Deep Video Decaptioning

Pengpeng Chu*[1]
2191420@mail.dhu.edu.cn
Weize Quan*[23]
qweizework@gmail.com
Tong Wang†[1]
wangtong@dhu.edu.cn
Pan Wang[4]
dixian.wp@alibaba-inc.com
Peiran Ren[4]
peiran.rpr@alibaba-inc.com
Dong-Ming Yan[23]
yandongming@gmail.com

[1] Donghua University,
Shanghai, China

[2] NLPR, Institute of Automation,
Chinese Academy of Sciences,
Beijing, China

[3] University of Chinese Academy of
Sciences,
Beijing, China

[4] Alibaba Group,
Hangzhou, China

* Equal contribution
† Corresponding author

Abstract

Video decaptioning aims to remove subtitles from and repair occluded areas in videos. However, recent deep-learning-based inpainting methods mostly require the masks indicating the corrupted parts, and these masks are unavailable for the input subtitled videos. Moreover, useful information hidden in the background of subtitles might be lost when these masked areas are directly regarded as invalid as the common setting of inpainting methods. In addition, existing blind video decaptioning methods often suffer from incomplete subtitles removal. In this paper, we propose a generic framework for video decaptioning, which consists of a caption mask extraction network and a frame-attention-based decaptioning network. The former is trained with supervision information using our proposed automatic annotation method, and predicts the position of the subtitle and background. The latter adopts an encoder-decoder architecture with the skip connection. The encoder extracts the features of all input frames. Then, multiple frame attention modules are used to aggregate these features from the spatial and temporal dimensions. Finally, the fused features are reconstructed into a target frame using the decoder. Extensive experiments demonstrate that our proposed method can accurately remove subtitles from videos in real time (60+ FPS), and outperforms the state-of-the-art approaches. Code is available at https://github.com/Linya-lab/Video_Decaptioning.

1 Introduction

Video inpainting is a challenging task that aims to restore the damaged regions in videos with plausible content, and has been widely studied by the computer vision community. Video inpainting technology is often used in video editing applications, such as super-resolution [10, 19, 28], special effects production [22, 25, 27], and object removal [7, 14, 33]. In this paper, we mainly focus on the removal of subtitles in videos and attempt to apply it in real-world video processing tasks.

Video decaptioning is a video inpainting task that aims to remove subtitles and their background. A straightforward approach is the use of an image restoration algorithms to remove the subtitles frame-by-frame. However, this approach ignores the time coherence of the video, often resulting in the unnatural connections between frames. Furthermore, the occluded parts of one frame may be intact in other frames of the video, and dealing with a single frame fails to take advantage of these valid parts. Another way of removing captions is to extend the 2D image inpainting method to the 3D video scenario, that is, replacing 2D convolution with 3D convolution. Although this approach can guarantee a certain time continuity, 3D operations usually consume a significant percentage of memory and may cause error accumulation.

The greatest challenge in video decaptioning is the absence of caption masks in the current public dataset. Consequently, the existing image/video inpainting methods cannot be directly applied to video decaptioning. Although the text detection method can be used to mark out the captions in the video, many subtitles have backgrounds with varying transparency (as shown in Fig. 1). At the same time, directly considering these regions as invalid according to the obtained caption mask will result in the loss of the underlying information hidden in the background of the subtitles. However, when no mask is used, direct blind video decaptioning always leads to incomplete subtitles removal.

To overcome the above challenges, we propose a generic video decaptioning framework with two stages. The first stage is a caption mask extraction network, which mainly obtains the mask of the caption and its background. The second stage is a decaptioning network, which removes the captions and obtains the final results. Unlike common image/video inpainting methods, the proposed method does not directly delete the content of the subtitle region according to the obtained mask. Instead, the mask and the video frames are taken together as input to the video decaptioning network, and the gated convolution is utilized to allow the network to adaptively learn what is valid in the video. This approach not only preserves the information hidden in the background of the captions but also enables the network to deal with various shapes of subtitles. Our model takes multiple reference frames as input and finally recovers the central target frame. To aggregate the effective features of the input frames to the target frame, we propose a frame attention module, which is embedded in the decaptioning network. The frame attention module searches for coherent contents from all the frames along both the spatial and temporal dimensions to fill the occluded regions in the target frame. Moreover, the frame attention module stacks multiple layers, allowing the improvement of the attention results for the occluded regions on the basis of the updated region features. Our contributions are summarized as follows:

- We propose a caption mask extraction network to predict the mask of the caption and its background, and this model is trained with the supervision information using our proposed automatic annotation tool.

- We propose an encoder-decoder decaptioning network with the skip connection, gated convolution, and embedded multiple frame attention modules to aggregate the effective features of all input frames to the target frame.

- Our proposed generic decaptioning framework exhibits state-of-the-art performance on the popular public video decaptioning dataset.

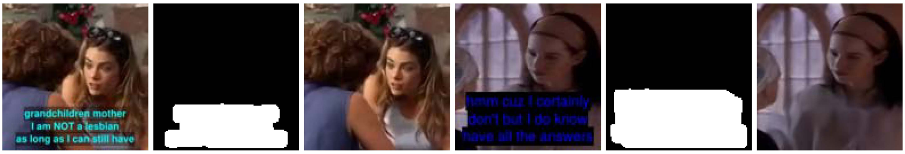


Figure 1: Selected video decaptioning results of our method. Each group includes subtitled frame, mask predicted by our caption mask extraction network, and subtitle removal result.

2 Related Work

2.1 Image Inpainting

Various approaches have been proposed for image inpainting. Diffusion-based algorithms [2, 5, 6] mainly depend on low-level image features to propagate appearance information from the boundaries to the missing regions. However, these methods often suffer from blur artifacts for relatively large holes. Patch-based methods [1, 8, 34] tend to solve this problem by searching the matching patch from the valid regions and copying them to the missing regions. These methods work well on the stationary texture regions but often fail in the non-stationary images.

With the advancement of deep learning, *i.e.*, *convolutional neural networks* (CNNs), recent image inpainting methods have been proposed for learning the mapping of corrupted images to completed images. Context Encoders [24] was first introduced in image inpainting by employing a CNN model with the adversarial loss [9]. Iizuka *et al.* [12] utilized global and local discriminators to enhance the consistency around the boundary of missing regions. However, their method fails to handle irregular holes. Zhang *et al.* [40] solved the image inpainting problem by providing the location of artifacts with a detector. To better collect contextual information, the attention mechanism [35] was introduced to image inpainting. Various attention computation methods [17, 32] have been proposed to improve the visual quality of inpainting results. Moreover, stacked mechanisms [3, 21, 36] and progressive strategies [15, 16, 38] have also been designed in image inpainting.

2.2 Video Inpainting

Video inpainting can be regarded as the extension of image inpainting in the time dimension. Many early video inpainting approaches [11, 20, 31] borrowed ideas from traditional image inpainting methods. Wexler *et al.* [31] synthesized missing contents by sampling similar spatial or spatial-temporal patches from valid regions on the basis of a global optimization. Optical flow estimation was also applied in video inpainting [11, 20]. These methods usually assume that the missing parts might appear in other frames and often suffer from high computational complexity. Similar to image inpainting, recent works have also begun to devise deep-learning-based methods to fill the missing areas in videos. Chang *et al.* [7] proposed a 3D gated convolution and temporal SN-PatchGAN for free-form video inpainting. Kim *et al.* [14] adopted the recurrent network and optical flow to ensure temporal coherence. Xu *et al.* [33] completed video frames and their optical flows together, and combined them to achieve good results. Although these methods can fuse some features in several neighboring frames, they are limited by the range of the convolutional receptive field and cannot use the effective information of distant locations. To solve this problem, some recent works proposed the attention/transformer-based methods. Oh *et al.* [23] recurrently calculated the

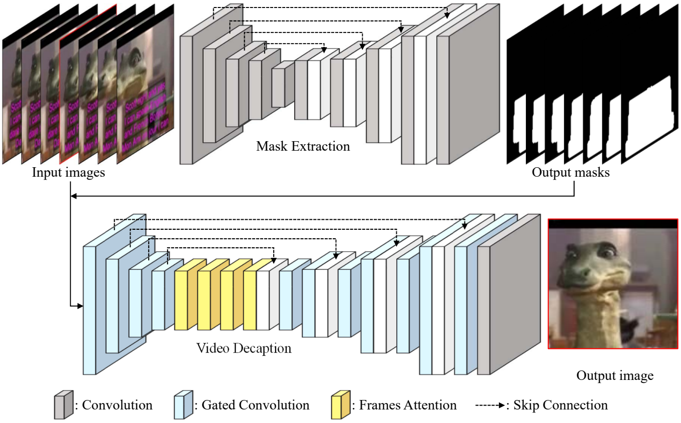


Figure 2: Overview of our video decaptioning framework. It removes the subtitles in two steps. The first step is to extract the subtitle masks, and the second step is to combine the input frames with the extracted masks to accurately remove the subtitles in the video.

attention scores between the target and reference frames, and eventually filled holes from the boundary. Zeng *et al.* [37] adopted a "multi-to-multi" inpainting strategy, and calculated the attention scores at different dimensions between the different channels of the image feature to adapt to various shapes of holes. Very recently, Liu *et al.* [18] proposed a decoupled spatial-temporal transformer for video inpainting. Their networks consist of multiple interweaving stacks of temporally- and spatially-decoupled transformer blocks.

Video decaptioning, which can be considered a special video inpainting task, has recently gained increasing attention in the computer vision community. The ECCV ChaLearn 2018 Inpainting Challenge presented the corresponding track for video decaptioning and shared a public dataset. The BVDNet proposed by Kim *et al.* [13] achieved the best result. However, their method fails to address captions with a solid background. Moreover, their method cannot completely remove subtitles when the video frames change quickly.

3 Proposed Method

In this work, we propose a novel two-stage network for video decaptioning. Fig. 2 presents an overview of our network. The first stage is a caption mask extraction network, which can accurately detect the subtitle regions in the video frame. The second stage is a decaptioning network, which removes the subtitles with an effective attention modules using the previously predicted masks. Given a target frame with subtitles I_t, our network erases the subtitles with the explicitly extracted location information, through fusing the information of several neighboring subtitled frames $\{I_{t-n:t+n}\}$ from the spatial and temporal dimensions.

3.1 Caption Mask Extraction

In this stage, the main challenges come from two aspects: (1) the subtitles in the video contain backgrounds with different transparencies, and (2) the current public dataset does not include annotated subtitle masks (*i.e.*, covering the text and its background). Consequently,

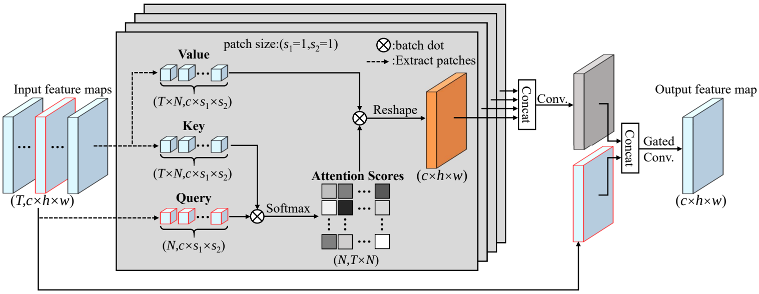


Figure 3: Overview of the frame attention module. Before calculating the attention score, the frames have been reduced from images of 128 × 128 to feature maps of 16 × 16 via an encoder. In order to remove subtitles of various sizes, we parallelly calculate attention scores in four different patch sizes (1 × 1, 2 × 2, 4 × 4, and 8 × 8).

conventional text detection methods cannot achieve complete subtitle masks, while deep learning-based methods have no mask label for training. Therefore, we propose a caption mask extraction network (the first row of Fig. 2) based on U-Net [26], and devise a simple and effective automatic annotation method. This annotation method provides the mask supervision information for network training. Given a pair of the original frame I_o and the subtitled frame I, we obtain the ground truth annotation of the subtitles via difference, filtering, and thresholding operations as follows:

$$M_{GT} = \mathcal{F}(\text{avg}(|I_o - I|)) > \varepsilon, \tag{1}$$

where $|\cdot|$ is the abstraction, avg computes the average value in the color channels, $\mathcal{F}(\cdot)$ is the mean filter with a kernel size of 5×5, and ε is the threshold (we set $\varepsilon = 0.05$). Filtering and thresholding operations can make the extracted mask completely cover the area affected by the subtitles. The ground truth mask can only be accessed in the training phase.

3.2 Video Decaptioning

After obtaining the subtitle masks, we do not directly remove the occluded regions in the video. Instead, we concatenate the extracted masks and the subtitled frames together as input for the decaptioning network. This network consists of three components, namely, an encoder, several frame attention modules, and a decoder (the second row of Fig. 2). The encoder extracts the hierarchical texture and structure representations, the frame attention modules predict the appropriate contents in the missing regions, and the decoder progressively reconstructs the inpainted video frame. In the encoder and the decoder, we employ gated convolution, which can adaptively perceive the location of corrupted regions and choose the information from the background of the subtitles. Gated convolution is formulated as follows:

$$G_{x,y} = \sum\sum W_g \cdot I,\ F_{x,y} = \sum\sum W_f \cdot I,\ O_{x,y} = \phi(F_{x,y}) \odot \psi(G_{x,y}), \tag{2}$$

where ψ is the sigmoid function to transform gate to value between 0 (invalid) and 1 (valid), ϕ is the original activation function (*e.g.*, LeakyReLU), and W_g and W_f are two different convolution kernels.

The frame attention module is mainly based on the organic combination of Patch-Match [1] and attention [29]. This module aims to fill the occluded part of the target frame by searching for the best matching patches. As shown in Fig. 3, we use $F_t \in R^{c\times h\times w}$ to denote the target feature map encoded from the frame-level encoder. Unlike the common attention mechanism in the transformer [29], we omit the embedding operation and directly divide F_t into $N = h/s_1 \times w/s_2$ patches $Q_i \in R^{c\times s_1\times s_2}$ according to the set size (s_1, s_2). We use the same division operation to process all input reference feature maps $F_{t-n:t+n}$ to obtain $T \times N$ patches $K_i \in R^{c\times s_1\times s_2}$, where $T = 2n+1$. Specifically, we reshape the extracted patches into 1-dimensional vectors, so that patch-wise similarity can be calculated via the dot production. The similarity between the target feature map and the valid (non-subtitled) regions of the reference feature maps is denoted as:

$$s_{i,j} = \frac{Q_i \cdot K_j}{\sqrt{s_1 \times s_2 \times c}}, \tag{3}$$

where Q_i is the i-th patch extracted from the target feature map F_t, K_j is the j-th patch extracted from the feature maps $F_{t-n:t+n}$ outside the subtitle region, (s_1, s_2) is the patch size and c is the depth of the feature map. Given that our caption mask extraction network can predict a mask indicating the location of subtitles, we calculate the attention score of each valid patch of all reference frames, avoiding the interference from the subtitle regions. This is written as $\alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_j \exp(s_{i,j})}$. Finally, the i-th patch of the target frame is updated via the weighted summation of all the non-subtitled patches in $F_{t-n:t+n}$: $\bar{Q}_i = \sum_j \alpha_{i,j} K_j$. Each patch of the target frame is subjected to the same process, and we update the full feature map of the target frame. The above computation is based on the patch. Considering that the subtitles have different shapes and sizes, we simultaneously calculate the attention and update the feature maps in the multiple patch sizes. In addition, since we adopt a "multi-to-one" decaptioning method, it is not necessary to divide the input features along with the channel dimension and then calculate the attention score to save computational memory as in the classical transformer. Instead, we calculate the results under different patch sizes and concatenate them, and then we obtain the attentive feature maps via a convolution operation. In this way, we can obtain more complete feature information. Finally, we use gated convolution to effectively fuse the original and attentive feature maps of the target frame. As shown in Fig. 2, we stack four frame attention modules to exploit the inpainting capacity of frame attention. We also visualize the attention maps learned by frame attention in the last layer as reported in Fig. 4. For the part of the target frame that is covered by the subtitles, frame attention can search for the effective area in all reference frames to reconstruct it.

3.3 Network Training

In this work, we first train the caption mask extraction network and then the decaptioning network. We sequentially process the input frames in a sliding window manner (window size: $T = 2n+1 = 7$, frame stride: $s = 3$) and produce the decaptioning results in real time. With the automatically annotated subtitle masks as the supervision information, the caption mask extraction network is trained with the binary cross entropy loss.

As with most inpainting tasks, we optimize the decaptioning model by taking the original video frames as ground truths without any other labels. We apply the L_1 losses calculated between the generated and original frames to ensure pixel-wise reconstruction accuracy. The

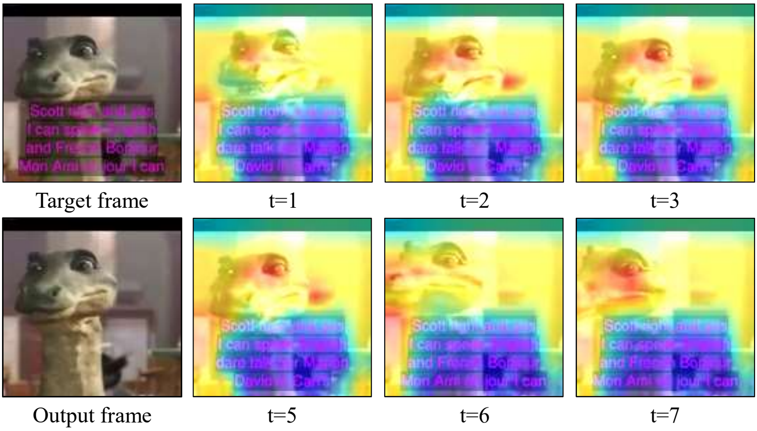


Figure 4: Illustration of the attention maps for subtitled regions learned by frame attention. In order to recover the part of the target frame that is occluded by the subtitles, our model can find the most suitable area from the non-subtitled area in the reference frames for repairing. Attention regions are highlighted with bright red.

L_1 losses for the subtitled and non-subtitled regions are respectively expressed as follows:

$$L_{sub} = \frac{\left\|\hat{M} \odot (I_o - \hat{I})\right\|_1}{\left\|\hat{M}\right\|_1}, L_{non\_sub} = \frac{\left\|(1-\hat{M}) \odot (I_o - \hat{I})\right\|_1}{\left\|1-\hat{M}\right\|_1}, \tag{4}$$

where $\hat{M}$ is the extracted mask, $\hat{I}$ and I_o respectively indicate the predicted and target ground-truth frames, $\odot$ denotes the element-wise multiplication, and the values are normalized by the region sizes. Similar to [13], we also apply the SSIM loss with a small patch window:

$$L_{ssim} = \frac{(2\mu_{\hat{I}}\mu_{I_o} + c_1)(2\sigma_{\hat{I}I_o} + c_2)}{(\mu_{\hat{I}}^2 + \mu_{I_o}^2 + c_1)(\sigma_{\hat{I}}^2 + \sigma_{I_o}^2 + c_2)}, \tag{5}$$

where μ and σ denote the average and the variance, respectively; and c_1 and c_2 are two constants for stabilizing the division ($c_1 = 0.01^2$, $c_2 = 0.03^2$). To this end, the final objective for training the decaptioning network is as follows:

$$L_{total} = \lambda_{sub} \cdot L_{sub} + \lambda_{non\_sub} \cdot L_{non\_sub} + \lambda_{ssim} \cdot L_{ssim}. \tag{6}$$

We set $\lambda_{sub} = 6$, $\lambda_{non\_sub} = 1$, and $\lambda_{ssim} = -1$ in all our experiments.

4 Experimental Results

4.1 Dataset and Experimental Settings

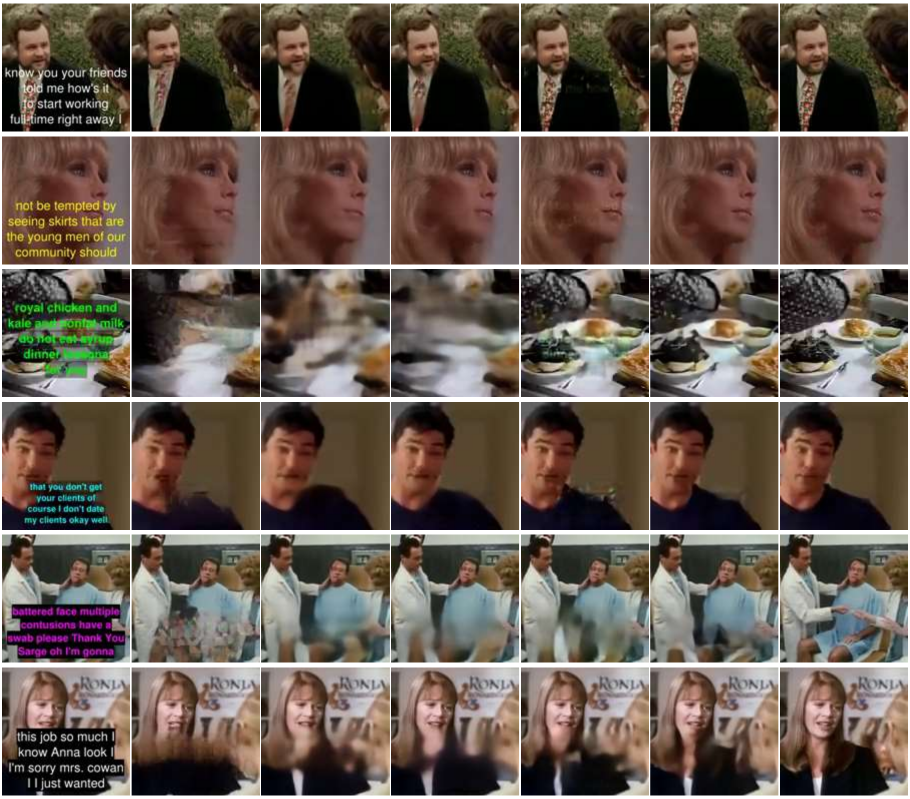
A public large-scale dataset, that is, the ECCV Chalearn 2018 LAP Video Decaptioning Challenge dataset, is used for training and testing. Each sample is a 5 seconds MP4 video clip, including 125 RGB frames of 128 × 128, and each frame pair contains encrusted subtitles and without subtitles. The captions in this dataset have different sizes, colors, positions,

Method	MSE↓	PSNR↑	DSSIM↓	LPIPS↓	VFID↓
Onion-Peel	0.0035	28.8437	0.0672	0.0769	1.1805
DSTT	0.0023	30.5102	0.0562	0.0939	1.0684
STTN	0.0022	30.7003	0.0553	0.0928	1.0177
BVDNet	0.0013	34.1275	0.0365	0.0529	0.8001
Ours	**0.0011**	**35.0251**	**0.0317**	**0.0497**	**0.6995**

Table 1: Quantitative comparisons of our method with Onion-Peel [23], DSTT [18], STTN [37], and BVDNet [13]. ↑: higher is better; ↓: lower is better.

and shadows. Following the default splitting, we use 70K samples for training and 5K samples for testing. All video clips are converted to PNG images for training and testing.

Our model is implemented with PyTorch v1.8.1, and run on a NVIDIA RTX 3090 GPU. The Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used to train the model. The batch size is 64, and the learning rate is set to 1×10^{-4}. During training, we also use random horizontal flipping for data augmentation and early-stopping to avoid over-fitting.



(a) Input frame (b) Onion-Peel (c) DSTT (d) STTN (e) BVDNet (f) Ours (g) Ground Truth

Figure 5: Visual results of subtitles removal with different backgrounds. From left to right: input subtitled frames, Onion-Peel [23], DSTT [18], STTN [37], BVDNet [13], our method, and ground truth.

4.2 Quantitative and Qualitative Comparisons

We compare the proposed model with state-of-the-art video inpainting models, Onion-Peel [23], DSTT [18], and STTN [37], and the blind video decaptioning method BVDNet [13]. Following the common setting of video inpainting, for Onion-Peel, DSTT, and STTN, we remove the subtitle and their background to simulate missing regions and use the extracted mask (the output of our caption mask extraction network) to indicate the missing regions.

For quantitative evaluation, we choose the metrics frequently used in inpainting tasks, including mean square error (MSE), peak signal-to-noise ratio (PSNR), structural dissimilarity (DSSIM), and learned perceptual image patch similarity (LPIPS) [39]. Furthermore, we calculate the video-based Fréchet Inception Distance (VFID) [7] with the I3D [4] pre-trained video recognition CNN as Vid2vid [30]. The corresponding results are reported in Table 1. Our method significantly outperforms the three advanced video inpainting methods mainly because video inpainting directly removes subtitles, which leads to information loss, especially for subtitles with transparent backgrounds. Our method exhibits better decaptioning performance than BVDNet.

Fig. 5 shows the qualitative results of the above five methods. Our method clearly achieves the best results. Onion-Peel, DSTT, and STTN often suffer from blur artifacts, and BVDNet displays the subtitle remnant phenomenon, and fills the smooth content in the subtitle regions with a solid background (see last two rows). Given that the caption mask extraction network provides the exact location of subtitles and decaptioning network applies the frame attention modules to effectively use the valid features in the reference frames, our method can achieve reasonable semantics and rich textures.

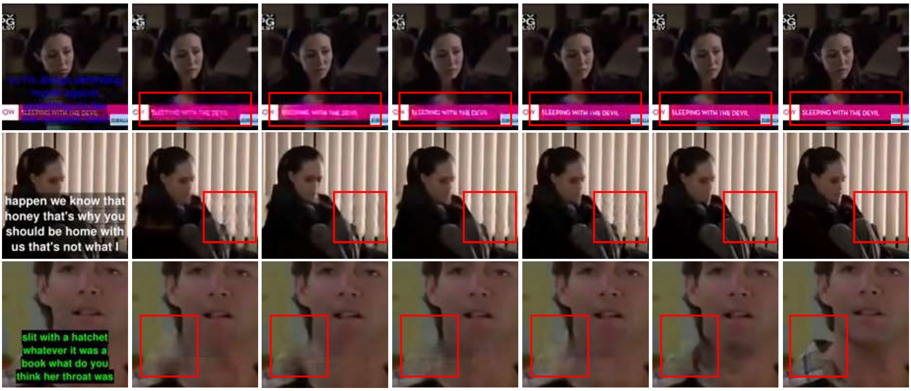
4.3 Ablation Study

We conduct ablation studies to evaluate the effectiveness of the different components of our model, and the corresponding results are reported in Table 2 and Fig. 6. The full model achieves the best performance. With mask guidance, the decaptioning network can mainly focus on the subtitle regions and prevent information loss. The gated convolution can better extract effective features by adaptively perceiving the valid regions in the video. In addition, the original details can be better preserved with the skip connection (*e.g.*, the first row of Fig. 6). With multiple frame attention modules, plausible structures and textures are obtained (*e.g.*, the second and third rows of Fig. 6).

Other network details, visual results, and ablation studies are presented in the *supplementary material*.

Exp	Mask	Gated.	Skip Con.	Frame Att.	MSE↓	PSNR↑	DSSIM↓
1					0.0014	32.6915	0.0416
2	✓				0.0013	32.9352	0.0400
3	✓	✓			0.0013	33.1307	0.0384
4	✓	✓	✓		0.0012	34.3264	0.0356
5	✓	✓	✓	✓	**0.0011**	**35.0251**	**0.0317**

Table 2: The ablation studies on mask extraction, gated convolution, skip connection, and frame attention. We evaluate on ChaLearn 2018 LAP Inpainting Track2 validation set.



(a) Input frame (b) Conv. (c) Mask+Conv. (d) Mask+Gated. (e) Mask+Gated. +Skip Con. (f) Mask+Gated. +Skip Con.+Frame Att. (g) Ground Truth

Figure 6: The visual evaluation of different components of our model. (b)-(f) separately corresponds to the settings (Exp 1-5) in Table 2.

5 Conclusions and Future Work

In this paper, we propose a generic framework for video decaptioning with two stages. The first stage is a caption mask extraction network, marking the position of subtitles and their backgrounds in the video. The second stage is a decaptioning network, which adopts an encoder-decoder structure with the skip connection to ensure the details of the results. We propose a simple and effective automatic annotation method to provide supervision for the training of the caption mask extraction network. For the model to make full use of the information of the input reference frames, the frame attention module is also proposed to fuse the features from the spatial and temporal dimensions to reasonably reconstruct the occluded part of the target frame. The final model can automatically remove subtitles from the video and restore the original contents in real time.

For very large caption, our results sometimes are slight smooth, and we would like to improve it by introducing the distant key frames into the process window. In our framework, we predict the binary mask of the subtitle and its background to assist the subsequent decaptioning network, which achieves the good performance. In the future, we would like to improve the cases with translucent backgrounds by providing more exact subtitle and background mask or non-binary mask. In addition, for a fair comparison, we mainly conduct experiments and comparisons on the only public video decaptioning dataset. We would also like to collect a large-scale high-resolution video dataset with diverse subtitle styles and motion in the area of the subtitles, and extend our method for high-resolution scenarios.

6 Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62102418 and 62172415), the National Key R&D Program of China (2019YFB2204104), and the Alibaba Group through Alibaba Innovative Research Program.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):24, 2009.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000.

[3] Xuewei Bian, Chaoqun Wang, Weize Quan, Juntao Ye, Xiaopeng Zhang, and Dong-Ming Yan. Scene text removal via cascaded text stroke detection and erasing. *Computational Visual Media*, 2021. to appear.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Tony F Chan and Jianhong Shen. Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation*, 12(4):436–449, 2001.

[6] Tony F Chan and Jianhong Shen. Variational image inpainting. *Communications on Pure and Applied Mathematics*, 58(5):579–619, 2005.

[7] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9066–9075, 2019.

[8] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[11] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6): 1–11, 2016.

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[13] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep blind video decaptioning by temporal aggregation and recurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4263–4272, 2019.

[14] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[15] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5961–5970, 2019.

[16] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7757–7765, 2020.

[17] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4169–4178, 2019.

[18] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021.

[19] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019.

[20] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(7):1150–1163, 2006.

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edge-Connect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, 2019.

[22] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7 (4):1993–2019, 2014.

[23] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4403–4412, 2019.

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[25] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmío. Video inpainting under constrained camera motion. *IEEE Transactions on Image Processing*, 16(2): 545–553, 2007.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[27] Timothy K Shih, Nick C Tan, Joseph C Tsai, and Hsing-Ying Zhong. Video falsifying by motion interpolation and inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[28] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, 2017.

[30] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[31] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time video completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004.

[32] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8858–8867, 2019.

[33] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019.

[34] Zongben Xu and Jian Sun. Image inpainting by patch propagation using patch sparsity. *IEEE Transactions on Image Processing*, 19(5):1153–1165, 2010.

[35] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[36] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[37] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Proceedings of the European Conference on Computer Vision*, pages 528–543. Springer, 2020.

[38] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In *Proceedings of the ACM International Conference on Multimedia*, page 1939–1947, 2018.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[40] Ruisong Zhang, Weize Quan, Baoyuan Wu, Zhifeng Li, and Dong-Ming Yan. Pixel-wise dense detector for image inpainting. *Computer Graphics Forum*, 39(7):471–482, 2020.