information

Article

Joint Subtitle Extraction and Frame Inpainting for Videos with Burned-In Subtitles

Haoran Xu [1], Yanbai He [2], Xinya Li [3], Xiaoying Hu [4], Chuanyan Hao [3,*] and Bo Jiang [3,*]

1 School of Electronic and Optical Engineering & School of Microelectronics, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; haoranxu2000@gmail.com
2 School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; heyanbai1999@gmail.com
3 School of Educational Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; lxya960616@gmail.com
4 School of Computer Engineering, Tongda College of Nanjing University of Posts and Telecommunications, Yangzhou 225127, China; jiayic259@gmail.com
* Correspondence: hcy@njupt.edu.cn (C.H.); jiangbo@njupt.edu.cn (B.J.)

Abstract: Subtitles are crucial for video content understanding. However, a large amount of videos have only burned-in, hardcoded subtitles that prevent video re-editing, translation, etc. In this paper, we construct a deep-learning-based system for the inverse conversion of a burned-in subtitle video to a subtitle file and an inpainted video, by coupling three deep neural networks (CTPN, CRNN, and EdgeConnect). We evaluated the performance of the proposed method and found that the deep learning method achieved high-precision separation of the subtitles and video frames and significantly improved the video inpainting results compared to the existing methods. This research fills a gap in the application of deep learning to burned-in subtitle video reconstruction and is expected to be widely applied in the reconstruction and re-editing of videos with subtitles, advertisements, logos, and other occlusions.

Keywords: subtitle extraction; burned-in subtitles; image inpainting; text region detection; text recognition

Citation: Xu, H.; He, Y.; Li, X.; Hu, X.; Hao, C.; Jiang, B. Joint Subtitle Extraction and Frame Inpainting for Videos with Burned-In Subtitles. *Information* **2021**, *12*, 233. https://doi.org/10.3390/info12060233

Academic Editors: Zhaoqing Pan and Yuan Tian

Received: 7 May 2021
Accepted: 27 May 2021
Published: 29 May 2021

1. Introduction

As an important clue to the semantics of a video, subtitles use text to emphasize, supplement, or explain the non-visual content. As video becomes a mainstream medium for information interaction, subtitles play an increasingly important role as they enrich the on-screen information, e.g., subtitles may imply commentaries or thoughts from the creator. In addition, subtitles effectively compensate for simultaneous sound and enhance the understanding of the video for viewers with hearing impairments.

For more convenient transmission, subtitles exist mainly in the form of burned-in video frames, especially in most short and old videos. However, the language-specific burned-in subtitles pose great challenges for the re-editing and communication of the video between different languages [1], e.g., the translation of the video. Hence, subtitle extraction has been gaining attention, and some techniques have emerged for the automatic recognition of subtitles to facilitate the understanding and transcription of videos [2,3]. On the other hand, a video is seriously damaged after the extraction and removal of the burned-in subtitles, while an intact, subtitle-free video is desired, e.g., for re-adding the translated subtitles. Hence, the inpainting of the subtitle-removed frames is of great value for the reuse of the video.

The reconstruction of the burned-in subtitle video is realized by the combined subtitle removal and video restoration, which can be generally divided into two stages: text detection and frame inpainting. Existing video reconstruction techniques are based on traditional text detection and texture reconstruction approaches and have achieved some

success. However, there are still problems. Previous video reconstruction methods employed a traditional text detection pipeline [4–8], which consists of a series of steps, such as stroke filtering, positioning, segmentation, and verification.

The performance of the methods heavily relies on character detection, while the complex steps result in the propagation of errors and, hence, poor robustness and reliability [9]. Few works could generate the subtitle files directly. Furthermore, traditional frame inpainting approaches are generally diffusion-based or patch-based [10–14]. However, both the diffusion method based on differential operators and the patching method based on similar source image filling do a poor job of inpainting heavily damaged and complex details [15].

Recently, deep learning methods have achieved remarkable success in text recognition and image inpainting [2,3,9,15,16]. For example, Yan et al. [3] used a residual neural network for subtitle recognition, and Nazeri et al. [15] used a generative adversarial model for image restoration, and both showed excellent performance. Hence, deep learning methods open up wide prospects and provide powerful tools for the reconstruction of burned-in subtitle videos.

However, video reconstruction is a complex task that demands several deep modules. Thus, how to realize the seamless collaboration among the modules becomes a vital issue. For now, there is still a lack of an effective method that employs deep learning approaches to solve the burned-in subtitle video reconstruction challenge.

In this paper, we propose a novel pipeline for burned-in subtitle video reconstruction, based on deep learning. The pipeline unites subtitle extraction and frame inpainting and consists of three stages: (1) text detection; (2) text recognition; and (3) frame inpainting, and is implemented by three state-of-the-art deep neural networks (CTPN [9], CRNN [16], and EdgeConnect [15], respectively). An intermediate-process as well as a post-process are designed to implement the coupling of the models and the transformation of the results. Our contributions are four-fold:

- The inverse conversion of the burned-in subtitle video to an independent subtitle file and subtitle-free video.
- A novel framework for burned-in subtitle video reconstruction based on deep learning.
- The first application of the state-of-the-art deep learning techniques for burned-in subtitle video reconstruction with significantly enhanced subtitle extraction and frame inpainting.
- A general pipeline can be applied in the reconstruction and re-editing of videos with subtitles, advertisements, logos, and other occlusions.

The rest of the paper is structured as follows. Section 2 introduces the related work. Section 3 describes the framework and methodology for burned-in subtitle video reconstruction in detail. Section 4 presents and discusses the experimental results. Section 5 concludes our work and looks forward to future work.

2. Related Work

Over the past decade, a few works have addressed the challenge of burned-in subtitle video reconstruction. In 2010, Favorskaya et al. [4] first proposed a hybrid method based on contour and color information from sequential frames for text detection, and reconstructed the texture by statistical analysis in the time-space domain. Then, a priority-based matching algorithm was proposed by Khodadadi et al. [5] for reconstruction in areas with texture variation.

Subsequently, Favorskaya et al. [17] proposed a neural network based on time-space parameters for inpainting small-area damage of videos. In 2016, Vuong et al. [18] proposed a reconstruction system capable of detecting and extracting burned-in subtitles in the form of text, avoiding the waste of the original subtitles.

Previous burned-in subtitle video reconstruction methods were based on traditional text detection and texture reconstruction, which still have many problems despite some success, e.g., poor robustness due to the complex text detection pipeline (see Section 2.1 for details), poor generality due to the lexicon-based text recognition (see Section 2.2 for details),

and the loss of high-frequency information for image restoration (see Section 2.3 for details). We summarize the pipeline of burned-in subtitle video reconstruction into three subtasks: (1) text detection, (2) text recognition, and (3) frame inpainting. The following introduces the related work in each of these three subtasks.

2.1. Text Detection

Previously, there were two common approaches for text detection in videos or images with complex backgrounds. The primitive methods are based on low-level properties of the frame such as the contour, color, or gradient, including the gradient method, stroke filtering, color threshold segmentation, etc. [4–8]. Text detection is implemented through a series of filtering components, which leads to the transfer and accumulation of errors, resulting in low accuracy and robustness, especially when dealing with complex backgrounds.

With the development of CNN, character-based text detection methods emerged [19–22], which detect candidate characters by densely moving a multi-scale window through an image. The content in the window is judged by a pre-trained classifier. However, dense window sliding imposes a huge computational overhead, which severely limits the detection speed. In addition, precise text line positioning is difficult for the above methods. The Connectionist Text Proposal Network (CTPN) [9] is a mature text detection framework that combines CNN and Long Short-Term Memory (LSTM) deep networks to greatly improve the localization accuracy through a vertical anchor mechanism, while overcoming the inefficiency of sliding window methods.

2.2. Text Recognition

Traditional text recognition is based on character recognition and word recognition. The primitive approaches crop and detect individual characters from a word image by sliding a window, and then recombine all characters into a complete word [23,24]. These approaches require a powerful character detector and strongly rely on a fixed lexicon to synthesize words. Subsequently, word-based approaches emerged [25], which treat text recognition as a word image classification task, assigning a category label to each word.

Despite the impressive results achieved by these methods, they require an ultra-multi-classification model, are seriously confined by the number of classes, and have poor generalizability. CNN and RNN are important branches of the deep neural network family, specializing in image feature extraction and sequence analysis, respectively [26–28]. Shi et al. proposed a novel network called the Convolutional Recurrent Neural Network (CRNN) [16] that integrates CNN and RNN into the text recognition task, to solve the problems that exist in traditional methods. Compared with previous text recognition systems, CRNN is end-to-end trainable, able to handle sequences of arbitrary length, and not limited by any predefined lexicon. It is also an efficient but small model that is well-suited to real-life scenes.

2.3. Image Inpainting

Previous video frame restoration techniques can be divided into two perspectives: spatial and temporal domains, and three basic approaches: overlaying (as a temporal algorithm), diffusing, and patching (as spatial algorithms). The overlaying methods cover the missing texture region on the current frame by the real texture fragment of the previous or next frame without texture smoothing and compositing [4]. It is difficult to solve the micro-displacement or out-of-tune state of the texture fragments on the image. The diffusion methods propagate local background information to the missing regions [10–12].

However, such methods do not take full advantage of the global information and, thus, cannot recover meaningful structures in the missing regions and poorly handle a large missing region. Meanwhile, the diffusion methods require a significant time overhead to reach appreciable inpainting effects, which is unacceptable for the inpainting of videos. With the application of deep learning to image inpainting, the patch-based

methods have emerged [13,14], which implement inpainting by copying similar regions from the image set.

Such methods strongly rely on the image set and, thus, are suitable for highly patterned scenes but have difficulty in inpainting unique patterns. Recently, generative adversarial networks (GANs) have achieved impressive performances in inpainting [15,29–31]. EdgeConnect [15] is a new GAN-based inpainting method, inspired by the creative idea of "lines first, color next", achieving coherence in the inpainting content and refinement of details by global edge-connecting with high time efficiency.

3. Method

As shown in Figure 1 (model diagram) and Figure 2 (processing flow), the entire pipeline of the proposed method contains three main modules plus an intermediate process and a post process. Given a video frame with burned-in subtitles as input, a text detection network is first adopted to precisely locate the subtitle text region. By taking the text region bounding box, an intermediate process is conducted to separate the processed video frame into two parts.

The cropped subtitle image is fed into a text recognition network to recognize the subtitle character contents, and the video frame together with the subtitle character mask are sent to an image inpainting network to fill up the missing pixels inside the region of subtitle characters. After the subtitle recognition and frame inpainting, a post-process is required to construct the subtitle text file and assemble the inpainted frames into a video file. The following subsections depict the technical details of each module.

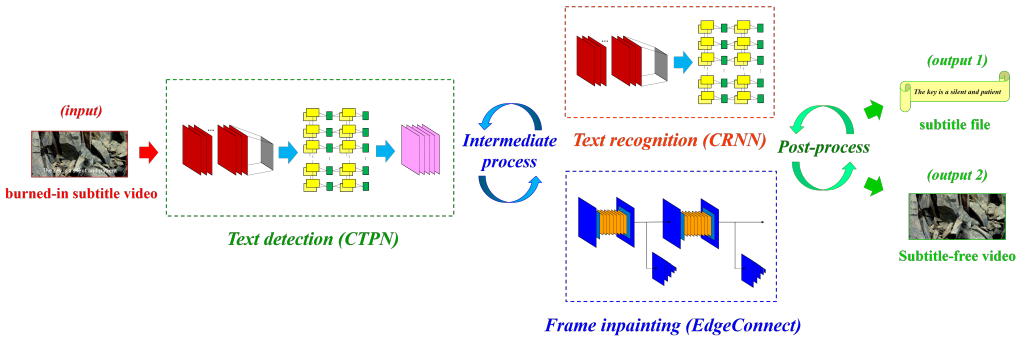

Figure 1. Model diagram of the entire deep-learning-based system for burned-in subtitle video reconstruction, which consists of three joint deep neural networks: CTPN (for text detection), CRNN (for text recognition), and EdgeConnect (for video inpainting).

3.1. Text Detection

The subtitle text region detection module employs the CTPN method [9], which utilizes a seamless combination of CNN and RNN to achieve the high-accuracy detection of horizontal text in complex scenes. CTPN enables the input video frame of an arbitrary size ($H \times W \times 3$) for text detection. At the beginning of detection, a CNN based on VGG-16 is first adopted to extract the deep features of input raw images. The feature map of layer conv5 is obtained as the last layer of VGG-16, with the total stride and receptive field fixed as 16 and 228 pixels, respectively.

Then, a 3×3 sliding window with a step size of 1 is performed on this feature map to obtain 256-D feature vectors. A RNN based on the bi-directional LSTMs (BiLSTMs) is used to learn feature sequences and predict the position of text according to the preceding and following texts. The feature vectors corresponding to all windows are fed into a BiLSTM network, consisting of two 128-D forward and inverse LSTMs. The output of the BiLSTM network is then fed into three regression layers through a 512-D fully connected layer.

Among the three regression layers, the 2 k vertical coordinates and k side-refinement are obtained to locate the k proposals (fixed-width, slender rectangular boxes), while 2 k scores are obtained to determine whether the proposal is text. Finally, every two adjacent

proposals with scores > 0.7 are merged to obtain the bounding box of the subtitle text region. The network configuration summary of CTPN is detailed in Table A1.

3.2. Text Recognition

The recognition module used to recognize the characters in the video subtitles is mainly based on CRNN [16]. The architecture of CRNN consists of three components from the bottom to top, including the convolutional layer (CNN, for extracting features), the recurrent layer (RNN, for predicting distributions) and the transcription layer (CTC, for synthesizing sequences), to achieve accurate recognition of indefinitely long text sequences. At the beginning of recognition, The gray-scale image of the subtitle text region is sent to the CRNN, and the image is deflated to $32 \times W$ and then fed into a CNN based on the VGG network.

After a series of convolution, pooling, and batch normalization operations on the image, the CNN extracts a $512 \times 1 \times 40$ feature map and converts it into 40×512-D feature vectors for the prediction in recurrent layers. On top of the convolutional layer, a BiLSTM-based RNN is built, which uses a 256-D BiLSTM network to learn feature vectors and predict the probability distribution of the labels.

At the end of the RNN, the propagated sequence is concatenated again into a map and fed back to the CNN, implementing a custom network layer called "Map-to-Sequence", which serves as a bridge between the CNN and RNN. On top of the recurrent layer, the transcription layer converts the label probability distribution from the RNN into an indefinitely long text sequence by de-duplication and integration, as the final output result. The network configuration summary of CRNN is detailed in Table A2.

3.3. Frame Inpainting

The inpainting of subtitle-removed frames is based on an adversarial edge learning image inpainting network named EdgeConnect [15]. EdgeConnect consists of two GAN cascades, including an edge generator and an image completion network, to generate hallucinated edges and inpaint the missing pixels by edge-guiding, via adversarial learning. Each GAN follows the adversarial model, consisting of a generator and discriminator.

For the GAN of EdgeConnect, the generator consists of an encoder, eight residual blocks, and a decoder, and the discriminator consists of five convolution layers. In the generator of the first-stage GAN (edge generator), the gray-scale map of the subtitle-removed frame and subtitle mask are used as pre-conditions to predict the edge map of the masked area. The input image is down-sampled twice by the encoder and fed into the residual blocks for dilated convolutions with a dilation factor of 2, resulting in a receptive field of 205 at the final residual layer. The final map is up-sampled twice by the decoder and resized to its original scale.

Similar to the first stage, the generator of the second-stage GAN (image completion network) takes the RGB map of the subtitle-removed frame and the predicted edge map as pre-conditions to complete the image by combining the background area of the ground truth edges with the predicted edges in the damaged area. For discriminators, a 70×70 PatchGAN architecture is used, which determines whether or not overlapping image patches of size 70×70 are real.

The discriminator of the edge generator discriminates whether the generated edge map is real with a joint loss as the training goal, including an adversarial loss and feature-matching loss. The discriminator of the image completion network discriminates whether the inpainted color map is real, with a joint loss as the training goal, including an $\mathcal{L}1$ loss, adversarial loss, perceptual loss, and style loss. The network configuration summary of EdgeConnect is detailed in Table A3.

3.4. Intermediate-Process

As shown in Figure 2, an intermediate-process stage was designed to connect the text region detection stage and the following text recognition and frame inapinting stages. This

process takes the bounding box (bbox) of the subtitle text area obtained from CTPN and the original video frame as input, and consists of three main steps. In the first step, the original frame is copied and cropped by the bbox. The cropped subtitle text image is fed into the CRNN, achieving end-to-end recognition of the subtitle by CRNN.

In the second step, the contour of the original frame in the bbox is extracted and expanded. The subtitle mask is obtained, which ensures the complete removal of subtitle text at the cost of minimal information loss. In the third step, the original frame is corroded by the subtitle mask. And the subtitle-removed frame is fed into EdgeConnect along with the subtitle mask. The whole process improves the accuracy of subtitle recognition through the precise-segmentation of the subtitle text area, and minimizes the error of frame inpainting through the careful-removal of subtitle text.

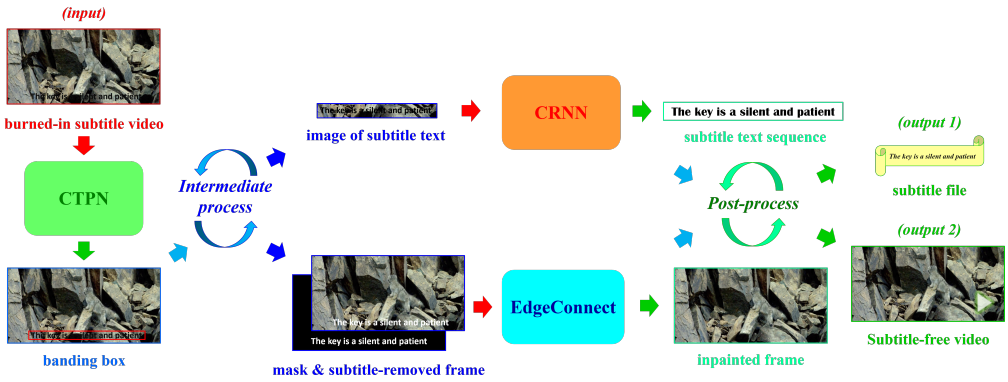

Figure 2. The processing flow of the entire deep-learning-based pipeline for burned-in subtitle video reconstruction, mainly consisting of two processes: an intermediate-process (for network coupling) and a post-process (for output conversion).

3.5. Post-Process

In order to obtain the final subtitle file and the subtitle-free video file, a post-process stage is needed. The subtitle text sequences are output by CRNN, while the inpainted frames are output by EdgeConnect. Hence, a post-process is required to synthesize the outputs into the subtitle file and video. As shown in Figure 2, the post-process takes all the inpainted frames and subtitle text sequences as input, where each inpainted frame and each sequence has an index corresponding to its position in the original video. The post-process consists of two parallel steps.

In the first step, all subtitle text sequences are sorted by the indices and the beginning and end of each subtitle in the time domain are calculated according to the original frame rate. Then, each subtitle is time-stamped and synthesized to a subtitle file. In the second step, all the inpainted frames are sorted by the indices and are assembled into a video at the original frame rate. Finally, the entire pipeline is completed with the post-process to convert the burned-in subtitle video to an integral subtitle file and video in reverse.

4. Results and Discussions

The proposed burned-in subtitle video reconstruction algorithm was implemented based on python programs. The three deep neural networks in the system were each trained on different training sets by adopting different strategies. Among them, CTPN was trained end-to-end on 3000 natural images by standard error back-propagation and stochastic gradient descent (SGD), with a learning rate of 10^{-3} for the initial 16 K iterations and 10^{-4} for the subsequent 4K iterations, using 0.9 momentum and 0.0005 weight decay. CRNN was also trained end-to-end on the Synth dataset [32] by back-propagation and SGD, using ADADELTA [33] to automatically calculate the learning rate for each dimension and iterating until convergence.

EdgeConnect uses the Adam optimizer [34] to optimize the model, with $\beta 1 = 0$ and $\beta 2 = 0.9$. The generators were trained end-to-end until convergence on the Places2 [35]

dataset with learning rates set to 10^{-4}, 10^{-5}, and 10^{-6}, gradually, while the discriminator's learning rate was one-tenth of the generator's. Finally, the network was fine-tuned by removing the discriminator of the first-stage GAN. The entire pipeline of burned-in subtitle video reconstruction was tested on 2186 video frames, and the experimental results of each stage are discussed in detail next.

4.1. Text Detection

As shown in Figure 3, frames with both Chinese and English subtitles are input to CTPN for text detection. The bboxes of the Chinese and English subtitle text areas are obtained at the output side of CTPN, and the Intersection over Union (IoU) [36] is calculated to measure the accuracy of subtitle detection.

$$IoU = \frac{The\ overlapping\ area\ of\ prediction\ and\ ground-true\ bounding\ boxes}{The\ union\ area\ of\ prediction\ and\ ground-true\ bounding\ boxes}$$

After testing, the IoUs of the output Chinese and English subtitle detection were 91.9% and 91.1%, respectively. As the input-processing-layer of the joint deep networks, CTPN achieved the precise positioning of multilingual subtitles in detection, ensuring accurate extraction and removal of the burned-in subtitles.

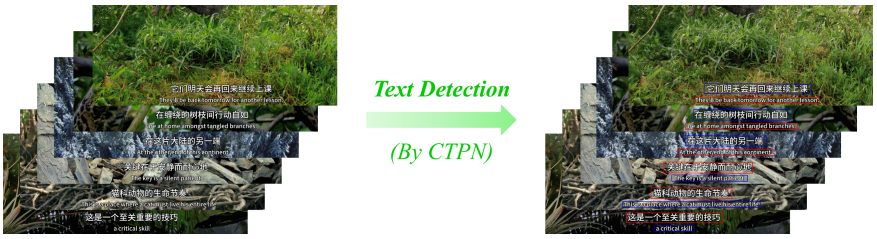

Figure 3. Precise bboxes were obtained with CTPN text detection.

4.2. Intermediate-Process

As shown in Figure 4, the bboxes of the Chinese/English subtitles from the CTPN and the original video frame were fed to the intermediate-process pipeline, and two groups of outputs were obtained: (1) images of the text area of the Chinese/English subtitle, and (2) subtitle masks and the subtitle-removed frames. The first group was input to the subtitle recognition network, and the second group was input to the frame inpainting network.

The whole process is based on the precise positioning of the subtitle text area, and the end-to-end recognition, and the minimal removal of subtitles is achieved by subtitle area segmentation and text contour extraction, which enhances the recognition accuracy of the subtitle text and the inpainting effect of the subtitle-removed frames.

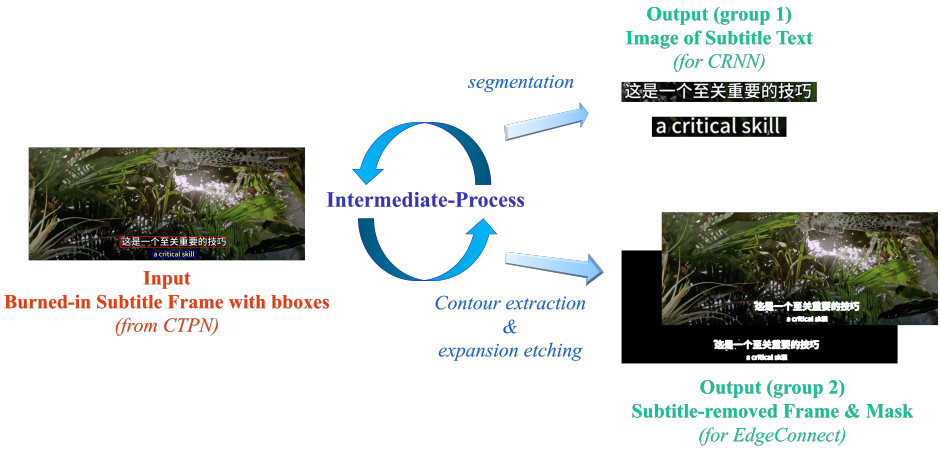

Figure 4. The burned-in subtitle frame with bboxes to subtitle text images and the subtitle-removed frame and mask, via the intermediate-process.

4.3. Text Recognition

The text images of Chinese and English subtitles were input to the CRNN for text recognition. As shown in Figure 5, the recognized texts of Chinese and English subtitles were output by the CRNN, respectively. Recognition was also performed on the entire frames without processing, as a comparison to demonstrate the advantages of the end-to-end recognition strategy. The recognition accuracy was calculated for the numerical evaluation.

Figure 5. The recognized results of Chinese and English subtitles output by the CRNN.

Table 1 lists the accuracies of the entire frame recognition and end-to-end recognition for Chinese/English subtitles, which indicates that the end-to-end recognition strategy significantly improved the dual-recognition accuracy of Chinese/English subtitles by minimizing the interference of irrelevant background information. Despite the acceptable result obtained by CRNN, it can still be seen that some non-negligible errors existed in the recognition of subtitles, due to the complex video image background. Fortunately, the proposed joint deep networks are partially modifiable; hence, the boosted text recognition network can be used to replace the existing text recognition part in the future.

$$Accuracy = \frac{Number\ of\ words\ correctly\ recognized}{Total\ number\ of\ recognized\ words}$$

Table 1. Accuracy of Chinese and English subtitle recognition under different strategies.

	Accuracy	
Strategy	**Chinese**	**English**
Entire frame recognition	70.5%	72.1%
End-to-end recognition	**81.6%**	**79.3%**

4.4. Frame Inpainting

Subtitle masks and the subtitle-removed frames were input to the EdgeConnect for inpainting, and the inpainted frames were obtained at the output of the EdgeConnect network. A traditional method and a state-of-the-art deep learning method were also tested as a comparison. As a representative diffusion-based inpainting method, the Fast Marching Method (FMM) [37], which utilizes existing domain pixels for gradient estimation to achieve fast marching of missing pixels, is suitable for video processing with high inpainting efficiency among the traditional methods.

As a representative GAN-based inpainting method, Globally and Locally Consistent Image Completion (GLCIC) [31] uses a fully convolutional network as a generator to inpaint pixels in arbitrarily shaped missing regions and discriminates the global and local consistency of the inpainted content by means of two discriminators. Hence, these two methods are used as traditional and state-of-the-art deep learning inpainting strategies, respectively, compared with our strategy.

In order to make an objective comparison between other existing methods and our method in terms of frame inpainting, the image quality metrics: Peak Signal-to-Noise Ratio (PSNR) [38], Structural SIMilarity (SSIM) [39], Normalized Root Mean Square Error (NRMSE), and Fréchet Inception Distance (FID) [40] were calculated for the entire inpainted frames to evaluate the inpainting performance. Figures 6 and 7 show the inpainting effects

of the traditional method (FMM), state-of-the-art deep learning method (GLCIC), and our method (EdgeConnect).

It can be seen that the textures inpainted by FMM method are blurred with insufficient details, while the textures inpainted by the GLCIC method are far from the ground true texture, albeit with more details. The inpainted textures of our method are significantly more vivid than those of the other existing methods and fit excellently with the ground true frames with higher fineness and realism. The frames inpainted by EdgeConnect are visually coherent and were produced faster than the FMM and GLCIC methods, making the video reconstruction system ideal for real-life applications.

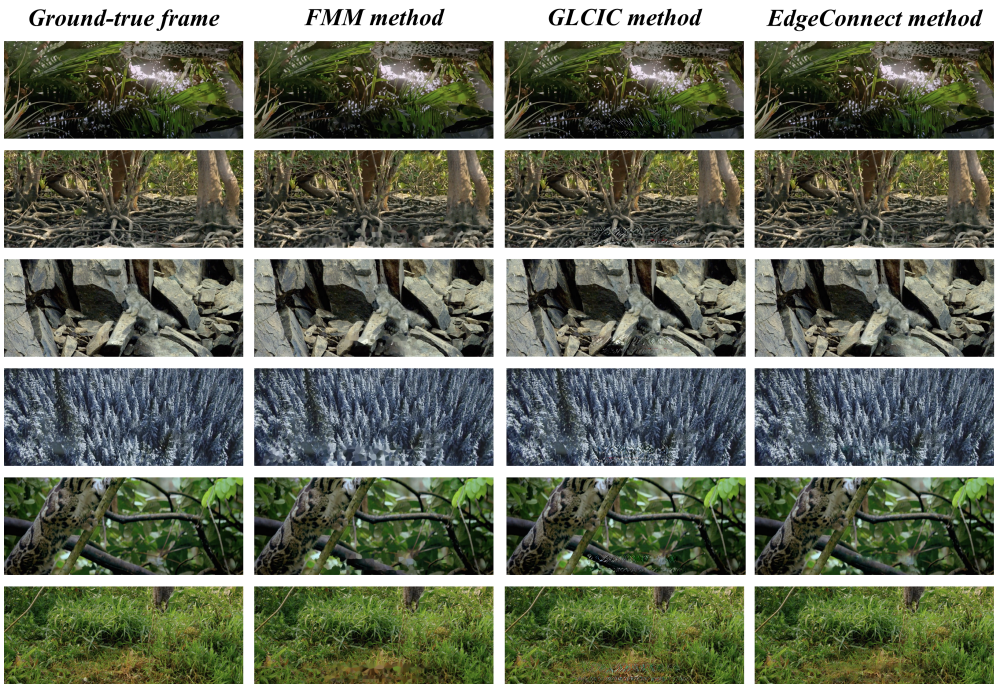

Figure 6. Comparison of the original frames and the inpainting results of other existing methods as well as our method.

Table 2 lists the evaluation metrics of the traditional method (FMM), state-of-the-art deep learning method (GLCIC), and our method (EdgeConnect). PSNR was used to measure the distortion, SSIM was used to measure the similarity, NRMSE was used to measure the pixel error, and FID was used to measure the feature vector distance between the ground-truth frames and the inpainted frames, using a pre-trained Inception-V3 model. Our method recovered the lost high-frequency information by edge-connecting based on adversarial learning, outperforming other existing methods in all the evaluation metrics; thus, the inpainted frames from our method demonstrated higher realism and more information.

Table 2. Evaluation metrics of the traditional method and our method.

	Evaluation Metrics			
Method	**PSNR**	**SSIM**	**NRMSE**	**FID**
FMM	28.804	0.945	0.110	0.280
GLCIC	29.241	0.960	0.099	0.142
EdgeConnect	**34.129**	**0.975**	**0.059**	**0.035**

4.5. Post-Process

The outputs of the joint deep networks were fed to a pipeline for post-processing. As shown in Figure 8, the Chinese/English subtitle text sequences from the CRNN were

synthesized into Chinese and English subtitle files, while the inpainted frames from the EdgeConnect were assembled into a video, during the post-process. The post-process finally realized the reconstruction of the burned-in subtitle video to the independent Chinese/English subtitle file and subtitle-free video, achieving the completeness of the entire reconstruction pipeline and facilitating users' re-editing.

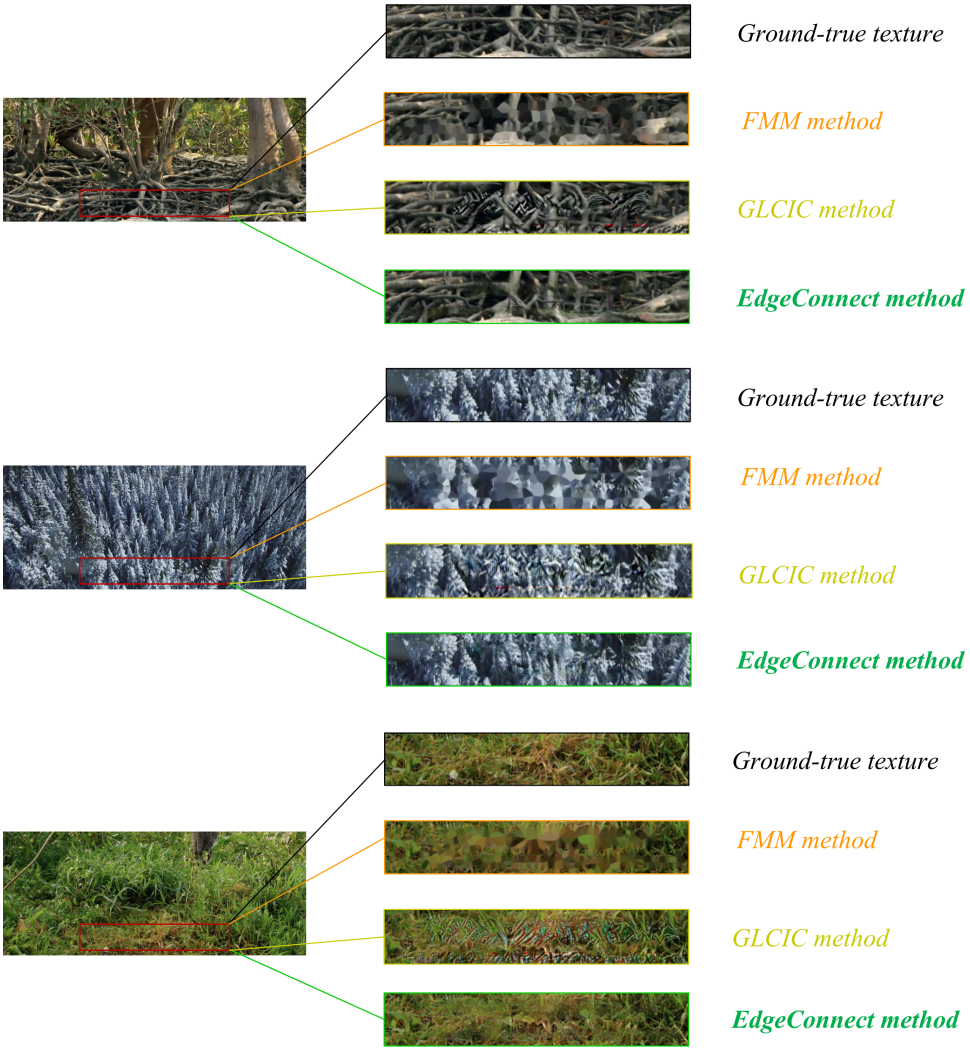

Figure 7. Zoomed-in comparison of the local texture details between the original frames and the inpainting results of other existing methods as well as our method.

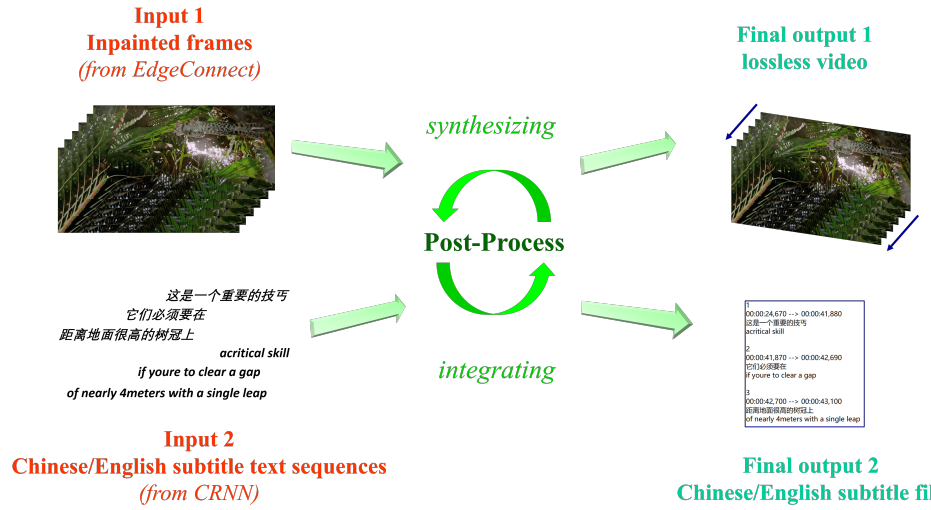

Figure 8. The inpainted frames and Chinese/English subtitle text sequences to Chinese/English subtitle file and subtitle-free video via post-processing.

5. Conclusions and Future Work

In this paper, we performed a deep-learning-based intelligent reconstruction system for burned-in subtitle videos. The novel system realized the seamless integration of CTPN, CRNN, and EdgeConnect through a well-designed intermediate-process. High-accuracy text extraction and high-quality frame restoration were achieved through joint deep neural networks. Finally, the system completed the inverse conversion from the burned-in subtitle video to the independent subtitle file and subtitle-free video by post-processing.

We evaluated the performance of the system, and found that the deep learning approach achieved high accuracy detection and recognition of subtitles and significantly enhanced the video inpainting compared to existing methods. This result is expected to be widely used in the field of reconstruction and re-editing of digital videos with subtitles, advertisements, logos, and other occlusions.

Future work can be continued in two aspects. The first is to improve the sub-net of the joint deep networks, especially for the text recognition network. According to the experimental results, both the text detection network (CTPN) and the frame inpainting network (EdgeConnect) achieved excellent performance; however, the accuracy of the text recognition network (CRNN) was still hindered by the complex video image background. We plan to combine audio recognition or a grammar checking network to enhance the subtitle recognition accuracy.

The second aspect is to polish the intermediate-processing steps for the coupling of deep networks, in particular for contour extraction. We plan to use a more accurate method for contour extraction to achieve the perfect removal of burned-in subtitles with minimal information loss.

Author Contributions: Conceptualization, B.J. and C.H.; Data curation, H.X., Y.H., and X.L.; Funding acquisition, B.J. and C.H.; Investigation, C.H.; Methodology, B.J., H.X., and Y.H.; Project administration, B.J.; Resources, C.H., X.H.; Software, H.X., Y.H., and B.J.; Visualization, Y.H. and X.L.; Writing—original draft, H.X., Y.H., X.L., and X.H.; Writing—review and Editing, H.X. and B.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61907025, 61702278), the Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 19KJB520048) and NUPTSF (Grant No. NY219069).

Data Availability Statement: Not Applicable, the study does not report any data.

Acknowledgments: The authors would like to thank all the anonymous reviewers for their valuable suggestions to improve this work.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Network configuration summary of CTPN. 'k', 's', and 'p' stand for the kernel size, stride size, and padding size, respectively.

Type	Configuration
Input	input raw image
Convolution	#maps:64, k:3 × 3, s:1, p:1
Convolution	#maps:64, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:128, k:3 × 3, s:1, p:1
Convolution	#maps:128, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:512, k:3 × 3, s:1, p:1
Convolution	#maps:512, k:3 × 3, s:1, p:1
Convolution	#maps:512, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:512, k:3 × 3, s:1, p:1
Convolution	#maps:512, k:3 × 3, s:1, p:1
Convolution	#maps:512, k:3 × 3, s:1, p:1
Map-to-Sequence	#maps:512, k:3 × 3, s:1, p:1
Bidirectional-LSTM	#hidden units:128
Bidirectional-LSTM	#hidden units:128
FullConnection	#dimension:512
Output 1	vertical coordinates
Output 2	side-refinement
Output 3	text/non-text scores

Table A2. Network configuration summary of CRNN. 'k', 's', and 'p' stand for the kernel size, stride size, and padding size, respectively.

Type	Configuration
Input	input gray-scale image
Convolution	#maps:64, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:128, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
MaxPooling	Window:1 × 2, s:2
Convolution	#maps:512, k:3 × 3, s:1, p:1
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
BatchNormalization	-
MaxPooling	Window:1 × 2, s:2
Convolution	#maps:512, k:2 × 2, s:1, p:0
Map-to-Sequence	-
Bidirectional-LSTM	#hidden units:256
Bidirectional-LSTM	#hidden units:256
Transcription	text sequence

Table A3. Network configuration summary of EdgeConnect. 'k', 's', and 'p' stand for the kernel size, stride size, and padding size, respectively.

EdgeGenerator	
Type	**Configuration**
Input	mask + edge + gray-scale map
Convolution	#in_channels:3, out_channels:64, k:7 × 7, p:0
Convolution	#in_channels:64, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:256, k:4 × 4, s:2, p:1
ResnetBlock×8	#dimension:256, dilation = 2
Convolution	#in_channels:256, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:64, k:4 × 4, s:2, p:1
Convolution	#in_channels:64, out_channels:1, k:7 × 7, p:0
EdgeDiscriminator	
Type	**Configuration**
Convolution	#in_channels:1, out_channels:64, k:4 × 4, s:2, p:1
Convolution	#in_channels:64, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:256, k:4 × 4, s:2, p:1
Convolution	#in_channels:256, out_channels:512, k:4 × 4, s:1, p:1
Convolution	#in_channels:512, out_channels:1, k:4 × 4, s:1, p:1
InpaintGenerator	
Type	**Configuration**
Input	edge map + RGB map
Convolution	#in_channels:4, out_channels:64, k:7 × 7, p:0
Convolution	#in_channels:64, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:256, k:4 × 4, s:2, p:1
ResnetBlock×8	#dimension:256, dilation = 2
Convolution	#in_channels:256, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:64, k:4 × 4, s:2, p:1
Convolution	#in_channels:64, out_channels:3, k:7 × 7, p:0
InpaintDiscriminator	
Type	**Configuration**
Convolution	#in_channels:3, out_channels:64, k:4 × 4, s:2, p:1
Convolution	#in_channels:64, out_channels:128, k:4 × 4, s:2, p:1
Convolution	#in_channels:128, out_channels:256, k:4 × 4, s:2, p:1
Convolution	#in_channels:256, out_channels:512, k:4 × 4, s:1, p:1
Convolution	#in_channels:512, out_channels:1, k:4 × 4, s:1, p:1

References

1. Liqin, J.I.; Jiajun, W. Automatic Text Detection and Removal in Video Images. *Chin. J. Image Graph.* **2008**, *13*, 461–466.
2. Xu, Y.; Shan, S.; Qiu, Z.; Jia, Z.; Shen, Z.; Wang, Y.; Shi, M.; Eric, I.; Chang, C. End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble. *Signal Process. Image Commun.* **2018**, *60*, 131–143. [CrossRef]
3. Yan, H.; Xu, X. End-to-end video subtitle recognition via a deep Residual Neural Network. *Pattern Recognit. Lett.* **2020**, *131*, 368–375. [CrossRef]
4. Favorskaya, M.N.; Zotin, A.G.; Damov, M.V. Intelligent inpainting system for texture reconstruction in videos with text removal. In Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems, Moscow, Russia, 18–20 August 2010; pp. 867–874.
5. Khodadadi, M.; Behrad, A. Text localization, extraction and inpainting in color images. In Proceedings of the 20th Iranian Conference on Electrical Engineering (ICEE2012), Tehran, Iran, 15–17 May 2012; pp. 1035–1040.
6. Jung, C.; Liu, Q.; Kim, J. A new approach for text segmentation using a stroke filter. *Signal Process.* **2008**, *88*, 1907–1916. [CrossRef]
7. Zhang, D.Q.; Chang, S.F. Learning to detect scene text using a higher-order MRF with belief propagation. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004; p. 101.
8. Wolf, C.; Jolion, J.M.; Chassaing, F. Text localization, enhancement and binarization in multimedia documents. In Proceedings of the Object Recognition Supported by User Interaction for Service Robots, Quebec City, QC, Canada, 11–15 August 2002; Volume 2, pp. 1037–1040.

9. Tian, Z.; Huang, W.; He, T.; He, P.; Qiao, Y. Detecting text in natural image with connectionist text proposal network. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 56–72.
10. Esedoglu, S.; Shen, J. Digital inpainting based on the Mumford–Shah–Euler image model. *Eur. J. Appl. Math.* **2002**, *13*, 353–370. [CrossRef]
11. Liu, D.; Sun, X.; Wu, F.; Li, S.; Zhang, Y.Q. Image compression with edge-based inpainting. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1273–1287.
12. Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; Verdera, J. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* **2001**, *10*, 1200–1211. [CrossRef] [PubMed]
13. Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D.B.; Sen, P. Image melding: Combining inconsistent images using patch-based synthesis. *Acm Trans. Graph. (TOG)* **2012**, *31*, 1–10. [CrossRef]
14. Huang, J.B.; Kang, S.B.; Ahuja, N.; Kopf, J. Image completion using planar structure guidance. *Acm Trans. Graph. (TOG)* **2014**, *33*, 1–10. [CrossRef]
15. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
16. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2298–2304. [CrossRef]
17. Favorskaya, M.N.; Damov, M.V.; Zotin, A.G. Intelligent method of texture reconstruction in video sequences based on neural networks. *Int. J. Reason. Based Intell. Syst.* **2013**, *5*, 223–236. [CrossRef]
18. Vuong, T.L.; Le, D.M.; Le, T.T.; Le, T.H. Pre-rendered subtitles removal in video sequences using text detection and inpainting. In Proceedings of the International Conference on Electronics, Information and Communication, Danang, Vietnam, 27–30 January 2016; pp. 94–96.
19. Jaderberg, M.; Vedaldi, A.; Zisserman, A. Deep features for text spotting. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 512–528.
20. Busta, M.; Neumann, L.; Matas, J. Fastext: Efficient unconstrained scene text detector. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1206–1214.
21. Huang, W.; Qiao, Y.; Tang, X. Robust scene text detection with convolution neural network induced mser trees. In *Proceedings of the European Conference on Computer Vision*; Springer, Berlin/Heidelberg, Germany, 2014; pp. 497–511.
22. Yin, X.C.; Yin, X.; Huang, K.; Hao, H.W. Robust text detection in natural scene images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 970–983.
23. Wang, T.; Wu, D.J.; Coates, A.; Ng, A.Y. End-to-end text recognition with convolutional neural networks. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba Science City, Japan, 11–15 November 2012; pp. 3304–3308.
24. Bissacco, A.; Cummins, M.; Netzer, Y.; Neven, H. Photoocr: Reading text in uncontrolled conditions. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 785–792.
25. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [CrossRef]
26. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
28. Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 855–868. [CrossRef] [PubMed]
29. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
30. Yeh, R.A.; Chen, C.; Yian Lim, T.; Schwing, A.G.; Hasegawa-Johnson, M.; Do, M.N. Semantic image inpainting with deep generative models. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5485–5493.
31. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *Acm Trans. Graph.* **2017**, *36*, 107:1–107:14. [CrossRef]
32. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *arXiv* **2014**, arXiv:1406.2227.
33. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, arXiv:1212.5701.
34. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
35. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef] [PubMed]
36. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
37. Telea, A. An image inpainting technique based on the fast marching method. *J. Graph. Tools* **2004**, *9*, 23–34. [CrossRef]

38. Sara, U.; Akter, M.; Uddin, M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *J. Comput. Commun.* **2019**, *7*, 8–18. [CrossRef]
39. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
40. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv* **2017**, arXiv:1706.08500