Doctoral Dissertation

Parametric Loss Based Super-Resolution for Scene Text Recognition

Supatta Viriyavisuthisakul

Supervisor Nguyen Le Minh

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2023

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY (JAIST)
SCHOOL OF INFORMATION SCIENCE

DISSERTATION

BY

MS. SUPATTA VIRIYAVISUTHISAKUL

ENTITLED

PARAMETRIC LOSS BASED SUPER-RESOLUTION FOR SCENE TEXT RECOGNITION

was approved as partial fulfillment of the requirements for the degree
of Doctor of Philosophy (Information Science)

on September 22, 2023

Thesis Title	PARAMETRIC LOSS BASED SUPER-RESOLUTION FOR SCENE TEXT RECOGNITION
Author	Ms. Supatta Viriyavisuthisakul
Degree	Doctor of Philosophy (Information Science)
Faculty/University	Japan Advanced Institute of Science and Technology
Thesis Advisor	Professor Minh Le Nguyen, Ph.D
Academic Years	2023

ABSTRACT

In this study, we propose the application of multiple parametric regularizations and parametric weight parameters to the loss function of the scene text image super-resolution (STISR) method to improve scene text image quality and text recognition accuracy. STISR is regarded as the process of improving the image quality of low-resolution (LR) scene text images to improve text recognition accuracy. In a previous study, a text attention network (TATT) was introduced to reconstruct high-resolution scene text images; the backbone method involved the convolutional neural network (CNN)-based and transformer-based architecture. Although it can deal with rotated and curved-shaped texts, it still cannot properly handle images containing improper-shaped texts and blurred text regions. This can lead to incorrect text predictions during the text recognition step. Parametric regularization in the single-image super-resolution (SISR) model has recently been proposed to deal with artifacts and restore the unseen texture in the natural image domain. However, unlike STISR, SISR does not focus only on text information. Here, we design and extend it into three types of methods: adding multiple parametric regularizations, modifying parametric weight parameters, and combining parametric weights and multiple parametric regularizations. Experiments were conducted and compared with state-of-the-art of STISR models. The results showed a significant improvement for every proposed method. Our methods achieved the best text recognition accuracy of 80.4% for the easy set, 64.1% for the medium set, and 46.5% for the hard set of Textzoom. Moreover, our methods generated clearer and sharper edges than the baseline with a better-quality image score.

Keywords: Scene Text, Image Reconstruction, Trainable parameter, Parametric, Regularization, Loss function

ACKNOWLEDGMENTS

First of all, I would like to thank the SIIT-JAIST-NSTDA scholarship for giving and supporting me the opportunity to study for the Ph.D. degree. I would like to thank the following people, without whom I would not have been able to complete this research. Japan Advanced Institute of Science and Technology, especially to my supervisor Professor Dr. Minh Le Nguyen, for supporting me when I studied in Japan. Sirindhorn International Institute of Technology, especially to my co-supervisor Assistant Prof. Dr. Natsuda Kaothanthong, whose pushing and advising the knowledge into the subject matter steered me through this research. And special thanks to the Senior lecturer Teeradaj Racharak, who took care of and gave advice in both research and life in Japan. My colleague, Associate Prof. Dr. Parinya Sanguansat at Panyapiwat Institute of Management, who have a lot supported, provided assistance, and put up with my stresses for the past four years of study. National Science and Technology Development Agency, Dr. Choochart Haruechaiyasak, believed in me and accepted to be my co-advisor. My professor at The University of Tokyo, Professor Dr. Toshihiko Yamasaki, supported and fulfilled me in my academic experiences and gave me a lot of advice to be a good professor in the future. Finally, thanks to my family for all the support you have shown me through this research.

Ms. Supatta Viriyavisuthisakul

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Overviews

Scene text recognition is fundamental to detecting text regions in complex backgrounds and labels. Textual information from different sources, such as scanned documents, the label of products on shops, or vehicle number plates, has risen to a large extent in the digitalization era Chen et al. (2022a). To make text detection and understanding, advanced computer vision is applied to convert the text present in the image or scanned documents to a machine-readable format that can be processed later, called optical character recognition (OCR) Chen et al. (2022a). To make efficient text recognition, a high-resolution (HR) scene text image is required. However, HR images need a high capacity to store. Although imaging devices and techniques are proposed, this kind of approach has limitations, flexibility, and cost when applied in practical applications.

Detection and recognition of text from natural scene images are challenging works. Since the text in the natural scene can be varied such as different languages, fonts, sizes, orientations, and shapes. We need to deal with diversity and variability. Sometimes, the natural scene may have pattern backgrounds with a shape that is extremely like any text which creates a problem of text detection. Moreover, disrupted images such as low-resolution (LR) images are common problems that directly affect the text detection, and recognition process Wang et al. (2019). The poor resolution of scene text images can lead to the wrong prediction of text recognizers because of losing the detailed content information as shown in Figure 1.1.

To alleviate this addressed the problem, the super-resolution (SR) technique is applied as a pre-processing step as shown in Figure 1.2 to enhance the quality of

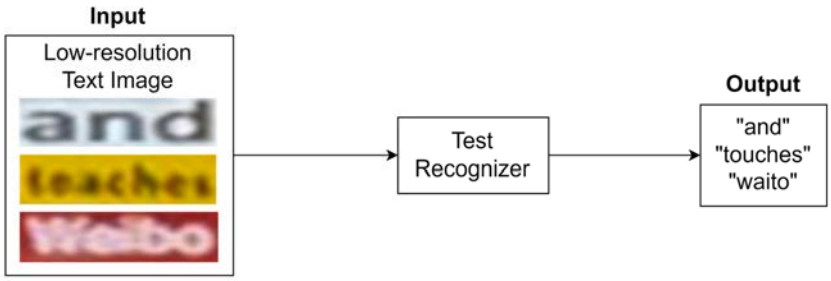

Figure 1.1 LR text image fed to a text recognizer directly

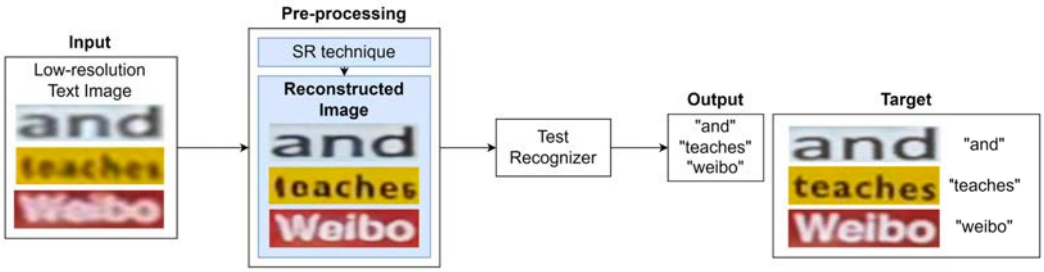

Figure 1.2 LR text images passed through the super-resolution method as a pre-processing before feeding to a text recognizer

scene text images Cai et al. (2019) before feeding them to the text recognizer. In the first era, SR methods were applied to enhance the scene text image Dong et al. (2016). However, SR in text scenes is different from traditional SR in the natural image. Since traditional SR methods only focus on reconstructing the detail of texture that satisfies human perceptual, SR in text scene images contains the content. Which, before and after characters have semantic information related to each other.

To develop a STISR model, paired LR-HR images are needed. Normally, it has two main approaches for building the pair dataset which is synthetic and realistic technique. The synthetic one is the process of downscaling the HR image and adding some noise to get the LR image which displays in Figure1.3. While the realistic technique, the LR and HR images are taken at different focal lengths of the camera as shown in Figure 1.4. Many SR methods trained by synthesis and realistic images dataset are proposed Chen et al. (2021a, 2022b); Wang et al. (2019); Zhang et al. (2019). However, the gap in the SR performance between synthesis and realistic data was found in practical applications, the trained models with synthesis datasets tend to

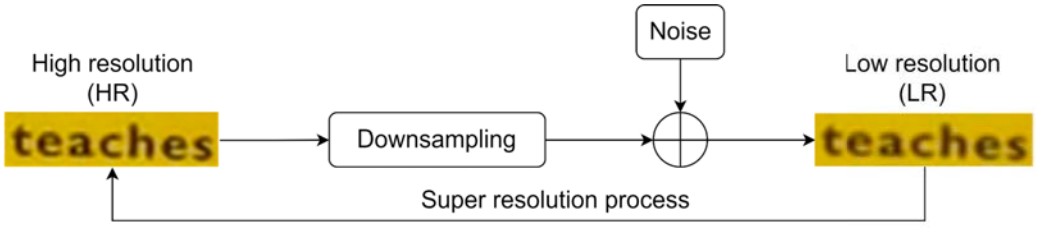

Figure 1.3 The building the LR-HR pair images dataset by synthesis technique

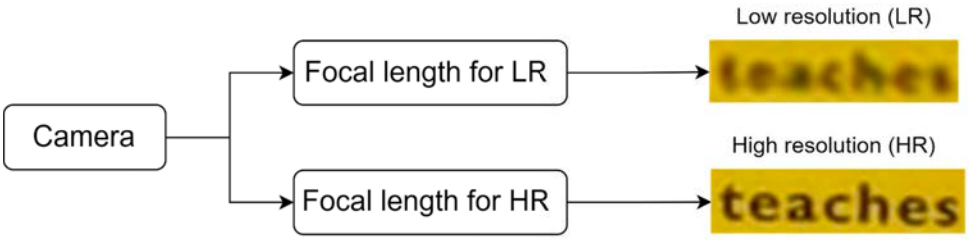

Figure 1.4 The building the LR-HR pair images dataset by realistic technique

drop the performance on real-world images Wang et al. (2020). To make it practical, some research has shifted to focus on the real-world single image.

For the realistic scene text dataset, it was proposed on ECCV-2020, namely, Textzoom Wang et al. (2020). Both of train-set and test-set of the Textzoom dataset are gathered from two SISR datasets: RealSR Cai et al. (2019) and SRRAW Zhang et al. (2019). It contains various natural scenes such as shops, street views, and documents. Comparing the synthesis and realistic LR images, the realistic one is much more challenging because the shape, luminance, and background are various, as shown in Figure 1.5. When we consider a small patch of the images, real LR

Figure 1.5 The examples of Textzoom dataset Wang et al. (2020) that compared between synthesis LR (syn LR), realistic LR (real LR), and ground truth (HR) images.

(a) Example images of easy subset.

(b) Example images of medium subset.

(c) Example images of hard subset.

Figure 1.6 Comparison between easy, medium, hard dataset of Textzoom dataset Wang et al. (2020)

images provide more server blur texture, while a synthesis LR still keeps remaining the original texture which makes it easier to restore the information.

The Textzoom consists of three subsets according to difficulty levels divided by the focal length camera as an easy, medium, and hard subset with annotation as shown in Figure 1.6. Moreover, they proposed the scene text image super-resolution, namely Text Super-Resolution Network (TSRN) Wang et al. (2020). To the best of my knowledge, TSRN is the first STISR method. A core architecture was based on SRResNet Ledig et al. (2017a). TSRN proposed to replace, sequential residual blocks in residual blocks, a boundary-aware loss, and a central alignment module. A gradient profile loss was proposed to be another loss for enhancing the boundary-aware character of the model. It was calculated by using the L1 norm of the gradient field of HR and generated images. The generated scene images have experimented on three text recognizers, ASTER Shi et al. (2018), MORAN Luo et al. (2019), and CRNN Shi et al. (2017). The result reported that TSNR outperformed in image quality and text recognition accuracy compared to state-of-the-art methods. However, the TSRN method focuses on every pixel in the image that can be disturbed from the background. It might affect upsampling performance on text images.

Since a transformer achieved huge success in national language processing (NLP), it was employed to develop the STISR method. To deal with the complicated background problem, a Transformer-Based Super-Resolution Network (TB-

SRN) Chen et al. (2021a) was proposed. It contains a self-attention module for extracting the sequential information, which this approach is robust to handle with the arbitrary orientation. TBSRN also employed a Position-Aware Module to highlight character regions with the reference of high-resolution images. It introduced to the use of the combination of the loss function in the method, such as a position-aware module using the L1 for computation of the attention map and a content-aware module taking a weight cross-entropy loss. TBSRN can achieve text recognition accuracy when compared with TSNR, at least 1% in the easy and hard subset and around 3% for the medium subset on Textzoom. However, it still remained the text-generating problem.

Text Gestalt Chen et al. (2022b) contains two modules, namely a Pixel-wise Supervision Module (PSM) to recover the color and contour of text image and a Stoke-Focused Module (SFM) to highlight the details of the stoke region. The interesting idea is the stroke region that tried to mimic the human commonsense for recovering detail process from a blurred image inspired by Gestalt psychology. The SFM consists of the multi-head self-attention blocks and norm layers. The loss function is calculated pixel-wise, and the attention map between HR and generated image is multiplied with a balanced parameter.

The idea of adding the knowledge to the model for the prediction of the texture of scene text was continued in Text Prior Guided Super Resolution (TPGSR) Ma et al. (2021a). Text prior (TP) was designed to guide the useful information to enhance the network for producing high-quality scene text. The generator of TP applied the CRNN Ma et al. (2021a) model to calculate the probability prediction. The loss function used the L1 norm and the KL divergence to measure the similarity between the TP of LR and HR images. Since the STISR models were applied to the Convolutional Neural Network (CNN) and transformer architectures, combining the different loss functions to improve the reconstructed scene text image, the generated image was much better than the previous one. However, this approach can increase the complexity of the model, unavoidable that it is caused by the overfitting problem.

Normally, regularization is the technique to reduce overfitting problems by

adding the coefficient parameter and penalty term. It can enhance the performance and simplify the model. In SISR, regularization techniques were applied. SRW-GANTV Shao et al. (2021) proposed the total variational regularization in Generative Adversarial Network (GAN) model to stabilize the network training and improve the quality of generated images. While Lipschitz Continuity Condition (LCC) Gouk et al. (2021) was employed to regularize the GAN model by mapping the image space to a new optimal latent space Zhong and Zhou (2021). Recently, L1 and L2 regularization terms were used in the GAN-based SISR model Viriyavisuthisakul et al. (2022c). The result reported that adding regularization terms can generate better detail than without regularization. However, the regularization parameters need to be fixed in the regularization term. A regularization parameter plays an important role in balancing the fidelity and regularization parameters that directly affect the reconstruction process. To overcome the limitation, Multiple Parametric Regularization (MPR) Viriyavisuthisakul et al. (2022a) was proposed. It is allowed the regularization parameters and degree of regularization term can be adjusted as an adaptive parameter in every training iteration. It was found that MPR can improve image quality in both image quality assessment (IQA) scores and human perception. Moreover, it could save computational time.

1.1.1 Research Problems and Contributions

STISR methods are special tasks which different from SR methods. Since traditional SR methods only focus on reconstructing the detail of texture and satisfying human visual perception. While STISR contains the semantic information that before and after characters have information related to each other. Therefore, the quality of the image is important, it can be affected by text recognizers such as missing characters, which cause wrong predictions.

In this research, we focus on improving the visual quality of realistic low-resolution scene text images by introducing the novel of adaptive weights and parametric regularization, it can employ the loss of many neural network architectures of super-resolution reconstruction as a pre-processing of the text recognition. The pro-

posed method should enhance the quality of images and achieve the highest accuracy when compared with the state-of-the-art. Therefore, we set the main contributions as below:

The main contributions of this study are as follows:

- We propose a novel adaptive framework for the loss function of STISR models that all parameters in the framework can be learnable.
- We propose three methods in the loss function of studied models: (1) Multiple Parametric Recognition (MPR), (2) Parametric Weights (PW), (3) Parametric Weights and Multiple Parametric Recognition (PW+MPR).
- The proposed method can improve the text recognition accuracy, visual comparison, and image quality assessment (IQA) than the state-of-the-art models.

The innovation point of this research is the adaptive learnable parameter in the parametric weights and multiple parametric regularization. A Parametric weights framework is different from normal balanced parameters that need to be fixed as constant values. However, it is not a state-forward approach to indicate the weight that should be. Most of the research was conducted by trial and error to find suitable values or follow the previous research. In PW, it allows the network can adjust those parameters by its gradient in every iteration. For multiple parametric regularization, it is the most general form; the regularization parameters and regularization terms can be added as much as needed, and the degree of the term can be any number followed by the gradient of the network. We tested the adaptive learned parameter regularization with Single Image Super Resolution (SISR) methods on the natural image domain.

CHAPTER 2
LITERATURE REVIEW

2.1 Scene Text Recognition

Scene Text Recognition (STR) aims to detect and recognize the text that appears on the scene images. There are the main approaches that involve the machine learning algorithm to recognize individual characters in an image or video and convert them into machine-readable text. It can be applied to read road signs, billboards, product labels, or documents. STR is more challenging than ordinary Optical Character Recognition (OCR) because it includes various text styles, shapes, orientations, and illuminations. It involves detecting and localizing text in an image or video, and then using OCR algorithms to recognize the individual characters.

STR algorithms are typically trained using supervised learning techniques, which involve labeling large datasets of images with the correct text to be recognized. These algorithms can be improved by using techniques such as data augmentation Atienza (2021), which involves generating new training data by applying transformations to existing images or finding the shape similarity Dai et al. (2022). The traditional methods usually adopted a bottom-up approach to recognize text images Jaderberg et al. (2014); Pan He et al. (2016). The text is detected and classified by the separated characters, then composed into text lines with the guidance of language models. While some approaches recognize the top-down manner Jaderberg et al. (2016). In the deep learning era, CRNN Shi et al. (2017) proposed to combine Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) in the encoder to obtain the sequential features of the text images. A Connectionist Temporal Classification (CTC) Baek et al. (2019) based decoder employs to maximize the

probability of paths that can reach a ground truth.

Since the scene text contains many shapes and different backgrounds, a Multi-Object Rectified Attention Network (MORAN) Luo et al. (2019) was proposed. To handle complex deformations, a MORAN was proposed to transform the scene text image for easier processing. The rectified image and character sequence were recognized by using an attention-based sequence recognition network. While ASTER Shi et al. (2018) introduced a Spatial Transformer Network (STN) to rectify irregular text images in an unsupervised manner for better recognition. FAN Cheng et al. (2017) focused on the attention drift problem and introduced a focusing network to rectify attention regions.

Scene text recognition has the potential to revolutionize a wide range of industries, including education, media, and advertising. It can also enable new applications such as automated translation of street signs, other public information, and assistive technologies for the visually impaired. However, the accuracy of scene text recognition algorithms is still limited by a number of challenges, including variations in font, style, and layout and the presence of noise and distortion in the images.

2.2 Single Image Super-Resolution (SISR)

Single Image Super Resolution (SISR) is an ill-posed problem. It is the technique to enhance the detail of images to be a high-resolution (HR) image by using its low-resolution (LR) image. Super-Resolution (SR) can be used to improve the quality of images or videos for various applications, such as medical imaging Armanious et al. (2020), surveillance Müller et al. (2020); Pejman et al. (2016), and entertainment Mengyu et al. (2020). It can also be used to reduce the amount of data required to transmit or store images and videos, by increasing the resolution of the data without increasing its size.

To train the model, the paired LR-HR dataset is required. The simplest way to build the paired LR image is a downscale HR image and add the noise. The degraded information technique is the general approach to creating the dataset such as DIV2K

Agustsson and Timofte (2017), Set5 Bevilacqua et al. (2012), Set14 Zeyde et al. (2012), or BSD200 Martin et al. (2001). However, the real LR-HR images dataset is considered. Since the SISR performance that trained on synthetic dataset world decreased performance significantly on real-world images. Given an LR images as I_{lr}, it can be expressed as Equation 2.1

$$I_{lr} = D(I_{hr}, \theta) \tag{2.1}$$

where $D()$ represents the downsampling process by the parameter of θ. The downscaling parameter θ is unknown in the real scenario. I_{hr} is HR images.

To obtain the estimate of the potential HR image, the process is reversed in Equation 2.1, which can be represented as below.

$$I_{sr} = F(I_{sr}, \theta_p), \tag{2.2}$$

where I_{sr} is generated image and $F()$ is super resolution process while θ_p represents corresponding parameter.

The traditional approaches leverage the detail by taking the average of neighboring pixels, the interpolation technique. These methods use algorithms to estimate the missing pixels in an image or video, based on the information in the surrounding pixels. However, the interpolation technique has a limitation in giving the image detail because of the data processing inequality Beaudry and Renner (2011). To improve the quality of image, machine learning based method was applied. The convolutional neural network was proposed to apply in SISR. These methods use machine learning algorithms to learn patterns in high-resolution images and use this information to generate enhanced versions of low-resolution images or videos. The limitation in the interpolation technique is solved by learning from the large dataset that the CNN network can learn to hallucinate the detail; namely, SRCNN Dong et al. (2016).

Recently, a Generative Adversarial Network (GAN) has become more pop-

ular in SISR, and SRGAN Ledig et al. (2017b) is the first pioneer work. In 2018, Enhanced SRGAN or ESRGAN Wang et al. (2018) was proposed to improve the architecture of SRGAN and loss function.

2.3 Scene Text Super-Resolution

A Scene Text Super-Resolution STISR aims to improve the resolution quality of text image and reconstructs semantically correct text which is different from SISR methods. In the part of improve the visual quality, those SISR methods can be directly adopted for STISR process. In the first era, the deep learning-based models are conducted many text super-resolution methods. SRCNN Dong et al. (2016) was applied in text image and achieved the best performance in ICDAR 2015 competition Peyrard et al. (2015). While, CNN layers were employed for feature extraction in binary document image by transposed convolution Pandey et al. (2018). In Lai et al. (2017), a Laplacian-pyramid backbone was applied and Gradient Difference Loss (GDL) with L1/L2 loss was proposed to enhance edges in super-resolution image. However, those SR methods were proposed for natural images super-resolution which not suitable for handling scene text images. Since, it directly used generic SR frameworks and ignored text-specific properties such as the character-level details and text layouts. Moreover, most of the models were trained by using the synthesis LR-HR image dataset which the performance would degrade significantly on real-world images.

To improve the performance of STISR Wang et al. (2020) created as a real-world STISR image dataset, namely Textzoom. Moreover, TSRN Wang et al. (2020) models proposed to use the central alignment module and sequential residual block (SRB) to take the semantic information in internal features. In STISR, the character edges are important. Therefore, many models tried to propose a technique that can reconstruct the shape edge. A Gradient Profile Prior (GPP) loss was proposed in TSRN to generate the shape edge. Text Gestalt Chen et al. (2021c) proposed a pixel-wise supervision module (PSM) to recover the color and a stroke-focused mod-

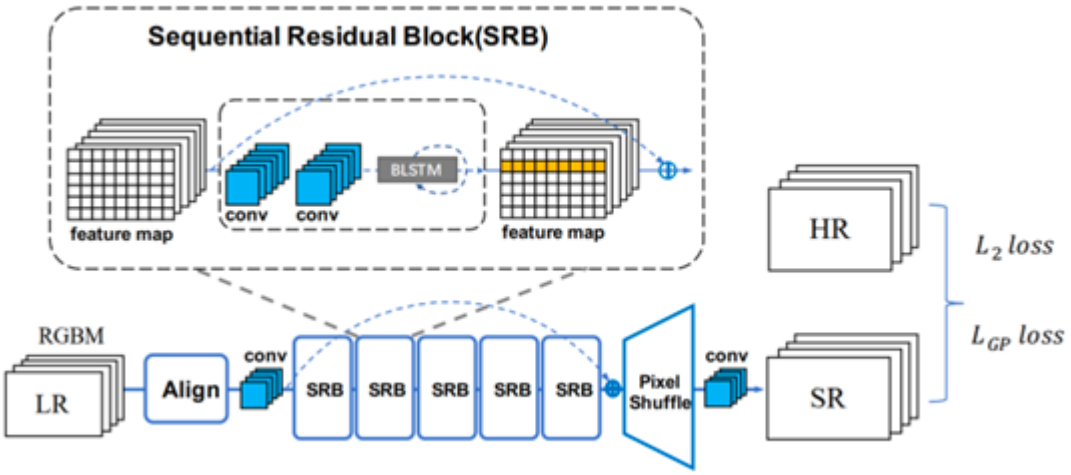

Figure 2.1 The overview architecture of TSRN Wang et al. (2020).

ule (SFM) to highlight the detail of stroke regions. While Transformer-Based Super Resolution Network (TBSRN) Chen et al. (2021b) contains a position-aware module and a content-aware module providing text prior to tackling the text detail properties. TPGSR Ma et al. (2021a) is the first method that introduces the categorical text prior information into the model learning process. TATT Ma et al. (2022) proposed combining the CNN and text attention network because only CNN-based methods are ineffective in dealing with spatially-deformed text images, including rotation and curved shapes.

2.3.1 Text Super-Resolution Network

The architecture of Text Super-Resolution Network (TSRN) is a CNN-based STISR method that improved from SRResNet Ledig et al. (2017a) architecture. The overview of TSRN is shown in Figure 2.1. TSRN proposed that sequential residual blocks (SRBs) replace the traditional ones in SRResNet. Second, gradient profile loss was introduced to enhance the edge of characters by using the gradient between HR and generated images. Moreover, they found that alignment between LR-HR paired images can affect the reconstructed images. Therefore, the central alignment method was proposed to connect the position of pixels.

From Figure 2.1 the RGB scene text images were rectified and aligned by

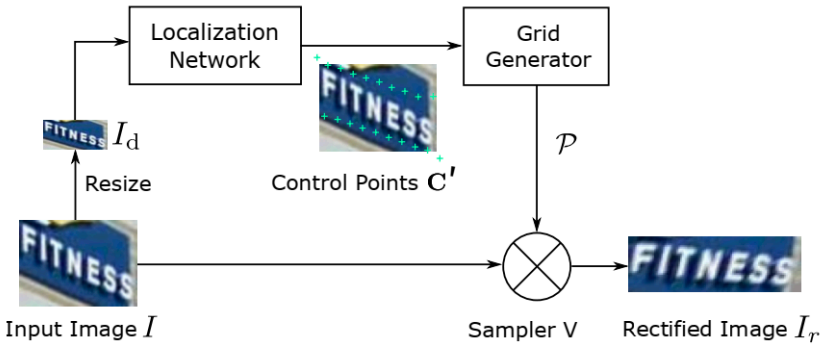

Figure 2.2 Structure of the rectification network Shi et al. (2018).

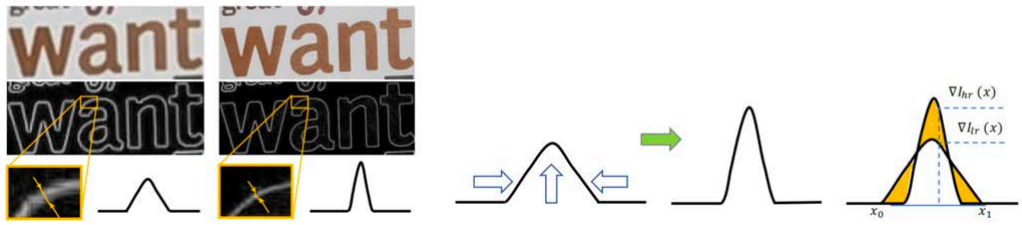

Figure 2.3 The illustration of gradient field and Gradient Prior Loss Shi et al. (2018).

the central alignment model as an input. Noted that the RGB rectified image is the process of transforming the text in the image to be easier to process by using the localization and control point of Thin-Plate-Spline (TPS) Dong et al. (2014) as shown in Figure 2.2. Then, the RBG rectified images are changed to the grayscale image and fed to the pipeline. CNN layers extract the shallow feature and pass through the chained SRBs. Here SRBs can extract deeper and sequential dependent features and do shortcuts like ResNet. It was modified to add the Bi-directional LSTM (BLSTM) mechanism, which can take the horizontal and vertical convolutional features as sequential input and update the weight by backpropagation. The loss function is calculated by MSE loss and Gradient profile loss which can be expressed below.

$$L_{TS} = L_2 + L_{GP}, \tag{2.3}$$

where L_2 represents MSE loss and L_{GP} is Gradient profile loss.

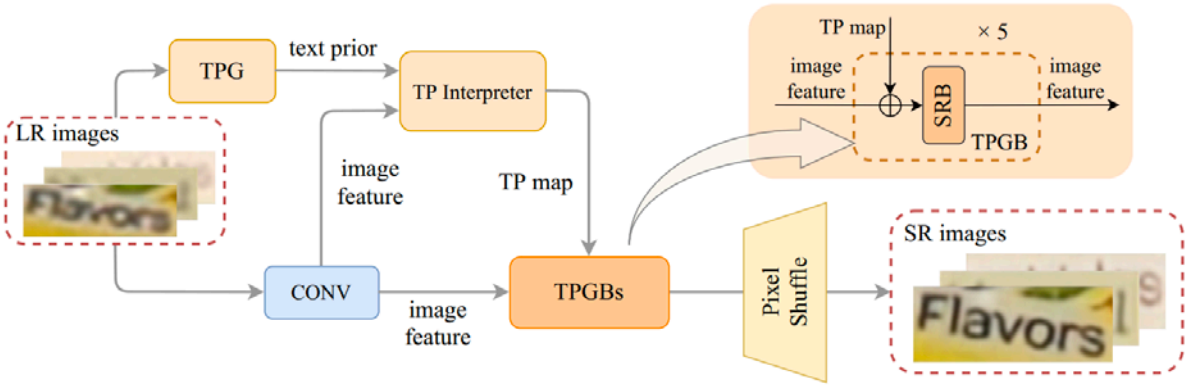

Figure 2.4 Architecture of TATT network for STISR Ma et al. (2022).

$$L_{GP} = \|\nabla I_{hr}(x) - \nabla I_{sr}(x)\|_1 \tag{2.4}$$

where $\nabla I_{hr}(x)$ and $\nabla I_{sr}(x)$ represent the gradient field of HR image and SR images respectively.

Gradient Profile loss is designed to encourage the model to generate a sharper edge in SISR as shown in Figure 2.3. The idea of this loss function takes advantage of the color of text characters in images that contrast with the background. In this case, sharpening the boundaries rather than smoothing the world make the character more obvious.

2.3.2 Text Attention Network (TATT)

Text Attention Network (TATT) was proposed to solve the problem of spatially deformed text in reconstructed text images, especially rotated and curved-shaped images. It was caused by the CNN-based methods that are adopted in locality-based operations. The architecture of TATT combined the ability of the CNN and Transformer to extract the text prior information and semantic guidance of text prior to the recognition process. To refine the visual appearance by imposing structural consistency on the reconstructions of regular and deformed text, a text structure consistency loss is proposed.

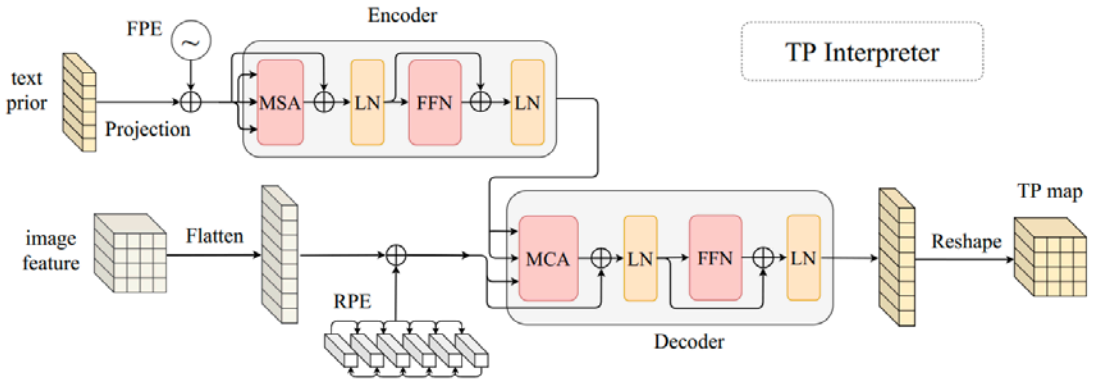

Figure 2.5 Architecture of TP Interpreter Ma et al. (2022).

The architecture of TATT is shown in Figure 2.4. The LR images are fed as the input to the Text Prior Generator (TPG) for finding the categorical probability vectors with 37 classes in total, *a* to *z*, *0* to *9*, a blank class, while LR images are passed through 9×9 convolution layer to extract space the image features. The text prior from TPG and image feature is sent to the Text Prior Interpreter (TP Interpreter) for mapping the correlation of text prior and image feature, then assigning the semantic guideline in the text prior sequence as TP map to a corresponding location with spatial domain for reconstructed image process. The TP map and image feature is the input in Text Prior Guided Blocks (TPGBs) that are composed of five blocks. Each TPGB fuses the TP map and image feature with element-wise attention to Sequential Recurrent Blocks (SRBs) to reconstruct the high-resolution image features.

In the proposed method TATT, TP Interpreter (TPI) is one of the crucial parts that aims to interpret the text prior to TPG and image features. It is designed to do the semantics guidance to the correlated spatial position in the image feature domain. The main idea of TP Interpreter is to enlarge the text prior to TPG to the shape of the image feature and merge them by convolution. Since the convolution operation is limited to a small effective range, the semantic of text prior cannot be assigned to the distant spatial location. Hence, TATT applies a Transformer-based attention mechanism in this unit to enforce the global correlation between text prior and image features.

In Figure 2.5, the TP Interpreter consists of encoder and decoder. Using the

correlation between the semantics of each character in the text, the Encoder encodes the text prior to outputting its context-enhanced features. As part of the semantic information to be translated into image features, the decoder applies cross attention between the context-enhanced feature and the image feature to it. In the next section, the loss function of TATT is described in detail.

Text Structure Consistency Loss

A Text Structure Consistency (TSC) loss is proposed to improve the visual appearance. Since the CNN model has a limitation on the deformed text feature representation from regular text features and the reconstructed text images have weaker character structures with relatively low contrast. A TSC loss minimizes the distance of three types of images, the deformed version of SR text images $D\mathcal{F}(Y)$, the SR version of the deformed LR text image $\mathcal{F}(DY)$, and the deformed ground truth $D(X)$. This loss is extended the structure-similarity index measure (SSIM) as in Equation 2.5 to the triplex SSIM (TSSIM) as Equation 2.6.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.5}$$

$$TSSIM(x,y,z) = \frac{(\mu_x\mu_y + \mu_y\mu_z + \mu_x\mu_z + C_1)(\sigma_{xy} + \sigma_{yz} + \sigma_{xz} + C_2)}{(\mu_x^2 + \mu_y^2 + \mu_z^2 + C_1)(\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + C_2)} \tag{2.6}$$

where μ_x, μ_y, μ_z and $\sigma_x, \sigma_y, \sigma_z$ represent the mean and standard deviation of triplet x, y and z, respectively. While, σ_{xy}, σ_{yz} and σ_{xz} denote the correlation coefficient between $(x,y),(y,z)$ and (x,z), respectively. C_1 and C_2 are small constants to the stability of the dividing values close to zero.

$$\mathcal{L}_{TSC} = 1 - TSSIM(D\mathcal{F}(Y), \mathcal{F}(DY), D(X)) \tag{2.7}$$

Finally, TSC loss $\mathcal{L}_{TSC}$ is used to measure the mutual structure difference

among $D\mathcal{F}(Y)$,$\mathcal{F}(DY)$ and $D(X)$ as in Equation 2.7, when D denotes the random deformation.

TATT loss function

The overall loss function of TATT consists of super-resolution loss $\mathcal{L}_{SR}$, a text prior loss $\mathcal{L}_{TP}$ and TSC loss $\mathcal{L}_{TSC}$. The SR loss $\mathcal{L}_{SR}$ calculates the difference between SR output and the ground-truth HR image by adopting the L_2 norm. The TP loss $\mathcal{L}_{TP}$ measures between the text prior that extracted from the LR image and those from the ground truth by taking L_2 norm and KL Divergence. The $\mathcal{L}_{SR}$ and $\mathcal{L}_{TP}$ are summed up with TSC loss $\mathcal{L}_{TSC}$ as the Equation 2.8.

$$\mathcal{L}_{TATT} = \mathcal{L}_{SR} + \alpha\mathcal{L}_{TP} + \beta\mathcal{L}_{TSC} \tag{2.8}$$

where α and β are balancing parameters that set to 1.0 and 0.1 respectively.

2.4 Regularization

To the best of our knowledge, the regularization term is applied to reduce the overfitting problems by adding the coefficient parameter and penalty term. It can enhance the performance and simplify the model. There are three types of regularization methods: L1 regularization, L2 regularization, and Elastic Net regularization. L1 regularization or Least Absolute Shrinkage and Selection Operator regression (Lasso regression) can estimate the median of the data distribution while calculating loss. L2 regularization or Ridge regression is performed by adding the square of the magnitude of nonzero coefficients. Each of these regularization techniques has its own advantage and disadvantage. L_1 regularization trends to produce spare models, whereas L_2 regularization trend to produce model with small weights. Meanwhile, Elastic Net regularization tried to merge both L1 and L2 regularization terms. Therefore, it can take benefit from both, good learning of complex data and robust to outlier data. However, adding the regularization term need to fix the regularization parame-

ter. The regularization parameter is important when the loss functions are combined. It is the weight coefficient of the total loss function to balance the fidelity term and regularization term. To make the regularization term more efficient, the parametric regularizations for the loss function are proposed in the next topic.

2.4.1 L1 and L2 regularization

L1 regularization or Least Absolute Shrinkage and Selection Operator regression (Lasso regression) is the sum of the absolute values of the weight parameters. For the loss function, the L1 regularization term is multiplied with the regularization parameter and added to the loss function as Equation 2.9. While calculating loss, L1 regularization estimates the median of the data distribution. It is robust to outliers but weak when learning complex patterns.

$$\mathcal{L}_{\ell 1reg} = \mathcal{L}_{total} + \lambda \sum_{\forall j} \left\| W_j \right\|, \tag{2.9}$$

$$\mathcal{L}_{\ell 2reg} = \mathcal{L}_{total} + \lambda \sum_{\forall j} \left(W_j \right)^2, \tag{2.10}$$

where

- $\mathcal{L}_{total}$ is the traditional loss of ground-truth and predicted result.
- λ is the hyper-regularization parameter of the weight coefficient that must be manually tuned.
- W_j is the value of the weight parameters.

L_2 regularization or Ride Regression or weight decay is the most common type of all regularization techniques. The L_2 regularization term of L2 is defined as the Euclidean Norm of the weight matrices. It is the sum overall squared weight value of a weight matrix. To be the loss function, L2 performs adding the penalty which

is equivalent to the square of the magnitude of nonzero coefficients as expressed in Equation 2.10. Despite its limitations, it is effective when learning complex patterns. It reduces complexity without reducing variables but is not robust to outliers data. Due to the squared difference, the error would be much larger.

To make the model more efficient and stable, the L1 and L2 regularization techniques were applied to the GAN-based SISR model Viriyavisuthisakul et al. (2022c). Results showed that adding regularization terms produces better detail in both image quality score and perception.

2.4.2 Elastic Net regularization

To make the Elastic Net regularization robust against outliers and good for learning complex patterns, it mixes both L1 and L2 regularization terms.

$$\mathcal{L}_{elastic\_reg} = \mathcal{L}_{total} + \alpha \sum_{\forall j} \left\| W_j \right\| + \beta \sum_{\forall j} \left(W_j \right)^2, \tag{2.11}$$

where

- $\mathcal{L}_{tatal}$ is the traditional loss of ground-truth and predicted result.
- α and β are the regularization parameters that need to be manually tuned.
- W_j is the value of the weight parameters.

System performance is strongly affected by parameter selection for regularization. However, this network parameter can be very difficult to optimize. We propose an approach in the next section that can help find the optimal parameter for the network.

2.5 Adaptive learned parameter (Parametric)

In an adaptive learning system, an activation function called Parametric Rectified Linear Unit (PReLU) He et al. (2015) has been proposed to rectify the neural

network. PReLU is designed to overcome the limitation of Rectified Linear Unit (ReLU) Xu et al. (2018) or Leaky Rectified Linear Unit (LeakyReLU) Maas et al. (2013) in a negative part. Since ReLU fixes a value as a zero to control the negative part as Equation 2.12 that can deactivate and activate some neurons at the same time. Compared to sigmoid/tanh functions, it greatly accelerates stochastic gradient descent (SGD). However, ReLU can be fragile during training, it is not zero-centric. It can have a dead neuron which is the biggest problem, due to the non-differentiable at zero. To fix the problem, instead of the function is zero, the LeakyReLU requires a small positive slope constant to avoid zero gradients as in Equation 2.13. This may reduce the occurrence of the detrimental zigzag effect that is noted in the linked thread.

Instead of fixing the slope parameter as in LeakyReLU, PReLU allows it to be adjustable. Since each layer learns a single slope parameter in the feed-forward network, it can be a learnable parameter. It can simply explain as the below Equation 2.14. Figure 2.6 shows the ReLU, LeakyReLU, and PReLU activation functions.

$$f(x) = \max(0, x) \tag{2.12}$$

$$f(x) = 1(x < 0)(\alpha x) + 1(x \geq 0)(x) \tag{2.13}$$

when α is a small constant such as 0.01, or so.

$$f(x) = \begin{cases} x, & if \quad x > 0 \\ ax, & if \quad x \leq 0 \end{cases} \tag{2.14}$$

Given x is any input and a is the negative slope which is a learnable parameter.

- if $a = 0$, f becomes ReLU

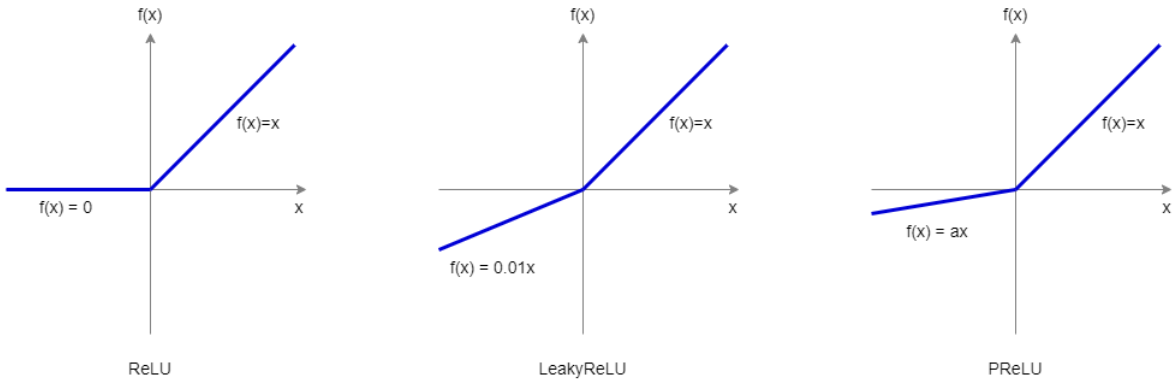

Figure 2.6 The activation function of ReLU, LeakyReLU, and PReLU function

- if $a > 0$, f becomes leaky ReLU
- if a is a learnable parameter, f becomes PReLU

From above formula can also be written as equation 2.15

$$f(x) = \max(0, x) + a \min(0, x) \tag{2.15}$$

To alleviate the problem of the suitable parameters in regularization term, an adaptive learned parameter regularization, Parametric L1 regularization, Parametric L2 regularization, Parametric elastic net regularization, and Multiple parametric regularizations were proposed Viriyavisuthisakul et al. (2022a). The parametric regularization function can work in both CNN-based and GAN-based SISR models. It can encourage the network to produce more detail and converge faster.

2.5.1 Image Quality Assessment

An automated Image Quality Assessment (IQA) has been developed to determine the suitability of an image for diagnostic purposes. However, it has emerged more recently Barman et al. (2019). Image Quality Assessment (IQA) is important in many fields, including photography, digital imaging, and computer vision. IQA algorithms are based on generic image quality parameters such as illumination, contrast, and sharpness, while some algorithms are based on storing image quality parameters

such as main position anatomical features within images. By evaluating the quality of images, it is possible to improve the accuracy and reliability of image-based systems and to ensure that the images produced by these systems are of the highest possible quality. This is typically done by comparing the original image to a reference image. There are many methods for assessing image quality, each with its own strengths and weaknesses. One common method is Peak-Signal-to-Noise-Ration (PSNR). It takes the mean squared error (MSE), for calculating the average of the squares of the pixel-wise differences between the original and reference images. This method is useful for determining the level of distortion in an image, but it can be sensitive to noise. It may not always provide an accurate assessment of image quality. The structural similarity index (SSIM) measures the structural similarity between two images by looking at the degradation in the image due to processing. Another method

IQA is a kind of objective evaluation method that is widely used in the super-resolution task. It can be classified into three groups such as Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) or objective blind, as shown in Figure 2.7.

Full-reference IQA

Full-Reference (FR) image quality assessment involves comparing an original, pristine image with a modified version of that same image (e.g., a compressed version) and using a metric to quantify the difference between the ground truth and generated image. This allows for a direct comparison of the quality of the modified image to the original Soundararajan et al. (2020).

Many different full-reference image quality assessment metrics have been developed, including:

- Peak Signal-to-Noise Ratio (PSNR) is a full-reference image quality assessment metric that measures the ratio of the peak signal strength to the noise present in an image. It is commonly used to compare the quality of lossy images and video compression.

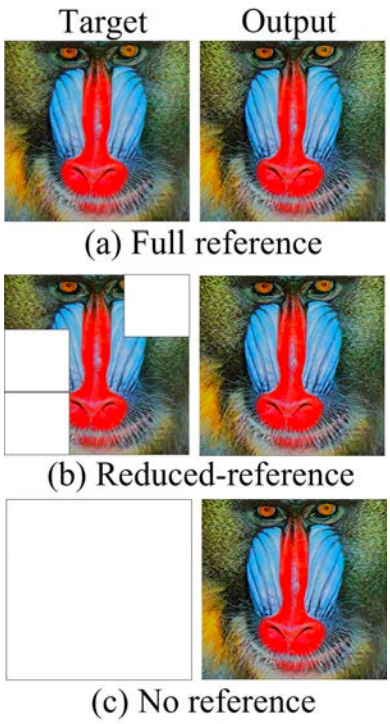

Figure 2.7 The list of metrics. (a) Full-Reference (FR) requires target and generated image or output to calculate the IQA score, (b) Reduced-Reference (RR) requires some part of the target image along with output, (c) No-Reference or Objective-blind is not require any target.

To calculate PSNR, the difference between the original and modified images is first calculated using Mean Squared Error (MSE) as Equation 2.16. The MSE is then converted to PSNR using the following formula:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{2.16}$$

where MAX is the maximum possible pixel value (e.g., 255 for 8-bit grayscale images or 65,535 for 16-bit images).

The resulting PSNR value is expressed in decibels (dB). A higher PSNR value indicates a higher quality image, with a maximum value of infinity for an exact copy of the original image. A PSNR value of 20 dB or higher is generally considered to be good quality.

PSNR is a simple and easy-to-compute metric, but it has some limitations. It

is sensitive to noise and can give overly optimistic results for images with high noise levels. It is also not a good predictor of perceived image quality, as it does not take into account the way the human visual system processes images.

- Structural SIMilarity index (SSIM) is a full-reference image quality assessment metric that compares local patterns in the original and modified images to determine their similarity. It is based on the idea that small local changes in an image are typically more noticeable to the human eye than global changes.

  To calculate SSIM, the original and modified images are first divided into smaller image windows and the mean, variance, and covariance of the pixel intensities within these windows are calculated. These values are then used to calculate the SSIM index using the following formula:

$$SSIM = \frac{(2 \cdot \mu_1 \cdot \mu_2 + c_1) \cdot (2 \cdot \sigma_{12} + c_2)}{(\mu_1^2 + \mu_2^2 + c_1) \cdot (\sigma_1^2 + \sigma_2^2 + c_2)} \tag{2.17}$$

  where μ_1 and μ_2 are the means of the pixel intensities in the original and modified images, σ_1^2 and σ_2^2 are the variances, and σ_{12} is the covariance. c_1 and c_2 are constants that ensure that the SSIM index is bounded between -1 and 1.

  The resulting SSIM index is a value between -1 and 1, with a value of 1 indicating a perfect match between the original and modified images. SSIM is considered to be a more reliable predictor of perceived image quality than PSNR, as it takes into account the way the human visual system processes images. However, it is more computationally expensive to calculate than some other image quality assessment metrics.

- Learned Perceptual Image Patch Similarity (LPIPS) Johnson et al. (2016) is a no-reference image quality assessment metric that uses machine learning techniques to learn the features of high-quality images and use these features to predict the quality of a given image. It is based on the idea that the human

visual system is more sensitive to certain image features than others, and that these features can be learned from a dataset of high-quality images.

To calculate LPIPS, the original and modified images are first divided into smaller image patches and the features of these patches are extracted using a convolutional neural network (CNN). The feature vectors for the patches in the original and modified images are then compared using a distance metric, such as the Euclidean distance or the cosine similarity. The overall LPIPS score for the images is calculated as the average distance between the patches.

LPIPS is considered to be a more reliable predictor of perceived image quality than some other IQA metrics, as it takes into account the way the human visual system processes images. However, it requires a large dataset of high-quality images for training and is computationally expensive to calculate.

- Mean squared error (MSE) Sheikh and Bovik (2006a); Zhang and Kaveh (2015): This measures the average squared difference between the original and modified images. It is a simple metric that is easy to compute, but it can be sensitive to noise.

- Wavelet-based image quality metrics Sheikh and Bovik (2005, 2006b): These metrics use wavelet transforms to decompose the images into different frequency bands and compare the energy in these bands between the original and modified images.

- Learned image quality metrics Bendale and Boult (2016); Ma et al. (2018): These use machine learning techniques to learn the features of high-quality images and use these features to predict the quality of a given image.

In general, full-reference image quality assessment is considered to be more reliable and accurate than no-reference or reduced-reference methods, as it takes into account the entire original image when evaluating quality. However, it is also more computationally expensive and requires access to the original image, which may not always be available.

In this research, we apply the well-known IQA, PSNR, and SSIM to measure the quality of images.

Reduced-Reference IQA

Reduced-Reference (RR) image quality assessment involves evaluating the quality of a modified image using only partial information about the original image. This can be useful in situations where the original image is not available or is too large to use for comparison.

Several different reduced-reference image quality assessment metrics have been developed, including:

- Feature-based metrics: These measure the similarity between features extracted from the original and modified images. The features could be based on color, texture, or other image characteristics.
- Noise-based metrics: These estimate the noise present in the modified image and compare it to the noise present in the original image.
- Spatial pooling-based metrics: These divide the original and modified images into smaller regions and compare the average pixel values within each region.

Reduced-reference image quality assessment is generally less accurate than full-reference methods, as it does not take into account the entire original image. However, it can be more practical in some situations, such as when the original image is unavailable or too large to compare.

No-Reference IQA

No-Reference (NR) image quality assessment involves evaluating the quality of an image without using any information about the original image. This can be useful in situations where the original image is not available or when it is not practical to use a full-reference, or reduced-reference method Zhang and Kankanahalli (2016).

Several different no-reference image quality assessment metrics have been developed, including:

- Natural image statistics (NISQ) Field (2002): These metrics use statistical models of natural images to predict the perceived quality of a given image.
- Blur and noise estimation: These metrics estimate the amount of blur and noise present in an image and use this information to predict the perceived quality.
- Human visual system (HVS) models: These models simulate how the human visual system processes images and uses this information to predict the perceived quality of an image.
- Mean Opinion Score (MOS): This method obtains from human individual opinion scores. However, the MOS score is employed objective metrics and is costly which can have bias, but it can reflect the quality of the image by human perception. However, the MOS score is employed objective metrics and is costly which can have bias, but it can reflect the quality of the image by human perception.

CHAPTER 3
METHODOLOGY

In this section, we will explain the proposed methods in STISR such as CNN-based and Transformer-based methods in detail, followed by the design of the loss function.

3.1 Dataset

The dataset used in this study is Textzoom, which was introduced by Wang et al. in their work ”TextZoom: A Magnification-Aware Dataset for Real-World Scene Text Super-Resolution” (2020) Wang et al. (2020). Textzoom is a novel dataset that addresses the need for real-scene text super-resolution (SR) by leveraging the RealSR and SR-RAW datasets. The Textzoom dataset comprises a total of 21,740 high-quality images, which are divided into a training set and a test set. The training set consists of 17,367 images, while the test set contains 4,373 images. Each image pair in Textzoom consists of a low-resolution (LR) text image and its corresponding high-resolution (HR) counterpart. Notably, each LR-HR image pair is meticulously annotated with information about its content, direction, and focal length.

To evaluate the performance of SR algorithms on the Textzoom dataset, the test set is further divided into three subsets based on the camera length: easy, medium, and hard. The easy subset contains 1,619 samples, the medium subset consists of 1,411 samples, and the hard subset comprises 1,343 samples. This division allows for a comprehensive assessment of SR algorithms across different difficulty levels, providing insights into their robustness and generalization capabilities. The availability of well-annotated LR-HR image pairs in the Textzoom dataset makes it a valuable resource for researchers and practitioners working on scene text super-resolution. By

leveraging this dataset, researchers can develop and evaluate novel SR algorithms that aim to enhance the quality and legibility of text in real-world scenes.

3.2 CNN-based STISR model with parametric weights

In Section 2.3.1, we introduced the Text Super-Resolution Network (TSRN) as a method to enhance the performance of super-resolution for scene text images. Our proposed approach builds upon the TSRN model by incorporating additional parameters weights into the traditional loss function, as illustrated in Figure 3.1. To improve the accuracy of text super-resolution, we utilize paired LR (low-resolution) and HR (high-resolution) images as inputs to the TSRN model. These image pairs are carefully annotated with corresponding text content, direction, and focal length, which are essential factors in accurately reconstructing SR scene text images.

During the optimization process, the loss function plays a crucial role in guiding the training of the TSRN model. In our approach, the loss function is calculated using two main components: the Mean Squared Error (MSE) and the gradient profile loss. The MSE measures the pixel-wise difference between the predicted HR image and the ground truth HR image. By minimizing this error, the TSRN model learns to generate higher-resolution text images that closely resemble the ground truth. In addition to the MSE loss, we introduce the gradient profile loss, which focuses on preserving the structural details of the text during the super-resolution process. By incorporating this loss component, the TSRN model is encouraged to generate images with sharper edges and clearer text boundaries, leading to improved legibility and visual quality. By combining these loss components and incorporating parameter weights into the optimization process, our proposed TSRN model demonstrates enhanced performance in reconstructing SR scene text images. The additional weights allow us to fine-tune the importance of different loss components, enabling better control over the optimization process and ultimately improving the overall super-resolution results.

In our study, we have made significant advancements in the trade-off weight

formulation between the Mean Squared Error (MSE) loss, denoted as L_2, and the gradient profile loss, denoted as L_{GP}, as shown in Equation 2.3. Instead of employing a constant value for this trade-off weight, we have introduced adaptive parameters that dynamically adjust to align with the gradients of the network. This modification enhances the flexibility and adaptability of the TSRN model during the training process.

To achieve this, we have incorporated the concept of parametric weights into the TSRN model, which enables the adjustment of the trade-off weight in each training iteration. These parametric weights are updated by the Stochastic Gradient Descent (SGD) optimizer, allowing the network to optimize and fine-tune the importance of the MSE loss and the gradient profile loss based on the specific characteristics of the input data. The new value is updated by the SGD optimizer in each training iteration, called parametric weight as in the yellow area in Figure 3.1

The adaptive nature of the parametric weights offers several benefits. Firstly, it allows the network to dynamically prioritize different loss components based on the current training stage and the complexity of the scene text images being processed. This adaptability ensures that the model can effectively balance the reconstruction accuracy and the preservation of fine details, leading to improved super-resolution results.

Furthermore, by aligning the parametric weights with the gradients of the network, the optimization process becomes more efficient and effective. The network can focus its learning on areas where the gradients are more prominent, enabling better convergence and facilitating the extraction of relevant features from the input data.

$$L_{TSP} = (W_p^1)L_2 + (W_p^2)L_{GP}, \tag{3.1}$$

where L_2 represents MSE loss and L_{GP} is gradient profile loss. While W_p^1 and W_p^2 are the parametric weight of MSE and gradient profile loss.

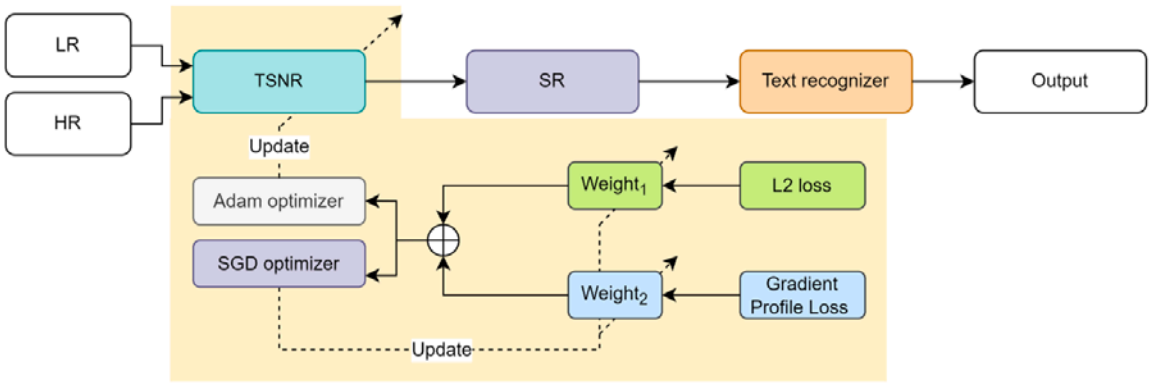

Figure 3.1 The proposed method of TSRN and weight parameter

3.2.1 CNN and Transformer-based STISR model with Parametric framework

Through our experiments, we have observed that traditional super-resolution (SR) techniques designed for natural images have limitations when applied to scene text image reconstruction. Recognizing this limitation, we have explored the integration of Convolutional Neural Networks (CNNs) and transformer techniques to overcome the challenges specific to scene text SR. CNNs have demonstrated their effectiveness in extracting meaningful features from images, making them an essential component in image processing tasks. However, when it comes to scene text SR, the transformer technique, with its attention mechanism and ability to capture sequential information, proves to be a better fit. This is because scene text SR requires not only generating high-quality textures but also accurately locating the pixels corresponding to the text regions. The attention maps generated by transformers aid in capturing the intricate details and relationships among text pixels, contributing to more precise and context-aware super-resolution results.

Furthermore, one of the common challenges faced by many STISR (Scene Text Image Super-Resolution) models is the problem of overfitting. Overfitting occurs when the model learns to predict characters in a scene text image that resemble other characters with similar shapes. To address this issue, regularization techniques are commonly employed to prevent the model from becoming overly sensitive to minute variations and noise in the training data. Regularization methods, such as L1 or L2 regularization, introduce penalties to the loss function during training. These

penalties discourage the model from assigning excessive importance to individual pixels and encourage it to learn more generalized representations. By promoting smoother and more robust predictions, regularization helps prevent overfitting and improves the generalization capability of the STISR models.

Multiple Parametric Regularization (MPR) technique, introduced by Viriyavisuthisakul et al. in their work "Multiple Parametric Regularization for Single Image Super-Resolution" (2022) Viriyavisuthisakul et al. (2022a), has shown promising performance in the field of image super-resolution. MPR utilizes multiple regularization parameters to enhance the regularization process and improve the quality of the super-resolved images. This technique can be effectively combined with both CNN-based and GAN-based models to achieve superior results.

However, it is important to recognize that text images present unique challenges compared to general image super-resolution tasks. Text images often have specific characteristics and requirements that differ from those of regular natural images. For example, accurately preserving the sharpness and legibility of text, ensuring the correct reconstruction of fine details, and precisely locating text regions are crucial factors in scene text image super-resolution. While CNN or GAN-based models have demonstrated success in various image-related tasks, including image super-resolution, the particularities of text images require additional considerations. Models designed specifically for scene text super-resolution must incorporate mechanisms that address the intricacies of text, such as the attention to character shape, alignment, and readability.

Therefore, when applying the MPR technique to text image super-resolution, it is essential to adapt and tailor the approach to the unique requirements of text reconstruction. This could involve incorporating attention mechanisms, utilizing text-specific loss functions, or incorporating additional contextual information during the super-resolution process. By accounting for these specific considerations, the MPR technique, when combined with CNN or GAN-based models, has the potential to deliver notable improvements in text image super-resolution.

Then we are eager to know how MPR works in the STISR methods. As the

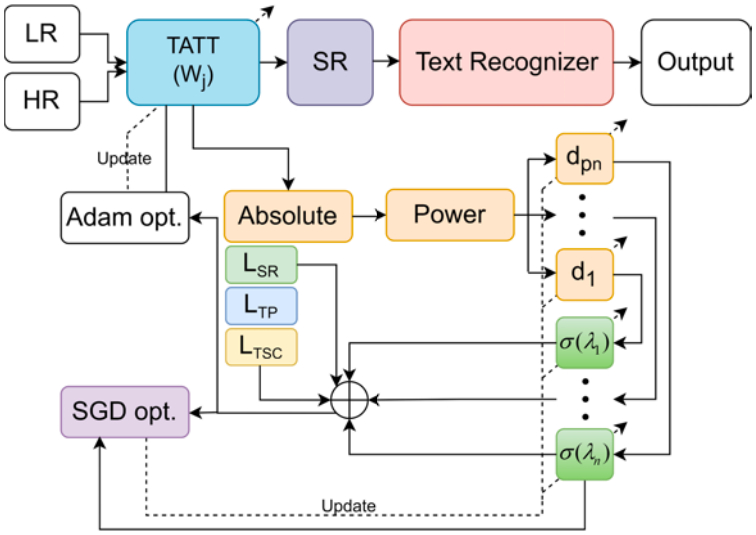

Figure 3.2 The architecture of our proposed method. The multiple parametric regularizations are added in the loss function. It was accumulated to be the new loss function.

MPR framework can work compatible with any type of model since it is applied in the loss function. Therefore, we aim to modify the loss function of the TATT model that has a CNN and Transformer as the backbone. We observe the loss function of TATT, it combines the three types of loss together, text prior, super-resolution, and text structure consistency loss as described in Section 2.3.2. Based on the idea of MPR, it is designed to adjust the value of the regularization parameters and the degree of regularization terms which it is freely adjusted and can be added as needed.

The method presented in this study is summarized in Figure 3.2. The input to the model consists of low-resolution (LR) and high-resolution (HR) text images, which are processed by the Text Attention-Aware Transformation (TATT) module. The reconstructed text image is then passed to the ASTER text recognizer, which predicts the text contained in the image.

The loss function used in the TATT module comprises three components: the super-resolution loss ($\mathcal{L}_{SR}$), the text prior loss ($\mathcal{L}_{TP}$), and the text structure consistency (TSC) loss ($\mathcal{L}_{TSC}$), as defined in Equation 2.8. These loss components contribute to the optimization of the model by guiding it to generate high-quality super-resolved text images.

In addition to the traditional loss function, the Multiple Parametric Regular-

ization (MPR) techniques is proposed to further enhance the TATT model. MPR introduces a set of regularization parameters, denoted as λ_n, to control the strength of the regularization terms. To ensure non-negative values for these parameters, a sigmoid function (σ) is applied.

During the weight optimization process, all the parametric parameters, including the regularization parameters, are updated using the Stochastic Gradient Descent (SGD) optimizer. On the other hand, the Adam optimizer is specifically used for optimizing the network gradients, ensuring efficient and effective training of the TATT model. By incorporating MPR and utilizing both SGD and Adam optimizers, the TATT model benefits from enhanced regularization capabilities and optimized network gradients, leading to improved performance in super-resolution and text recognition tasks.

$$\mathcal{L}_{TATT}^{MPR} = \mathcal{L}_{TATT} + \sum_{\forall n} \lambda_n \sum_{\forall j} \left\| W_j \right\|^{d_n}, \tag{3.2}$$

where $\mathcal{L}_{TATT}$ indicates the traditional loss function of TATT. λ_n is the parametric regularization parameters that passed to the sigmoid function σ, which becomes $\sigma(\lambda_n)$. The parametric degree represents d_n, and W_j is the weight parameters in the network.

3.2.2 CNN and Transformer-based STISR model with adding parametric weight and combining parametric weight and multiple parametric regularizations

In the context of the balancing parameters in text prior and text structure consistency loss, it is crucial to determine suitable values for the fixed parameters α and β in Equation 2.8. These parameters play a significant role in controlling the influence of each loss component and ultimately affect the performance of the super-resolution model.

Traditionally, determining the optimal values for these parameters involves conducting multiple experiments or referring to the state-of-the-art in the field. Re-

searchers often compare the performance of the super-resolution model with different parameter values to find the most suitable configuration. However, it is important to note that the ideal values may vary depending on the specific characteristics and requirements of the input data.

To address this challenge, we propose modifying these fixed parameters into parametric weights (PW), allowing them to be adjusted dynamically during the training process. The use of parametric weights enables the model to adaptively learn the importance of each loss component based on the characteristics of the input data and the specific training stage.

In Section 4.1, we conducted experiments on a CNN-based super-resolution method with the introduction of parametric weights. These experiments aimed to evaluate the impact of the parametric weight approach on text recognition accuracy and image quality. By incorporating the parametric weights, the model gained the ability to dynamically adjust the influence of different loss components, leading to improved performance and enhanced adaptability.

The introduction of parametric weights not only addresses the challenge of determining suitable fixed parameter values but also allows the model to learn the optimal balance between the text prior and text structure consistency loss components. This adaptability is particularly valuable in the context of scene text super-resolution, as it enables the model to effectively capture and reconstruct fine text details while maintaining the overall structure and coherence of the text region.

$$\mathcal{L}_{TATT}^{PW} = \mathcal{L}_{SR} + \widehat{\alpha}\mathcal{L}_{TP} + \widehat{\beta}\mathcal{L}_{TSC}, \tag{3.3}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the parametric weight parameters, $\hat{\alpha}$, and $\hat{\beta}$. It is multiplied with $\mathcal{L}_{TP}$ and $\mathcal{L}_{TSC}$, then it is summed up with $\mathcal{L}_{SR}$. It can be described in Equation 3.3.

Text recognition accuracy can be significantly improved by modifying balancing parameters. Nevertheless, the reconstructed text image shows some inconsistency. For example, in Figure 4.7(B), we compare the word *lles* in naturelles between

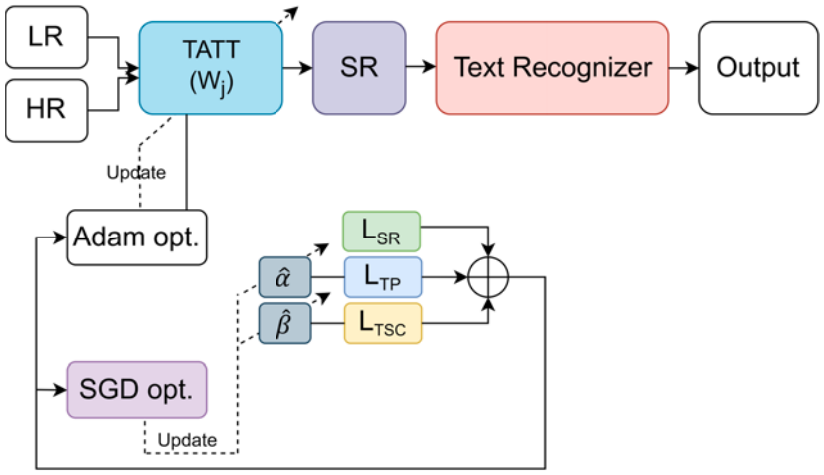

Figure 3.3 Architecture of our proposed method. The balancing parameters are modified to be the adaptive weight parameters or parametric weight parameters and multiplied with other losses.

adding MPR and PW which combining PW seems to give an inconsistency texture in this region, while the result of adding MPR is better. Further, the PW model tends to give blurrier results than MPR, such as Figures 4.7(a) and 4.7(e), but PW can increase the text recognition accuracy when we compare with baseline methods surprisingly. Here, we take advantage of both worlds, MPR and PW. We extend previously proposed methods by utilizing the MRP and PW together as Equation 3.4.

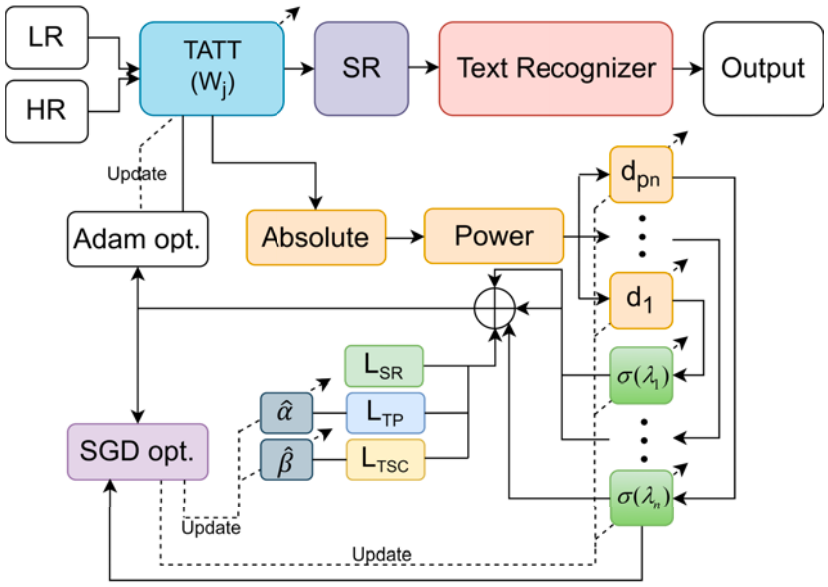

Figure 3.4 The overview of our proposed method. The parametric weight parameters are multiplied with text prior loss $\mathcal{L}_{TP}$ and text structure consistency $\mathcal{L}_{TSC}$. Then, it is calculated with multiple parametric regularizations.

$$\mathcal{L}_{TATT}^{PW\_MPR} = \mathcal{L}_{SR} + \widehat{\alpha}\mathcal{L}_{TP} + \widehat{\beta}\mathcal{L}_{TSC} + \sum_{\forall n} \lambda_n \sum_{\forall j} \left\| W_j \right\|^{d_n}, \tag{3.4}$$

The overview is shown in Figure 3.4. To address the challenges encountered in the previous method, we propose a modified approach where the traditional fixed balancing parameters are replaced with adaptive parameters called parametric weights (PW). These parametric weights are designed to be adjustable throughout the training process, allowing for greater flexibility in controlling the regularization parameters and the degree of regularization.

By incorporating parametric weights, we establish a direct relationship between the Multiple Parametric Regularization (MPR) terms and the adaptive regularization parameters. This adjustment enables the model to dynamically adapt the regularization strength based on the specific requirements of the training data and the current stage of training.

During each iteration of the training process, the Stochastic Gradient Descent (SGD) optimizer is employed to update all the parametric parameters, including the regularization parameters and the parametric weights. This adjustment ensures that the model continually learns and optimizes the regularization process according to the evolving characteristics of the data. On the other hand, the Adam optimizer is utilized to handle the gradient of the network, optimizing the network parameters and facilitating efficient and effective training.

Adaptive parameters and utilizing both SGD and Adam optimizers, our proposed approach provides increased flexibility in adjusting the regularization parameters and controlling the degree of regularization. This adaptability allows the model to effectively learn and generalize from the training data, leading to improved performance in terms of super-resolution and text recognition.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Experiment on CNN-based STISR model with parametric weights

Our proposed method is compared the performance with the baseline model. The parametric weight parameters are added to MSE and gradient profile loss. We investigate the impact of parametric weight on training time, text recognition accuracy, and visual perception. All evaluation is performed on the real-world STISR dataset Textzoom. The text recognization is performed by ASTER Shi et al. (2018).

In Table 4.1, the text recognition accuracy of the proposed method in medium and hard are increased at 56.91% and 40.21%, respectively while the accuracy of easy level is degraded down to 71.59%. However, our method can accelerate the model to converge faster. Since TSRN requires 500 epochs or around 54.39 hrs. but adding parametric weight can decrease the number of epochs down to 130 epochs. We visualize the SR images by comparing them with high resolution (HR), low resolution (LR), outputs of TSRN, and outputs of our proposed method in each subset as shown in Figures 4.1-4.3. The wrong prediction characters of text recognition are represented in red color. Compared with the baseline, adding parametric weight can boost the important detail of the text. However, the proposed method still has room for improvement. In Figure 4.4, it is some examples of wrong prediction because the SR images are not clear enough for text recognizer.

Table 4.1 SR text recognition performance of competing between TSRN and TSRN with parametric weight.

Model	Loss	Required epochs/times	Accuracy		
			Easy	Medium	Hard
TSRN	$L_2 + L_{GP}$	500	**75.1**	56.3	40.1
Ours	$(W_p^1)L_2 + (W_p^2)L_{GP}$	**130**	71.59	**56.91**	**40.21**

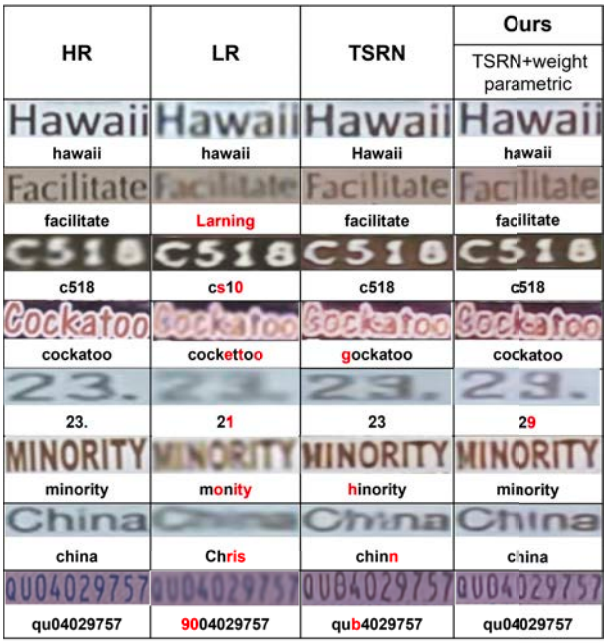

Figure 4.1 Visual comparison between TSRN and proposed method on Textzoom in *easy* subset.

Figure 4.2 Visual comparison between TSRN and proposed method on Textzoom in *medium* subset.

Figure 4.3 Visual comparison between TSRN and proposed method on Textzoom in *hard* subset.

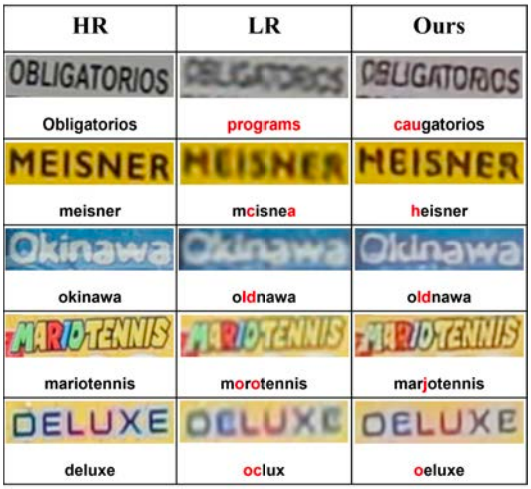

Figure 4.4 Example of wrong prediction in proposed method

Table 4.2 Average PSNR/SSIM/LPIPS comparison between the baseline and our proposed methods on the Textzoom dataset. Med. stands for the medium, which is one of the test sets.

Method	PSNR			SSIM			LPIPS		
	easy	med.	hard	easy	med.	hard	easy	med.	hard
TATT	24.15	**18.96**	**20.23**	0.892	0.681	0.766	0.106	0.205	0.176
MPR	19.82	17.61	18.15	0.774	0.610	0.662	0.132	0.202	0.204
PW	24.3	18.79	19.66	**0.897**	**0.694**	**0.768**	0.102	0.214	0.214
PW+MPR	**24.77**	18.87	20.09	**0.897**	0.692	**0.768**	**0.098**	**0.150**	**0.140**

4.2 Experiment on CNN and Transformer-based STISR method with multiple parametric regularization and parametric weights

In this section, we compare the performance of the proposed methods in Section 3.2.2 that modify the loss function with state-of-the-art on visual quality and text recognition accuracy. Image quality assessment (IQA) metrics are used to display the quantitative criteria of the generated images. To make a fair comparison, we take the pre-trained text recognizer as ASTER to predict the answer from the output.

We conduct experiments to demonstrate the effectiveness of the proposed methods in the loss function of TATT, adding MPR, modifying balancing parameters to be PW, and merging between MPR and PW. Based on the TATT model, we use the symbols ✗ and ✓ to represent our setting which ✗ means no adding and ✓ is adding in the loss function in Table 4.3 -4.5.

4.2.1 Quantitative measurement

To examine the performance of our three proposed methods, the well-known IQA metrics are applied, Peak-Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Typically, higher is better in PSNR and SSIM, while LPIPS prefers a lower score. Our proposed methods, MPR, PW, and PW+MPR, are compared with TATT on Textzoom dataset as in Table 4.2. TATT can achieve better performance on PSNR in medium and hard set, but PW+MPR gets a higher score in the easy set. While PW obtains further performance improvement in SSIM, PW+MPR can get the best LPIPS scores in every set. Tables 4.3-4.5 shows the IQA scores of each method compared with the

Table 4.3 Average PSNR comparison between the baseline and our proposed methods on the Textzoom dataset. Multiple parametric regularizations are MPR, and parametric weight is PW.

No. of MPR	Using PW	PSNR		
		easy	medium	hard
✗	✗	24.15	**18.96**	**20.23**
1	✗	18.53	17.50	17.80
2	✗	18.87	17.12	16.99
3	✗	19.07	17.24	17.77
4	✗	19.82	17.61	18.15
5	✗	18.79	17.26	17.26
✗	✓	24.30	18.79	19.66
1	✓	**24.61**	18.87	20.09
2	✓	23.81	18.57	20.09
3	✓	24.77	18.75	20.00
4	✓	22.94	18.20	19.21
5	✓	19.88	16.91	17.06

Table 4.4 Average SSIM comparison between the baseline and our proposed methods on the Textzoom dataset. Multiple parametric regularizations are MPR, and parametric weight is PW.

No. of MPR	Using PW	SSIM		
		easy	medium	hard
✗	✗	0.892	**0.681**	**0.766**
1	✗	0.732	0.592	0.630
2	✗	0.753	0.610	0.646
3	✗	0.745	0.593	0.642
4	✗	0.774	0.604	0.662
5	✗	0.729	0.582	0.623
✗	✓	0.897	0.694	0.768
1	✓	**0.897**	**0.692**	**0.768**
2	✓	**0.895**	0.685	**0.766**
3	✓	0.896	0.683	0.766
4	✓	0.887	0.680	0.759
5	✓	0.868	0.669	0.732

Table 4.5 Average LPIPS comparison between the baseline and our proposed methods on the Textzoom dataset. Multiple parametric regularizations are MPR, and parametric weight is PW.

No. of MPR	Using PW	LPIPS		
		easy	medium	hard
✗	✗	0.106	0.205	0.176
1	✗	0.207	0.227	0.259
2	✗	0.132	0.220	0.204
3	✗	0.171	0.218	0.217
4	✗	0.150	0.257	0.223
5	✗	0.163	0.223	0.247
✗	✓	0.102	0.214	0.214
1	✓	**0.098**	**0.150**	**0.140**
2	✓	0.115	0.184	0.172
3	✓	0.102	0.200	0.153
4	✓	0.115	0.214	0.161
5	✓	0.112	0.226	0.174

baseline in each test set.

Even though the IQA score is used to measure the quality of images from the reference images, it cannot express human perception. Since most IQA metrics calculate all of the pixels in the images, but scene text image focuses on the text region that can affect the text prediction process.

4.2.2 Qualitative measurement

To verify the superiority of our proposed framework, we use several popular SISR and state-of-the-art STISR models, including SRCNN Dong et al. (2016), SRResNet Ledig et al. (2017a), HAN Niu et al. (2020), TSRN Wang et al. (2020), PCAN Zhao et al. (2021), TG Chen et al. (2021c), TSRGAN Fang et al. (2021), TB-SRN Chen et al. (2021b), TPGSR Ma et al. (2021b), TRSRT Honda et al. (2022), and TATT Ma et al. (2022) along with our proposed methods. The loss of each method is shown by pre-trained text recognizers, ASTER, MORAN, and CRNN in Table 4.6, which the later STISR models tend to combine the different types of loss to make the model more robust.

The results of text recognition accuracy are presented in Tables 4.9-4.11. We found that our proposed methods can significantly improve text recognition accuracy.

Table 4.6 Loss function of the state-of-the-art STISR methods and our proposed methods. TRSRT-EDSR and TRSRT-BLSTM indicate that the backbones are Resblock of EDSR and Resblock with BLSTM. Each parameter is defined in Table 4.8.

Method	Loss
LR	-
Bicubic	-
SRCNN Dong et al. (2016)	$\mathcal{L}_2$
SRResNet Ledig et al. (2017a)	$\mathcal{L}_2 + \mathcal{L}_{TV} + \mathcal{L}_P$
HAN Niu et al. (2020)	$\mathcal{L}_2$
TSRN Wang et al. (2020)	$\mathcal{L}_2 + \mathcal{L}_{GP}$
PCAN Zhao et al. (2021)	$\mathcal{L}_2 + \mathcal{L}_{GP}$
TG Chen et al. (2021c)	$\mathcal{L}_2 + \mathcal{L}_{SFM}$
TSRGAN Fang et al. (2021)	$\mathcal{L}_{REC} + \lambda\mathcal{L}_G + \lambda\mathcal{L}_{WAV}$
TBSRN Chen et al. (2021b)	$\mathcal{L}_{PSM} + \lambda\mathcal{L}_{PA} + \lambda\mathcal{L}_C$
TPGSR Ma et al. (2021b)	$\mathcal{L}_2 + \mathcal{L}_{TP}$
TRSRT-EDSR/BLSTM Honda et al. (2022)	$\lambda\mathcal{L}_{SR} + \lambda\mathcal{L}_{REC} + \lambda\mathcal{L}_{Feat}$
TATT Ma et al. (2022)	$\mathcal{L}_2 + \mathcal{L}_{TP} + \mathcal{L}_{TSC}$
Ours (MPR)	$\mathcal{L}_2 + \mathcal{L}_{TP} + \mathcal{L}_{TSC} + MPR_n$
Ours (PW)	$\mathcal{L}_2 + \widehat{\alpha}\mathcal{L}_{TP} + \widehat{\beta}\mathcal{L}_{TSC}$
Ours (PW+MPR)	$\mathcal{L}_2 + \widehat{\alpha}\mathcal{L}_{TP} + \widehat{\beta}\mathcal{L}_{TSC} + MPR_n$
HR	-

Taking TATT as an example, comparing MPR with it, the accuracy on the easy set is equal and drops 0.3% on the medium set but increases 0.4% on the hard set. However, when we compare the accuracy between MPR and PW, it shows that the easy set is improved from 78.9% to 79.5% (increasing 0.6%), from 63.1% to 63.9% (increasing 0.8%) on medium set, and the hard set is dropped 0.1% 4 According to the output image from MRP method and the text recognition accuracy on PW method, we combine two methods to be PW+MPR. Not only generate more text detail when compared with other methods but also boost the performance in every test set, from 78.9% to 80.4% (increasing 1.5%) on easy, from 63.4% to 64.1% (increasing 0.7%), from 45.4% to 46.5% (increasing 1.1%) when compared with TATT. This indicates the effectiveness and advantages of our methods. In Figure 4.7, our proposed methods can reconstruct the detail better than the baseline, especially when parametric weight and multiple parametric regularizations are added.

Figure 4.5 compares the reconstructed images around the edge regions. Among our proposed methods, we found that PW+MPR5 reconstructs pixel-level charac-

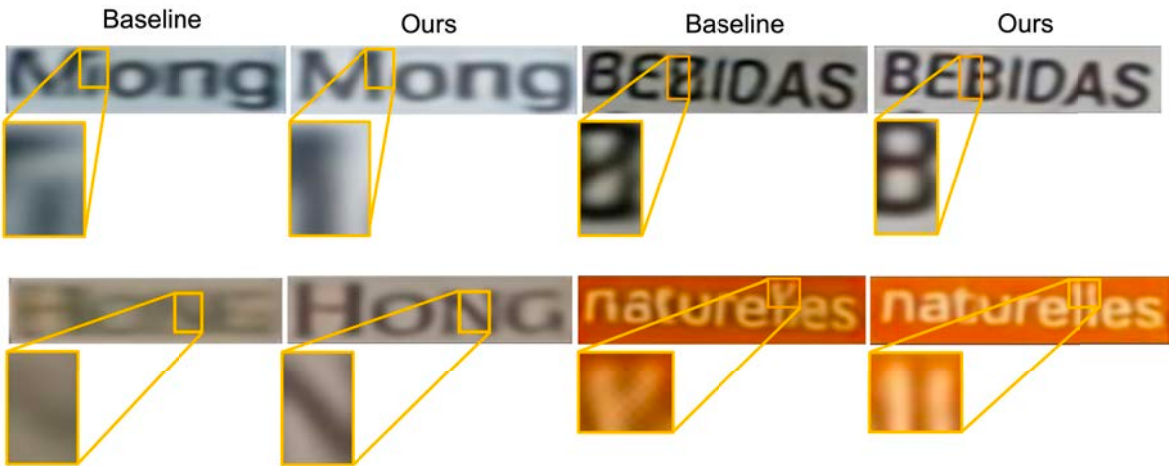

Figure 4.5 Comparison of the reconstructed images around the edge regions.

Table 4.7 Comparison of the performance of each method.

No. of MPR	Using PW.	Avg time (seconds)/epoch	Baseline ratio
✗	✗	61.2	1.00
1	✗	67.4	1.10
2	✗	78.2	1.28
3	✗	87.6	1.43
4	✗	102.6	1.68
5	✗	116.4	1.90
✗	✓	61.4	1.00
1	✓	79.4	1.30
2	✓	90.0	1.47
3	✓	102.8	1.68
4	✓	122.0	1.99
5	✓	129.0	2.11

ter outlines and color with more readable character strokes, resulting in correct text recognition.

According to training time, it is summarized in Table 4.7. The average time per epoch increases when the MPR terms are added, while the parametric weight method does not require more time to train when we compare it with the baseline ratio. The number of parameters is the same in every method, 15.94 M parameters which the two-parameter is added when we use the parametric wights because of alpha α and *beta*. Meanwhile, MPR requires two more parameters when one term of MPR is added because each MPR uses a regularization parameter and degree of the term. Here, we do the experiment on five MPR terms, which means the ten parameters are

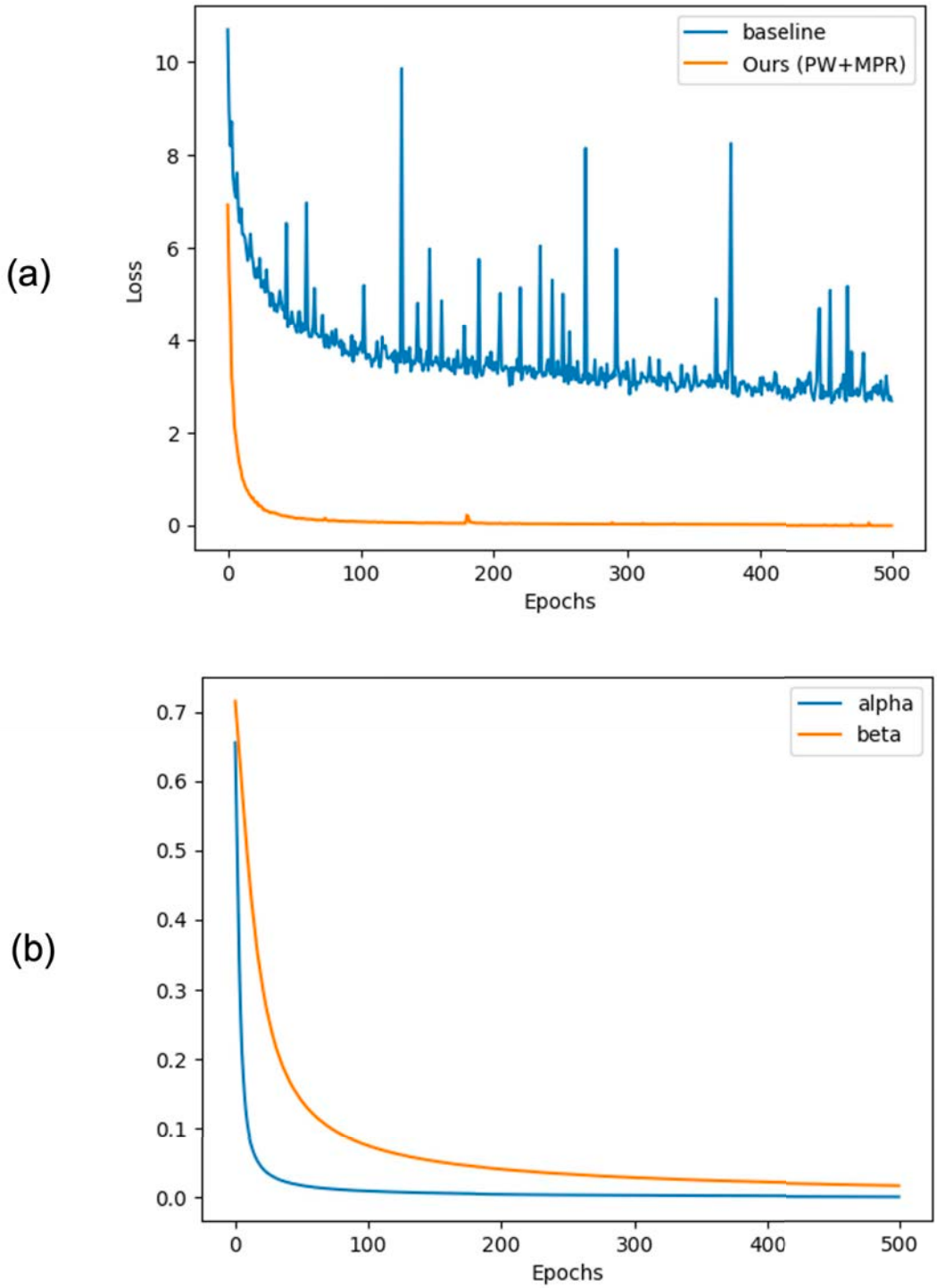

Figure 4.6 The demonstrated graph to show the ability of our proposed method (a) The performance comparison of loss function between baseline and our proposed method (PW+MPR) and (b) The value of α and β in parametric weight.

Table 4.8 Parameters description that displayed in Table 4.6.

Parameter	Discription
λ	Hyperparameter
$\mathcal{L}_{TV}$	Total Variation Loss
$\mathcal{L}_{P}$	Perceptual Loss
$\mathcal{L}_{GP}$	Gradient Profile Loss
$\mathcal{L}_{C}$	Content-aware Loss
$\mathcal{L}_{PA}$	Position-aware Loss
$\mathcal{L}_{TP}$	Text Prior Loss
$\mathcal{L}_{TSC}$	Text Structure Consistency Loss
$\mathcal{L}_{EG}$	Calculated by the squared term expectation of the difference between the function of the network output SR images and HR images
$\mathcal{L}_{G}$	Generator Loss
$\mathcal{L}_{WAV}$	Wavelet Loss
$\mathcal{L}_{SFM}$	Stroke-Focused Module Loss
$\mathcal{L}_{PSM}$	Pixel-wise Supervision Loss
$\mathcal{L}_{Feat}$	Feature-driven Loss

added to the computational graph.

The convergence speed enhancement using our proposed method is demonstrated in Figure 4.6. We can observe that our proposed method can achieve much smoother and quicker convergence than the baseline method. Figure 4.6(b) shows the time evolution of α and β. In the early stage, α and β are both large, and they are quickly reduced to small values. The total loss is decreased drastically by the assist of α and β. It is clearly shown that setting proper α and β lets the L_{SR} become smaller smoothly.

Table 4.9 Comparing the accuracy between the state-of-the-art STISR methods and our proposed methods in Textzoom test set by ASTER.

Method	**ASTER**		
	easy	**med**	**hard**
Bicubic	64.7	42.4	31.2
SRCNN Dong et al. (2016)	69.4	43.4	32.2
SRResNet Ledig et al. (2017a)	69.6	47.6	34.3
HAN Niu et al. (2020)	71.1	52.8	39
TSRN Wang et al. (2020)	75.1	56.3	40.1
TSRGAN Fang et al. (2021)	75.7	57.3	40.9
TBSRN Chen et al. (2021b)	75.7	59.9	41.6
TG Chen et al. (2021c)	77.9	60.2	42.4
PCAN Zhao et al. (2021)	77.5	60.7	43.1
TPGSR Ma et al. (2021b)	77	60.9	42.4
TRSRT-EDSR Honda et al. (2022)	72.1	55.6	39.5
TRSRT-BLSTM Honda et al. (2022)	74.8	59.5	42
TATT Ma et al. (2022)	78.9	63.4	45.4
Ours (MPR)	78.9 (n=3)	63.1 (n=1)	45.8 (n=3)
Ours (PW)	79.5	63.9	45.7
Ours (PW+MPR)	**80.4 (n=5)**	**64.1 (n=3)**	**46.5 (n=2)**
HR	94.2	87.5	76.2

Table 4.10 Comparing the accuracy between the state-of-the-art STISR methods and our proposed methods in Textzoom test set by MORAN.

Method	**MORAN**		
	easy	**med**	**hard**
Bicubic	60.6	37.9	30.8
SRCNN Dong et al. (2016)	63.2	39.0	30.2
SRResNet Ledig et al. (2017a)	60.7	42.9	32.6
HAN Niu et al. (2020)	67.4	48.5	35.4
TSRN Wang et al. (2020)	70.1	53.3	37.9
TSRGAN Fang et al. (2021)	72	54.6	39.3
TBSRN Chen et al. (2021b)	74.1	57.0	40.8
TG Chen et al. (2021c)	75.8	57.8	41.4
PCAN Zhao et al. (2021)	73.7	57.6	41.0
TPGSR Ma et al. (2021b)	72.2	57.8	41.3
TRSRT-EDSR Honda et al. (2022)	69.8	54.3	37.9
TRSRT-BLSTM Honda et al. (2022)	72.5	57.2	40.2
TATT Ma et al. (2022)	72.5	60.2	43.1
Ours (MPR)	**76.5 (n=2)**	60.9 (n=5)	44.0 (n=3)
Ours (PW)	75.5	60.3	43.5
Ours (PW+MPR)	75.2 (n=4)	**61.0 (n=4)**	**44.2 (n=4)**
HR	91.2	85.3	74.2

Figure 4.7 Comparing the reconstruction text region and text prediction between TATT and our proposed methods. The red characters below the image are the wrong prediction.

Table 4.11 Comparing the accuracy between the state-of-the-art STISR methods and our proposed methods in Textzoom test set by CRNN.

Method	**CRNN**		
	easy	**med**	**hard**
Bicubic	36.4	21.1	21.1
SRCNN Dong et al. (2016)	38.7	21.6	20.9
SRResNet Ledig et al. (2017a)	39.7	27.6	22.7
HAN Niu et al. (2020)	51.6	35.8	29
TSRN Wang et al. (2020)	52.5	38.2	31.4
TSRGAN Fang et al. (2021)	56.2	42.5	32.8
TBSRN Chen et al. (2021b)	59.6	47.1	35.3
TG Chen et al. (2021c)	61.2	47.6	35.5
PCAN Zhao et al. (2021)	59.6	45.4	34.8
TPGSR Ma et al. (2021b)	61.0	49.9	36.7
TRSRT-EDSR Honda et al. (2022)	51.7	39.6	31.2
TRSRT-BLSTM Honda et al. (2022)	57.3	45.6	35.4
TATT Ma et al. (2022)	62.6	53.4	39.8
Ours (MPR)	64.1 (n=4)	54.3 (n=1)	39.9 (n=2)
Ours (PW)	**64.2**	**54.6**	**40.9**
Ours (PW+MPR)	64.1 (n=4)	54.4 (n=3)	40.1 (n=3)
HR	76.4	75.1	64.6

CHAPTER 5
CONCLUSION AND FUTURE WORK

Applying scene text image super-resolution techniques has shown promising results in enhancing the quality of low-resolution scene text images, leading to increased text recognition accuracy. Using advanced image processing algorithms, scene text images can be upscaled to higher resolutions, resulting in sharper and clearer text that is easier for text recognition systems to interpret.

However, our preliminary study revealed that most STISR models tend to overfit, reconstructing the same or different characters and blurring the texture edges. While regularization can be used, its fixed value poses a limitation. To address this issue, we propose a novel parametric framework that incorporates a customized loss function in deep learning-based methods, including convolutional neural network (CNN), generative adversarial network (GAN), transformer-based SR, and STISR.

Our parametric framework applies parametric weight and multiple parametric regularizations to CNN, GAN, transformer-based SR, and STISR methods. We found that STISR methods with parametric weight improved image quality and text recognition accuracy, while GAN and transformer SR models failed to enhance the image. Although the CNN-based STISR method worked well with parametric weight, its architecture was limited in local feature extraction. Thus, we propose to use parametric weights and multiple parametric regularizations with CNN and transformer-based STISR models, which surprisingly enhanced image quality and accuracy in different ways. Adding multiple parametric regularizations improved edge characters while modifying parametric weight achieved the best text recognition accuracy.

To prove our hypothesis, we incorporated the two proposed methods and compared the performance with 12 state-of-the-art methods in IQA score, human per-

Figure 5.1 Samples of visualization on misprediction of text recognizer and human perception.

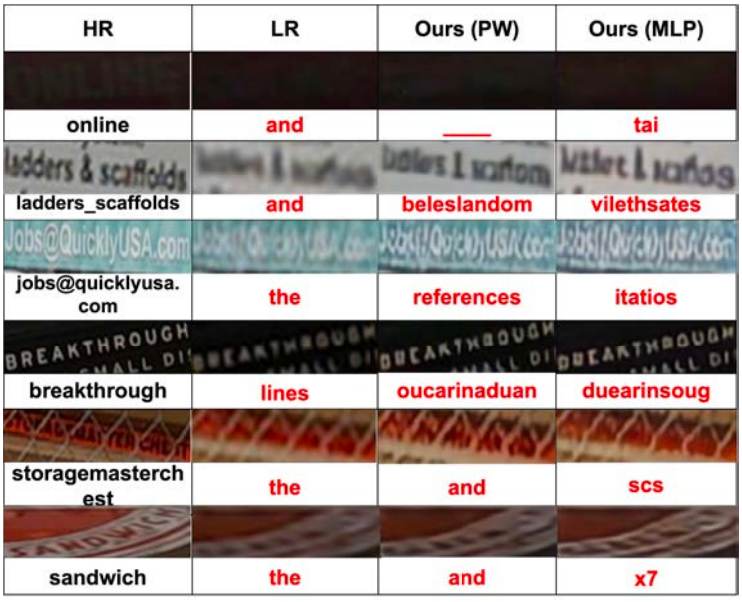

Figure 5.2 Example on the visualization result and text recognition on extremely dark, blurred, compressed, and unaligned text images.

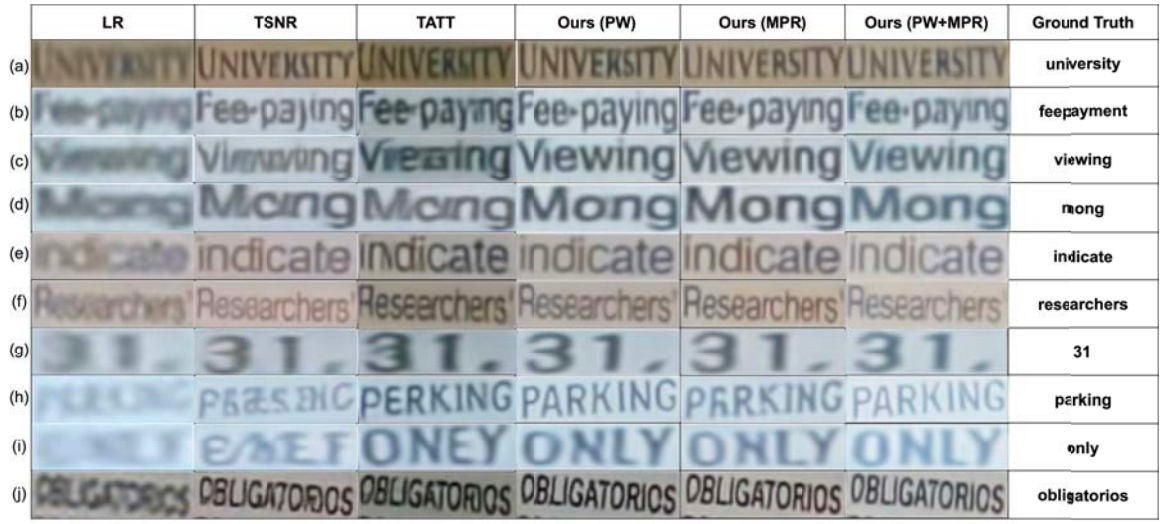

Figure 5.3 Samples of reconstructed images on real low-resolution images without the target images. It consists of low-resolution images as the input in the first column and the reconstructed results of baselines and our proposed methods. The last column shows the ground-truth texts.

ception, and text recognition accuracy. Our proposed STISR method effectively enhanced the quality of low-resolution scene text images, providing clear and easily recognizable text information crucial for human recognition and perception. However, the method had limitations in extremely dark text images, compressed or unaligned text, or text overlaid with noise as displayed in Figure 5.2. Additionally, errors could occur due to misprediction in text recognizer 5.1.

In Figure 5.3, our proposed method of STISR has been verified to enhance the quality of low-resolution scene text images effectively. The results of STISR show that the enhanced images provide clear and easily recognizable text information, which is crucial for human recognition and perception. By improving the quality of low-resolution images, STISR has the potential to improve the accuracy of text detection in real-world applications and provide valuable information.

In summary, our proposed parametric framework improves the performance of CNN, GAN, transformer-based SR, and STISR methods by optimizing the loss function parameters for each specific model, leading to higher accuracy and better visual quality in super-resolved images. The proposed STISR method enhances the quality of low-resolution scene text images, potentially improving the accuracy of text detection in real-world applications and providing valuable information.

For the potential directions in scene text image super-resolution, the current study mainly focuses on enhancing the quality of low-resolution scene text images with traditional training datasets. Future work could investigate the use of transfer learning techniques or the creation of more diverse and challenging datasets to improve the performance of STISR methods on real-world applications. The current study focuses on English text, STISR methods can potentially be applied to other languages or scripts that use different characters or fonts. This can be particularly useful in multilingual or cross-cultural settings, where accurate text recognition is crucial for effective communication. Moreover, it is important to consider the computational cost of STISR methods, particularly in real-time applications.

CHAPTER 6
APPENDIX

6.1 GAN-based SR model with multiple parametric regularization loss

According to the result of parametric weight described in Section 4.1 of our method in Section 3.2, the text recognizer still predicted the wrong answer in some samples. It is possible that some important textures are missing or not clear enough to detect. Therefore, we concentrate on improving the resolution in the generated scene text image. Intuitively, the model should have more complexity and capacity to extract the feature from LR input. From our knowledge, SR GAN-based model performs well in generating the detail of images in the natural image domain. Hence, we replace the GAN-based model to do the super-resolution. Many loss functions are combined such as pixel loss, adversarial loss, and feature loss. Multiple parametric regularizations Viriyavisuthisakul et al. (2022b) can give the impressed result in the natural image domain. We also apply it to the network as another loss function. The proposed method is illustrated in Figure 6.1. The SR images are fed to the text recognition and predict the text inside. The overall loss function can be expressed as Equation 6.4.

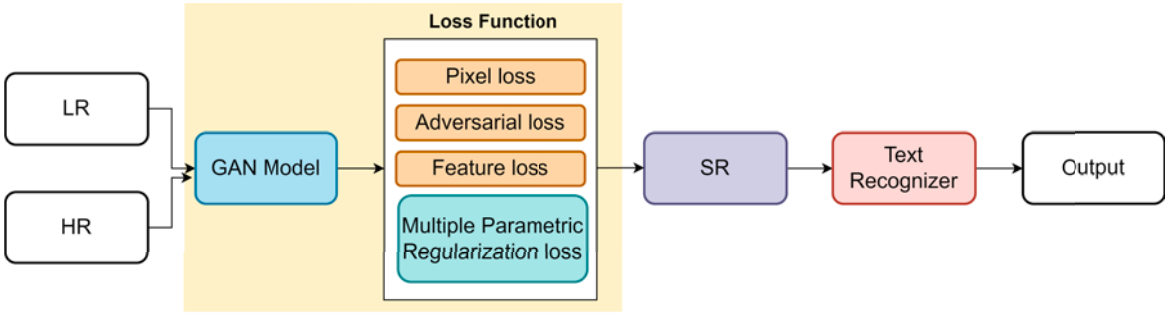

Figure 6.1 The proposed SR method with multiple parametric regularizations

A pixel loss is defined to use the $L1$-norm to calculate the distance between the ground truth and a reconstructed image as the Equation 6.1.

$$\mathcal{L}_{pixel} = \sum_{x=1}^{W} \sum_{y=1}^{H} \left\| I_{x,y}^{HR} - I_{x,y}^{SR} \right\|, \tag{6.1}$$

where $I_{x,y}^{HR}$ and $I_{x,y}^{SR}$ are a coordinate pixel of ground truth, and a generated $I_{x,y}^{SR} = G(I_{x,y}^{LR})$ image, reconstructed from the LR image $I_{x,y}^{LR}$ input by the generator. Moreover, the features of both images were also used to measure the difference by the VGG network as shown in Equation 6.2.

$$\mathcal{L}_{feature} = \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left\| \phi_{i,j}(I_{x,y}^{HR}) - \phi_{i,j}(I_{x,y}^{SR}) \right\|, \tag{6.2}$$

where $W_{i,j}$ and $H_{i,j}$ are the dimensions of the respected feature map ϕ within the VGG feature extractor network. Finally, the adversarial loss is applied to the relativistic discriminator to appraise the realistic probability of realness and fakeness in the generated images and HR images.

$$\mathcal{L}_{adv} = -\mathbb{E}_{I^{HR}}[\log(1 - D_{Ra}(I^{HR}, I^{SR}))] - \mathbb{E}_{I^{SR}}[\log(D_{Ra}(I^{HR}, I^{SR}))], \tag{6.3}$$

where D_{Ra} is relativistic discriminator. These models demonstrate the improvement of perceived quality and reconstructed accuracy of the generated image compared with the previous SR methods as it enables to produce more realistic and sharper details.

$$\mathcal{L}_{GAN\_MPR} = L_{pixel} + L_{ads} + L_{feature} + L_{mpr}, \tag{6.4}$$

where

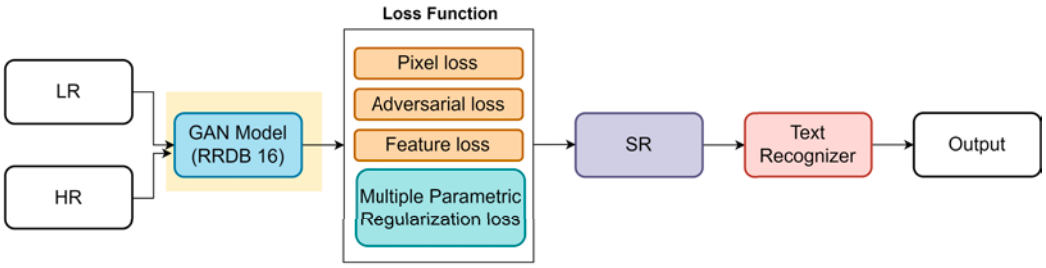

Figure 6.2 The proposed SR method with 16 RRDBs in the generator and multiple parametric regularization

- L_{pixel} is the pixel loss that calculates the distance between the corresponding pixel of HR and SR images.
- L_{ads} is the adversarial loss from relativistic discriminator
- $L_{feature}$ is the feature loss that calculates the distance between features of HR and SR images.
- L_{mpr} is the multiple parametric regularization method to balance the losses in the network.

6.1.1 GAN-based SR model with RRDB 16 and multiple parametric regularization loss

From the experiment of the proposed method in Section 6.3 of the proposed method in Section 6.1, the GAN-based model can restore the detail from the scene text image better the baseline when we compared by human perception. However, the text recognition accuracy is dropped. Since the GAN-based model treats the scene text image as the natural image, it tried to boost up the detail in every pixel as a result of the SR reconstruction module. It does not focus on the edge of the character. Therefore, it can come up with noise regarding scene text images.

In this proposed method, we aim to decrease the feature extraction process in the GAN model. Residual in Residual Dense Blocks (RRDBs) is the module that is applied for local feature extraction. The overview of the proposed method is displayed in Figure 6.2. Instead of adding the 23 RRDBs as the original of GAN-based

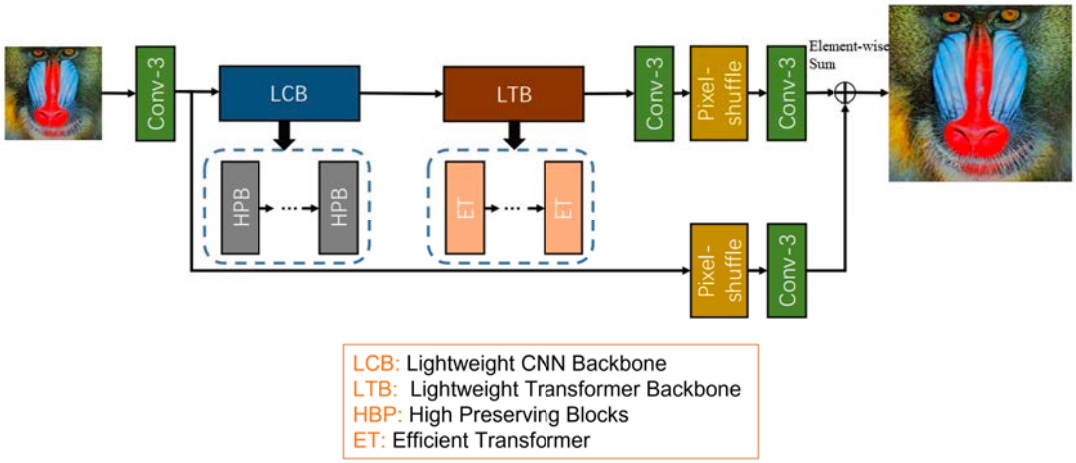

Figure 6.3 The architecture of ESRT model Lu et al. (2022)

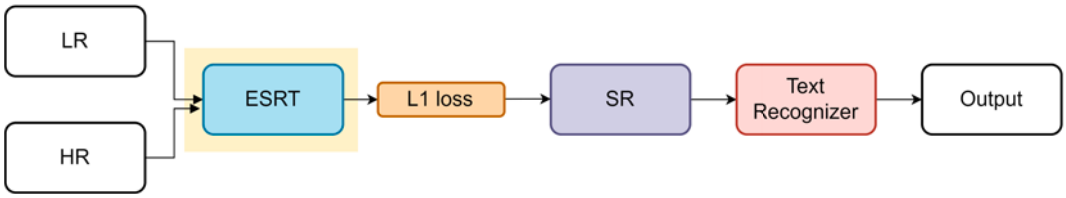

Figure 6.4 The proposed SR method with ESRT

SR model to the generator, it is modified to 16 RRDBs same as in SRGAN as in the yellow area. We still combined the pixel, adversarial, feature loss, and multiple parametric regularizations in the loss function following Equation 6.4. For other settings, we keep maintaining all datasets and parameters the same as the previous method.

6.2 Transformer-based SR model in scene text images recognition

Recently, the transformer performed an outstanding result in the Natural Language Processing (NLP) task. It is also explored in the application of Transformer in computer vision tasks Han et al. (2022). One of the well-known architectures in Vision Transformer (ViT) proposed for image recognition Dosovitskiy et al. (2020). It can be beaten with state-of-the-art CNN models. However, with the heavy computational cost and high GPU memory occupation of the vision transformer, the network cannot be designed too deep.

To overcome this problem, Efficient Transformer for Single Image Super-Resolution (ESRT) Lu et al. (2022) is designed for fast and accurate image super-

resolution images. ESRT is a hybrid method between transformer and CNN-based SR networks. ESRT composes of two backbones as in Figure 6.3, lightweight CNN backbone (LCB), and lightweight Transformer backbone (LTB). LCB is a lightweight SR network that dynamically resizes the feature map to obtain deep SR features using low computational resources. It allows the model to gain initial SR capability and extract the latent SR features in advance. An LCB is composed of series of High Preserving Blocks (HPBs). A characteristic of HPB is the reduction of shape and size of processing features. However, the results in relativity unnatural SR images due to the loss of image details. HPB solves this problem by creatively preserving high-frequency information while reducing the feature map size using the High-frequency Filtering module (HFM). Meanwhile, LTB consists of a series of Efficient Transformers (ET). A low-cost ET algorithm can capture long-term dependence between similar local regions within an image, In this section, we proposed to apply the transformer architecture in the SR process. Here, ESRT is employed as an SR module in pre-processing. The overview of the method is shown in Figure 6.4.

6.3 Experiment on GAN-based SR model with multiple parametric regularization loss

In this section, we do the experiment of the proposed method that is explained in Section 6.1. We focus on restoring the detail of the text in the text image by using a GAN-based SR model and multiple parametric regularizations. The model is trained in 300K iterations. From the experiment of the multiple parametric regularizations, they reported that the model could converge faster at 100K iterations. Therefore, we inspect the result on 100k and 300k iterations to compare the effectiveness of the model. We compare the result between HR images, LR images, the generated images from TSRN, the generated image from the proposed method at 100k, and 300k iterations as shown Figures 6.5-6.7.

The SR images generated from our proposed method are clearer than the baseline in every subset. For example, in Figure 6.6, our proposed method at 100k itera-

Figure 6.5 Visual comparison between TSRN and proposed method at 100k and 300k iterations on Textzoom in *easy* subset.

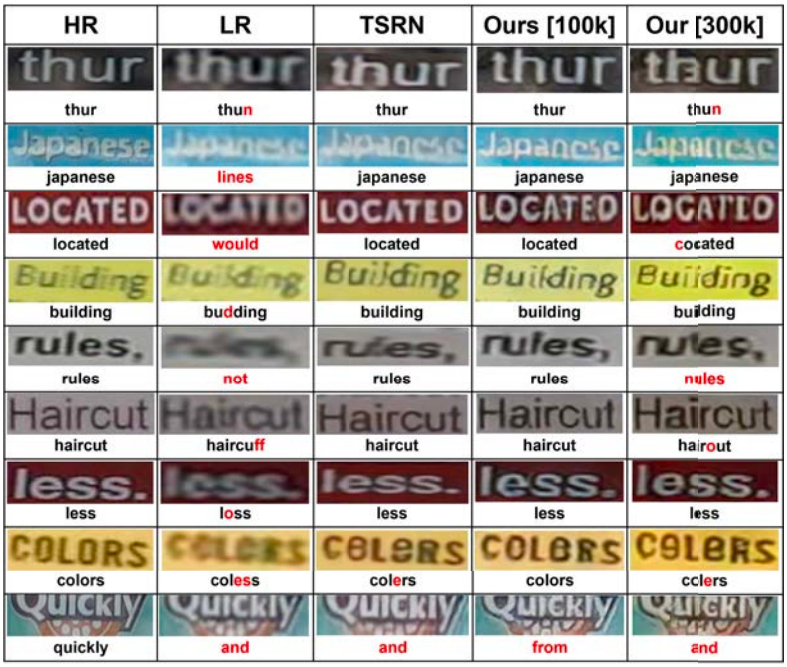

Figure 6.6 Visual comparison between TSRN and proposed method at 100k and 300k iterations on Textzoom in *medium* subset.

Figure 6.7 Visual comparison between TSRN and proposed method at 100k and 300k iterations on Textzoom in *hard* subset.

Table 6.1 SR text recognition performance of competing between TSRN and the proposed method at 100k and 300k iterations.

Model	Loss	Training epochs	Accuracy		
			Easy	Medium	Hard
TSRN	$L_2 + L_{GP}$	500	**75.1**	**56.3**	**40.1**
Ours	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	100	61.58	48.97	34.48
Ours	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	300	60.59	47.20	31.87

tions can generate the texture of the word *building* better than TSRN same as the word *rules*. However, the text recognition accuracy of our proposed method is dropped as displayed in Table 6.1.

6.3.1 Experiment on GAN model with RRDB 16 and multiple parametric regularization loss

From the proposed method in Section 6.1.1, we aim to decrease the feature extraction process by dropping the RRDBs from 16 to 23 blocks. The GAN model and multiple parametric regularizations are applied like the proposed method in Section 6.1.

We compare the efficiency of the model between the baseline, the proposed

Figure 6.8 Visual comparison between TSRN and proposed method in the proposed method in Section 6.1 (Ours_1) and proposed method in Section 6.1.1 (Ours_2) at 100k on Textzoom in *easy* subset.

method in Section 6.1, and this proposed method in both text recognition accuracy and quality of images. Figures 6.8-6.10, shows the visual comparison of HR, LR, baseline (TSRN), the proposed method on Section 6.1 (Ours_1), and the proposed method in this Section 6.1.1 (Ours_2). The wrong text prediction is displayed in red text.

We found that RRDB has affected directly with reconstruction process. Decreasing in the number of RRDB, it results in missing character details in generated image. For example, the word *Beef* in Figure 6.8, Ours_1 can generate the horizontal line of the *f* character in 100k training iterations while Ours_2 is unable and predicts the wrong answer to be *Beel*. Another example, it is the result in Figure 6.9, Ours_2 cannot generate the clear edge of text in the word *building* but it is *suilding*. For the text recognition accuracy, the performance is dropped around 1-2% after removing some RRDBs from the generator as in Table 6.2, but it can save a lot of computational time.

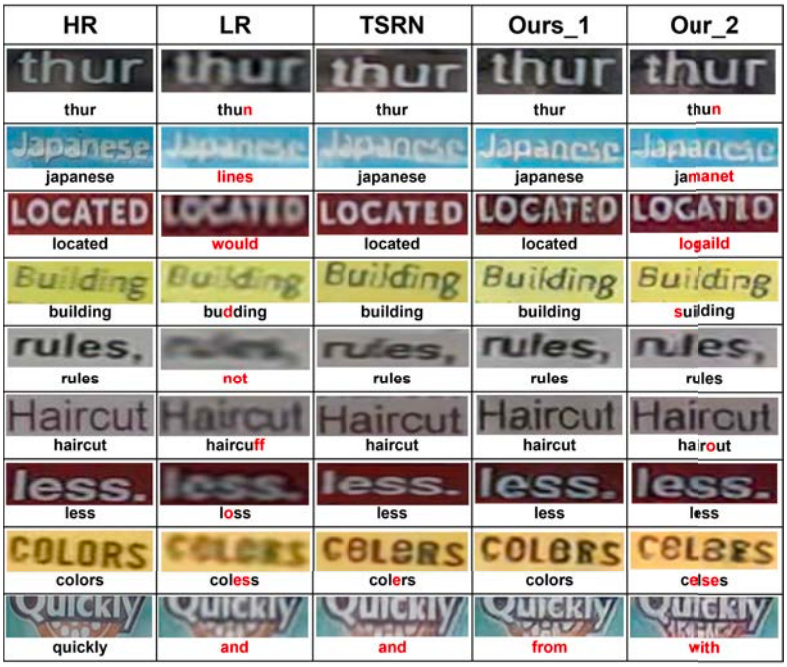

Figure 6.9 Visual comparison between TSRN and proposed method in the proposed method in Section 6.1 (Ours_1) and proposed method in Section 6.1.1 (Ours_2) at 100k on Textzoom in *medium* subset.

Figure 6.10 Visual comparison between TSRN and proposed method in the proposed method in Section 6.1 (Ours_1) and proposed method in Section 6.1.1 (Ours_2) at 100k on Textzoom in *hard* subset.

Table 6.2 SR text recognition performance of competing between TSRN and the proposed method in Section 6.1 (Ours_1) and the proposed method in Section 6.1.1 (Ours_2)

Model	Loss	Training epochs	Accuracy		
			Easy	Medium	Hard
TSRN	$L_2 + L_{GP}$	500	**75.1**	**56.3**	**40.1**
Ours_1 (RRDB23)	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	100	61.58	48.97	34.48
Ours_2 (RRDB16)	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	100	60.59	47.98	32.09

Figure 6.11 Visual comparison between TSRN and proposed method in ESRT on Textzoom in *easy* subset.

6.4 Experiment on Efficient Transformer for Single Image Super-Resolution (ESRT)

We investigate the result of the proposed method in Section 6.2. The visual quality of the generated images is illustrated in Figures 6.11-6.13. The ESRT is designed to support the SISR task using CNN and a transformer. However, we found that the transformer is a data greedy method. While Textzoom is insufficient for its. Therefore, the reconstructed images are unsuitable for scene text recognition and lead to the wrong prediction, as shown in Table 6.3.

Figure 6.12 Visual comparison between TSRN and proposed method in ESRT on Textzoom in *medium* subset.

Figure 6.13 Visual comparison between TSRN and proposed method in ESRT on Textzoom in *hard* subset.

Table 6.3 SR text recognition performance of competing between TSRN and the proposed method that applied in transformer-based SR method.

Model	Loss	Accuracy		
		Easy	Medium	Hard
TSRN	$L_2 + L_{GP}$	**75.1**	**56.3**	**40.1**
Ours	L_1	50.96	37.42	24.94

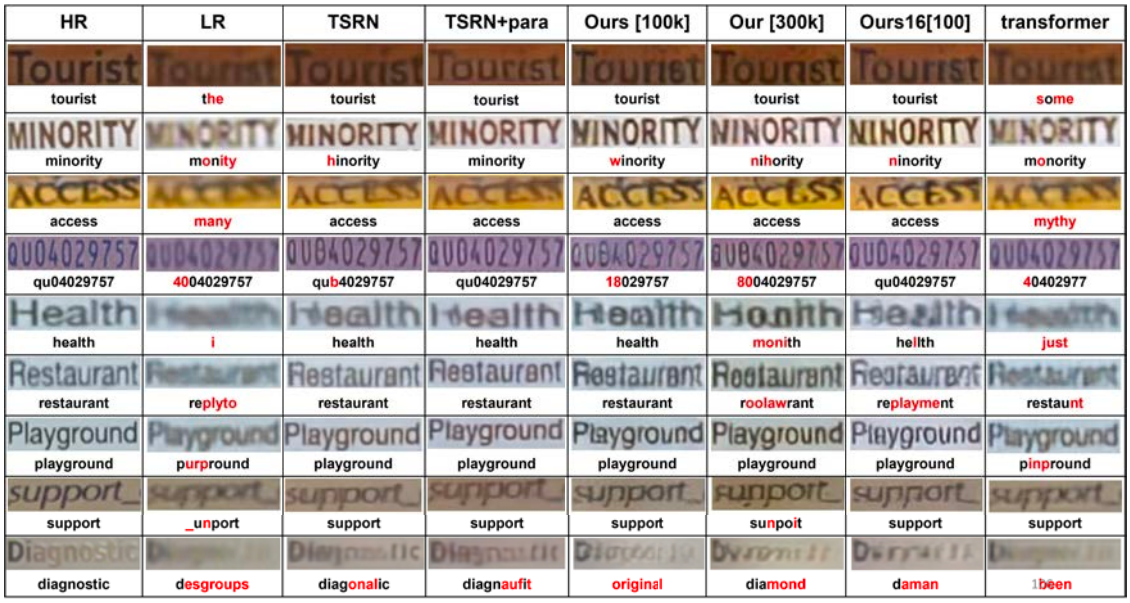

Figure 6.14 Visual comparison between TSRN and proposed methods on Textzoom in *easy* subset. TSRN+para indicates a CNN-based method with parametric weight. Ours[100K] and Ours[300K] are GAN-based SR methods with multiple parametric regularizations at 100K and 300K iterations, respectively. Ours16 represents the result from the GAN-based SR method with reduced RRDB to 16 blocks at 100K iterations. Transformer is the result of a Transformer-based SR model in scene text image recognition.

6.5 Summarizing

We summarize the performance of our four proposed method from Section 4.1-6.4 as follow.

- CNN-based SR model with parametric weights
- GAN-based SR model and multiple parametric regularization loss
- Transformer-based SR model in scene text image recognition

All proposed methods focus on improving image quality by using a super-resolution process and text recognition accuracy. We compare the result of the methods with HR, LR, TSNR, and the above three proposed methods on easy, medium, and hard on Textzoom dataset as shown in Figures 6.14-6.16. The red characters below the images are the wrong predictions.

From Table 6.4, we have compared the results of our proposed methods with the baseline model. All evaluation is performed on Textzoom. The text recogni-

Figure 6.15 Visual comparison between TSRN and proposed methods on Textzoom in *medium* subset. TSRN+para indicates a CNN-based method with parametric weight. Ours[100K] and Ours[300K] are GAN-based SR methods with multiple parametric regularizations at 100K and 300K iterations, respectively. Ours16 represents the result from the GAN-based SR method with reduced RRDB to 16 blocks at 100K iterations. Transformer is the result of a Transformer-based SR model in scene text image recognition.

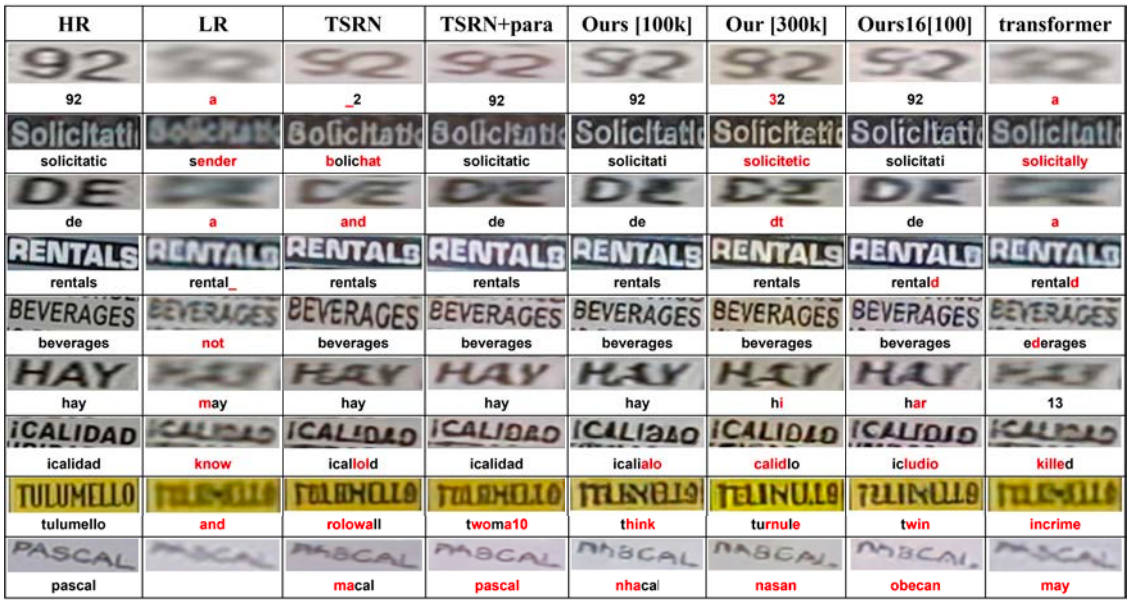

Figure 6.16 Visual comparison between TSRN and proposed methods on Textzoom in *hard* subset. TSRN+para indicates a CNN-based method with parametric weight. Ours[100K] and Ours[300K] are GAN-based SR methods with multiple parametric regularizations at 100K and 300K iterations, respectively. Ours16 represents the result from the GAN-based SR method with reduced RRDB to 16 blocks at 100K iterations. Transformer is the result of a Transformer-based SR model in scene text image recognition.

Table 6.4 The summary of the text recognition accuracy between baseline and proposed methods

Model	Architecture		Training epochs	Accuracy		
	Engine	Loss		Easy	Meduim	Hard
TSRN (Baseline)	TSRN	$L_2 + L_{GP}$	500	**75.1**	56.3	40.1
Ours (Method 1)	TSRN	$(W_p^1)L_2 + (W_p^2)L_{GP}$	130	71.59	**56.91**	**40.21**
Ours (Method 2)	nESRGAN+ (RRDB 23)	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	100	61.58	48.97	34.48
Ours (Method 2)	nESRGAN+ (RRDB 23)	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	300	60.59	47.20	31.87
Ours (Method 3)	nESRGAN+ (RRDB 16)	$L_{pixel} + L_{ads} + L_{feature} + L_{mpr}$	100	60.59	47.98	32.09
Ours (Method 4)	Efficient transformer	L_1	400	50.96	37.42	24.94

tion is performed by ASTER Shi et al. (2018). Our first proposed method archives the best performance in text prediction and visual image when we compared it with others. It archives text recognition accuracy in the medium and hard levels are increased up to 56.91%, and 40.21%, respectively while the accuracy of the easy level is degraded down to 71.59%. Moreover, TSRN requires 500 epochs for training but adding parametric weight can decrease the number of epochs down to 130 epochs. It can highlight the performance of multiple parametric regularizations to accelerate the model to converge faster.

REFERENCES

Agustsson, E. and Timofte, R. (2017). NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., and Yang, B. (2020). MedGAN: Medical Image Translation Using GANs. *Computerized Medical Imaging and Graphics*, 79:101684.

Atienza, R. (2021). Data augmentation for scene text recognition. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1561–1570.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character region awareness for text detection.

Barman, S. A., Welikala, R. A., Rudnicka, A. R., and Owen, C. G. (2019). Image quality assessment. *Computational Retinal Image Analysis*, pages 135–155.

Beaudry, N. J. and Renner, R. (2011). An intuitive proof of the data processing inequality. *Quantum Information and Computation*, 12(5-6):432–441.

Bendale, A. and Boult, T. E. (2016). Neural image assessment. In *European Conference on Computer Vision*, pages 191–207. Springer, Cham.

Bevilacqua, M., Roumy, A., Guillemot, C., and Morel, M. L. A. (2012). Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *Proceedings of the British Machine Vision Conference*, pages 1–10.

Cai, J., Zeng, H., Yong, H., Cao, Z., and Zhang, L. (2019). Toward Real-World Single Image Super-Resolution: A New Benchmark and a New Model. In *the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–6.

Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., and Zhu, C. (2022a). Real-world single image super-resolution: A brief review. *Information Fusion*, 79:124–145.

Chen, J., Li, B., and Xue, X. (2021a). Scene Text Telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12021–12030. IEEE Computer Society.

Chen, J., Li, B., and Xue, X. (2021b). Scene Text Telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12021–12030. IEEE Computer Society.

Chen, J., Yu, H., Ma, J., Li, B., and Xue, X. (2021c). Text Gestalt: Stroke-Aware Scene Text Image Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 285–293. Association for the Advancement of Artificial Intelligence (AAAI).

Chen, J., Yu, H., Ma, J., Li, B., and Xue, X. (2022b). Text Gestalt: Stroke-Aware Scene Text Image Super-Resolution. In *36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, pages 1–14.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., and Zhou, S. (2017). Focusing Attention: Towards Accurate Text Recognition in Natural Images. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 5086–5094. Institute of Electrical and Electronics Engineers Inc.

Dai, P., Li, Y., Zhang, H., Li, J., and Cao, X. (2022). Accurate scene text detection via

scale-aware data augmentation and shape similarity constraint. *IEEE Transactions on Multimedia*, 24:1883–1895.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

Fang, C., Zhu, Y., Liao, L., and Ling, X. (2021). TSRGAN: Real-world Text Image Super-Resolution Based on Adversarial Learning and Triplet attention. *Neurocomputing*, 455:88–96.

Field, D. (2002). Natural image statistics and ecological perception. *Journal of the Optical Society of America A*, 19(3):659–665.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. (2021). Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., and Tao, D. (2022). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Honda, K., Fujita, H., and Kurematsu, M. (2022). Improvement of Text Image Super-Resolution Benefiting Multi-task Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13343:275–286.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2016). Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*, 116(1):1–20.

Jaderberg, M., Vedaldi, A., and Zisserman, A. (2014). Deep features for text spotting. *Computer Vision – ECCV 2014*, 8692:512–528.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham. Springer International Publishing.

Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, pages 624–632.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017a). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 4681–4690.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017b). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 4681–4690.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. (2022). Transformer for Single Image Super-Resolution. In *CVPR workshop 2022*, pages 457–466.

Luo, C., Jin, L., and Sun, Z. (2019). MORAN: A Multi-Object Rectified Attention Network for scene text recognition. *Pattern Recognition*, 90:109–118.

Ma, C., Li, Y., and Liu, Y. (2018). Deep learning-based no-reference image quality assessment. *IEEE Access*, 6:68818–68828.

Ma, J., Guo, S., and Zhang, L. (2021a). Text Prior Guided Scene Text Image Super-resolution. pages 1–19.

Ma, J., Guo, S., and Zhang, L. (2021b). Text Prior Guided Scene Text Image Super-resolution. In *Computer Vision and Pattern Recognition*, pages 1–19.

Ma, J., Liang, Z., and Zhang, L. (2022). A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1–10.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *30 th International Conference on Machine Learning*, volume 28, pages 1–6, Atlanta, Georgia, USA.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423.

Mengyu, C., You, X., Jonas, M., Laura, L.-T., and Nils, T. (2020). Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation. *ACM Transactions on Graphics*, 39(4):75:1–75:13.

Müller, M. U., Ekhtiari, N., Almeida, R. M., and Rieke, C. (2020). Super-Resolution of Multispectral Satellite Images Using Convolutional Neural Networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5(1):33–40.

Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., and Shen, H. (2020). Single Image Super-Resolution via a Holistic Attention Network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12357:191–207.

Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang (2016). Reading scene text in deep convolutional sequences . In *the 13th AAAI Conference on Artificial Intelligence*, pages 3501–3508.

Pandey, R. K., Vignesh, K., Ramakrishnan, A. G., and B, C. (2018). Binary Document Image Super Resolution for Improved Readability and OCR Performance. In *Computer Vision and Pattern Recognition*.

Pejman, R., Tõnis, U., Sergio, E., and Gholamreza, A. (2016). Convolutional Neural Network Super Resolution for Face Recognition in Surveillance Monitoring. In *Articulated Motion and Deformable Objects*, pages 175–184, Cham. Springer International Publishing.

Peyrard, C., Baccouche, M., Mamalet, F., and Garcia, C. (2015). ICDAR2015 competition on Text Image Super-Resolution. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, volume 2015-November, pages 1201–1205. IEEE Computer Society.

Shao, J., Chen, L., and Wu, Y. (2021). SRWGANTV: Image Super-Resolution through Wasserstein Generative Adversarial Networks with Total Variational Regularization. *2021 IEEE 13th International Conference on Computer Research and Development, ICCRD 2021*, pages 21–26.

Sheikh, H. R. and Bovik, A. C. (2005). A universal image quality index. *Signal Processing Letters, IEEE*, 12(3):293–296.

Sheikh, H. R. and Bovik, A. C. (2006a). *Image and Video Quality Assessment*. Springer, New York.

Sheikh, H. R. and Bovik, A. C. (2006b). Multiwavelet-based visual information fidelity. *Signal Processing*, 86(6):1468–1478.

Shi, B., Bai, X., and Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2298–2304.

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2018). ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(c):1.

Soundararajan, R., Kankanahalli, M., and Bovik, A. (2020). A survey of image quality assessment: Traditional and learned approaches. *IEEE Access*, 8:141082–141112.

Viriyavisuthisakul, S., Kaothanthong, N., Sanguansat, P., Nguyen, M. L., and Haruechaiyasak, C. (2022a). Parametric regularization loss in super-resolution reconstruction. *Machine Vision and Applications*, 33(5):71.

Viriyavisuthisakul, S., Kaothanthong, N., Sanguansat, P., Nguyen, M. L., and Haruechaiyasak, C. (2022b). Parametric regularization loss in super-resolution reconstruction. *Machine Vision and Applications*, 33(5):1–21.

Viriyavisuthisakul, S., Kaothanthong, N., Sanguansat, P., Racharak, T., Le Nguyen, M., Haruechaiyasak, C., and Yamasaki, T. (2022c). A Regularization-Based Generative Adversarial Network for Single Image Super-Resolution. In *The Eleventh International Workshop on Image Media Quality and its Applications (IMQA)*, pages 43–49, Campus Plaza Kyoto, Kyoto, Japan.

Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., and Bai, X. (2020). Scene Text Image Super-Resolution in the Wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12355 LNCS:650–666.

Wang, W., Xie, E., Sun, P., Wang, W., Tian, L., Shen, C., and Luo, P. (2019). TextSR: Content-Aware Text Super-Resolution Guided by Recognition.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C. C., and Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV) Workshops*, pages 1–23.

Xu, J., Chae, Y., Stenger, B., and Datta, A. (2018). Residual Dense Network for Image Super Resolution. *Proceedings - International Conference on Image Processing, ICIP*, pages 71–75.

Zeyde, R., Elad, M., and Protter, M. (2012). On Single Image Scale-Up Using Sparse-Representations. In *Curves and Surfaces 2010*, volume 6920, pages 711–730. Springer, Berlin, Heidelberg.

Zhang, L. and Kankanahalli, M. (2016). No-reference image quality assessment: A literature review. *Signal Processing: Image Communication*, 47:34–46.

Zhang, L. and Kaveh, M. (2015). A comprehensive survey on image quality assessment: Evolution, performance evaluation, and applications. *IEEE Access*, 3:882–908.

Zhang, X., Chen, Q., Ng, R., and Koltun, V. (2019). Zoom to learn, learn to zoom. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 3757–3765. IEEE Computer Society.

Zhao, C., Feng, S., Zhao, B. N., Ding, Z., Wu, J., Shen, F., and Shen, H. T. (2021). Scene Text Image Super-Resolution via Parallelly Contextual Attention Network. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 2908–2917, New York, NY, USA. Association for Computing Machinery.

Zhong, S. and Zhou, S. (2021). Optimizing Generative Adversarial Networks for Image Super Resolution via Latent Space Regularization. *Image and Video Processing*, pages 1–11.

BIOGRAPHY

Name	Ms. Supatta Viriyavisuthisakul
Date of Birth	August 13, 1990
Education	2009: Bachelor of Science Information and Technology Thai-Nichi Institute of Technology 2010: Master of Science Information Technology Panyapiwat Institute of Management

Publications

- Journal

S. Viriyavisuthisakul, N. Kaothenthong, P. Sanguansat, Minh Le Nguyen, C.Haruechaiyasak. Parametric regularization loss in super-resolution reconstruction. Machine Vision and Applications 33, 71 (2022). https://doi.org/10.1007/s00138-022-01315-9

S. Viriyavisuthisakul, N. Kaothanthong, P. Sanguansat, T. Racharak, Minh Le Nguyen, C.Haruechaiyasak , and T. Yamasaki. Parametric Loss based Super-Resolution for Scene Text Recognition. Machine Vision and Applications, 34, 61 (2023). https://doi.org/10.1007/s00138-023-01416-z

- International Conference

S. Viriyavisuthisakul, N. Kaothanthong, P. Sanguansat, T. Racharak, Minh Le Nguyen, C.Haruechaiyasak , and T. Yamasaki (2023). Parametric Regularization Loss in Super-Resolution Reconstruction. Woman in Computer Vision (WiCV) in CVPR. Vancouver, Canada.

S. Viriyavisuthisakul, N. Kaothenthong, P. Sanguansat, T. Racharak, Minh Le Nguyen, C.Haruechaiyasak , and T. Yamasaki (2022).A Regularization-based Generative Adversarial Network for Single Image Super Resolution. *Proceedings of* 11^{th} *International Workshop on Image Media Quality and Its Applications (IMQA)* (pp. 43-49). Kyoto, Japan.

S. Viriyavisuthisakul, N. Kaothenthong, P. Sanguansat, C.Haruechaiyasak, Minh Le Nguyen, S.Sarampakhul, T. Chansumpao and D. Songsaeng (2020) . Evaluation of Window Parameters of Noncontrast Cranial CT Brain Images for Hyperacute and Acute Ischemic Stroke Classification with Deep Learning. *Proceedings of* 11^{th} *Annual International Conference on Industrial Engineering and Operations Management (IEOM)* (pp. 179-188). Singapore.

- Domestic Conference

S. Viriyavisuthisakul, P. Sanguansat, and T. Yamasaki (2023). Subjective evaluation of Super-Resolution Reconstructed by Trainable Regularization. The 26th Meeting on Image Reconstruction and Understanding. Hamamatsu, Japan.

S. Viriyavisuthisakul, P. Sanguansat, T. Racharak, Minh Le Nguyen, and T. Yamasaki (2023). Text Recognition in Low Resolution Images Using Trainable Regularization. The 37th Annual Conference of the Japanses Society for Artificial Intelligence (JSAI). Kumamoto, Japan.

S. Viriyavisuthisakul, N. Kaothenthong, P. Sanguansat, T. Racharak, Minh Le Nguyen, C.Haruechaiyasak , and T. Yamasaki (2022). Lasso , Ridge , and Elastic Net Regularized Generative Adversarial Networks for Single Image. *Proceedings of* 11^{th} *The 25th Meeting on Image Recognition and Understanding (MIRU)* (pp. 1-4). Kyoto, Japan.

S. Viriyavisuthisakul, P. Sanguansat, T. Racharak, Minh Le Nguyen, and T. Yamasaki (2022). Text Scene Image Super-Resolution with Parametric Weight Loss. *Proceeding of the* 25^{th} *Information-Based Induction Sciences Workshop (IBIS)*. Tsukuba, Japan.

S. Viriyavisuthisakul, P. Sanguansat, T. Racharak, Minh Le Nguyen, and T. Yamasaki (2022). Text Scene Image Super-Resolution with Parametric Weight Loss. *Proceeding of the* 37^{th} *Picture Coding Symposium of Japan and the* 27^{th} *Image Media Processing Symposium (PCSJ/IMPS)*. Gotemba, Japan.

S. Viriyavisuthisakul, P. Sanguansat, T. Racharak, Minh Le Nguyen, and T. Yamasaki (2023). Parametric Weight Method for Scene Text Image Super-Resolution. *The Institute of Electronics Information and Communication Engineers.* (pp. 165-168), Hokkaido, Japan.