특집논문 (Special Paper)
방송공학회논문지 제28권 제2호, 2023년 3월 (JBE Vol.28, No.2, March 2023)
https://doi.org/10.5909/JBE.2023.28.2.178
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

텍스트 인식률 개선을 위한 한글 텍스트 이미지 초해상화

권 준 형[a)], 조 남 익[a)‡]

Korean Text Image Super-Resolution for Improving Text Recognition Accuracy

Junhyeong Kwon[a)] and Nam Ik Cho[a)‡]

요 약

카메라로 촬영한 야외 일반 영상에서 텍스트 이미지를 찾아내고 그 내용을 인식하는 기술은 로봇 비전, 시각 보조 등의 기반으로 활용될 수 있는 매우 중요한 기술이다. 하지만 텍스트 이미지가 저해상도인 경우에는 텍스트 이미지에 포함된 노이즈나 블러 등의 열화가 더 두드러지기 때문에 텍스트 내용 인식 성능의 하락이 발생하게 된다. 본 논문에서는 일반 영상에서의 저해상도 한글 텍스트에 대한 이미지 초해상화를 통해서 텍스트 인식 정확도를 개선하였다. 트랜스포머에 기반한 모델로 한글 텍스트 이미지 초해상화를 수행하였으며, 직접 구축한 고해상도-저해상도 한글 텍스트 이미지 데이터셋에 대하여 제안한 초해상화 방법을 적용했을 때 텍스트 인식 성능이 개선되는 것을 확인하였다.

Abstract

Finding texts in general scene images and recognizing their contents is a very important task that can be used as a basis for robot vision, visual assistance, and so on. However, for the low-resolution text images, the degradations, such as noise or blur included in text images, are more noticeable, which leads to severe performance degradation of text recognition accuracy. In this paper, we propose a new Korean text image super-resolution based on a Transformer-based model, which generally shows higher performance than convolutional neural networks. In the experiments, we show that text recognition accuracy for Korean text images can be improved when our proposed text image super-resolution method is used. We also propose a new Korean text image dataset for training our model, which contains massive HR-LR Korean text image pairs.

Keyword : Scene Text Image Super-Resolution, Korean Text, Transformer

a) 서울대학교 전기·정보공학부 뉴미디어통신공동연구소(Department of ECE, INMC, Seoul National University)
‡ Corresponding Author : 조남익(Nam Ik Cho)
E-mail: nicho@snu.ac.kr
Tel: +82-2-880-8420
ORCID: https://orcid.org/0000-0001-5297-4649

※This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023. And this research was supported by LG AI Research.
· Manuscript January 16, 2023; Revised March 13, 2023; Accepted March 13, 2023.

Ⅰ. 서 론

간판, 표지판 등을 포함한 야외 촬영 이미지 내에 존재하는 텍스트들은 다양한 분야에 활용될 수 있는 유용한 정보들을 담고 있다. 하지만 검출한 텍스트 영역이 크기가 매우 작은 저해상도 이미지인 경우에는 해당 텍스트 이미지에 포함된 블러나 노이즈에 의한 글자 왜곡이 일반적인 크기의 이미지보다 더 두드러지기 때문에, 해당 텍스트 내용을 인식하는 데 어려움을 겪게 된다.

최근 딥러닝 기술이 비약적으로 발전하면서 영상 초해상화, 영상 디블러링, 영상 내 노이즈 제거 등의 다양한 영상 품질 향상 방법들의 성능도 향상되어 왔다. EDSR[1]은 SRResNet[10] 구조에 기반한 여러 개의 잔차 블록(Residual Block)으로 구성된 네트워크로, 배치 정규화 레이어(Batch Normalization Layer) 대신 상수 스케일링 레이어(Constant Scaling Layer)를 사용하여 네트워크의 학습 과정을 안정화하였다. RDN[2]은 기존의 잔차 블록 구조에 레이어 간의 연결 구조를 추가하여 비교적 얕은 네트워크로도 피쳐 간의 장거리 종속성(Long-Range Dependency)을 잡아낼 수 있게 하였다. 그러나 이러한 기존의 영상 품질 향상 기법들[1,2]은 텍스트 이미지가 아닌 일반적인 영상의 품질 향상을 목표로 하므로, 텍스트 인식률을 높이는 데에는 크게 도움이 되지 않는다. 텍스트 이미지가 아닌 일반 영상의 초해상화는 PSNR이나 SSIM과 같은 이미지 품질 평가 척도를 향상시키는 것이 목적이므로, 영상을 구성하는 배경과 전경을 모두 고려하여야 한다. 하지만 이와 달리 텍스트 이미지는 그것보다는 텍스트 인식 성능을 올리는 것이 주요 목적이기 때문에, 이미지를 구성하는 배경보다 텍스트 영역의 세부 정보를 잘 살려서 복원하는 것이 더 중요하다.

이러한 문제를 해결하기 위해 최근 다양한 텍스트 이미지 초해상화 방법들이 연구되고 있다. 텍스트 이미지 초해상화 방법들은 일반적인 이미지 초해상화 방법과는 달리 고해상도 텍스트 이미지를 복원하기 위해 저해상도 텍스트 이미지와 텍스트 인식 결과를 함께 이용한다. 그 중 TATT[3]는 텍스트 이미지 초해상화 분야에서 가장 좋은 성능을 보이는 방법 중 하나로, 텍스트 인식기를 통해 주어진 텍스트 이미지에서 텍스트 정보를 추출하고, 추출한 텍스트 정보를 트랜스포머[4]에 기반한 모듈을 이용해 텍스트 이미지 정보와 융합하여 이미지 초해상화를 수행한다. 하지만 한글 텍스트 이미지를 초해상화하기 위해 이러한 기존 텍스트 이미지 초해상화 모델을 그대로 사용하는 데에는 두 가지 문제점이 존재한다. 우선, 기존 텍스트 이미지 초해상화 모델들이 텍스트 정보를 추출하기 위해 사용한 텍스트 인식기는 상대적으로 오래되고 단순한 구조를 가진 텍스트 인식기인 CRNN[5]으로, 최근에 텍스트 인식 분야에서 많이 사용되는 다른 텍스트 인식기에 비해서는 다소 떨어지는 인식 성능을 보이는 모델이다. 그리고, 이러한 텍스트 인식기는 영어 벤치마크 데이터셋[6]에서 학습되었기 때문에, 한글 텍스트 이미지를 제대로 인식할 수 없다는 문제점이 있다.

본 논문에서는 TATT 모델 내부의 텍스트 인식기를 CRNN보다 좋은 성능을 가지는 인식기인 CDistNet[7]으로 대체하여 텍스트 이미지 초해상화를 수행하였다. 또한, 한글 텍스트 초해상도 시스템을 학습하기 위한 기존 데이터셋이 없기 때문에, 본 논문에서는 한글 텍스트 이미지로 구성된 고해상도-저해상도 이미지 쌍을 직접 구축하였다. 마지막으로 모델의 성능을 확인하기 위해 제안한 한글 텍스트 데이터셋에 대하여 PSNR, SSIM과 같은 이미지 품질 평가 척도와 텍스트 인식 성능을 측정하였다.

Ⅱ. 제안하는 방법

1. 모델 구조

TATT[3]는 텍스트 이미지 초해상화 분야에서 가장 성능이 좋은 모델 중 하나로, 모델 구조를 그림 1에 나타내었다. TATT는 입력 이미지의 텍스트 사전 정보를 추출하는 텍스트 인식 모듈(Text Prior Generator)과 이 사전 정보를 트랜스포머[4]의 인코더-디코더 구조(Text Prior Interpreter)를 통해 텍스트 이미지 복원 과정에 전달하는 모듈로 구성된다. 이러한 트랜스포머 구조의 도움을 받아 TATT는 회전이나 휘어짐 등의 다양한 공간적 왜곡이 가해진 텍스트 이미지들도 효과적으로 처리할 수 있게 한다.

TATT에 사용된 텍스트 인식 모듈인 CRNN[5]은 딥러닝을 사용한 텍스트 인식 모델 중 초창기 모델이다. CNN을

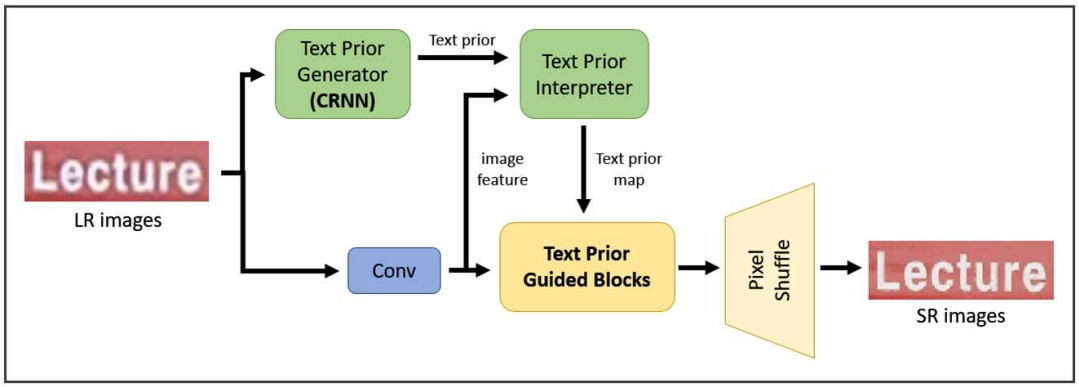


그림 1. TATT[3]의 모델 구조
Fig. 1. Architecture of TATT[3]

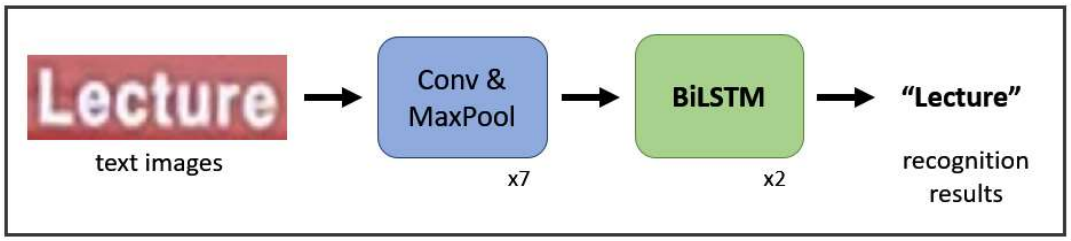


그림 2. CRNN[5]의 모델 구조
Fig. 2. Architecture of CRNN[5]

통해 입력 이미지의 피쳐를 추출하고, 이를 bidirectional LSTM에 통과시켜 최종 인식 결과 벡터를 출력하는 간단한 구조로 구성되어 있다. 모델 구조는 그림 2와 같다. CRNN은 발표된 당시에는 텍스트 인식 분야에서 좋은 성능을 보였으나, 최근 트랜스포머를 사용하여 텍스트 인식을 수행하는 다양한 모델에 비해서는 인식 성능이 떨어지는 편이다.

어텐션 기반의 인코더-디코더 구조로 이루어진 트랜스포머를 사용한 텍스트 인식 모델들은 서로 다른 도메인인 텍스트 이미지 정보와 텍스트 의미 정보를 융합하는 데에 강

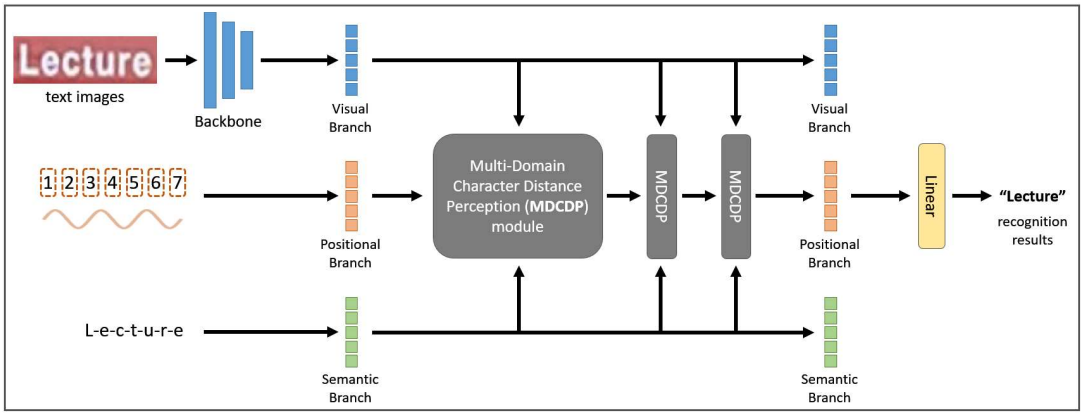


그림 3. CDistNet[7]의 모델 구조
Fig. 3. Architecture of CDistNet[7]

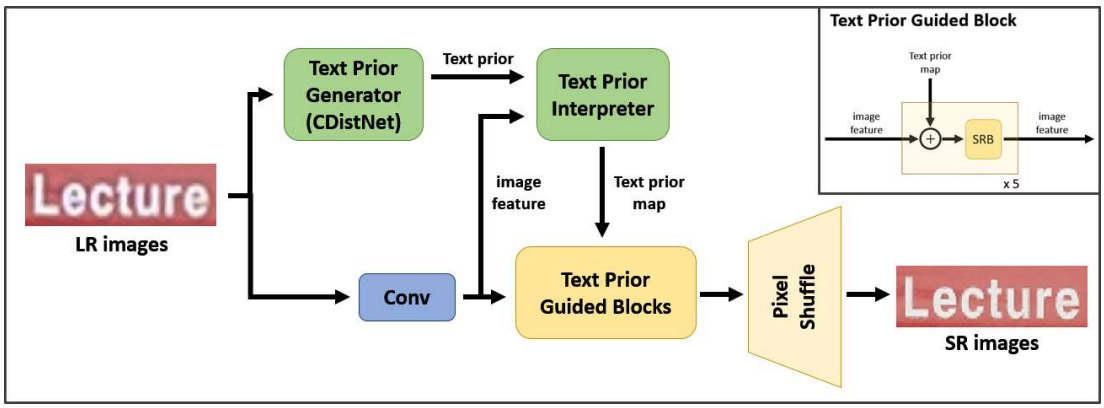


그림 4. 사용한 모델 구조
Fig. 4. Architecture of proposed network

이미지 정보(Visual information)와 텍스트 의미 정보(Semantic information) 간의 어텐션이 어긋나는 경우가 있다. 이러한 문제를 해결하기 위해 그림 3과 같은 CDistNet[7]은 텍스트 이미지 정보, 텍스트 의미 정보에 텍스트 위치 정보(Positional information)를 추가로 사용하고, 이 세 가지 도메인의 정보들을 트랜스포머를 이용하여 융합하는 모듈(Multi-Domain Character Distance Perception module, MDCDP)을 제안하였다.

본 논문에서는 초해상화 성능을 더 높이기 위해, TATT의 텍스트 인식 모듈을 기존 CRNN보다 더 좋은 인식 성능을 보이는 CDistNet으로 교체하였다. 그 외에 이미지 정보와 텍스트 정보를 융합하여 초해상화 이미지를 만드는 모듈(Text Prior Guided Blocks)은 TATT와 동일한 구조를 사용하였다. 사용한 모델 구조는 그림 4에 나타내었다.

2. 손실 함수

우선 원본 이미지 X와 입력 이미지 Y를 초해상화한 결과 이미지 $F(Y)$ 간의 차이를 측정하는 초해상화 손실 함수 L_{SR}는 두 이미지 사이의 L_2 norm으로 정의된다.

$$L_{SR}=\frac{1}{N}\sum_{i=1}^{N}\| F(Y_i)-X_i \|_2^2 \quad (1)$$

그리고 텍스트 이미지에서 올바른 텍스트 사전 정보(text prior information)를 추출하기 위해, 텍스트 사전 정보 손실 함수 (Text Prior Loss)를 사용한다. 이는 저해상도 이미지와 고해상도 원본 이미지에서 추출한 사전 정보 사이의 쿨백-라이블러 발산과 L_1 norm의 합으로 정의된다.

$$L_{TP}=\frac{1}{N}\sum_{i=1}^{N}\| p_i^Y-p_i^X \|_1+D_{KL}(p_i^Y,p_i^X) \quad (2)$$

마지막으로, 텍스트 이미지 구조의 일관성을 위한 텍스트 구조 일관성 손실 함수 (Text Structure Consistency Loss)를 사용한다. 이는 이미지에 가해지는 회전 변환(rotation), 밀림 변환 (shearing) 등을 포함하는 왜곡 변환을 $\boldsymbol{D}$라 했을 때, $\boldsymbol{D}F(Y)$, $F(\boldsymbol{D}Y)$, $\boldsymbol{D}X$ 사이의 구조적 유사도를 측정하는 손실 함수이다.

$$L_{TSC}=1-TSSIM(\boldsymbol{D}F(Y),F(\boldsymbol{D}Y),\boldsymbol{D}X) \quad (3)$$

여기서 TSSIM[3]은 Triplex Structure-Similarity Index Measure의 약자로, 두 이미지 간의 구조적 유사도를 측정하는 척도인 SSIM을 세 이미지에 적용할 수 있도록 확장한 척도이다. 이상에서의 손실함수들을 조합한, 모델 학습에 사용한 전체 손실 함수는 다음과 같다.

$$L=L_{SR}+\alpha L_{TP}+\beta L_{TSC} \quad (4)$$

여기서, α와 β는 각 손실 함수값 간의 균형을 결정하는

하이퍼파라미터로, 각각 1과 0.1을 사용하였다.

Ⅲ. 한글 텍스트 데이터셋

한글 텍스트 이미지 초해상화 모델 학습을 위해서는 고해상도-저해상도 한글 텍스트 이미지 쌍으로 구성된 데이터셋 구축이 필수적이다. 이를 위해 AI Hub[8]에서 제공하는 인공지능 학습용 데이터 중 하나인 AI Hub 야외 실제 촬영 한글 이미지 데이터셋[9]을 이용하였다. AI Hub 야외 실제 촬영 한글 이미지 데이터셋은 다양한 폰트의 한글 텍스트들이 다수 포함된 간판, 책표지 등의 실내외 이미지들과, 한글 텍스트 영역 및 텍스트 내용 등이 포함된 라벨로 구성되어 있다.

AI Hub 야외 실제 촬영 한글 이미지 데이터셋을 한글 텍스트 이미지 초해상화 모델 학습 및 테스트에 사용하기 위해, 전체 이미지에서 한글 텍스트 영역만을 자른 후 이를 가로 128픽셀, 세로 32픽셀의 크기로 bicubic down-sampling하여 고해상도 한글 이미지를 만들었다. 이 고해상도 한글 이미지를 가로 64픽셀, 세로 16픽셀의 크기로 다시 bicubic downsampling 하여 저해상도 한글 이미지를 만들었다. 이런 식으로 고해상도-저해상도 한글 이미지 쌍으로 구성된 데이터셋을 구축하였다. 모델 학습을 위해 10만여 장, 테스트를 위해 약 1만여 장을 제작하여 사용하였다.

Ⅳ. 실험 결과 및 분석

AI Hub 야외 실제 촬영 한글 이미지 데이터셋으로 만든 고해상도-저해상도 한글 이미지 쌍의 테스트 셋에 대하여 제안한 텍스트 이미지 초해상화 모델의 성능을 측정하였다. 이미지 초해상화 분야에서 널리 사용되는 이미지 품질 평가 척도인 PSNR 및 SSIM을 측정하였으며, 제안한 모델의 우수성을 보이기 위해 단순 bicubic upsampling 방식, EDSR[1], TATT[3]의 성능도 함께 측정하였다. 사용한 모든 모델은 제작한 한글 이미지 데이터셋으로 학습되었다. 또한 제안한 방법으로 초해상화한 텍스트 이미지에 대해 텍스트 인식 정확도를 측정하여 텍스트 이미지 초해상화가 실제로 텍스트 인식 성능 개선에 도움을 주는지를 확인하였다. 한글 텍스트 인식 정확도 측정을 위해서는 한글 이미지 쌍의 학습셋을 이용하여 학습한 CDistNet[7]을 사용하였다.

표 1. 1만장의 한글 데이터 테스트 셋에 대한 PSNR/SSIM 및 텍스트 인식 정확도
Table 1. PSNR/SSIM metrics for proposed method on Korean text image test set (10,000 test images) and the text recognition accuracy

	PSNR	SSIM	Rec. result (CDistNet)
Bicubic	22.59dB	0.8197	83.75%
EDSR[1]	25.26dB	0.8933	87.54%
TATT[3]	27.54dB	0.9321	90.89%
Proposed	**27.76dB**	**0.9367**	**92.12%**

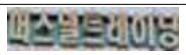
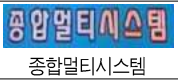
Bicubic	 너스널트레이닝	 종합얼티시스템	 부동로	 치킨	 이보와
EDSR[1]	 퍼스널트레이딩	 송압멀티시스템	 무흥로	 시민	 티보와
TATT[3]	 퍼스널트레이닝	 송합멀티시스템	 무흥로	 시민	 티보와
Proposed	 퍼스널트레이닝	 종합멀티시스템	 부흥로	 시민	 티보와
HR	 퍼스널트레이닝	 종합멀티시스템	 부흥로	 시민	 티보와

표 2. 제안한 모델로 복원된 한글 텍스트 이미지 및 텍스트 인식 결과 예시
Table 2. Visualization of Korean text images recovered by proposed method and the SR text recognition results

제안한 모델을 포함한 다양한 초해상화 모델에 대한 한글 텍스트 이미지 초해상화의 결과를 표 1에 나타내었다. 제안한 텍스트 초해상화 방법은 EDSR[1]과 같은 일반적인 초해상화 방법과 달리 전체 이미지에서 텍스트 부분에 특히 집중하여 초해상화를 수행하므로, 이미지 품질 평가 척도인 PSNR과 SSIM 그리고 텍스트 인식 정확도가 가장 높은 것을 확인할 수 있다. 또한 기존 TATT 모델과 텍스트 인식 정확도를 비교해 보면, 텍스트 인식 모듈을 기존의 CRNN에서 CDistNet으로 바꾸었을 때 텍스트 인식 정확도가 약 1.23% 상승하였다. 이는 제안한 방법대로 텍스트 인식 모듈으로 CDistNet을 사용했을 때 이미지 초해상화에 필요한 텍스트 의미 정보를 더 정확하게 만들어낸다는 것을 보여준다. 초해상화 결과와 텍스트 인식 결과를 표 2에 나타내었으며, 표 2의 초해상화 결과 이미지를 보면, 제안한 방법으로 초해상화한 결과 이미지가 bicubic 이미지, EDSR 결과 이미지 및 TATT 결과 이미지에 비해 한글 텍스트 부분을 더 선명하게 복원하는 것을 육안으로도 관찰할 수 있다.

V. 결 론

본 논문에서는 야외 일반 이미지에 포함된 한글 텍스트 중 해상도가 낮아 인식이 어려운 텍스트에 이미지 초해상화를 적용하여 텍스트 인식 정확도를 높이는 방법을 제안하였다. 한글 텍스트를 초해상화하기 위해 AI Hub 야외 실제 촬영 한글 이미지 데이터셋으로부터 텍스트 영역만을 남겨서 초해상화 모델 학습을 위한 데이터셋을 구축하였다. 그 결과, 한글 텍스트 이미지 초해상화 후 텍스트 인식 정확도 및 PSNR, SSIM 수치가 큰 폭으로 상승하였고, 제안한 초해상화 모델이 텍스트 부분에 집중하여 초해상화를 수행하는 것을 확인하였다.

참 고 문 헌 (References)

[1] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced Deep Residual Networks for Single Image Super-Resolution,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136-144, 2017.
doi: https://doi.org/10.1109/CVPRW.2017.151

[2] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual Dense Network for Image Super-Resolution,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472-2481, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00262

[3] J. Ma, Z. Liang, and L. Zhang, “A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-Resolution,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911-5920, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00582

[4] A. Vaswani et al. “Attention is All You Need,” *Advances in Neural Information Processing Systems*, 30, 2017.

[5] B. Shi, X. Bai, and C. Yao, “An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” *IEEE transactions on pattern analysis and machine intelligence*, Vol.39, No.11, pp.2298-2304, 2016.
doi: https://doi.org/10.1109/TPAMI.2016.2646371

[6] W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, C. Shen, and X. Bai, “Scene Text Image Super-Resolution in the Wild,” *European Conference on Computer Vision*, Springer, Cham, 2020.
doi: https://doi.org/10.1007/978-3-030-58607-2_38

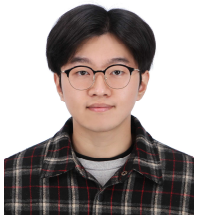
[7] T. Zheng, Z. Chen, S. Fang, H. Xie, and Y. G. Jiang, “Cdistnet: Perceiving Multi-Domain Character Distance for Robust Text Recognition,” *arXiv preprint arXiv:2111.11011*, 2021.
doi: https://doi.org/10.48550/arXiv.2111.11011

[8] AI Hub, https://aihub.or.kr (accessed Dec. 28, 2022.)

[9] Outdoor images including Korean texts, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=105 (accessed Dec. 28, 2022.)

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, and W. Shi, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4681-4690, 2017.
doi: https://doi.org/10.1109/CVPR.2017.19

저 자 소 개



권 준 형

- 서울대학교 전기정보공학부 석박통합과정
- ORCID : https://orcid.org/0009-0003-6383-8775
- 주관심분야 : 영상 복원, 영상 초해상화, 컴퓨터 비전

조 남 익

- 서울대학교 전기정보공학부 교수
- ORCID : https://orcid.org/0000-0001-5297-4649
- 주관심분야 : 디지털 신호처리, 영상처리, 컴퓨터 비전