Journal of KIIT. Vol. 22, No. 11, pp. 181-196, Nov. 30, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.11.181

Development of a Named Entity Recognition-based Text Masking System for Preventing Personal Information Exposure in Shipping Labels

Ha-Eun Kim*, Soo-Yong Kim**, Myeong-Seop Kim***, Sang-Ho Kim****, Ye-Jin Cho*****, Eun-Sun Choi******, and Gil-Sang Yoo*******

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2023-00246191)

Abstract

The widespread use of social media and video platforms has frequently and unintentionally exposed sensitive personal information, such as shipping labels, in shared images and videos. Masking entire shipping labels is not only time-consuming and labor-intensive but it can also lead to video distortion. This paper proposes a system based on named entity recognition and text recognition technologies to automatically detect and mask personal information within shipping labels. Experimental results demonstrate that the system achieved an accuracy of 81.2% in recognizing and masking personal information. The proposed technology selectively masks the relevant personal information, minimizing video distortion, and can be effectively applied to social media platforms and video production environments where there is a high risk of exposing sensitive information.

요 약

소셜 미디어와 비디오 플랫폼의 광범위한 사용으로 인해 공유된 이미지와 비디오에서 배송 라벨과 같은 민감한 개인정보가 의도치 않게 자주 노출되고 있다. 운송장 전체를 마스킹하는 것은 시간과 인력이 많이 소요될 뿐만 아니라 영상 왜곡을 초래할 수 있다. 본 논문에서는 개체명 인식과 텍스트 인식 기술에 기반하여 택배 운송장 내의 개인정보를 자동으로 탐지하고 마스킹하는 시스템을 제안한다. 실험 결과, 본 시스템은 개인정보 인식 및 마스킹에서 81.2%의 정확도를 보였다. 제안된 기술은 개인정보만을 선택적으로 마스킹함으로써 영상 왜곡을 최소화할 수 있으며, 민감한 개인정보가 노출될 위험이 있는 소셜 미디어 플랫폼과 영상 제작 환경에 효과적으로 적용될 수 있다.

Keywords

image segmentation, scene text detection, named entity recognition, de-identification, text masking

* MS candidate, Department of Electrical and Computer Engineering, Korea University
- ORCID: https://orcid.org/0009-0007-2836-3160
** MS degree, Department of Artificial Intelligence, Seoul National University
- ORCID: https://orcid.org/0009-0001-7111-1301
*** BS degree, Department of Mathematics, Korea University
- ORCID: https://orcid.org/0009-0003-4723-0444
**** MS degree, Department of Business, Hankuk University of Foreign Studies
- ORCID: https://orcid.org/0009-0008-4342-1207
***** MS candidate, Department of Computer and Artificial Intelligence, Dongguk university
- ORCID: https://orcid.org/0009-0002-0333-6506
****** Undergraduate student, Department of Computer Science and Engineering, Korea University
- ORCID: https://orcid.org/0009-0006-1832-1103
******* Professor, Department of Creative Informatics & Computing Institute, Korea University
- ORCID: https://orcid.org/0009-0002-1085-5355

· Received: Aug. 27, 2024, Revised: Oct. 11, 2024, Accepted: Oct. 14, 2024
· Corresponding Author: Gil-Sang Yoo
Creative Informatics & Computing Institute, 145 Anam-ro, Seongbuk-gu, Seoul, S.Korea
Tel.: +82-2-3290-1674, Email: ksyoo@korea.ac.kr

Ⅰ. Introduction

After the pandemic, the global growth of e-commerce and the increase in online shopping have led to a significant increase in parcel shipping. Essential to logistics, shipping labels contain sensitive personal information such as names, addresses, and contact details. The exposure of such sensitive data, like addresses and contact numbers, on shipping labels can infringe on personal privacy and potentially lead to criminal risks. Hence, the importance of processing images containing sensitive information is becoming increasingly recognized. Moreover, while various post-processing techniques are employed to obscure sensitive information, masking the entire area containing such data can diminish the authenticity and natural feel of an image.

In recent years, Optical Character Recognition (OCR) has been widely employed as a method to detect personal information from images. However, the effectiveness of OCR is significantly influenced by factors such as brightness, contrast, and distortion of the image, resulting in inconsistent recognition rates, particularly for the Korean language. To address these challenges, Park (2017) proposed a system that incorporates digit-based fixed pattern recognition and Convolutional Neural Network(CNN) to detect personal information from identification cards, offering an improved method for identification and data masking[1]. Despite its advancements, this approach is limited to detecting simple numerical patterns and faces difficulties in recognizing uncorrected or distorted images. This limitation underscores the need for more robust solutions capable of handling diverse image conditions and complex data patterns.

Therefore, this paper addresses the limitations of traditional masking techniques that indiscriminately mask entire areas in images exposing postal and parcel shipping labels. We propose an algorithm based on the KoELECTRA model, a form of Named Entity Recognition(NER) technology. Prior to entity recognition, the TrOCR text recognition algorithm is utilized to identify text within images, and the KoELECTRA model is then used to determine if the recognized text contains personal information, ultimately masking only the sensitive text. Thus, our system locally masks text containing personal information on the shipping labels within images, preventing data leakage while preserving the flow of the footage.

The structure of this paper is as follows: Section 2 discusses the prior studies on image segmentation and text recognition techniques employed in our system. Section 3 introduces the architecture of the proposed system and the models used. Section 4 presents the results of the proposed models with images and statistical data displayed in tables. Finally, Section 5 discusses the expected outcomes of the research and future study directions.

Ⅱ. Related Work

2.1 Image segmentation

Image segmentation, the process of extracting specific regions from an image, is applicable in various fields such as autonomous driving, medical image analysis, and augmented reality. A prominent method for image segmentation is machine learning-based segmentation[2]. This approach consists of extracting features from images and training a segmentation model, with numerous recent studies being conducted. Notably, among machine learning-based methods, CNN have gained significant attention for their effective feature extraction capabilities in image segmentation.

One of the representative models in CNN-based image segmentation is the Fully Convolutional Network(FCN). FCN is capable of delivering segmentation results for every pixel within an image.

Recently, various studies have been conducted to address the limitations of FCN, such as reduced resolution and slow processing speeds. Notable improvements include models like U-Net[3], DeepLabv3[4], and DINOv2[5]. U-Net builds on the architecture of FCN but enhances accuracy in the segmentation process by symmetrically arranging the encoder and decoder. It incorporates skip connections between the encoder and decoder layers to preserve high-resolution information. DeepLabv3 utilizes atrous convolutions to achieve more precise object boundaries and segmentation. It effectively handles objects of varying sizes within the network and incorporates a module in the final segmentation phase that leverages image-level features to achieve refined object boundaries. DINOv2, a model based on self-supervised learning, employs contrastive learning. It takes advantage of a broader dataset compared to traditional FCN and can flexibly respond to variations in images. The system proposed in this paper performs image segmentation using the DINOv2 model, which has been pre-trained on approximately 142 million images, enabling effective learning of features from a large-scale shipping label image dataset.

2.2 Optical character recognition

OCR is a technology used to identify text within images or videos, consisting of two main processes: text detection and text recognition. Text detection involves extracting regions containing text from images or videos, serving as a preprocessing step necessary for text recognition. Text recognition identifies and converts text from these detected regions into digital text.

Among deep learning-based OCR technologies, methods based on CNN have gained prominence for effectively extracting features from images and videos. Notable models for CNN-based text detection include Efficient and Accurate Scene Text detector(EAST) and Character Region Awareness for Text Detection(CRAFT)[6]. Particularly, CRAFT utilizes both CNN and Recurrent Neural Network(RNN) models to enhance the accuracy of text detection. It extracts features using CNN and processes each pixel sequentially using RNN, specialized for sequential data, to accurately detect the boundaries of text regions.

In terms of text recognition, the Convolutional Recurrent Neural Network(CRNN)[7] represents a leading technology that, like CRAFT, combines CNN with RNN to improve the accuracy of text recognition. Recently, several studies have been conducted to improve the performance of CRNN, with notable advancements including SENet and TrOCR[8]. TrOCR, in particular, has improved upon the CRNN framework by using a Transformer-based network architecture instead of an RNN. This Transformer architecture processes the information all at once, considering the entire sequence, which enables efficient parallelized learning and inference. The CNN extracts feature from images or videos, and the Transformer module, taking these features as input, recognizes text. During this process, the Transformer uses an Attention module to effectively transfer information between CNN and the Transformer, and a Multi-Head Attention module to adapt the system to a variety of texts. This paper employs the TrOCR model in the OCR process.

2.3 Named entity recognition

Named Entity Recognition(NER) is a crucial area of study in the field of natural language processing, involving the identification of named entities within sentences[9]-[11]. For example, the task entails recognizing the category to which a word representing a name belongs. In the sentence "Eugene entered Korea University in 2024," the entities would be recognized as Eugene(name), 2024(time), and Korea University(organization).

Therefore, NER is utilized in various applications, including document summarization, text mining, and analysis. Recently, deep learning-based NER technologies have gained attention. These technologies enable flexible and accurate named entity recognition through automatic feature extraction and extensive data training[12].

A prominent method in deep learning-based NER is the Transformer-based approach. Transformers specialize in processing sequence data, dividing sentences into tokens, and classifying whether each token represents an entity. Transformer-based models like Bidirectional Encoder Representations from Transformers(BERT)[13][14][15] and Efficiently Learning an Encoder that Classifies Token Replacements Accurately(ELECTRA)[16] are effective for named entity recognition because they consider the word information in sentences simultaneously. BERT, a pre-training model for language modeling, employs Masked Language Modeling(MLM) and Next Sentence Prediction(NSP) during its training phase. MLM involves masking random tokens in a sentence and restoring them, while NSP predicts the connectivity between two sentences. This approach demonstrates excellent performance in various natural language processing tasks through transfer learning, although it has the drawback of low training efficiency.

ELECTRA introduces the concept of Generative Adversarial Network(GAN) to natural language processing as a pre-training model. It utilizes a structure where a generator and discriminator networks compete and learn from each other, making the pre-training process efficient. Compared to BERT, ELECTRA operates with a smaller model size and a relatively simpler model structure, leading to faster inference speeds and efficiency.

KoELECTRA[17] is a natural language processing model specialized for the Korean language, based on ELECTRA. Like ELECTRA, it uses a GAN-based learning method, allowing for quick inference with a simple model structure and pre-training adapted to the characteristics of Korean text, effectively handling the context and structure of Korean sentences. In this paper, KoELECTRA is utilized during the named entity recognition stage for masking written Korean on shipping labels.

III. Text Identification and Masking Model Based on Named Entity Recognition

This study proposes a text recognition algorithm model for identifying and masking personal information in texts. As shown in Figure 1, the modeling process involves several steps: data collection, image segmentation, image quality enhancement, text location detection, character recognition, and named entity recognition, culminating in generating the final masked image. This chapter details the specific image processing procedures performed at each stage of the proposed model.

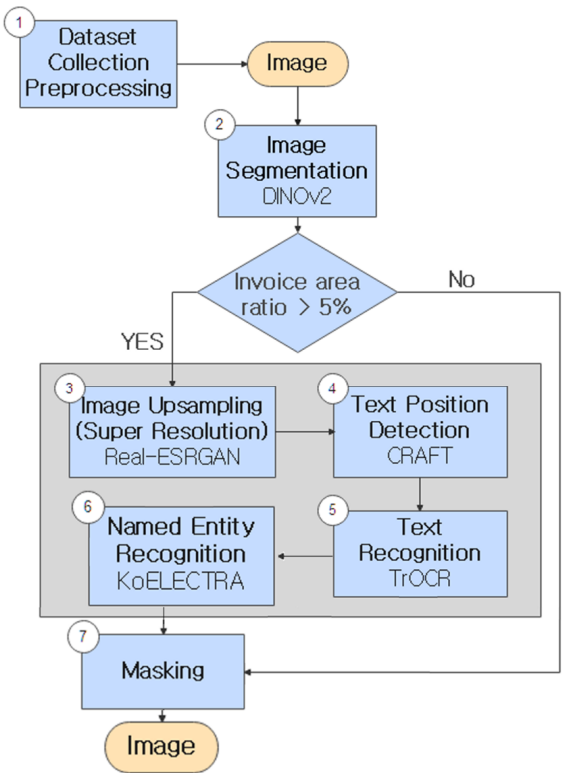


Fig. 1. Overall flowchart of the proposed model

3.1 Dataset collection and preprocessing

To collect datasets for model training, we performed web crawling for shipping label images using search engines such as Naver and Google. We specifically collected labels from seven major courier companies in South Korea, which dominate the parcel delivery market, as well as from convenience store courier services. These companies, in order of preference, include Korea Post, CJ Logistics, Coupang, Hanjin Express, Market Kurly, Lotte Logistics, Logen Logistics, and convenience store-based courier services. The collected image dataset was categorized into three types: intact shipping labels, damaged labels that were torn or obscured, and labels with partial masking. Ultimately, the number of shipping label images used for training, arranged in order of company preference, was 255, 194, 193, 96, 68, 56, 36, and 230, respectively.

Prior to training, we generated labeled images for use in image segmentation using the open data annotation platform, Computer Vision Annotation Tool(CVAT)[18]. Intact shipping label images were classified as Type 1, damaged labels as Type 2, and partially masked labels as Type 3. Figure 2 illustrates the proportion of the area occupied by the shipping labels in the dataset, with shipping labels appearing on average in 39% of the images across all types.

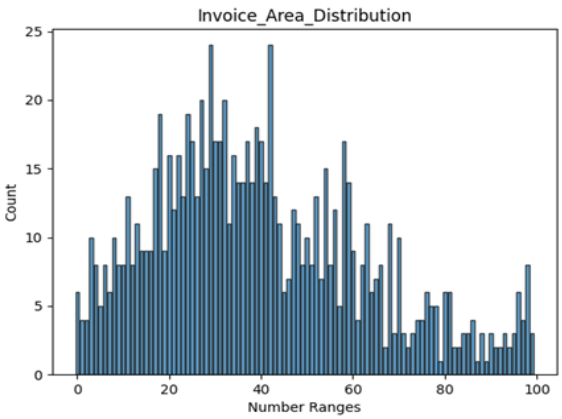


Fig. 2. Rate of shipping label area distribution

Next, to improve the accuracy of image segmentation training, we augmented the collected image dataset and its boxed version by setting hyperparameters as shown in Table 1. We employed various augmentation techniques including horizontal and vertical flips, rotations, and scaling and cropping methods. As a result, as illustrated in Figure 3, when the original number of images was augmented by three to four times, the highest performance was observed, with a validation dice score of 0.81 after 3000 steps.

Table 1. Hyperparameter of augmentation

Epochs	5
Batch size	1
Learning rate	1e-05
Training size	822
Validation size	91
Checkpoints	True
Device	CPU
Images scaling	0.5
Mixed precision	False

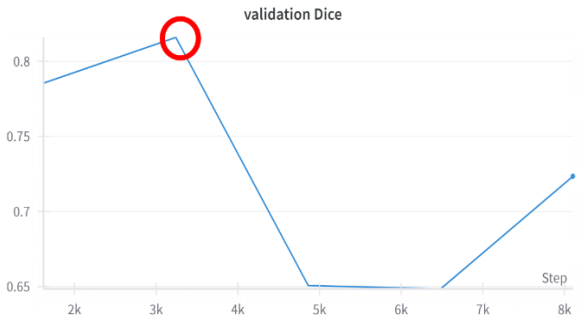


Fig. 3. Graph of validation dice

3.2 Parcel shipping label object detection

Following the preprocessing outcomes in Section 3.1, the parcel shipping label objects within the images were detected using the DINOv2 image segmentation model. We employed a segmentation model rather than an object detection model to precisely determine the boundaries and area ratios of the parcel shipping labels.

Segmentation models, compared to object detection models, can detect the boundaries of objects more finely at the pixel level by dividing them into multiple areas, allowing for accurate calculation of area ratios. The detected area ratios of the parcel shipping label objects are subsequently used as criteria for image classification in the processing stages.

In this study, we considered three segmentation models: YOLOv8[19][20], U-Net, and DINOv2. YOLOv8 is suitable for real-time video segmentation, where processing speed is crucial. Experimental results showed that DINOv2 provided more precise detection of parcel shipping label areas and boundaries compared to U-Net. Therefore, DINOv2 was ultimately utilized.

(a) Original image

(b) Result of DINOv2

Fig. 4. Shipping label area detection using DINOv2

DINOv2 is a semantic segmentation model that assigns all objects of the same class to the same area. However, this model faced a challenge when multiple shipping labels appeared in an image, as it segmented them into the same area without distinction. To address this, we utilized OpenCV's Contour detection to identify disconnected pixels in the segmentation results, treating them as independent shipping labels. The outcomes of the shipping label area detection in the images are presented in Figure 4.

3.3 Enhancing the quality of parcel shipping labels

Upsampling is the process of increasing the sampling frequency of an image or data, thereby expanding or augmenting its size. In particular, upsampling techniques within images can enhance the resolution and enable better discernment of fine details. Specifically, super-resolution technology transforms low-resolution images into high-resolution ones without compromising the details of the original image. In this paper, we utilize the Enhanced Super Resolution Generative Adversarial Networks (Real-ESRGAN)[21], a prominent deep learning-based super-resolution method that employs GAN model.

Real-ESRGAN mimics the real-world photo degradation process more closely through high-order degradation techniques. This involves repeatedly applying the degradation process D, as described in Equation (1), where each iteration of the process D modifies only the hyperparameters of the same degradation sequence. Additionally, the degradation process, as shown in Equation (2), sequentially applies blur, resize, noise, and JPEG compression to the ground-truth input image y.

$$x = D^n(y) = (D_n \circ \cdots \circ D_2 \circ D_1)(y) \tag{1}$$

$$x = D(y) = [(y \circledast k) \downarrow_r + n]_{JPEG} \tag{2}$$

In the proposed system, Real-ESRGAN is used to enhance the resolution of the shipping label images, thereby facilitating smoother text detection and extraction of positional coordinates. The process is as follows: Initially, the minimum and maximum x and y coordinates of the shipping label areas detected by DINOv2 are used to define a rectangular region enclosing the label. Subsequently, this rectangular area is extracted from the entire image, and the cropped shipping label image is input into the Real-ESRGAN model to generate a high-resolution output. Figure 5 compares the original shipping label images with those improved using super-resolution technology.

(a) Original image

(b) Result of super-resolution x2 with real-ESRGAN

Fig. 5. Super-resolution x2 using real-ESRGAN

3.4 Text location detection in shipping labels

Prior to text extraction from the high-quality cropped images of shipping labels obtained through Real-ESRGAN, we used the CRAFT model to detect the locations of text. CRAFT enhances the accuracy of character detection by combining CNN and RNN technologies, where the CNN extracts feature from the image and the RNN identifies the boundaries of character regions.

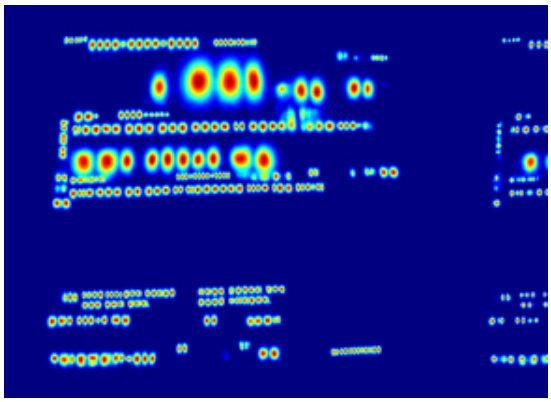
The application process of the CRAFT algorithm is as follows: Given a specific word sample 'w' from the training data, the area of the word is denoted as '$R(w)$', and the length of the word as '$l(w)$'. By segmenting the word into characters, the boxes for each character and their total length '$lc(w)$' are estimated. Subsequently, a confidence score for each word sample 'w' is calculated using the following equation (3).

$$sconf(w) = \frac{l(w) - \min(l(w), |l(w) - lc(w)|)}{l(w)} \tag{3}$$

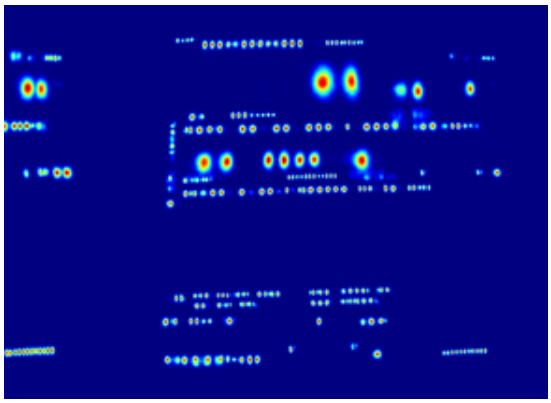
In the given image, a confidence score '$sconf(w)$' is assigned to each pixel 'f' within the word region '$R(w)$', and different values are assigned to all other pixels to generate a confidence map 'Sc'. The overall objective 'L' is to compute the sum of the values of the confidence map '$Sc(p)$' for each pixel 'p', and the product of the sum of the squared differences between this and the ideal area score '$Sr(p)$' and affinity score '$Sa(p)$'.

$$L = \sum_p Sc(p) \cdot (\|Sr(p) - S_r^*(p)\|_2^2 + \|Sa(p) - S_a^*(p)\|_2^2) \tag{4}$$

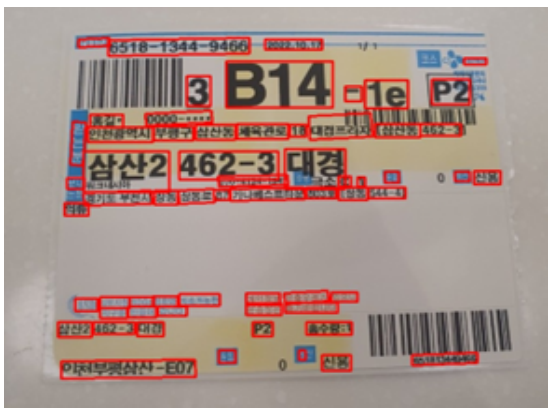
In equation (4), '$S_r^*(p)$' and '$S_a^*(p)$' represent the ideal area and affinity scores, respectively, while '$Sr(p)$' and '$Sa(p)$' denote the model's predictions for these scores. When training with synthetic data, since the actual ground truth is available, the confidence map '$Sc(p)$' for every pixel 'p' is set to 1. The coordinates of text regions obtained through the CRAFT algorithm are then used in the text recognition and masking stages. The heatmap showing the regions, locations, and boundaries of characters detected by CRAFT, along with the extracted text areas, is presented in Figure 6.



(a) Region score



(b) Affinity score

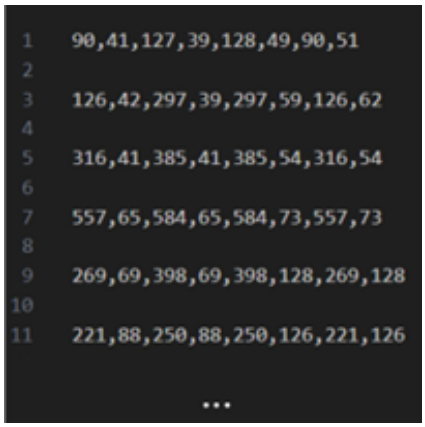


(c) Result image of CRAFT

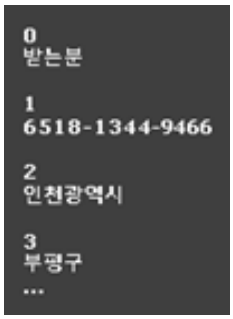
Fig. 6. Text recognition using CRAFT

3.5 Text recognition within parcel shipping labels

In this study, we considered two text recognition models: CRNN and TrOCR. Both models were trained using 9,000 images of printed Korean text and 1,000 images of handwritten Korean text provided by AI-Hub[22]. TrOCR demonstrated superior performance; therefore, it was selected for use in this research. Figure 7(a) illustrates text locations obtained using CRAFT, which TrOCR utilizes to identify text positions and extract each text as shown in Figure 7(b). While cropped images generated by Real-ESRGAN showed excellent performance in confirming text locations, the quality alterations in these cropped images led to decreased character recognition rates when used for text recognition. Therefore, original images were used for text recognition purposes[23].



(a) Text location with CRAFT



(b) Extracted text with TrOCR

Fig. 7. Text Extraction using CRAFT and TrOCR

TrOCR consists of an encoder that converts images into vectors and a decoder that translates these vectors into words. It takes an image containing text as input (3*H0*W0), resizes it, and then splits it into patches of size p*p. Next, as described in Equation (5), each patch is embedded and then linearly projected to the Transformer's hidden size D. The D vectors are fed into the decoder, where the word corresponding to each vector h_i is identified. Equation (6) represents

the probability that the $i-th$ vector's word is the $j-th$ word among V possible words.

$$h_i = Proj(Emb(T_i)) \tag{5}$$

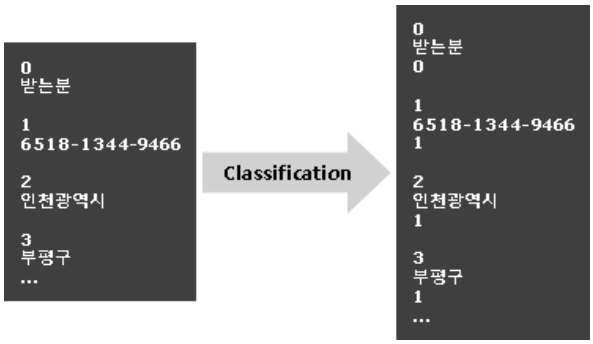
$$\sigma(h_{ij}) = \frac{e^{h_{ij}}}{\sum_{k=1}^{V} e^{h_{ij}}} \text{ for } j = 1, 2, ..., V \tag{6}$$

3.6 Named entity recognition and personal information identification

To determine whether the text information obtained from the TrOCR text recognition model contains personal data such as names, addresses, and phone numbers, we applied the fine-tuned named entity recognition model, KoELECTRA-small-v3-modu-ner. This model was pretrained on a Korean dataset categorized into 15 tag sets according to the broad classification standards of the Korea Information and Communication Technology Association(TTA). It takes text information as input and assigns classes to words(named entities) corresponding to people, roads, buildings, etc., according to predefined classification standards, as shown in Figure 8(a). In our system, the named entity recognition model classifies personal information in the text of shipping labels—such as names, addresses, and phone numbers—as Class 1, while all other text is classified as Class 0. For names and addresses, the model uses tags for people(PS), locations(LC), and areas(AR) from its broad classification criteria to classify them as Class 1. Additionally, numerical information like phone numbers and shipping label numbers is classified as Class 1 through a separate definition for masking purposes. Figure 8(b) illustrates the results of the text recognition model TrOCR and its subsequent classification using the KoELECTRA named entity recognition model. Using the class information obtained from this process along with the text locations previously determined by CRAFT, masking is performed on the original images.

분류	표기	정의
ARTIFACTS	AF	사람에 의해 창조된 인공물로 문화재, 건물, 악기, 도로, 무기, 운송수단, 작품명, 공산품 명이 모두 이에 해당
ANIMAL	AM	사람을 제외한 짐승
CIVILIZATION	CV	문명/문화
DATE	DT	기간 및 계절, 시기/시대
EVENT	EV	특정 사건/사고/행사 명칭
STUDY_FIELD	FD	학문 분야, 학과 및 유파
LOCATION	LC	지역/장소와 지형/지리 명칭 등을 모두 포함
MATERIAL	MT	원소 및 금속, 암석/보석, 화학물질
ORGANIZATION	OG	기관 및 단체 명칭
PERSON	PS	인명 및 인물의 별칭 (유사 인물 명칭 포함)
PLANT	PT	꽃/나무, 육지식물, 해초류, 버섯류, 이끼류
QUANTITY	QT	수량/분량, 순서/순차, 수사로 이루어진 표현
TIME	TI	시계상으로 나타나는 시/시각, 시간 범위
TERM	TM	타 개체명에서 정의된 세부 개체명 이외의 개체명
THEORY	TR	특정 이론, 법칙 원리 등

(a) 15 classes of tag set from TTA



(b) Result image of KoELECTRA (0 or 1)

Fig. 8. Classification of extracted texts using KoELECTRA

IV. Implementation Results

In this chapter, we present the results of fine-tuning the DINOv2 and TrOCR models aimed at enhancing the performance of the proposed system, as well as the final outcomes of the masking process.

4.1 Image segmentation

Image segmentation takes the images processed during data collection and preprocessing as input. The models considered for the image segmentation include YOLOv8, U-Net, and DINOv2. To compare the performance of these models, we compared the original images(ground truth) with the mask images predicted by the three models.

The hyperparameters used were as follows: all three models employed the SGD optimizer; YOLOv8 used a default learning rate of 0.01, while U-Net and DINOv2 used 4e-5. The batch sizes were set at 32, 16, and 8 for YOLOv8, U-Net, and DINOv2, respectively. YOLOv8 utilized a model pretrained on the COCO 2017(Common Objects in Context 2017) dataset[24], U-Net was used as a basic model without pre-training, and DINOv2 was simulated with a pretrained backbone through fine-tuning.

Performance comparison among the three models was conducted using metrics such as dice score and mAP(mean Average Precision), as outlined in Table 2. The dice score, which measures the similarity between the mask generated by the model and the actual image mask, is calculated by dividing the size of the overlapping area by the total area size. Scores closer to 1 indicate superior model performance. The dice scores for YOLOv8, U-Net, and DINOv2 were 0.74, 0.68, and 0.82, respectively, indicating that DINOv2 performed the best. mAP, which calculates the average over all classes by considering the area under the curve of precision and recall for each class, also indicates better object detection performance as it approaches 1. For a more accurate performance comparison, mAP was measured incrementally from an IoU of 0.5 to 0.95 by 0.05, alongside mAP 50, which assumes detection only if IoU is above 0.5. The IoU indicates the degree of overlap between the predicted bounding box and the ground truth bounding box. As shown in Table 2, the mAP scores for YOLOv8, U-Net, and DINOv2 were 0.68, 0.64, and 0.74, respectively, and the mAP 50 scores were 0.81, 0.77, and 0.86, respectively, with DINOv2 showing the highest performance in both metrics. Additionally, as depicted in Figure 9, DINOv2 was able to extract the borders of the shipping labels most clearly in the images. Therefore, DINOv2, having demonstrated the highest performance across all metrics, was used for image segmentation in this system.

Table 2. Full and PCA projected size of the feature sets

	Dice score	mAP	mAP 50
YOLOv8	0.74	0.68	0.81
U-Net	0.68	0.64	0.77
DINOv2	0.82	0.74	0.86

(a) Original image

(b) Result of U-Net

(c) Result of DINOv2

Fig. 9. Comparison of image segmentation between U-Net and DINOv2

4.2 Character recognition

For the character recognition model, simulations were conducted using both the CRNN and TrOCR models on images resulting from the text area detection model. To compare the performance of the two models, we analyzed three images(ground truth) as shown in Figure 10, and compared the results applied by each model. The hyperparameters utilized in the experiment were as follows: both CRNN and TrOCR utilized the Adam optimizer with a learning rate set at 1e-3. Additionally, the batch sizes were set at 80 for CRNN and 20 for TrOCR, with only the TrOCR model undergoing fine-tuning. Experimental results, as shown in Table 3, indicated that TrOCR exhibited superior character recognition accuracy across the three images compared to CRNN. Consequently, the TrOCR model was selected for text recognition.

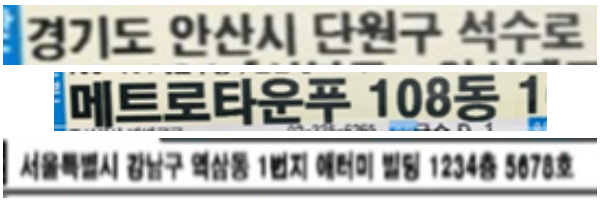


Fig. 10. Input images of text recognition model

Table 3. Full and PCA projected size of the feature sets

Ground truth	Fig. 10 (a)	경기도 안산시 단원구 석수로
	Fig. 10 (b)	메트로타운푸 108동 1
	Fig. 10 (c)	서울특별시 강남구 역삼동 1번지 애터미 빌딩 1234층 5678호
CRNN	Fig. 10 (a)	경기도 안산시단원구석수로
	Fig. 10 (b)	로타운푸 108동 1
	Fig. 10 (c)	서땅ITp의치 바지외필땡뚤15545ttet
TrOCR	Fig. 10 (a)	경기도 만산시 단원구 석주로
	Fig. 10 (b)	메트로타운푸108동1
	Fig. 10 (c)	서물목킬시 감님구 1비지 매비미 밀된

4.3 Named entity recognition

To determine whether the text extracted by TrOCR contains personal information, named entity recognition was performed using the BERT and KoELECTRA models through binary classification. Both models were trained using Binary Cross Entropy Loss(BCE Loss). Upon measuring the validation accuracy, as shown in Table 4, BERT achieved an accuracy of 0.68, while KoELECTRA achieved a higher accuracy of 0.82.

Table 4. Results of named entity recognition with BERT and KoELECTRA

Model	Accuracy
BERT	0.68
KoELECTRA	0.82

4.4 Final model

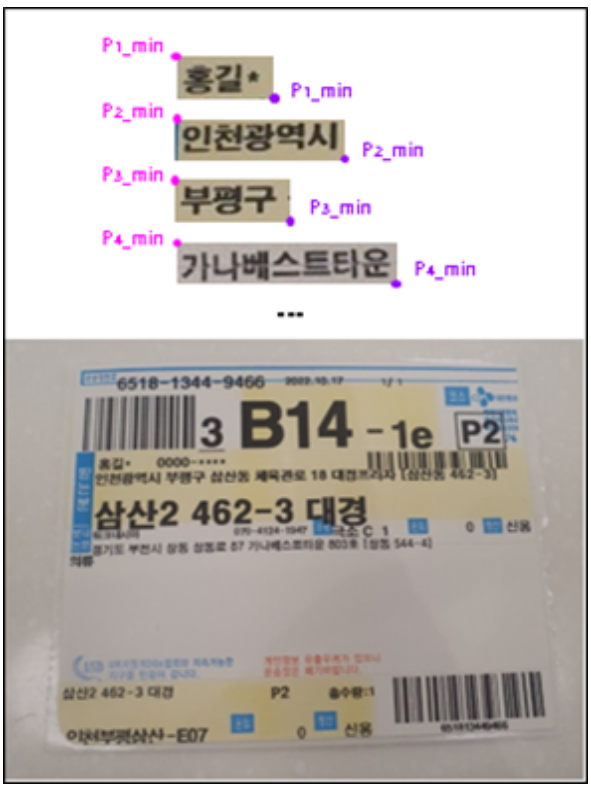
To measure the performance of the final model, we conducted simulations integrating the models introduced from Sections 4.1 to 4.3 and verified their accuracy. Simulations were performed on six shipping label images containing personal information. During the image segmentation, image quality enhancement, and text detection stages, the models DINOv2, Real-ESRGAN, and CRAFT were applied based on experimental and model characteristics. For text recognition and named entity recognition, it was crucial for our model to accurately recognize the preprocessed images and texts and identify personal information. Therefore, we combined two text recognition models and two named entity recognition models, conducting a total of four simulations. Accuracy was measured based on the ratio of personal information correctly identified from the text images, as extracted by the CRAFT model. As shown in Table 5, the combination of TrOCR and KoELECTRA as the text recognition and named entity recognition models, respectively, achieved the highest accuracy of 0.8124 among the four scenarios.

Table 5. Results of named entity recognition model with text recognition model

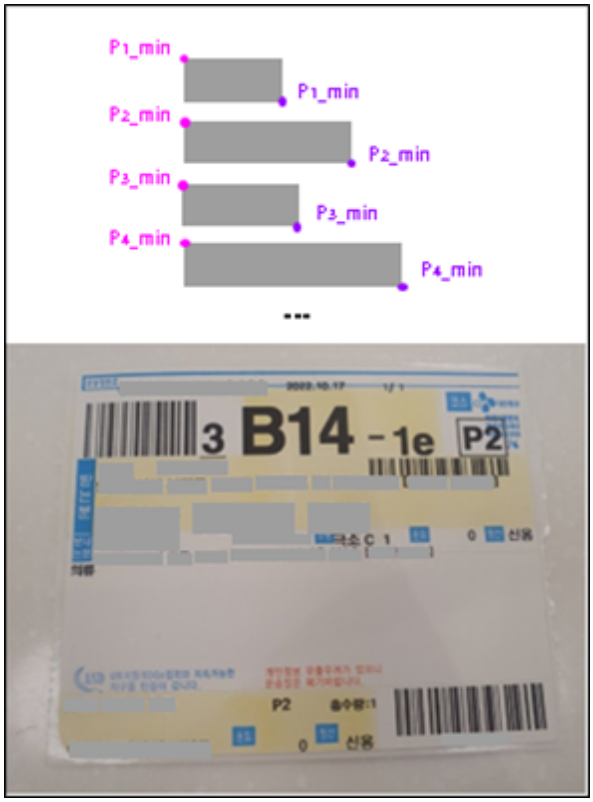
Model	Accuracy
TrOCR+KoELECTRA	0.81
TrOCR+BERT	0.68
CRNN+KoELECTRA	0.62
CRNN+BERT	0.51

4.5 Personal information masking

Personal information masking, or de-identification, involves techniques such as mosaicking and blurring to obscure personal text data. In this study, instead of distorting images with a fixed-size pixel mosaic, blurring was applied to preserve the natural appearance of the images. The Gaussian Blur used in our system averages the pixel values around each pixel in the input image by assigning weights to the surrounding pixels, thus effectively masking the data. Figure 11 demonstrates the experimental results of identifying the positions of personal information text within shipping labels and applying masking based on these locations.



(a) Text position coordinates extracted by CRAFT



(b) Example image of final results

Fig. 11. Masking stage

4.6 Comparative analysis

The proposed system offers more functionalities compared to existing systems from a technical perspective, as shown in Table 6. First, it provides a clear classification of the recognized text according to the type of personal information (e.g., name, phone number, address). Second, the system applies the CRAFT technique to accurately detect the position of unstructured, variable data and adapt to different text sizes. Third, to enable precise correction and detection from images distorted at various angles, the system incorporates Real-ESRGAN generation technology and semantic segmentation. Overall, the system demonstrated a 19.9% improvement in recognition rate for distorted images compared to Tesseract OCR, with a successful classification and masking rate of 81.2% for various types of personal information.

Table 6. Comparison and evaluation results with similar models

Category	Tesseract OCR	Park[1]	Proposed system
Types of personal information	Basic English recognition (high), basic Korean recognition (medium)	ID, email addresses, contact numbers, account numbers, vehicle numbers	Recognizes names, email addresses, and contact numbers
Recognition of image distortion	Low recognition rate based on image quality	Fixed pattern recognition based on digit position	Recognition based on individual characters in an image Character Region Awareness for Text detection
Applied technology	TesseractOCR	TesseractOCR + CNN	Transformer-based OCR, CNN, NER, Real-ESRGAN, CRAFT
Recognition rate	Rrecognition depending on preprocessing about 61.3%	- For ID numbers 94% - 86.6% recognition of faces	- Accurate recognition of personal information 81.2%
Specialized area	No specialized model		Specialized model enabled

V. Conclusion

The convenience of online shopping and the explosive increase in delivery services following the pandemic have brought to the forefront issues related to the exposure of sensitive personal information, such as shipping label data. Furthermore, advancements in camera resolution technology and the increasing use of mobile devices such as smartphones and smartwatches, along with platforms like social media and video services, have significantly heightened the risk of personal information being leaked through media. This paper proposes a technical method for automatically detecting, extracting, and masking personal information texts from shipping labels exposed in everyday settings, such as SNS photos and YouTube videos.

The implementation results confirmed the smooth detection, extraction, and masking of personal information texts such as names, addresses, phone numbers, and shipping label numbers in images containing shipping labels. Additional fine-tuning improved the recognition rate of shipping label areas within images and the text recognition rate within those areas. Compared to existing similar technologies, the proposed system has been confirmed to detect a wider variety of personal information types from a technical perspective, accurately detect the location of unstructured and variable data, and adapt to various text sizes. Additionally, it showed a 19.9% higher recognition rate on distorted images compared to Tesseract OCR, and the system achieved an 81.2% success rate in classifying and masking various types of personal information. By locally masking only the parts of the image containing personal information texts, our system minimizes the disruption to user immersion and, given that shipping label information can be easily collected via web search as observed during our data collection process, is expected to mitigate the issue of personal information leakage through shipping labels as well. Moreover, the proposed technology has the potential to be applied to a wider range of documents containing personal information, such as identity cards and documents, allowing for diverse utilization of the system.

Future research tasks will include extending the masking scope to include elements like shipping label barcodes and developing algorithms to improve the accuracy and recognition rate of text at the word level.

References

[1] S. C. Park, "Design and Implementation of Personal Information Identification and Masking

System Based on Image Recognition", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 17, No. 5, pp. 1-8, Oct. 2017. https://doi.org/10.7236/JIIBC.2017.17.5.1.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", Computer Vision and Pattern Recognition(CVPR), Boston, MA, USA, pp. 3431-3440, Jun. 2015. https://doi.org/10.1109/cvpr.2015.7298965.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", Medical Image Computing and Computer-Assisted Intervention(MICCAI), Munich, Germany, Vol. 9351, pp. 234-241, Oct. 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[4] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation", arXiv:1706.05587v3, Dec. 2017. https://doi.org/10.48550/arXiv.1706.05587.

[5] M. Oquab, et al., "Dinov2: Learning robust visual features without supervision", arXiv:2304.07193v2, Feb. 2023. https://doi.org/10.48550/arXiv.2304.07193.

[6] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection", Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, pp. 9357-9366, Jun. 2019. https://doi.org/10.1109/cvpr.2019.00959.

[7] 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", Transactions on Pattern Analysis and Machine Intelligence(TPAMI), Vol. 39, No. 11, pp. 2298-2304, Nov. 2017. https://doi.org/10.1109/tpami.2016.2646371.

[8] M. Li, et al., "Trocr: Transformer-based optical character recognition with pre-trained models", AAAI Conference on Artificial Intelligence, pp. 13094-13102, Jun. 2023. https://doi.org/10.1609/aaai.v37i11.26538.

[9] C. Liu, Y. Yu, X. Li, and P. Wang, "Named Entity Recognition in Equipment Support Field Using Tri-Training Algorithm and Text Information Extraction Technology", IEEE Access, Vol. 9, pp. 126728-126734, Sep. 2021. https://doi.org/10.1109/access.2021.3109911.

[10] G. Kim, J. Son, J. Kim, H. Lee, and H. Lim, "Enhancing Korean Named Entity Recognition With Linguistic Tokenization Strategies", IEEE Access, Vol. 9, pp. 151814-151823, Nov. 2021. https://doi.org/10.1109/access.2021.3126882.

[11] M. Kutbi, "Named Entity Recognition Utilized to Enhance Text Classification While Preserving Privacy", IEEE Access, Vol. 11, pp. 117576-117581, Oct. 2023. https://doi.org/10.1109/ACCESS.2023.3325895.

[12] S. Jun, E. Yun, H. Lee, M. Kang, J. Kim, and G. Yoo, "Development of a Candidate Scoring Network based Speaker Recognition System Utilizing Textual Context Information", The Journal of Korean Institute of Information Technology, Vol. 22, No. 5, pp. 151-163, May 2024. https://doi.org/10.14801/jkiit.2024.22.5.151.

[13] H. Son, Y. Han, K. Nam, S. Han, and G. Yoo, "Development of a News Trend Visualization System based on KPF-BERT for Event Changes and Entity Sentiment Analysis", The Journal of Korean Institute of Information Technology, Vol. 22, No. 1, pp. 203-213, Jan. 2024. https://doi.org/10.14801/jkiit.2024.22.1.203.

[14] S. Han, D. Yu, B. W. On, and I. Lee, "Empirical Study on the Loss Functions of Contrastive Learning-based Multi-scale BERT model for Automated EssayScoring", The Journal of Korean Institute of Information Technology, Vol. 21, No. 9, pp. 51-63, Sep. 2023. https://doi.org/10.14801/jkiit.2023.21.9.51.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv:1810.04805v2, Oct. 2018. https://doi.org/10.

48550/arXiv.1810.04805.

[16] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators", arXiv:2003.10555v1, Mar. 2020. https://doi.org/10.48550/arXiv.2003.10555.

[17] S. S. Lee, S. M. Cha, B. Ko, and J. J. Park, "Extracting Fallen Objects on the Road From Accident Reports Using a Natural Language Processing Model-Based Approach", IEEE Access, Vol. 11, pp. 139521-139533, Dec. 2023. https://doi.org/10.1109/access.2023.3339774.

[18] M. Guillermo, et al., "Implementation of Automated Annotation through Mask RCNN Object Detection model in CVAT using AWS EC2 Instance", IEEE Region Conference (TENCON), Osaka, Japan, pp. 708-713, Dec. 2020. https://doi.org/10.1109/tencon50793.2020.9293906.

[19] S. Niu, et al., "Research on a Lightweight Method for Maize Seed Quality Detection Based on Improved YOLOv8", IEEE Access, Vol. 12, pp. 32927-32937, Feb. 2024. https://doi.org/10.1109/access.2024.3365559.

[20] Y. G. Lee and J. H. Kang, "Performance Analysis by the Number of Learning Images on Anti-Drone Object Detection System with YOLO", Journal of Korean Institute of Communications and Information Sciences, Vol. 49, No. 3, pp. 356-360, Mar. 2024. https://doi.org/10.7840/kics.2024.49.3.356.

[21] X. Wang, L. Xintao, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data", International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, Vol. 1, No. 2, pp. 1905-1914, Nov. 2021. https://doi.org/10.1109/iccvw54120.2021.00217.

[22] S. J. Kim, W. H. Son, J. H. Lee, H. H. Nguyen, and H. Y. Jeong, "Sampling-based Analysis of Labeling Errors in AI-Hub Traffic-Light Datasets", Journal of the Institute of Electronics and Information Engineers, Vol. 60, No 3, pp. 109-112, Mar. 2023. https://doi.org/10.5573/ieie.2023.60.3.109.

[23] H. J. Kim and J. H Nah, "Comparative analysis of the deep-learning-based super-resolution methods for generating high-resolution texture maps", Journal of the Korea Computer Graphics Society, Vol. 29, No. 5, pp. 31-40, Dec. 2023. https://doi.org/10.15701/kcgs.2023.29.5.31.

[24] T. Y. Lin, et al., "Microsoft coco: Common objects in context", European Conference on Computer Vision (ECCV), Zurich, Switzerland, pp. 740-755, Sep. 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

Authors

Ha-Eun Kim

2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University

2024. 2 : BS degree, School of Electrical Engineering, Korea University

2024. 3 ~ present : MS candidate, Department of Electrical and Computer Engineering, Korea University

Research interests : Deep Learning, Machine Learning, Networking and Distributed Computing

Soo-Yong Kim

2022. 8 : BS degree, School of Health and Environmental Science, Korea University

2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University

2024. 8 : MS degree, Department Artificial Intelligence, Seoul National University

Research interests : Multimodal, Deep Learning, Image Generation

Myeong-Seop Kim

2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University
2024. 8 : BS degree, Department of Mathematics, Korea University
Research interests: Deep Learning, Computer Vision, Natural Language Process, Data Science

Sang-Ho Kim

2021. 2 : BS degree, Department of Computer Engineering, Gachon University
2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University
2024. 2 : MS degree, Department of Business, Hankuk University of Foreign Studies
Research interests: Data Science, Data Analysis, CRM, Machine Learning, Deep Learning, Image Processing

Ye-Jin Cho

2023. 2 : BS degree, Departmentl of Multimedia Engineering, Dongguk University
2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University
2024. 3 ~ present : MS candidate, Department of Computer and Artificial Intelligence, Dongguk university
Research interests: Computer Vision, Deep Learning, Machine Learning, Data Science, Computer Graphics

Eun-Sun Choi

2023. 12 : Certitificate of Intelligent Information Software Academy, Korea University
2020. 3 ~ present : Undergraduate student, Department of Computer Science and Engineering, Korea University
Research interests: Deep Learning, Machine Learning, Computer Vision

Gil-Sang Yoo

2010. 2 : Phd degree, Department of Imaging Engineering, Chungang University
2010. 3 ~ Present : Director, Korea Computer Game Society
2011. 3 ~ Present : Professor, Department of Creative Informatics & Computing Institute, / Intelligent Information Software Academy, Korea University
2023. 3 ~ Present : Senior Vice President, Korea Media Art Industry Association
Research interests : Data Science, 3D Content, Machine Learning, Deep Learning, Computer Education