation of perception realism by user study

Ⅳ. 결 론

기존 선화로 스타일 전이를 하는 여러 모델은 원본 사진에 대해 쌍으로 이루어진 선화 사진이 필요했다. 그러나, 쌍으로 이루어진 선화 사진이 부족하다는 문제로 인해 기존 선화 모델의 경우 의미론적 정보와 기하학적 정보를 사용하여 해당 문제를 해결하였다. 해당 방법을 사용한 기존 선화 모델의 경우 CLIP과 깊이 추정 모델을 사용하여 학습하였다. 그러나 기하학적 정보를 위해 깊이 추정 모델만을 사용하였을 때 값이 부정확하며, 배경이나 음영에 대한 자세한 표현이 부족하다는 것이 확인되었다. 본 논문에서 제안한 모델은 기하학적 표현의 부족을 보완하기 위해 객체 분할 정보를 추가하였다. 그 결과, 기존 모델보다 제안 모델이 다양한 데이터 집합에서 음 음영이나 배경을 더 잘 표현함으로써 더 좋은 성능을 내는 것을 정성적으로 평가하였다. 또한, 생성한 선화 사진이 원본사진을 더 잘 담고 있는지 CLIP 점수를 통해 정량적으로 평가하여 더 좋은 성능을 내는 것을 확인하였다. 이를 통해, 객체 분할 정보는 선화를 생성하는 데 있어 기하학적인 정보를 추가함으로써 더 나은 성능을 내는 데 있어 영향을 주는 것을 확인하였다. 학습과 추론 소요 시간의 경우 비록 학습 시간은 객체 분할 정보의 추가로 인해 시간이 추가되었으나, 추론 시간의 경우 거의 동일하였다.

References

[1] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *Proc. IEEE/CVF Conf. CVPR*, pp. 7915-7925, 2022. (https://doi.org/10.1109/cvpr52688.2022.00776)

[2] A. Radford, et al., "Learning transferable visual models from natural language supervision," *Int. Conf. Mach. Learn.*, PMLR, pp. 8748-8763, 2021. (https://arxiv.org/abs/2103.00020)

[3] R. Yi, et al., "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proc. IEEE/CVF Conf. CVPR 2019*, pp. 10743-10752, 2019. (https://doi.org/10.1109/cvpr.2019.01100)

[4] S. M. H. Miangoleh, et al., "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *Proc. IEEE/CVF Conf. CVPR*, pp. 9685-9694, 2021.

(https://doi.org/10.1109/cvpr46437.2021.00956)
[5] M.-M. Cheng, et al., “Structure-preserving neural style transfer,” *IEEE Trans. Image Process.*, vol. 29, pp. 909-920, 2019. (https://doi.org/10.1109/tip.2019.2936746)
[6] J.-Y. Zhu, et al., “Unpaired image-to-image translation using cycle consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223-2232, 2017. (https://doi.org/10.1109/iccv.2017.244)
[7] X. Mao, et al., “Least squares generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2794-2802, 2017. (https://arxiv.org/abs/1611.04076)
[8] O. Ronneberger, et al., “U-Net: Convolutional networks for biomedical image segmentation,” *18th Int. Conf. MICCAI*, pp. 234-241, Munich, Germany, Oct. 2015. (https://arxiv.org/abs/1505.04597)
[9] T.-Y. Lin, et al., “Microsoft coco: Common objects in context,” in *ECCV 2014*, pp. 740-755, Springer, Zurich, Switzerland, Sep. 2014. (https://arxiv.org/abs/1405.0312)
[10] V. Bychkovsky, et al., “Learning photographic global tonal adjustment with a database of input/output image pairs,” in *CVPR 2011*, pp. 97-104, 2011. (https://doi.org/10.1109/cvpr.2011.5995332)
[11] D.-M. Kim, et al., “Generation of stage tour contents with deep learning style transfer,” *J. KIICE*, vol. 24, no. 11, pp. 1403-1410, 2020. (http://jkiice.org/)
[12] A. Geiger et al. “Vision meets robotics: The kitti dataset.” *The International Journal of Robotics Research* vol. 32 no. 11, pp. 1231-1237, 2013.
[13] R. Gal, et al., “Stylegan-nada: Clip-guided domain adaptation of image generators,” *ACM Trans. Graphics (TOG)*, vol. 41, no. 4, pp. 1-13, 2021. (https://doi.org/10.1145/3528223.3530164)
[14] T. Kim, *Anime sketch colorization pair*, https://www.kaggle.com/ktaebum/anime-sketch-colorization-pair, 2018.
[15] M. Li, Z. Lin, R. Mech, E. Yumer, and D. Ramanan, “Photo-sketching: Inferring contour drawings from images,” in *2019 IEEE WACV*, pp. 1403-1412, 2019. (https://doi.org/10.1109/wacv.2019.00154)

최 재 웅 (Jaewoong Choi)

2023년 3월~현재 : 국민대학교 인공지능 융합전공 석사과정
<관심분야> 생성형 인공지능, 스타일 전이
[ORCID:0009-0004-5596-9078]

이 재 구 (Jaekoo Lee)

2011년~2013년 : LG전자 CTO 부문, 주임 연구원
2018년 : SK 텔레콤 Data 기술원, 매니저
2018년 : 서울대학교 전기컴퓨터공학부 박사
2018년~현재 : 국민대학교 SW 융합대학 인공지능 조교수
<관심분야> 인공지능, 기계학습, 심층 신경망, 자율주행, 객체인지, 생체신호 분석
[ORCID:0000-0002-5947-5487]