Computer Vision and Image Understanding 238 (2024) 103880

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Disentangled generation network for enlarged license plate recognition and a unified dataset

Chenglong Li[a,b,c], Xiaobin Yang[b,d], Guohao Wang[b,d], Aihua Zheng[a,b,c,*], Chang Tan[e], Jin Tang[a,b,d]

[a] *Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, 230601, China*
[b] *Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei, 230601, China*
[c] *School of Artificial Intelligence, Anhui University, Hefei, 230601, China*
[d] *School of Computer Science and Technology, Anhui University, Hefei, 230601, China*
[e] *iFLYTEK Co., Ltd., Hefei, 230088, China*

ARTICLE INFO

Communicated by Arjan Kuijper

*Keywords:*
Text recognition
Enlarged license plate
Disentangled generation
Generative adversarial network
Benchmark dataset

ABSTRACT

License plate recognition plays a critical role in many practical applications, but license plates of large vehicles are difficult to be recognized due to the factors of low resolution, contamination, low illumination, and occlusion, to name a few. To overcome the above challenges, the transportation management department generally introduces the enlarged license plate behind the rear of a vehicle. However, enlarged license plates have high diversity as they are non-standard in position, size, and style. Furthermore, the background regions contain a variety of noisy information which greatly disturbs the recognition of license plate characters. In this work, we address the enlarged license plate recognition problem and contribute a dataset containing 9342 images, which cover most of the challenges of real scenes. However, the created data are still insufficient to train deep methods of enlarged license plate recognition, and building large-scale training data is very time-consuming and high labor cost. To handle this problem, we propose a novel data generation framework based on the Disentangled Generation Network (DGNet), which disentangles the generation of enlarged license plate data into the text generation and background generation in an end-to-end manner to effectively ensure the generation diversity and integrity, for robust enlarged license plate recognition. Extensive experiments on the created dataset are conducted, and we demonstrate the effectiveness of the proposed approach in three representative text recognition frameworks.

## 1. Introduction

License plate recognition is an important problem in the field of computer vision and plays a critical role in many practical applications, such as traffic safety, vehicle management, and urban security. However, as shown in Fig. 1(a), when encountering some scenarios such as low resolution, contamination, low illumination and occlusion, the license plates of large vehicles are difficult to be recognized. According to the requirements of the traffic management department, the rear of large vehicles needs to be painted with enlarged license plates to handle shortcomings of the standard license plates. In many practical scenarios, the enlarged license plates are easier to be captured by surveillance cameras, and thus the enlarged license plate recognition plays a significant role in identifying large vehicles.

However, enlarged license plate recognition is also a challenging problem. As shown in Fig. 1(b), on one hand, due to the lack of standard painting requirements in position, size, and style, enlarged license plates are highly diverse. On the other hand, the background contains a variety of noisy information, which has a serious impact on enlarged license plate recognition.

In this work, we standardize the task of enlarged license plate recognition, aiming to answer the following two questions: (1) How to create a unified benchmark dataset to promote the research and development of enlarged license plate recognition? (2) How to improve the performance of deep recognizers when training data are insufficient?

**Benchmark dataset**. To establish a unified benchmark dataset, we collect 9342 enlarged license plate images in 18 provinces in China. It covers most of the real challenges of enlarged license plate recognition such as low resolution, contamination, and occlusion, as shown in Fig. 1(a). Due to the special properties of enlarged license plates, it is hard to collect balanced data for all provinces. Therefore, the issue of long-tail distribution is a key challenging factor in enlarged license plate recognition. To facilitate the training and evaluation of different algorithms, we split the dataset into training and testing sets.

[*] Corresponding author at: School of Artificial Intelligence, Anhui University, Hefei, 230601, China.
*E-mail address:* ahzheng214@foxmail.com (A. Zheng).

https://doi.org/10.1016/j.cviu.2023.103880
Received 25 May 2023; Received in revised form 23 October 2023; Accepted 1 November 2023
Available online 4 November 2023

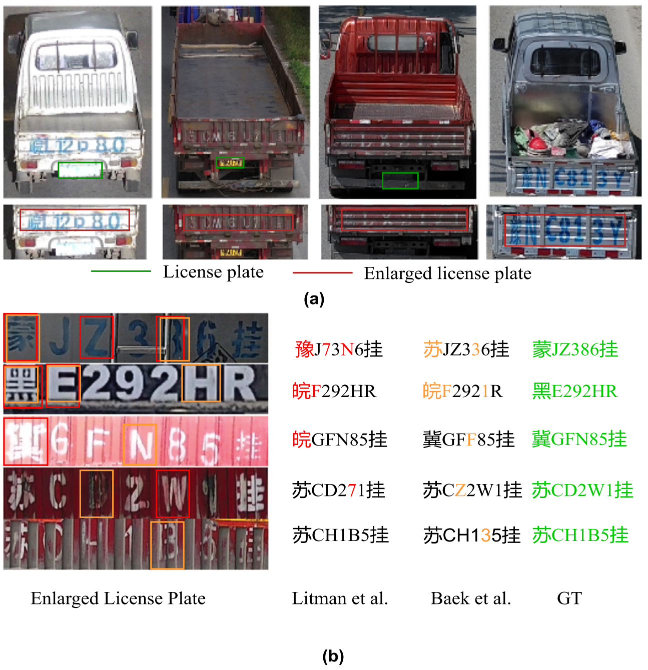


**Fig. 1.** (a) Comparison of enlarged license plates with standard license plates. Enlarged license plates are more suitable than standard license plates in identifying large vehicles under some challenging scenarios. (b) Recognition results of the enlarged license plates by the state-of-the-art recognition methods, including Litman et al. (2020) and Baek et al. (2019). All recognition models are trained using our proposed enlarged license plate dataset. The results show that the task of enlarged license plate recognition is challenging.

In specific, the testing set occupies about 20 % of the whole dataset and the remaining data are used as the training set. We adopt the enlarged license plate recognition accuracy and the character recognition accuracy as our evaluation indicators for different methods.

The huge and diverse challenges such as low illumination and occlusion seriously affect the performance of enlarged license plate recognition. To facilitate the challenge-based performance analysis, we annotate 10 challenges for each image, including inclined angle, abnormal illumination, different spacing, size variation, blur, abrasion, background clutter, non-standard character, double-row plate and occlusion.

**Disentangled generation network**. Enlarged license plates have high diversity as they are non-standard in position, size and style, and background regions contain a variety of noisy information. Therefore, the created dataset is hard to cover all real challenges and thus insufficient to train deep recognition networks. Moreover, building a large-scale training dataset is very time-consuming and high labor cost. To handle this problem, we propose to synthesize large-scale training data to simulate real scenarios.

Many researchers (Cheng et al., 2018; Gupta et al., 2016; Wang et al., 2017) propose to generate synthetic images in natural scenes to improve recognition performance. For example, Wang et al. (2017) introduce W-Distance (Arjovsky and Bottou, 2017) in the training process of CycleGAN, and can synthesize a large number of standard license plates. However, it is easy to result in the mode collapse issue, and thus difficult to generate enlarged license plates with various styles. Luo et al. (2021) introduce an additional recognizer to supervise the generator to ensure the integrity of generated characters. However, existing disentanglement generation algorithms usually extract the background and text information separately through encoders, but are hard to disentangle them well, making the generated enlarged license plates have a lot of noises and errors. The enlarged license plates usually have complex background and high diversity of text, and it thus is difficult to generate various enlarged license plates under the condition of unpaired datasets without the supervision information. Furthermore, it is a more difficult task to generate an enlarged license plate with the specified license plate text.

To handle these problems, we propose to use a task-level disentanglement generation strategy to decompose the enlarged license plate image generation into two sub-tasks, including background image generation and text image generation. Through improving the diversity of background generation and text generation respectively, we can not only generate high-diverse synthetic enlarged license plate images, but control their contents and styles, which are beneficial to boosting the recognition performance significantly. Therefore, we can obtain a variety of synthesized enlarged license plate images by inputting text and background of different styles, without being limited by the mutual influence of the text and background to generate images with blurred background or incomplete text structure.  It effectively ensures the diversity and integrity of generated enlarged license plate data.

The effectiveness of disentangled generation is validated on various tasks (Lee et al., 2020; Huang et al., 2018; Yi et al., 2020; Ning et al., 2021). Inspired by disentangled generation, we propose to disentangle the generation of enlarged license plate data into two processes, including text generation and background generation, to address the problem of mode collapse. In specific, existing disentangled generation networks (Lee et al., 2020; Huang et al., 2018) usually extract content and background from source and target domains respectively, and then combine the content and background to generate target image. However, enlarged license plates have high diversity, which makes it difficult to accurately separate the text and background from an image. Therefore, we propose a task-level disentangled generation network specifically designed for the task of enlarged license plate recognition to avoid the problem of incomplete separation.

For the text generation, we first collect a series of license plate character images. Then, we combine these images into a unified blue background image. To increase the diversity of text images, we augment these character images by changing the attributes of size, shape and position, etc. More importantly, the changes of these attributes are completely retained in synthesized enlarged license plates, and the diversity is thus enhanced. For the background generation, the complex and diverse backgrounds of enlarged license plates are the important factors that affect the performance of recognition methods. To better simulate the real data, we construct a complex and diverse background template set, which contains almost all the backgrounds of enlarged license plates in real scenes. Based on this set, we can randomly combine background templates and text images to generate high-quality synthesized enlarged license plates.

Instead of using recognition methods to supervise generation in existing works (Luo et al., 2021), we use a mask image to supervise the generation by introducing a mask constraint loss, which is calculated by the average absolute error between the mask and output images. Herein, the mask image is obtained by subtracting the blue background image from the text image. The designed loss helps us to effectively ensure the integrity of generated enlarged license plates.

**Contributions**. The main contributions of this work can be summarized as follows.

- To promote the research and development of enlarged license plate recognition, we contribute a dataset containing 9342 images, which cover most of the challenges of real scenes, for enlarged license plate recognition. For free academic usage, we have released this dataset to public.[1]
- To provide a large-scale training data while avoiding time consuming and high labor cost, we propose a novel task-level disentangled generation framework, which disentangles the enlarged license plate generation into the text generation and background generation in an end-to-end manner.

[1] https://github.com/mmic-lcl/Datasets-and-benchmark-code/blob/main/README.md

- We design a series of strategies to ensure the diversity and integrity of generated enlarged license plates. On one hand, we combine a set of augmented text images and a set of constructed background templates to enhance the diversity. On the other hand, we design a mask constraint loss based on the mask images to ensure the integrity.
- We evaluate the effectiveness of generated enlarged license plates using three representative text recognition methods on the created dataset, and the results demonstrate the effectiveness of the proposed approach.

## 2. Related work

In this section, we review the related works that are most relevant to us, including license plate recognition, natural scene text recognition and generative adversarial networks.

### 2.1. License plate recognition

Existing license plate recognition algorithms can be divided into two categories, including segmentation based methods (Gou et al., 2015; Guo and Liu, 2008) and segmentation-free based methods (Li and Shen, 2016). The segmentation-based methods need to segment license plate into individual characters, and then recognize them one by one. After license plate segmentation is completed, template matching (Rasheed et al., 2012) and learning based (Wen et al., 2011) algorithms are usually used to classify characters. However, the segmentation methods lose the internal information of license plates, and the segmentation performance would seriously affect the recognition accuracy. Li and Shen (2016) propose a cascaded framework based on CNN and LSTM for segmented free-based license plate recognition, which significantly improves the accuracy of standard license plate recognition.

Xu et al. (2018) contribute a very large dataset and annotate different challenges, which greatly facilitate researches in this field. Sun et al. (2021) and Zhang et al. (2020) respectively propose image generation networks to generate realistic standard license plate images, and assist training recognition models to obtain better recognition accuracy. Gong et al. (2022) present a dataset with Chinese multi-LP images, and propose an end-to-end trainable network to detect and recognize license plates, which are able to deal with a variety of application scenarios. However, different from standard license plates, the enlarged license plates are with high diversity in both backgrounds and texts, and existing methods of standard license plate recognition can thus not handle the enlarged license plate recognition well. Tao Wen and Wang (2022) propose to combine vision and rule evaluation for the detection and recognition of enlarged license plates, in which the character characteristics and naming rules of enlarged license plates are used for recognition. However, these characteristics and rules are difficult to define and cover various scenarios.

### 2.2. Scene text recognition

In recent years, many works have emerged to solve the task of irregular text recognition. Yang et al. (2017) and Li et al. (2019) use the two-dimensional attention mechanism for irregular text recognition. Liao et al. (2019) propose to use a semantic segmentation network to identify irregular scene text. In addition, Luo et al. (2019) propose a rectification network to convert irregular text images into regular text images that reduce background interference. Yang et al. (2019a) use the character-level supervision to be able to train the model accurately. These methods greatly improve the recognition accuracy of irregular texts. However, huge data support is the key to training these networks and building a large-scale training data is very time consuming and high labor cost.

There are some differences between the enlarged license plate recognition and scene text recognition need to be explained. First, their challenges are different. The main challenges of scene text recognition are huge changes in ratio, scale, and orientation. Scene text recognition is usually used to recognize the text of store signs, street signs and public places, etc., which are usually with high resolution. While the enlarged license plates are captured by traffic surveillance cameras, and the styles of license plates vary greatly, and the challenges include blur, abrasion, background clutter and occlusion. Second, their design rules are different. If it is roughly considered as a special version of scene text recognition, the prior knowledge of the enlarged license plate recognition is lost. Taking it as a new task in the follow-up research, it will cooperate with the standard license plate to further improve the accuracy of vehicle identification and promote the development of intelligent transportation.

### 2.3. Generative adversarial networks

With the widespread applications of Generative Adversarial Networks (GANs), Azadi et al. (2018), Cheng et al. (2019) and Yang et al. (2019b) have achieved amazing results on document images using adversarial learning methods. These methods focus on a single character and constrain the character to generate a single style. However, our goal is to generate enlarged license plates with various styles. This requires us to maintain character information while being able to generate a good background. The traditional binarization method works well on document images, but cannot maintain the performance when the appearance of text in natural images changes greatly. Therefore, it is still an open issue to coordinate the generations of the content and background.

Recently, several attempts in image translation have achieved a critical step. For example, Isola et al. (2017) achieve the generation of complex image pairs by using paired datasets and use pixel-level losses to generate complex image pairs. The CycleGAN is proposed by Zhu et al. (2017) and the cycle consistency loss is used to solve the problem of unpaired data. The DRIT proposed by Lee et al. (2020) uses different encoders to solve different tasks, and realizes the diverse generation of complex images by constraining the embedding space of different encoders. To make better use of the discriminator, Chen et al. (2020) propose the idea of reusing the discriminator encoder to generate more complex images with better quality. Some image generation models (Karras et al., 2019; Choi et al., 2018; Brock et al., 2018; Shao and Zhang, 2021; Yun et al., 2021; Emami et al., 2020) can generate better synthetic images, but they are unsupervised, and thus we cannot use them to generate enlarged license plates with specified text.

## 3. DGNet: Disentangled Generation Network

In this section, we first introduce a novel task-level disentanglement generation framework based on the Disentangled Generation Network (DGNet), which disentangles the generation into the text generation and background generation, for robust enlarged license plate recognition. Then we present a detailed description of the mask constraint loss, which effectively ensures the integrity of generated enlarged license plates.

### 3.1. Overview

Inspired by NICEGAN (Chen et al., 2020) and DRIT (Lee et al., 2020), we propose DGNet as shown in Fig. 2. DGNet consists of four parts, including text image generation, background template set construction, enlarged license plate image generation and text image reconstruction.

Text image generation and background template set construction helps us to generate highly diverse text images and background templates. Through this way, we can guarantee the diversity of generated enlarged license plates and control their contents and styles. Then, we

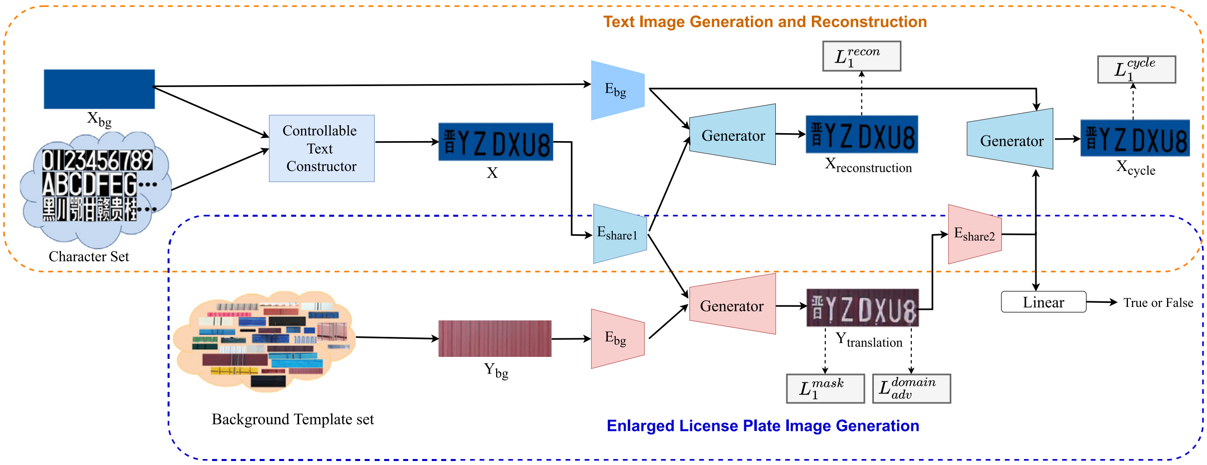


**Fig. 2.** Pipeline of the proposed framework. $X$ is the text image, $X_{bg}$ is unified background image, and $Y_{bg}$ is the background template sampled from the background template set. $X_{reconstruction}$ is the reconstructed image by the combination of text image and unified background image, $Y_translated$ is the synthetic enlarged license plate, and $X_{cycle}$ is the synthetic text image. $E_{bg}$ is the background encoder, and the discriminator consists of a shared encoder $E_{shared2}$ and a linear classifier.

use the text image and background template to generate the enlarged license plate by the module of enlarged license plate image generation, in which the mask constraint loss is introduced to ensure the integrity of the enlarged license plates. The problem of unpaired data is solved through the module of text image reconstruction.

In addition, we use the similar discriminator and generator to NICEGAN (Chen et al., 2020). To be specific, the discriminators consists of a shared encoder $E_{share2}$ and a linear classifier, and the generators consists of an encoder and a decoder. The above mentioned encoders all have the same structure, which contains six residual blocks for extracting features. These encoders shorten the domain translation path between low-dimensional hidden space vectors and promote the domain translation between high-dimensional images. The decoder is composed of two sub-pixel convolutional layers for up-sampling, and a normalization layer is applied for better learning the style and content information.

### 3.2. Controllable text image generation

As shown in Fig. 5, the characters of enlarged license plates lack uniform standards in font, position, and other attributes. To simulate real scenarios, we design a strategy of controllable text image generation.

In specific, we build a character set including all characters that existed in enlarged license plates, and combine some characters with a unified blue background image $X_{bg}$ by a controllable text constructor to form the text image $X$. For simplicity, we use one font as a prototype. In the controllable text constructor, we augment these character images by changing their sizes, shapes, and positions according to our requirements as follows. During this process, the shape of characters will be changed, and we use them to simulate different fonts of characters. First, the character images are randomly resized between 30 × 60 and 60 × 120. Second, to reshape these text images, we make the characters bigger by randomly expanding 0 to 2 pixels outward or make the characters smaller by randomly shrinking 0 to 2 pixels inward along the character edges. Finally, to simulate the character positions in real-world scenarios, we preset many sets of character positions, and the controllable text constructor will randomly choose from them. More importantly, the changes of these characters are completely retained in synthesized enlarged license plates, and the diversity is thus enhanced. We show some generated images using our proposed method, as shown in Fig. 6. We can see that the characters of the enlarged license plate have high diversity. It seems a visual gap between real and generated enlarged license plate image, which is caused by the huge intra-class difference of enlarged license plates. Such gap is a key challenge of enlarged license plate recognition.

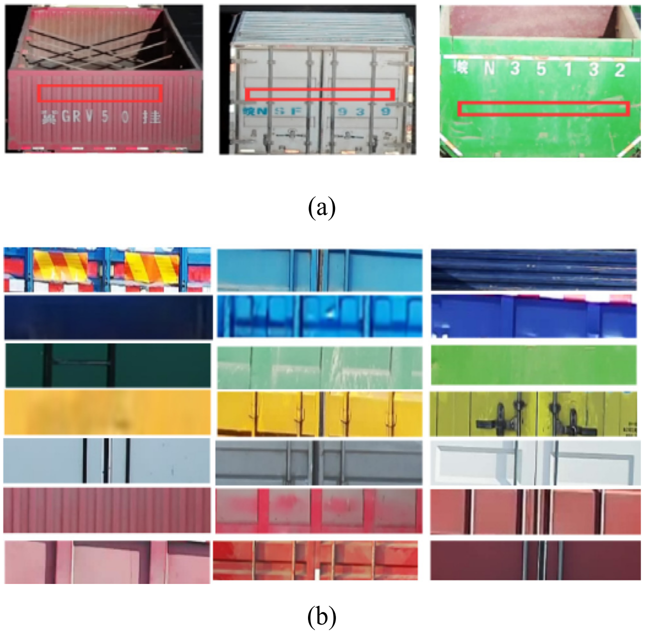


**Fig. 3.** (a) Visualization of constructed background templates. Herein, we randomly select the background templates around enlarged license plates. (b) Some samples from the background template set, which contains more than 1000 randomly cropped images from different regions.

In addition, we also show some images generated by random selection of text images and background templates. As shown in Fig. 4, we can see that the background style of the synthetic enlarged license plated can be controlled by different background templates, and the sizes and positions can be controlled by the augmented text images. In this way, we can control the size, shape and position of enlarged license plate.

### 3.3. Multi-style background image construction

The complexity of background in enlarged license plates is an important factor that affects the performance of recognition methods. Existing works (Lee et al., 2020; Huang et al., 2018) usually extract text and background information from a single image. However, enlarged license plates have high diversity and lots of noisy information, and it is thus difficult to accurately separate the text and background

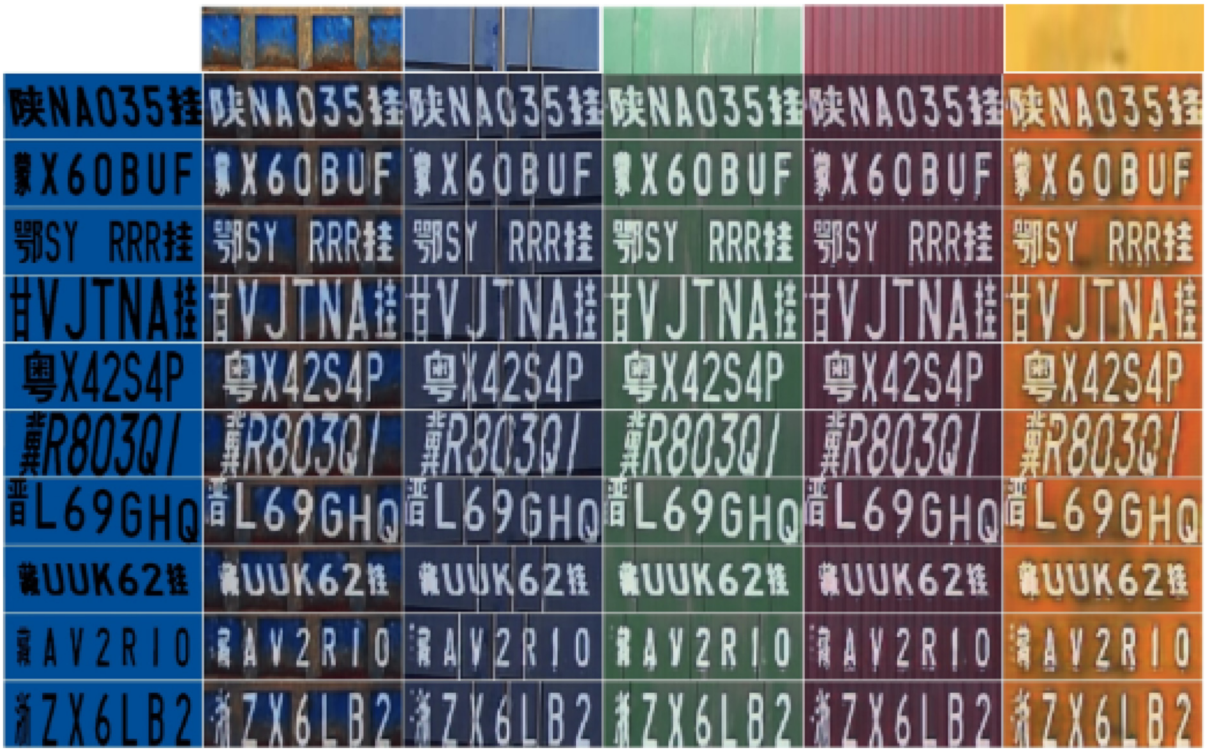


**Fig. 4.** Visualization of synthetic enlarged license plates with different text images and background templates. The first row indicates different background templates and the first column denotes different text images. Other images are the synthetic enlarged license plates generated by our DGNet.

from an image. Therefore, to better simulate the diverse background information, we propose to build a multi-style background template set. In specific, we extract the pure background template without character information around the enlarged license plates, as shown in Fig. 3(a). The background template set contains more than 1000 randomly cropped images from different regions, and has various styles of backgrounds, which cover almost all styles in the real scenes, as shown in Fig. 3(b). The background template can be randomly selected and combined with the text image to generate an enlarged license plate with the high-diversity background.

### 3.4. Enlarged license plate image generation

Given the text image $X$ and the background template $Y_{bg}$, we first extract text and background features by a two-stream encoder. The shared encoder $E_{share1}$ is used to extract the text features, and the background encoder $E_{bg}$ is used to extract the background features of the input enlarged license plate. Then, we send the combined text and background features to the generator to generate the high-quality synthesized enlarged license plate $Y_{translation}$.

Our training data are divided into real enlarged license plates (providing style information) and text masks (providing content information), and there is no one-to-one correspondence between them. Due to the imbalance of training data, the generator often generates frequently seen characters in training data. Therefore, we introduce the mask constraint loss to ensure the integrity of generated enlarged license plates. In specific, we use a mask image to supervise the generation by introducing a mask constraint loss, which is calculated by the average absolute error between the mask image and the output image. By subtracting the blue background image from the text image, we obtain the glyph image with a white background, and then add the glyph image and the input background template to compute the text mask image for supervising the generated images. In addition, the features extracted by the shared encoder $E_{share2}$ are used as the inputs of text image reconstruction and the linear classifier.

### 3.5. Text image reconstruction

To solve the problem of unpaired training data, we introduce the text image reconstruction module. On one hand, the cycle consistency loss (Zhu et al., 2017) is used to force the images between two domains to perform generation with each other, which effectively solve the unpaired training data problem. On the other hand, we combine the text features extracted by the shared encoder $E_{share1}$ with the background features extracted by the background encoder $E_{bg}$ as the input of the text image generator to obtain the reconstructed text image $X_{reconstruction}$ and the generated text image $X_{cycle}$. The text features of $X_{reconstruction}$ come directly from the shared encoder connected to the text image, and $X_{cycle}$ comes indirectly from this shared encoder.

### 3.6. Loss function

The overall loss consists of four parts, including generation adversarial loss, cycle consistency loss, reconstruction loss and mask constraint loss.

**Generation adversarial loss**. First, we make use of the least-square adversarial loss by Mao et al. (2017) due to its more stable training and high-quality generation. The generation adversarial loss is the key to improving the performance of the generator, which can be represented as follows:

$$L_{adv}^{domain} = \log D_Y(Y) + \log(1 - D_Y(G_{X\mapsto Y}(X, Y_{bg}))), \tag{1}$$

where $X$ is the text image, $Y$ is the enlarged license plate, $Y_{bg}$ is the background template, $G_{X\mapsto Y}$ is the generator of text image to enlarged license plate, and $D$ is Discriminator.

**Cycle consistency loss**. Inspired by CycleGAN (Zhu et al., 2017), we use the cycle consistency loss to solve unpaired data problem. Cycle-consistency loss can force the generators to be each others inverse and it is expressed as follows:

$$Y_{translation} = G_{X\mapsto Y}(E_{share1}(X), E_{bg}(Y_{bg})), \tag{2}$$

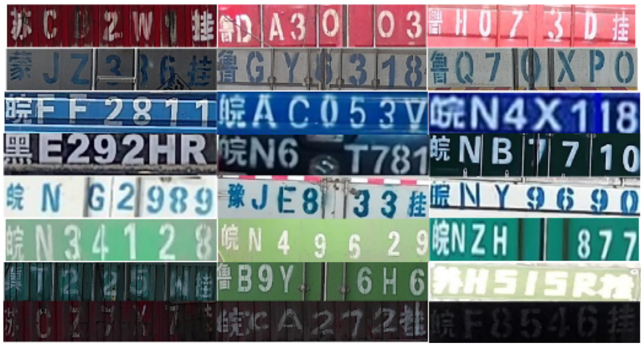


**Fig. 5.** Examples from ELPR Dataset. The license plates come from 18 different provinces, and all of them are captured from real traffic monitoring scenes.

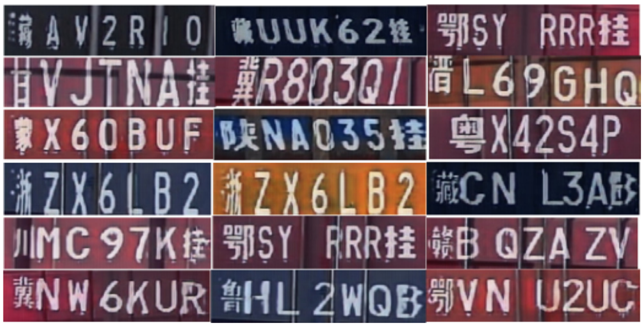


**Fig. 6.** Examples of synthetic images. The proposed image generation approach can effectively maintain text information and generate clear background information by the mask constraint loss.

$$L_1^{cycle} = |X - G_{Y\mapsto X}(E_{share2}(Y_{translation}), E_{bg}(X_{bg}))|_1, \tag{3}$$

where $|\cdot|$ denotes the $l_1$ norm, $E_{share2}$ is the shared encoder, $G_{Y\mapsto X}$ is the generator of enlarged license plate to text image, $E_{bg}$ is the background encoder, $X_{bg}$ is the unified blue background image, and $Y_{translation}$ is the synthesized enlarged license plate.

**Reconstruction loss**. In order to maintain the consistency of the generated background, we also introduce the reconstruction loss. Our reconstruction is based on the shared-latent space assumption (Chen et al., 2020). Reconstruction loss is to regularize the translation to be near an identity mapping when real samples' hidden vectors of the source domain are provided as the input to the generator of the source domain. It is expressed as follows:

$$L_1^{recon} = |X - G_{Y\mapsto X}(E_{share1}(X), E_{bg}(X_{bg}))|_1, \tag{4}$$

where $E_{share1}$ is the shared encoder.

**Mask constraint loss**. In order to effectively ensure the integrity of generated enlarged license plate, we design the mask constraint loss. It is expressed as follows:

$$L_1^{mask} = |I_{mask} - Y_{translation}|_1, \tag{5}$$

where $I_{mask}$ is the mask image used to supervise the generation.

**Full objective**. The objective function is given by

$$L = \lambda_1 \cdot L_{adv}^{domain} + \lambda_2 \cdot L_1^{recon} + \lambda_3 \cdot L_1^{cycle} + \lambda_4 \cdot L_1^{mask}, \tag{6}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ depict the hyper-parameters used to balance the trade-off between different supervisions, which are empirically set to 1.0, 10.0, 10.0, 15.0 respectively.

## 4. ELPR benchmark dataset

A large-scale dataset is crucial in enlarged license plate recognition because it can be used not only to train deep recognition models, but also to evaluate different recognition algorithms. Therefore, we provide a unified dataset for enlarged license plate recognition, called ELPR. We will introduce ELPR in detail.

### 4.1. Data creation

The current recognition field of enlarged license plate recognition lacks a unified dataset, and we thus create a large-scale ELPR dataset. Our goal is to provide a unified and highly diverse dataset to cover real-world scenes and challenges. Therefore, we use surveillance cameras in real traffic scenes to capture enlarged license plates of large vehicles. Because the data come from real scenes, our ELPR contains most real challenges such as occlusion, abnormal illumination, and blur. Fig. 5 shows typical examples of ELPR dataset, and we can see that many factors, such as complex background and various text styles, increase the difficulty of ELPR. By the way, we collect a total of 9342 images, including the license plates of 18 different provinces.

### 4.2. Annotation

In order to ensure the data quality, we train two professional annotators to label enlarged license plates one by one. In addition, we also ask professional checkers to prevent errors and inaccurate annotations. In addition to the special challenges of enlarging license plate such as complex background and high diversity in text size and position, some common challenges such as abnormal illumination, blur, and occlusion also seriously affect the recognition performance. Therefore, to better evaluate the performance of different recognition algorithms, we annotate each image with several challenges from the total 10 challenges, including inclined angle, abnormal illumination, different spacing, size variation, blur, abrasion, background clutter, non-standard character, double-row plate, and occlusion. The challenges are defined in Tables 1 and 2 shows the number distribution of challenges of ELPR dataset.

### 4.3. Data split

There is no other dataset for enlarged license plate recognition, and thus we divide it into a training set and a testing set to facilitate the training and evaluation of recognition methods. In specific, inspired by the standard license plate dataset CCPD (Xu et al., 2018), the testing set is randomly sampled, accounting for about 20 % of the whole dataset.

## 5. Evaluation

In this section, we will provide the details of experiments and report the experimental results on the benchmark dataset ELPR to validate the effectiveness of our DGNet against the state-of-the-art methods.

### 5.1. Evaluation metrics

Like some GAN evaluation methods, we adopt the Kernel Inception Distance (KID) (Bińkowski et al., 2018) and the Frechet Inception Distance (FID) (Heusel et al., 2017) to evaluate the quality of image generation. FID compares the statistics of generated data against real data, and fits a Gaussian distribution to the hidden activations of InceptionNet for each compared image set. Then, it computes the Frechet distance between those Gaussians. Lower FID is better, corresponding to more realistic generated images. KID is a metric similar to FID but uses the squared maximum mean discrepancy between Inception representations with a polynomial kernel. Unlike FID, KID has a simple unbiased estimator, making it more reliable especially when there are

**Table 1**
Descriptions of 10 different challenges in ELPR dataset.

| Challenge | Definition |
|---|---|
| Inclined angle | The images of enlarged license plates are with different inclination angles. |
| Abnormal illumination | The images are captured in high or low illumination conditions. |
| Different spacing | In a enlarged license plate, the spaces of different characters is different. |
| Size variation | In a enlarged license plate, the sizes of different characters are different. |
| Blur | The images are blur caused by fast motion and inaccurate camera focus. |
| Abrasion | The images are abraded due to the long-term use of vehicles, and some information of characters is missing. |
| Background clutter | The background contains many noises which disturbs the recognition of enlarged license plate. |
| Non-standard character | The characters are handwritten instead of standard printed. |
| Double-row plate | The enlarged license plate is painted in two rows. |
| Occlusion | Some characters are partially or completely occluded by other objects. |

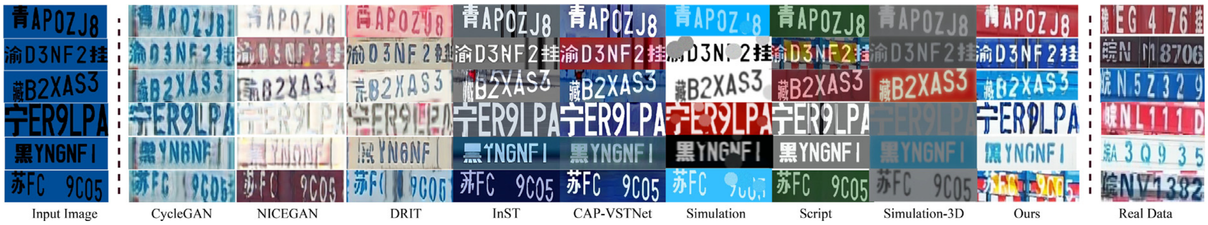


**Fig. 7.** Comparison of the generated images by CycleGAN, NICEGAN, DRIT, InST, CAP-VSTNet, Simulation, Script, Simulation-3D and our DGNet.

**Table 2**
Distribution of different challenges in ELPR dataset.

| Challenge | Total number (Train/Test) |
|---|---|
| Inclined angle | 760 (590/170) |
| Abnormal illumination | 1110 (807/303) |
| Different spacing | 1970 (1492/478) |
| Size variation | 347 (258/89) |
| Blur | 860 (650/210) |
| Abrasion | 2056 (1404/652) |
| Background clutter | 2971 (2435/536) |
| Non-standard character | 590 (485/105) |
| Double-row plate | 32 (10/22) |
| Occlusion | 814 (589/225) |

much more Inception feature channels than image numbers. Lower KID indicates high visual similarity between real and generated images.

In addition, our goal is to improve the recognition accuracy of enlarged license plates. Therefore, we adopt two extra indicators of enlarged license plate recognition accuracy (RA) and character recognition accuracy (CRA) for qualitative evaluation. Enlarged license plate recognition accuracy can be defined as:

$$RA = \frac{Number\ of\ correctly\ recognized\ license\ plates}{Number\ of\ all\ license\ plates}, \tag{7}$$

while character recognition accuracy can be defined as:

$$CRA = \frac{Number\ of\ correctly\ recognized\ characters}{Number\ of\ all\ characters}. \tag{8}$$

### *5.2. Implementation details*

Our network is implemented based on Pytorch and trained with a single Tesla P40 GPU. We use the Adam optimizer to optimize the proposed network with the learning rate 0.0001 and ($\beta_1$, $\beta_2$) = (0.5, 0.999) on Tesla P40 trained over 100K iterations. For the inputs, we resize all images to the size of 256 × 256. In addition, we compute the mean and standard deviation of all images in training set and normalize the inputs. The batch size is set to 1, and the training speed is about 1.5 iterations per second. In the testing phase, the generation of an enlarged license plate costs 2.0 ms on average.

### *5.3. Comparison with state-of-the-art methods*

In order to study the effectiveness of our synthesized enlarged license plates, we carry out experiments on several recent representative recognition algorithms, which are the models proposed by Litman et al. (2020), Yue et al. (2020) and Baek et al. (2019) respectively. Baek et al. propose a modular four-stage scene text recognition framework, in which each component is interchangeable and different algorithms can be integrated (this model is called as TPS-ResNet-BiLSTM-Attn). On this basis, Litman et al. propose the stacked encoding and decoding modules to achieve advanced performance. Yue et al. focus on context-free text recognition with location attention, and show advanced performance in semantic-free text recognition.

**Table 3**
Comparison of three representative recognition methods with the synthetic images generated by different methods.

| Baseline | GAN Method | Training data | | RA | CRA |
|---|---|---|---|---|---|
| | | Real | Synthetic | | |
| Litman et al. | Script | 7K | 20K | 76.82 | 94.40 |
| | NICEGAN | 7K | 20K | 76.65 | 93.90 |
| | CycleGAN | 7K | 20K | 76.40 | 93.92 |
| | DRIT | 7K | 20K | 76.98 | 94.45 |
| | InST | 7K | 20K | 76.48 | 94.32 |
| | CAP-VSTNet | 7K | 20K | 76.20 | 94.24 |
| | Simulation | 7K | 20K | 78.59 | 94.23 |
| | Simulation-3D | 7K | 20K | 78.50 | 94.44 |
| | DGNet (Ours) | 7K | 20K | **80.11** | **94.69** |
| Yue et al. | Script | 7K | 20K | 56.14 | 88.10 |
| | NICEGAN | 7K | 20K | 57.79 | 88.47 |
| | CycleGAN | 7K | 20K | 56.18 | 87.65 |
| | DRIT | 7K | 20K | 57.00 | 88.09 |
| | InST | 7K | 20K | 56.92 | 88.24 |
| | CAP-VSTNet | 7K | 20K | 57.50 | 88.56 |
| | Simulation | 7K | 20K | 58.94 | 89.05 |
| | Simulation-3D | 7K | 20K | 56.64 | 88.11 |
| | DGNet (Ours) | 7K | 20K | **60.62** | **89.12** |
| Baek et al. | Script | 7K | 20K | 61.94 | 90.17 |
| | NICEGAN | 7K | 20K | 55.57 | 88.18 |
| | CycleGAN | 7K | 20K | 58.36 | 89.05 |
| | DRIT | 7K | 20K | 58.57 | 88.94 |
| | InST | 7K | 20K | 59.43 | 89.22 |
| | CAP-VSTNet | 7K | 20K | 61.20 | 89.83 |
| | Simulation | 7K | 20K | 61.61 | 90.04 |
| | Simulation-3D | 7K | 20K | 60.05 | 89.01 |
| | DGNet (Ours) | 7K | 20K | **65.15** | **90.74** |

**Overall performance**. We use different methods to generate synthetic data for experimental verification. First, to ensure the integrity and diversity of the generated enlarged license plate, we use the image

**Table 4**
FID and KID × 100 for different algorithms. Herein, the lower values are better.

| Method | FID↓ | KID ×100 ↓ |
|---|---|---|
| CycleGAN | 283.38 | 35.88 |
| NICEGAN | 216.60 | 22.17 |
| DRIT | 156.41 | 13.79 |
| InST | 168.94 | 15.63 |
| CAP-VSTNet | 432.44 | 42.23 |
| Simulation | 189.40 | 15.42 |
| Simulation-3D | 181.81 | 15.24 |
| Script | 164.26 | 15.27 |
| DGNet (Ours) | **101.32** | **7.46** |

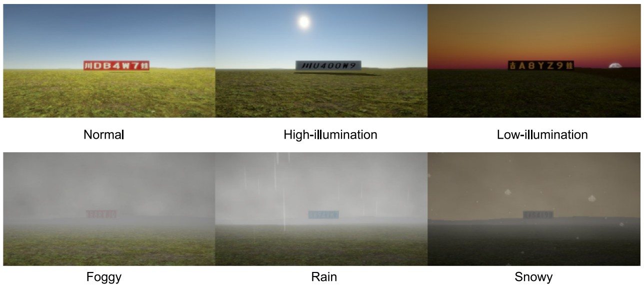

Normal
High-illumination
Low-illumination
Foggy
Rain
Snowy

**Fig. 8.** Some examples of synthetic images based on Unity, which can simulate different weather environments and illumination changes.

processing technology to extract the character mask from the text image and add it to the background template to obtain the synthetic image for comparison. We name this method as Script in our paper. What is more, based on the text images and virtual background images, we first generate the initial simulation images. Then, we use the traditional 2D and 3D simulation methods to simulate different attributes to generate the final simulation images. Specifically, in the 2D simulation experiment, the simulation of illumination and blur is based on OpenCV library, and the simulation of occlusion is achieved by covering randomly selected positions with pixel blocks of different sizes. We call this method as Simulation. In the 3D simulation experiment, we use UniStorm[2] to simulate various scenarios in the real world. It is a plug-in software based on Unity, which can simulate complex weather systems such as rainy days, snowy days and foggy days and time changes including early morning, noon, evening and midnight, etc. In specific, we put the initial simulation image into a set of random weather environments, and then randomly change the camera position to capture license plates to obtain the final simulation image. We name this method as Simulation-3D in our paper and some simulation environments are shown in Fig. 8.

The second one is to synthesize enlarged license plates using existing GAN methods, including CycleGAN (Zhu et al., 2017), NICEGAN (Chen et al., 2020), DRIT (Lee et al., 2020), InST (Zhang et al., 2023) and CAP-VSTNet (Wen et al., 2023). The final one is to use our proposed model to generate enlarged license plates. We use four representative methods to train with the above synthetic data respectively and evaluate them on testing set. In addition, we have paid attention to some methods for generating synthetic standard license plate images, such as Sun et al. (2021) and Zhang et al. (2020). However, the code sources of these methods are not open, and thus we are unable to use these methods to generate synthetic enlarged license plate images and compare them with our method.

As shown in Table 3, the results show that our method significantly improves the recognition accuracy of these recognition methods. In the best recognition method (Litman et al., 2020), our method achieves 3.46%, 3.71%, 3.29%, 3.13%, 3.63%, 3.91%, 1.52% and 1.61% higher than NICEGAN (Chen et al., 2020), CycleGAN (Zhu et al., 2017), Script, DRIT (Lee et al., 2020), InST (Zhang et al., 2023), CAP-VSTNet (Wen et al., 2023), Simulation and Simulation-3D respectively. Compared with our proposed method, it seems that the images generated by the simulation methods seem more realistic in visualization, but perform worse in different metrics. The results suggest that our method can generate more effective data for the training of different recognizers, while the generated data by other simulation methods might have bigger gap with real data in feature level.

In addition, in the recognition framework proposed by Litman et al. (2020) and Baek et al. (2019), Script achieves 0.42%, 0.17% and 3.58%, 6.37% higher than CycleGAN (Zhu et al., 2017) and NICEGAN (Chen et al., 2020) in RA, which can be explained that this method can generate higher diverse license plates than CycleGAN (Zhu et al., 2017) and NICEGAN (Zhu et al., 2017). However, our DGNet can generate enlarged license plates with high diversity and small domain gap, and thus our method outperforms Script in a large margin.

Note that the performance improvement of our synthesized data is obvious in the recognition framework proposed by Yue et al. (2020). The overall performance achieves 2.83%, 4.44%, 4.48%, 3.62%, 3.70%, 3.12%, 1.68% and 3.98% higher than NICEGAN (Chen et al., 2020), CycleGAN (Zhu et al., 2017), Script, DRIT (Lee et al., 2020), InST (Zhang et al., 2023), CAP-VSTNet (Wen et al., 2023), Simulation and Simulation-3D in RA, respectively. Yue et al. propose the recognition method uses the positional attention enhancement to extract text character features in images. Training data are not enough to cover the position diversity of real scenes, which affects the performance of enlarged license plate recognition. While our synthesized data show high diversity in position, size and other attributes and effectively make up for the shortage of training data, which can improve the robustness of this recognition framework.

The evaluation of synthetic image quality is shown in Table 4. The FID and KID scores of our proposed method are obviously smaller than those of CycleGAN (Zhu et al., 2017), NICEGAN (Chen et al., 2020), Script and other methods, which suggest that the enlarged license plates generated by our DGNet are more similar to real data.

The visualization of enlarged license plates generated by different methods is shown in Fig. 7. The first column is the input text image, which determines the text of the output, and the last column is real data and others are generated by different image translation methods. What is more, the real data are unpaired with the generated images, and it only serves as a reference for the style of the real enlarged license plates. It can be found that although the synthesized data generated by the Script have reliable text information, there is obvious domain difference with real enlarged license plates. CycleGAN (Zhu et al., 2017) and NICEGAN (Chen et al., 2020) lose the text structure information in the training process, which seriously affects the recognition performance. DRIT (Lee et al., 2020) has the ability to maintain text information, but it has a severe mode collapse and generates all images with white backgrounds. InST (Zhang et al., 2023) can accurately distinguish the character area and the background area and generate license plate images with more realistic backgrounds, but its ability to generate characters is insufficient. The wrong character structure will have negative impact when training the recognition model. CAP-VSTNet (Wen et al., 2023) transfers the background style of the enlarged license plate very well, but has almost no impact on the character part, which is very different from the styles in real enlarged license plates.

It is worth noting that our proposed method has additional background input compared with existing methods. It can be explained from the following aspects. First, CycleGAN (Zhu et al., 2017), NICEGAN (Chen et al., 2020), InST (Zhang et al., 2023) and CAP-VSTNet (Wen et al., 2023) use an unsupervised way to obtain the background and content information from input image. DRIT (Lee et al., 2020) extracts background and content information by disentangled representation. These methods does not need additional background input

[2] https://assetstore.unity.com/packages/tools/particles-effects/unistorm-volumetric-clouds-sky-modular-weather-and-cloud-shadows-2714

**Table 5**
Comparison results of RA/CRA scores(%) of different recognition methods on different challenges in ELPR dataset, where R/S refers to real or synthetic data.

| Baseline | Method | Training | NOR | IA | AI | DS | SV | BLU | ABR | BC | NSC | DRP | OCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R/S | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA | RA/CRA |
| Litman et al. | – | 7K/0 | 85.33/97.41 | 82.35/96.64 | 67.33/91.37 | 78.03/94.8 | 75.28/95.5 | 66.19/90.88 | 59.97/90.73 | 70.71/93.52 | 71.43/94.69 | 0/31.17 | 59.11/90.41 |
| | CycleGAN | 7K/20K | 88.80/97.99 | 85.29/97.05 | 71.29/91.56 | 88.29/96.68 | 82.02/96.79 | 66.19/90.88 | 60.28/90.05 | 77.43/94.43 | 69.52/94.29 | 0/22.08 | 68.44/92.19 |
| | NICEGAN | 7K/20K | 88.42/97.77 | **89.41/97.73** | 69.31/91.75 | 84.52/95.40 | 70.79/95.02 | 69.52/90.82 | 62.42/90.75 | 76.49/93.74 | 73.77/95.10 | 0/25.97 | 69.78/92.51 |
| | DRIT | 7K/20K | 89.19/98.12 | 85.89/97.39 | 69.64/92.17 | 85.98/96.23 | 78.65/96.15 | 69.05/91.56 | 63.19/**92.00** | 79.10/94.62 | 68.57/94.42 | 0/27.92 | 65.78/92.63 |
| | InST | 7K/20K | 89.58/98.07 | 84.71/97.23 | 68.65/91.65 | 87.24/**96.86** | 69.66/94.70 | 70.48/91.70 | 62.58/91.04 | 77.43/94.99 | 72.38/94.69 | 0/**32.47** | 65.78/92.38 |
| | CAP-VSTNet | 7K/20K | 88.80/97.90 | 84.71/97.06 | 68.32/91.61 | 86.19/96.74 | 75.28/95.51 | 69.05/**91.90** | 61.20/90.75 | 78.17/**95.12** | 72.38/94.42 | 0/27.27 | 63.56/92.19 |
| | Script | 7K/20K | 89.38/98.12 | 83.53/96.97 | 67.00/91.47 | 85.36/96.32 | 71.91/95.50 | 70.48/91.77 | 63.34/91.50 | 78.17/95.04 | 76.19/95.78 | 0/29.22 | 66.22/92.70 |
| | Simulation | 7K/20K | 90.15/97.99 | 85.29/97.31 | 68.65/91.70 | 89.12/96.83 | 76.40/95.99 | **73.33**/92.52 | 64.88/90.91 | 79.85/95.10 | 69.52/93.33 | 0/25.97 | **70.67**/92.70 |
| | Simulation-3D | 7K/20K | 92.66/ **98.51** | 86.47/97.14 | 68.32/91.56 | 87.66/96.44 | 82.02/96.95 | 68.57/91.36 | 64.11/90.95 | 78.92/94.99 | 74.29/94.97 | 0/28.57 | 70.22/**93.97** |
| | DGNet (Ours) | 7K/20K | **92.86**/98.43 | 88.24/**97.73** | **71.95/92.27** | **89.33**/96.56 | **83.15/97.43** | 70.00/91.29 | **65.64**/91.56 | **80.97**/94.94 | **78.10/95.92** | 0/24.68 | **70.67**/93.46 |
| Yue et al. | – | 7K/0 | 17.95/71.4 | 15.29/66.39 | 10.23/62.47 | 9.00/63.09 | 12.36/65.33 | 14.76/65.1 | 11.96/64.48 | 8.77/62.39 | 14.29/67.07 | 0/**24.03** | 10.22/62.6 |
| | CycleGAN | 7K/20K | 71.82/94.21 | 63.53/91.01 | 44.22/82.74 | 59.41/88.76 | 55.06/89.73 | 50.48/83.67 | 44.79/82.87 | 53.92/86.89 | 43.81/86.39 | 0/18.83 | 43.11/85.59 |
| | NICEGAN | 7K/20K | 73.17/95.01 | 63.94/91.85 | 46.21/83.22 | 62.55/90.71 | 57.30/91.01 | 48.57/84.97 | 46.47/**84.98** | 55.03/87.50 | 53.55/87.48 | 0/18.18 | 44.89/85.90 |
| | DRIT | 7K/20K | 73.75/94.43 | **65.88**/91.01 | 45.55/84.02 | 61.30/89.15 | 47.19/89.25 | 48.10/85.51 | 43.71/83.26 | 56.16/87.39 | 47.62/87.35 | 0/20.78 | 44.00/84.25 |
| | InST | 7K/20K | 74.52/94.65 | 59.41/89.83 | 45.87/84.30 | 59.00/89.48 | 48.31/88.44 | 50.00/**85.78** | 44.33/84.27 | 53.36/86.65 | 54.29/87.07 | 0/21.43 | 42.22/85.14 |
| | CAP-VSTNet | 7K/20K | 73.55/95.06 | 58.82/91.43 | 46.53/84.06 | 63.60/90.35 | 55.06/90.21 | 48.10/84.49 | 45.40/84.38 | 54.66/87.29 | 51.43/87.48 | 0/22.73 | 40.44/85.27 |
| | Script | 7K/20K | 72.59/94.68 | 56.47/90.59 | 44.22/82.46 | 60.46/89.12 | 56.18/91.49 | **54.29**/85.51 | 42.64/83.87 | 52.61/86.83 | **58.10/88.57** | 0/22.73 | 41.78/84.63 |
| | Simulation | 7K/20K | 74.52/95.06 | 62.94/91.68 | 46.20/83.92 | **65.06/91.36** | 53.93/89.73 | 53.33/86.33 | 46.01/84.97 | 55.41/88.19 | 48.87/85.85 | 0/22.08 | **47.11**/86.67 |
| | Simulation-3D | 7K/20K | 73.75/94.51 | 60.59/91.18 | 45.54/83.59 | 64.23/90.71 | 53.93/88.76 | 48.57/84.97 | 42.79/83.24 | 56.53/87.63 | 48.57/85.85 | 0/21.43 | 39.56/84.38 |
| | DGNet (Ours) | 7K/20K | **76.64/95.26** | **65.88/92.86** | **51.16/84.77** | **65.06**/90.68 | **65.17/92.46** | 50.00/85.17 | **46.78**/84.40 | **59.52/88.83** | 49.52/87.48 | 0/17.53 | 46.67/**86.98** |
| Baek et al. | – | 7K/0 | 66.41/92.2 | 56.47/89.16 | 42.9/82.84 | 52.09/86.28 | 50.56/88.93 | 53.81/85.85 | 42.95/83.11 | 50.37/85.79 | 48.57/88.57 | 0/22.73 | 37.78/82.67 |
| | CycleGAN | 7K/20K | 75.68/95.12 | 65.88/92.52 | 46.21/84.87 | 63.39/90.38 | 57.3/91.65 | 56.67/86.6 | 43.25/83.7 | 55.22/88.3 | 49.52/89.52 | 0/25.32 | 44/85.84 |
| | NICEGAN | 7K/20K | 72.01/94,43 | 66.47/92.94 | 44.88/83.4 | 59.41/89.15 | 49.44/90.21 | 50.95/85.17 | 42.49/83.39 | 54.29/87.87 | 42.86/88.03 | 0/27.92 | 39.11/82.6 |
| | DRIT | 7K/20K | 76.06/95.01 | 67.06/91.68 | 48.18/84.25 | 63.60/90.44 | 59.55/91.49 | 56.19/87.14 | 42.79/82.76 | 54.66/88.25 | 56.19/90.07 | 0/21.43 | 42.67/84.13 |
| | InST | 7K/20K | 76.06/95.15 | 67.06/91.85 | 45.87/83.69 | 65.27/90.82 | 55.06/91.17 | 53.81/87.35 | 44.17/84.33 | 57.28/88.91 | 51.43/88.98 | 0/**29.87** | 45.33/86.22 |
| | CAP-VSTNet | 7K/20K | 77.22/95.84 | 67.06/93.03 | 51.16/85.81 | 65.27/90.56 | 56.18/91.17 | 50.48/86.94 | 46.32/85.06 | 61.75/89.61 | 56.19/89.25 | 0/26.62 | **50.67/87.62** |
| | Script | 7K/20K | 77.22/95.64 | 68.24/92.61 | 51.49/85.81 | 69.87/91.93 | **66.29**/92.3 | **59.05/88.91** | 46.01/**85.45** | 60.08/89.69 | 50.48/88.57 | 0/28.57 | 48.89/87.49 |
| | Simulation | 7K/20K | 76.06/95.34 | 62.94/91.93 | **52.15**/84.91 | 67.15/91.27 | 62.92/92.13 | 57.62/87.21 | 47.55/84.86 | 61.38/89.66 | 54.29/88.44 | 0/25.32 | 47.56/86.10 |
| | Simulation-3D | 7K/20K | 78.76/95.84 | 67.06/92.35 | 46.53/84.44 | 69.67/91.81 | 56.18/91.01 | 54.76/87.41 | 41.41/82.73 | 58.02/89.29 | 57.14/90.88 | 0/25.97 | 42.67/86.22 |
| | DGNet (Ours) | 7K/20K | **79.92/96.25** | **74.71/94.45** | 51.49/**85.90** | **74.48/92.62** | **66.29/93.58** | **59.05**/88.50 | **48.62**/85.43 | **65.30/91.15** | **60.95/90.20** | 0/25.33 | 45.78/**87.62** |

**Table 6**
Comparison results of different recognition methods using different number of real data, where the number of synthetic images is fixed.

| Method | Training data | | RA | CRA |
|---|---|---|---|---|
| | Real | Synthetic | | |
| Litman et al. | 2K | 25K | 73.36 | 93.05 |
| | 5K | 22K | 77.47 | 94.42 |
| | 7K | 20K | **80.11** | **94.69** |
| Yue et al. | 2K | 25K | 47.88 | 84.39 |
| | 5K | 22K | 60.13 | 88.94 |
| | 7K | 20K | **60.62** | **89.12** |
| Baek et al. | 2K | 25K | 58.90 | 88.70 |
| | 5K | 22K | 64.04 | 90.51 |
| | 7K | 20K | **65.15** | **90.74** |

**Table 7**
Comparison results of different setting in our method in different recognition methods, including without synthetic data and mask constraint loss respectively.

| Baseline | Method | Training data | | RA | CRA |
|---|---|---|---|---|---|
| | | Real | Synthetic | | |
| Litman et al. | DGNet (Ours) | 7K | 0 | 72.13 | 93.42 |
| | DGNet (Ours) | 7K | 7K | 77.52 | 94.39 |
| | DGNet (Ours) | 7K | 14K | 78.05 | 94.59 |
| | DGNet w/o $L_{mask}$ | 7K | 20K | 74.56 | 94.02 |
| | DGNet (Ours) | 7K | 20K | **80.11** | **94.69** |
| Yue et al. | DGNet (Ours) | 7K | 0 | 12.82 | 65.82 |
| | DGNet (Ours) | 7K | 7K | 58.65 | 83.83 |
| | DGNet (Ours) | 7K | 14K | 60.42 | 89.76 |
| | DGNet w/o $L_{mask}$ | 7K | 20K | 51.38 | 86.03 |
| | DGNet (Ours) | 7K | 20K | **60.62** | **89.12** |
| Baek et al. | DGNet (Ours) | 7K | 0 | 52.53 | 86.68 |
| | DGNet (Ours) | 7K | 7K | 63.17 | 90.00 |
| | DGNet (Ours) | 7K | 14K | 63.34 | 91.00 |
| | DGNet w/o $L_{mask}$ | 7K | 20K | 48.66 | 86.00 |
| | DGNet (Ours) | 7K | 20K | **65.15** | **90.74** |

and also cannot leverage additional background information. The target domain images generated by these methods are uncontrollable and tend to generate a fixed style, as shown in Fig. 7. Second, by adding an extra-input background information, our generator is guided to generate relevant target domain images, which not only effectively improves the diversity, but also makes the background of generated enlarged license plates controllable.

It can be seen that the synthesized data generated by any method can improve the performance of the recognizer. It is because that even though their character structures have been destroyed and cannot work at the pixel level, they can still effectively train recognition models with the feature level information.

**Attribute-based performance**. To analyze the performance under different challenges faced by existing recognition methods, we evaluate our method against three algorithms on 11 challenge attributes including inclined angle (IA), abnormal illumination (AI), different spacing (DS), size variation (SV), blur (BLU), abrasion (ABR), background clutter (BC), non-standard character (NSC), double-row plate (DRP) and occlusion (OCC), as shown in Table 5. It can be seen from the results that our DGNet achieves better results in most attributes compared with all other image generation algorithms. Note that RA is more important evaluation metric. Under the attributes of IA, AI, DS, SV, ABR, and BC, our method is better than other methods in RA score, which further proves the diversity of our synthetic enlarged license plates and good simulation to real data.

There are two key points here need to be explained. First, under the attribute of DRP, the enlarged license plates are not recognized well, because these recognition methods are based on single-line text recognition, which will be invalidated in multi-line text recognition. There are some solutions to handle this case. On the one hand, multi-line text recognition is often converted to single-line text recognition tasks using detection networks. However, for license plate recognition, common detection networks usually require more computational resources and interface time and are difficult to apply to real scenarios. On the other hand, pre-processing is a better way to handle double-row plates. Specifically, when the input is double-row plates, we crop and split them into single-row license plates for recognition.

We set up an experiment to verify the effectiveness of our proposed method by using this pre-processing method. As shown in Table 9, when only real data are used for training, the performance is poor. If the synthetic data proposed by our method are added, the performance

**Table 8**
Comparison results of different Recognition methods only using a fixed number of synthetic images.

| Method | Training data | | RA | CRA |
|---|---|---|---|---|
| | Real | Synthetic | | |
| Litman et al. | 0 | 25K(CycleGAN) | 0.08 | 5.53 |
| | 0 | 25K(NICEGAN) | 0.37 | 30.48 |
| | 0 | 25K(InST) | 5.13 | 29.71 |
| | 0 | 25K(CAP-VSTNet) | 2.34 | 33.38 |
| | 0 | 25K(Script) | 1.03 | 16.61 |
| | 0 | 25K(Simulation) | 0.62 | 29.00 |
| | 0 | 25K(Simulation-3D) | 0.64 | 30.53 |
| | 0 | 25K(Ours) | **6.86** | **55.21** |
| Yue et al. | 0 | 25K(CycleGAN) | 0.12 | 10.55 |
| | 0 | 25K(NICEGAN) | 0.45 | 31.17 |
| | 0 | 25K(InST) | 2.75 | 30.10 |
| | 0 | 25K(CAP-VSTNet) | 2.01 | 30.18 |
| | 0 | 25K(Script) | 1.19 | 32.19 |
| | 0 | 25K(Simulation) | 0.37 | 28.21 |
| | 0 | 25K(Simulation-3D) | 0.35 | 28.05 |
| | 0 | 25K(Ours) | **4.40** | **45.55** |
| Baek et al. | 0 | 25K(CycleGAN) | 0.29 | 15.0 |
| | 0 | 25K(NICEGAN) | 1.89 | 45.00 |
| | 0 | 25K(InST) | 3.90 | 33.46 |
| | 0 | 25K(CAP-VSTNet) | 6.00 | 43.51 |
| | 0 | 25K(Script) | 2.80 | 30.00 |
| | 0 | 25K(Simulation) | 1.85 | 44.00 |
| | 0 | 25K(Simulation-3D) | 1.87 | 45.40 |
| | 0 | 25K(Ours) | **9.04** | **60.11** |

**Table 9**
RA/CRA results of our DGNet by using synthesizing data and pre-processing operation on DRP attribute.

| Method | Real | Synthetic | Pre-processing | RA | CRA |
|---|---|---|---|---|---|
| Litman et al. | ✓ | ✗ | ✗ | 0 | 31.17 |
| | ✓ | ✓ | ✗ | 0 | 24.68 |
| | ✓ | ✗ | ✓ | 22.73 | 70.78 |
| | ✓ | ✓ | ✓ | **81.82** | **91.91** |
| Yue et al. | ✓ | ✗ | ✗ | 0 | 24.03 |
| | ✓ | ✓ | ✗ | 0 | 17.53 |
| | ✓ | ✗ | ✓ | 18.18 | 58.44 |
| | ✓ | ✓ | ✓ | **40.91** | **77.27** |
| Baek et al. | ✓ | ✗ | ✗ | 0 | 22.73 |
| | ✓ | ✓ | ✗ | 0 | 25.32 |
| | ✓ | ✗ | ✓ | 4.55 | 57.14 |
| | ✓ | ✓ | ✓ | **22.73** | **72.08** |

is improved from 22.73%, 18.18% and 4.55% to 81.82%, 40.91% and 22.73% respectively. In addition, we also visualize the double-row plates generated by our method in Fig. 11. In this process, we just need to change the position and style of the text image and without extra training.

Second, our proposed method performs worse than other methods on some attributes such as IA, BLU and NSC. The main reason is that other methods have significant bias in training models under different attributes. As shown in Table 5, there is a large gap in the proportion of different attributes in the dataset, which causes the generators to be biased to generate some fixed attributes. However, compared with other methods, our proposed model achieves better results on most attributes, which further shows that our DGNet can generate highly diverse license plates. Other works tend to generate license plates with specific attributes due to mode collapse. They will provide more training data on these attributes and thus perform better than our proposed method on these attributes.

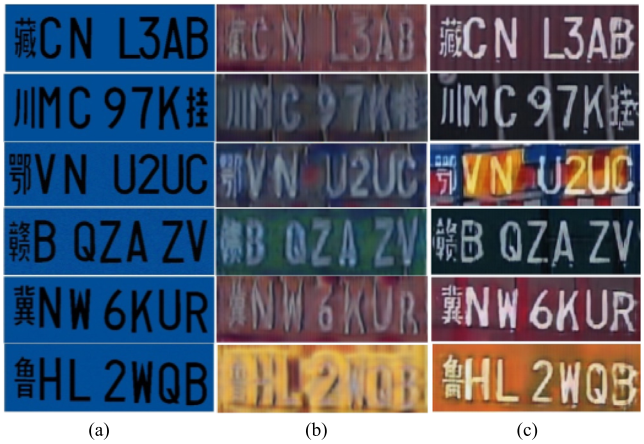


**Fig. 9.** Comparison of the generated images using the mask constraint loss. (a) Input text images. (b) Generated images without the mask constraint loss. (c) Generated images with the mask constraint loss.

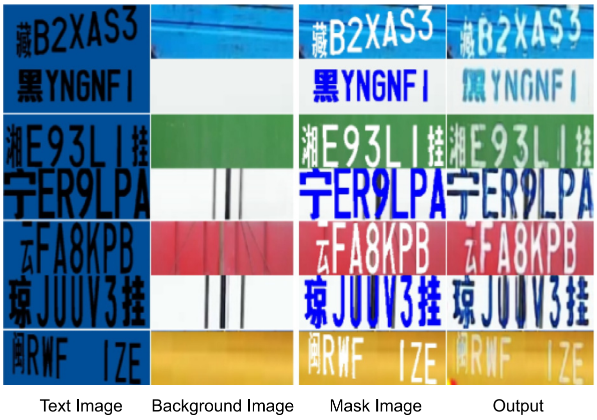


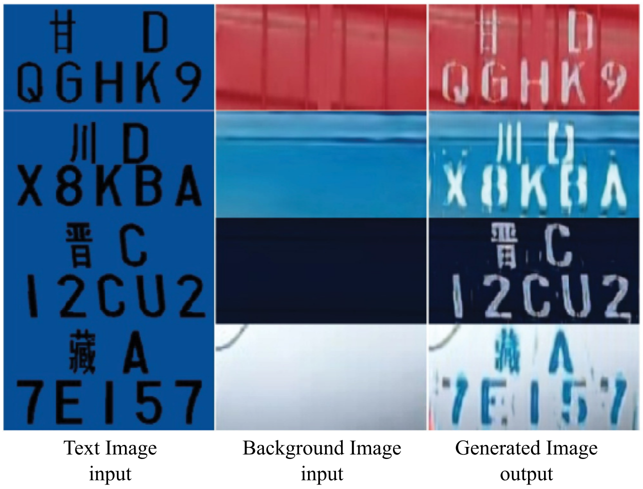
**Fig. 10.** Visualization samples of the mask images. The first and second columns denote the input images include text images and background images, and the third column denotes the generated mask images, and the last column denotes the generated enlarged license plates.

### *5.4. Ablation study*

In order to further study the role of real data, we set up an ablation experiment to gradually increase the proportion of real data, including 2k, 5K and 7K respectively, and evaluate them on testing set. As shown in Table 6, by increasing the proportion of real data step by step, the mixing of data reduce the domain gap. Thus, the recognition accuracy of the models is consistently improved. Through these experiments, it can be found that our synthetic data can effectively be validated by increasing the amount of real data.

Similarly, we also set up an ablation experiment to gradually increase the proportion of synthetic data for further study the role of synthetic data. The amount of synthetic images is set to 7K, 14k and 20K respectively and evaluate them on testing set. As shown in Table 7, by increasing the proportion of synthetic data step by step, the mixing of data will cover more real scenes, and the recognition accuracy of the models is thus consistently improved. Through these experiments, it can be found that our synthetic data can effectively improve the recognition performance of enlarged license plates. Although we have not directly set up relevant ablation experiments to verify the effectiveness of task-level disentanglement, the effectiveness can be fully verified by the comparisons with other image generation algorithms.

To verify the effectiveness of the mask constraint loss, we set up an ablation experiment to compare the generated images without and



Text Image input

Background Image input

Generated Image output

**Fig. 11.** Visualization samples of the double-row plates. The first and second columns are the input images include text images and background images, and the last column denotes the generated enlarged license plates.

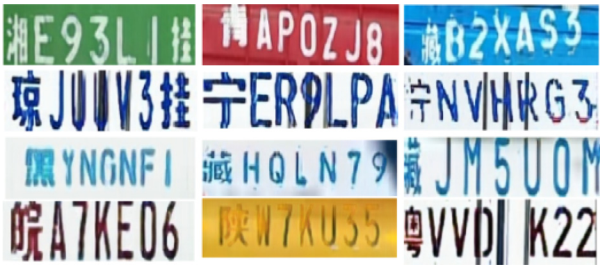


**Fig. 12.** Visualization samples of different colors of the text in generated enlarged license plates.

with the mask constraint loss as shown in Fig. 9. It can be found that the generated image quality is higher after adding the mask constraint loss, which proves that the mask constraint loss contributes greatly to the generated character quality and background. The mask constraint loss supervises the generation at the pixel level, and thus effectively maintains the structural information of text characters. As shown in Table 7, with the help of the mask constraint loss, our DGNet improves RA scores by 5.55%, 9.34%, and 16.49% respectively.

In addition, the mask images play an important role in the color, position, size and other styles of generated images. As shown in Fig. 10, the mask image is obtained by subtracting the blue background image from the text image and as supervisory information for generated images.

To further prove the effectiveness of our synthetic data, we set up an ablation experiment to train recognition methods only using synthetic data by different image translation methods including CycleGAN (Zhu et al., 2017), NICEGAN (Chen et al., 2020) and Script. As shown in Table 8, it can be found that for the recognition method proposed by Litman et al. our proposed network achieves 6.78%, 6.49% and 5.83% higher than CycleGAN (Zhu et al., 2017), NICEGAN (Chen et al., 2020) and Script in RA respectively, which shows that we can narrow the domain gap between real and generated data effectively. Of course, there is still a certain domain gap, and we will study this issue in future work.

In addition, we show some generated images using our proposed method, as shown in Fig. 6. The results show that we can generate complex backgrounds and multi-style texts. A problem in Fig. 6 is that the color of all the generated text is white. However, it does not mean that our DGNet can only generate this style. As shown in Fig. 12, we can generate other colors of texts like blue and red. These colors are often on a white background, which is the same as in the real world.

Note that although we have not set up relevant ablation experiments to verify the effectiveness of text and background image disentanglement, we can verify it by comparing our method with other image generation algorithms. Among these generation algorithms, DRIT is based on disentanglement, while CycleGAN and NICEGAN are not disentanglement. From the results it can be found that when using the synthetic images generated by DRIT to train these recognizers, its performance is better than those of CycleGAN and NICEGAN. In addition, our method is better than DRIT, because DRIT extracts the content and attribute information of the image through multiple encoders. However, due to the complexity and diversity of the text and background from enlarged license plates, these encoders cannot extract and completely separate the attribute and content information of enlarged license plates. These results demonstrate the effectiveness of text and background image disentanglement in enlarged license plate recognition.

## 6. Conclusion

In this work, we construct a unified enlarged license plate recognition dataset, which contains most of challenges in real scenes. We also propose a task-level disentanglement generation framework based on the disentangled generation network to effectively ensure the diversity and integrity of enlarged license plates, and thus greatly improve the recognition accuracy. Extensive experiments on the dataset demonstrate the effectiveness of our method in different recognition frameworks. By releasing this dataset, we believe that it will help the research and development of enlarged license plate recognition. In addition, the proposed method is also suitable for the tasks whose data can be decoupled and the decoupled data can be independently generated by different generative models. According to the property, it could be applied to some widely applicable scenarios such as medical image analysis and self-driving system.

Although we have explored the way of synthesizing enlarged license plates, there are still many potential problems to be solved. For example, the domain difference between synthesized enlarged license plates and real ones is large. How to use synthetic data to train the recognizer more efficiently and how to design an effective recognition model to address the special challenges of enlarging license plate recognition are two unsolved issues. In the future, we will study more effective recognition algorithms and image generation algorithms to further improve the performance of enlarged license plate recognition, and expand the dataset to cover more realistic scenes, e.g., changing the fonts of characters or randomly erasing characters in the controllable text constructor.

## CRediT authorship contribution statement

**Chenglong Li:** Conceptualization, Supervision, Writing – review & editing. **Xiaobin Yang:** Investigation, Methodology, Software, Writing – original draft. **Guohao Wang:** Formal analysis, Validation, Visualization. **Aihua Zheng:** Writing – review & editing, Supervision, Project administration. **Chang Tan:** Project administration, analysis and interpretation of data, Approving the final version of the article. **Jin Tang:** Funding acquisition, Project administration, Resources, Accepting for publication.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

We contribute an enlarged license plate dataset and have released this dataset to public.

## Acknowledgments

This work is partly supported by National Natural Science Foundation of China (No. 62376004), Natural Science Foundation of Anhui Province, China (No. 2208085J18), Natural Science Foundation of Anhui Higher Education Institution, China (No. 2022AH040014), The University Synergy Innovation Program of Anhui Province, China (No. GXXT-2022-033), and Anhui Provincial Colleges Science Foundation for Distinguished Young Scholars, China (No. 2022AH020093). The authors also acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

## References

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.

Azadi, S., Fisher, M., Kim, V.G., Wang, Z., Shechtman, E., Darrell, T., 2018. Multi-content gan for few-shot font style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7564–7573.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H., 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4715–4723.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Brock, A., Donahue, J., Simonyan, K., 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Chen, R., Huang, W., Huang, B., Sun, F., Fang, B., 2020. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Cheng, M.-M., Liu, X.-C., Wang, J., Lu, S.-P., Lai, Y.-K., Rosin, P.L., 2019. Structure-preserving neural style transfer. IEEE Trans. Image Process. 29, 909–920.

Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., Zhou, S., 2018. Aon: Towards arbitrarily-oriented text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5571–5579.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797.

Emami, H., Aliabadi, M.M., Dong, M., Chinnam, R.B., 2020. Spa-gan: Spatial attention gan for image-to-image translation. IEEE Trans. Multimed. 23, 391–401.

Gong, Y., Deng, L., Tao, S., Lu, X., Wu, P., Xie, Z., Ma, Z., Xie, M., 2022. Unified Chinese license plate detection and recognition with high efficiency. J. Vis. Commun. Image Represent. 86, 103541.

Gou, C., Wang, K., Yao, Y., Li, Z., 2015. Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines. IEEE Trans. Intell. Transp. Syst. 17 (4), 1096–1107.

Guo, J.-M., Liu, Y.-F., 2008. License plate localization and character segmentation with feedback self-learning and hybrid binarization techniques. IEEE Trans. Veh. Technol. 57 (3), 1417–1424.

Gupta, A., Vedaldi, A., Zisserman, A., 2016. Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2315–2324.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 6626–6637.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision. pp. 172–189.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410.

Lee, H.-Y., Tseng, H.-Y., Mao, Q., Huang, J.-B., Lu, Y.-D., Singh, M.K., Yang, M.-H., 2020. DRIT++: Diverse image-to-image translation viaDisentangled representations. Int. J. Comput. Vis. 1–16.

Li, H., Shen, C., 2016. Reading car license plates using deep convolutional neural networks and LSTMs. arXiv preprint arXiv:1601.05610.

Li, H., Wang, P., Shen, C., Zhang, G., 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8610–8617.

Liao, M., Zhang, J., Wan, Z., Xie, F., Liang, J., Lyu, P., Yao, C., Bai, X., 2019. Scene text recognition from two-dimensional perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8714–8721.

Litman, R., Anschel, O., Tsiper, S., Litman, R., Mazor, S., Manmatha, R., 2020. SCATTER: Selective context attentional scene text recognizer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11962–11972.

Luo, C., Jin, L., Sun, Z., 2019. Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognit. 90, 109–118.

Luo, C., Lin, Q., Liu, Y., Jin, L., Shen, C., 2021. Separating content from style using adversarial learning for recognizing text in the wild. Int. J. Comput. Vis. 129 (4), 960–976.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802.

Ning, H., Zheng, X., Lu, X., Yuan, Y., 2021. Disentangled representation learning for cross-modal biometric matching. IEEE Trans. Multimed. 24, 1763–1774.

Rasheed, S., Naeem, A., Ishaq, O., 2012. Automated number plate recognition using hough lines and template matching. In: Proceedings of the World Congress on Engineering and Computer Science. pp. 24–26.

Shao, X., Zhang, W., 2021. SPatchGAN: A statistical feature based discriminator for unsupervised image-to-image translation. arXiv preprint arXiv:2103.16219.

Sun, Y.-F., Liu, Q., Chen, S.-L., Zhou, F., Yin, X.-C., 2021. Robust Chinese license plate generation via foreground text and background separation. In: Proceedings of the International Conference on Image and Graphics. pp. 290–302.

Tao Wen, J.L., Wang, Z., 2022. Detection and recognition method of enlarged license plate based on combined vision and rule evaluation. J. Chinese Comput. Syst. 8, 1697–1702.

Wang, X., Man, Z., You, M., Shen, C., 2017. Adversarial generation of training examples: applications to moving vehicle license plate recognition. arXiv preprint arXiv:1707.03124.

Wen, L., Gao, C., Zou, C., 2023. CAP-VSTNet: Content affinity preserved versatile style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18300–18309.

Wen, Y., Lu, Y., Yan, J., Zhou, Z., von Deneen, K.M., Shi, P., 2011. An algorithm for license plate recognition applied to intelligent transportation system. IEEE Trans. Intell. Transp. Syst. 12 (3), 830–845.

Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., Huang, L., 2018. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In: Proceedings of the European Conference on Computer Vision. pp. 255–271.

Yang, M., Guan, Y., Liao, M., He, X., Bian, K., Bai, S., Yao, C., Bai, X., 2019a. Symmetry-constrained rectification network for scene text recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9147–9156.

Yang, X., He, D., Zhou, Z., Kifer, D., Giles, C.L., 2017. Learning to read irregular text with attention mechanisms. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 3280–3286.

Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., Guo, Z., 2019b. Controllable artistic text style transfer via shape-matching gan. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4442–4451.

Yi, Z., Chen, Z., Cai, H., Mao, W., Gong, M., Zhang, H., 2020. BSD-GAN: Branched generative adversarial network for scale-disentangled representation learning and image synthesis. IEEE Trans. Image Process. 29, 9073–9083.

Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W., 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In: Proceedings of the European Conference on Computer Vision. pp. 135–151.

Yun, X.-L., Zhang, Y.-M., Yin, F., Liu, C.-L., 2021. Instance GNN: a learning framework for joint symbol segmentation and recognition in online handwritten diagrams. IEEE Trans. Multimed. 24, 2580–2594.

Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C., 2023. Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10146–10156.

Zhang, L., Wang, P., Li, H., Li, Z., Shen, C., Zhang, Y., 2020. A robust attentional framework for license plate recognition in the wild. IEEE Trans. Intell. Transp. Syst. 22 (11), 6967–6976.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision.