# 텍스트 인식률 개선을 위한 한글 및 영어 텍스트 이미지 초해상화

*권준형 **조남익
서울대학교 전기정보공학부 뉴미디어통신공동연구소
*gjh8760@snu.ac.kr **nicho@snu.ac.kr

# Korean and English Text Image Super-Resolution for Improving Text Recognition Accuracy

*Kwon, Junhyeong **Cho, Nam Ik
Department of ECE, INMC, Seoul National University

요 약

야외 환경을 카메라로 촬영한 일반 영상에서 텍스트 이미지를 검출하고 인식하는 기술은 로봇 비전, 시각 보조 등의 기반이 되는 기술로 활용될 수 있어 매우 중요한 기술이다. 하지만 저해상도의 텍스트 이미지의 경우 텍스트 이미지에 포함된 노이즈나 블러 등이 더 두드러지기 때문에 텍스트 내용을 인식하는 것이 어렵다. 이에 본 논문은 일반 영상에서의 저해상도 한글 및 영어 텍스트에 대한 이미지 초해상화를 통해 텍스트 인식 정확도를 개선하였다. 트랜스포머에 기반한 모델로 한글 및 영어 텍스트에 대한 이미지 초해상화를 수행하였으며, 영어 및 한글 데이터셋에 대해 제안한 초해상화 방법을 적용했을 때 그렇지 않을 때보다 텍스트 인식 성능이 개선되는 것을 확인하였다.

## 1. 서론

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

최근 딥러닝의 발전에 힘입어 영상 내 노이즈 제거, 영상 초해상화, 영상 디블러링 등의 다양한 영상 품질 향상 방법들의 성능이 비약적으로 상승해 왔다. 그러나 이러한 기존 영상 품질 향상 기법들은 텍스트 영역 이미지가 아닌 일반적인 이미지의 품질 향상을 목표로 하고, 텍스트의 인식률을 높이는 데에는 초점이 맞춰져 있지 않다. 텍스트 이미지가 아닌 일반적인 영상의 초해상화를 위해서는, 영상을 구성하고 있는 배경과 전경 모두 고려해야 한다. 하지만 이와 달리 텍스트 이미지는 PSNR이나 SSIM과 같은 이미지 품질 평가 척도보다는 텍스트 인식 성능을 높이는 것이 더 중요하므로, 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. 그 중 TATT[1]는 가장 성능이 좋은 모델 중 하나로, 텍스트 인식기를 통해 주어진 텍스트 이미지에서 텍스트 정보를 추출하고, 이 텍스트 정보를 트랜스포머에 기반한 모듈을 이용하여 텍스트 이미지를 복원하는 데 사용한다. 하지만 TATT에서 텍스트 정보를 추출하기 위해 사용한 텍스트 인식기는 상대적으로 오래되고 단순한 구조를 가진 텍스트 인식기인 CRNN[2]으로, 최근에 많이 사용되고 있는 다른 텍스트 인식기에 비해서는 다소 떨어지는 인식 성능을 보인다.

본 논문에서는 TATT 모델의 텍스트 인식기를 CRNN보다 좋은 성능을 가지는 인식기인 CDistNet[3]으로 대체하여 텍스트 이미지 초해상화를 수행하였다. 또한, 영어 텍스트 데이터에 대해서만 이미지 초해상화를 수행한 기존의 방법과 달리, AI Hub 야외 실제 촬영 한글 이미지 데이터셋을 이용하여 한글 텍스트 데이터에 대해서도 이미지 초해상화가 가능하도록 하였다. 마지막으로 제안한 모델의 성능을 확인하기 위해 영어 텍스트 데이터셋과 한글 텍스트 데이터셋에 대해 PSNR, SSIM과 같은 이미지 품질 평가 척도와 텍스트 인식 성능을 측정하였다.

## 2. 사용한 데이터셋

영어 텍스트 이미지 초해상화 모델 학습을 위해 TextZoom[4] 데이터셋을, 한글 텍스트 이미지 초해상화 모델 학습을 위해 AI Hub 야외 실제 촬영 한글 이미지 데이터셋을 각각 사용하였다.

### 2.1. TextZoom

TextZoom 데이터셋은 텍스트 이미지 초해상화 분야에서 널리 사용되는 벤치마크 데이터셋으로, 야외 환경을 카메라의 초점 거리를 바꿔가며 촬영한 저해상도-고해상도 이미지에서 텍스트 영역만을 추려서 제작되었다. 총 21,740개의 저해상도-고해상도 텍스트 이미지 쌍과 해당 텍스트의 라벨로 구성되어 있고, 그 중 학습에 사용되는 데이터는 17,367개이다. 나머지 데이터는 테스트 용도이고, 카메라의 초점 거리에 따라 easy (1,619개), medium (1,411개), hard (1,343개)의 세 가지 하위 항목으로 나뉜다. 영어 텍스트 이미지 초해상화 모델의 학습 및 성

표 1. TextZoom 테스트 셋에서의 PSNR/SSIM 및 텍스트 인식률. Bicubic은 저해상도 이미지를 원본 이미지 크기로 bicubic upsampling 한 이미지를 의미한다.

| | PSNR | | | SSIM | | | 텍스트 인식률 (ASTER) | | |
|---|---|---|---|---|---|---|---|---|---|
| 방법 | easy | medium | hard | easy | medium | hard | easy | medium | hard |
| Bicubic | 22.99 | 19.57 | 19.61 | 0.7985 | 0.6319 | 0.6668 | 64.7% | 42.2% | 31.6% |
| TATT[1] | **24.72** | **19.02** | **21.52** | **0.9006** | **0.6911** | 0.7703 | 78.9% | 63.4% | 45.4% |
| 제안한 모델 | 24.55 | 18.97 | 20.03 | 0.8982 | 0.6794 | **0.7731** | **81.3%** | **64.6%** | **48.0%** |

표 2. AI Hub 야외 실제 촬영 한글 이미지 테스트 셋에서의 PSNR/SSIM 및 텍스트 인식률.

| | PSNR | SSIM | 인식률 (CDistNet) |
|---|---|---|---|
| Bicubic | 22.59 | 0.8197 | 83.75% |
| 제안한 모델 | 27.76 | 0.9367 | 92.12% |

능 측정 시 TextZoom 데이터셋의 학습 및 테스트 셋을 그대로 사용하였다.

### 2.2. AI Hub 야외 실제 촬영 한글 이미지

AI Hub 야외 실제 촬영 한글 이미지 데이터셋은 간판, 책표지 등 일상에서 접할 수 있는 다양한 한글 텍스트가 포함된 이미지와 해당 텍스트의 라벨로 구성되어 있다. 한글 텍스트 이미지 초해상화를 위해, AI Hub 야외 실제 촬영 한글 이미지 데이터셋에서 텍스트 영역만을 자르고, 이를 bicubic downsampling 하여 고해상도-저해상도 이미지 쌍을 구축하였다. 학습 데이터는 총 10만 장, 테스트 데이터는 총 1만장을 제작하여 사용하였다.

## 3. 제안하는 방법

### 3.1. 모델 구조

TATT[1]는 텍스트 이미지 초해상화 분야에서 가장 성능이 좋은 모델 중 하나이다. TATT는 입력 이미지의 텍스트 사전 정보를 추출하는 텍스트 인식 모듈과 이 사전 정보를 어텐션 매커니즘[6]을 통해 텍스트 이미지 복원 과정에 전달하는 트랜스포머 기반의 모듈로 구성된다. 이러한 트랜스포머 구조의 도움을 받아 TATT는 회전이나 휘어짐 같은 다양한 공간적 왜곡이 가해진 텍스트 이미지들도 효과적으로 처리할 수 있게 된다. 하지만 TATT에 사용된 텍스트 인식 모듈인 CRNN[2]은 최근 발표된 텍스트 인식 모델들과 비교했을 때 상대적으로 간단한 구조를 가지고, 더 낮은 텍스트 인식 성능을 보인다. 따라서 모델의 초해상화 성능을 보다 높이기 위해, TATT의 텍스트 인식 모듈을 기존의 CRNN에서 더 좋은 인식 성능을 보이는 CDistNet[3]으로 교체하였다.

### 3.2. 손실 함수

우선 원본 이미지 $X$와, 입력 이미지 $Y$를 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, 밀림 변환 (shearing) 등의 왜곡을 $\boldsymbol{D}$ 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[1]은 triplex Structure-Similarity Index Measure의 약자로, 두 이미지 간의 구조적 유사도를 측정하는 척도인 SSIM을 세 이미지에 적용할 수 있도록 확장한 척도이다.

모델 학습에 사용한 전체 손실 함수는 다음과 같다.

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

$\alpha$와 $\beta$는 각 손실 함수 값 간의 균형을 결정하는 하이퍼파라미터로, 각각 1과 0.1을 사용하였다.

그림 1. 제안한 영어 모델로 복원된 TextZoom 테스트 셋 샘플 이미지 및 텍스트 인식 결과 예시. HR은 고해상도 원본 이미지를 의미한다. 빨간색으로 쓰여진 글자는 틀리거나 빠진 글자를 의미한다.

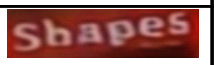
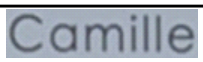
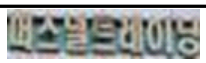
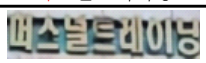
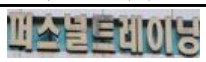
| | | | | | |
|---|---|---|---|---|---|
| Bicubic |  camillo |  wado |  droom |  treasons |  shorts |
| 제안한 모델 |  camille |  weibo |  choose |  treasure |  shapes |
| HR |  camille |  weibo |  choose |  treasure |  shapes |

그림 2. 제안한 한글 모델로 복원된 AI Hub 야외 실제 촬영 한글 이미지 데이터셋의 테스트 샘플 이미지 및 텍스트 인식 결과 예시. HR은 고해상도 원본 이미지를 의미한다. 빨간색으로 쓰여진 글자는 틀리거나 빠진 글자를 의미한다.

| | | | | | |
|---|---|---|---|---|---|
| Bicubic |  너스널트레이닝 |  종합얼티시스템 |  부동로 |  치킨 |  이보와 |
| 제안한 모델 |  퍼스널트레이닝 |  종합멀티시스템 |  부흥로 |  시민 |  티보와 |
| HR |  퍼스널트레이닝 |  종합멀티시스템 |  부흥로 |  시민 |  티보와 |

## 4. 실험 결과

영어 데이터셋인 TextZoom의 테스트 셋과 한글 데이터셋인 AI Hub 야외 실제 촬영 한글 이미지의 테스트 셋에 대하여 제안한 텍스트 이미지 초해상화 모델의 성능을 측정하였다. 우선 이미지 초해상화 분야에서 널리 사용되는 이미지 품질 평가 척도인 PSNR 및 SSIM를 측정하였고, 초해상화 시킨 텍스트 이미지에 대한 인식 성능을 측정함으로써 텍스트 이미지 초해상화가 실제로 인식 성능 개선에 도움을 주는지를 확인하였다. 영어 텍스트 인식 성능 측정을 위해서는 ASTER[5] 모델을, 한글 텍스트 인식 성능 측정을 위해서는 CDistNet 모델을 사용하였다.

제안한 모델과 기존 TATT 모델에 대한 TextZoom 데이터셋의 결과는 표 1과 같다. 텍스트 인식률의 경우 제안한 모델이 기존의 TATT 모델보다 더 뛰어난 성능 개선을 보이는 것을 확인할 수 있다. 그러나 이미지 품질 평가 척도인 PSNR과 SSIM의 경우 TATT 모델보다 제안한 모델을 사용했을 때 대체로 더 낮은 수치를 보이는데, 이는 제안한 모델이 텍스트가 아닌 배경 부분보다는 텍스트 부분에 집중하여 초해상화를 수행하였기 때문에 전체적인 품질 척도는 낮아진 것으로 생각된다. 하지만 PSNR이나 SSIM의 하락 폭이 미미하고, 텍스트 이미지 초해상화의 목적은 초해상화를 통한 텍스트 인식률의 상승이므로, 이러한 품질 척도 저하는 그렇게 큰 문제가 되지 않는 것으로 생각된다.

AI Hub 야외 실제 촬영 한글 이미지 데이터셋의 결과를 표 2에 나타내었다. 마찬가지로 이미지 품질 평가 척도와 텍스트 인식률 모두 bicubic 이미지에 비해 초해상화를 거친 이미지가 더 높은 성능을 보이는 것을 확인할 수 있다.

## 5. 결론

본 논문에서는 야외 이미지에 포함된 한글 및 영어 텍스트 중 해상도가 낮아 인식이 어려운 텍스트에 이미지 초해상화를 적용하여 텍스트 인식률을 높이는 방식을 제안하였다. 특히 한글 텍스트의 경우, 성공적인 초해상화를 위해 AI Hub 야외 실제 촬영 한글 이미지 데이터셋으로부터 텍스트 영역만을 남겨서 초해상화 모델 학습을 위한 데이터셋을 구축하였다. 그 결과, 텍스트 이미지 초해상화 후 텍스트 인식률이 한글과 영어 데이터 모두에 대해 큰 폭으로 상승하였고, 제안한 초해상화 모델이 텍스트 부분에 집중해서 초해상화를 수행하는 것을 확인하였다.

## 감사의 글

이 논문은 2022년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

본 연구는 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-01062).

## 참고문헌

[1] Ma, J., Liang, Z., & Zhang, L. (2022). A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5911-5920).

[2] Shi, B., Bai, X., & Yao, C. (2016). 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,* 39(11), 2298-2304.
[3] Zheng, T., Chen, Z., Fang, S., Xie, H., & Jiang, Y. G. (2021). Cdistnet: Perceiving multi-domain character distance for robust text recognition. *arXiv preprint arXiv:2111.11011.*
[4] Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., & Bai, X. (2020, August). Scene text image super-resolution in the wild. In *European Conference on Computer Vision* (pp. 650-666). Springer, Cham.
[5] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., & Bai, X. (2018). Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9), 2035-2048.
[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.