

Highlight Text Extraction Pipeline

Original → Detection → OCR

(a) Original Image
(Yellow Highlight)

합성 데이터 생성은 라벨링 비용을 절감하고 대규모 데이터셋을 구축합니다.



(b) Color Detection
(HSV Mask)



(c) OCR Extraction
(Tesseract)

Detected: "한국"

한국어 텍스트를 찾았습니다. 이를 통해 대규모 데이터셋을 구축합니다.

(d) Original Image
(Green Highlight)

이미지 전처리는 노이즈 제거, 미진화, 대비 확장 등을 포함합니다.



(e) Color Detection
(HSV Mask)



(f) OCR Extraction
(Tesseract)

Detected: "한국 일본"

한국어와 일본어 텍스트를 찾았습니다. 이를 통해 대규모 데이터셋을 구축합니다.