Construction of a Video Inpainting Dataset Based on a Subjective Study

Amine Mohamed Rezki, Mohamed-Chaker Larabi, Amina Serir

Construction of a Video Inpainting Dataset Based on a Subjective Study

Amine Mohamed Rezki[1], Mohamed-Chaker Larabi[2], Amina Serir[1]
[1] LTIR, Faculty of Electrical Engineering, USTHB, Algeria
[2] CNRS, Univ. Poitiers, XLIM, UMR 7252, France

***Abstract*—Video inpainting, the automated process of reconstructing missing or corrupted regions in video sequences, has gained significant attention in the fields of computer vision and image processing in recent years. However, a recurring remark in the literature has been the lack of a dedicated database specifically designed for video inpainting. As a result, existing inpainting studies have relied on locally created videos or datasets primarily intended for other applications. To address this limitation, this paper introduces the first publicly available video inpainting dataset accompanied by subjective scores, which closely aligns with real-world applications. The dataset covers three key inpainting scenarios: video hole completion, object removal, and post-stabilization inpainting. By providing this dataset, our goal is to facilitate the comparison and evaluation of both current and future video inpainting techniques. Moreover, we anticipate that it will serve as a solid foundation for the development of novel video inpainting assessment metrics in the future, thereby encouraging further advancements in this field. It is worth noting that, to the best of our knowledge, there is only one metric dedicated to video inpainting quality assessment, apart from those developed for image inpainting that can potentially be extended to videos. The dataset and the subjective scores are available on this link: https://github.com/rezkimed/VID-SS.**

***Index Terms*—Video inpainting, dataset, quality assessment**

I. Introduction

The absence of a dedicated dataset specifically designed for video inpainting has been a significant limitation in the field. Existing video inpainting works have either relied on locally created videos for result validation [1], [2], utilized videos from previous works for comparative analysis, or referenced datasets primarily intended for segmentation and facial recognition tasks [3]–[8]. Notably, the Davis dataset [9] and YouTube-VOS dataset [10] have been frequently employed in recent video inpainting studies for evaluation and training purposes. While they provide videos and annotated frames that can guide the inpainting process, they are not ideally suited for the inpainting problem due to two main reasons.

Firstly, they offer only one type of mask suitable for a specific video inpainting application, namely, object removal. This limitation restricts their applicability to other inpainting scenarios. Secondly, the videos in these datasets are either short in duration [9] (ranging from 2 to 4 seconds) or possess a low frame rate [10] (only 6 frames per second). The primary reason behind this choice, as stated by the authors of YouTube-VOS [10], was to reduce annotation effort, as they believed it would not significantly impact the segmentation process. However, unlike segmentation, the inpainting process requires intermediate frames to ensure temporal consistency. Therefore, using these databases for objective or subjective inpainting evaluation is not really practical and does not align closely with real-world video inpainting applications.

To address these limitations, we propose the development of a new dedicated dataset designed for video inpainting. This dataset will encompass a diverse range of inpainting scenarios, including video hole completion, object removal, and post-stabilization inpainting. By providing a comprehensive dataset, our aim is to facilitate the evaluation and comparison of existing and future video inpainting techniques. Additionally, we envision that this dataset will serve as a foundation for the development of novel video inpainting assessment metrics, thereby driving further advancements in this field.

The Free-form video inpainting (FVI) dataset [11] was an initial attempt to introduce a publicly available video inpainting dataset. Created by the authors to train and validate their proposed inpainting model, this dataset consists of videos sourced from the YouTube-BoundingBoxes [12] and YouTubeVOS [10] datasets. However, it should be noted that the FVI dataset mainly comprises very short video sequences, with an average frame rate of 36 frames per video and an average duration of 1.16 seconds per video (played back at 24 frames per second).

More recently, Szeto et al. presented the Diagnostic Evaluation of Video Inpainting on Landscapes (DEVIL) dataset [13]. This work aimed to benchmark seven video inpainting methods and assess their performance across five attributes associated with the video sources and applied masks. The DEVIL dataset focuses specifically on landscape videos, characterized by background content without foreground elements. The generated masks in this dataset are tailored for a specific type of inpainting application, which we refer to as video hole completion in this article. It is important to note that the DEVIL dataset primarily emphasizes the comparative evaluation of different inpainting methods, utilizing objective assessment metrics that may not be inherently suitable for the inpainting problem. The dataset's main focus is on objective evaluations rather than subjective assessments.

The primary objective of this work is to address the aforementioned gap by introducing the first publicly available video inpainting dataset accompanied by a comprehensive subjective study, closely aligned with real-world applications. This dataset encompasses three distinct video inpainting scenarios: object deletion, video completion or hole filling, and post-

stabilization video inpainting.

By providing this dataset, we aim to facilitate the evaluation and comparison of existing as well as future video inpainting methods. Furthermore, it opens up new possibilities for enhancing video inpainting quality assessment techniques. The dataset is supplemented with a detailed study of the subjective quality of inpainting, leveraging Mean Opinion Scores (MOS) collected from subjective psycho-physical experiences.

Through this combined objective and subjective evaluation framework, we can gain valuable insights into the effectiveness and perceptual quality of various video inpainting techniques, thus contributing to the advancement of this field.

II. PROPOSED DATASET AND SUBJECTIVE STUDY

A. Dataset collection

The proposed dataset consists of a total of 186 videos, collectively containing 29,358 frames. These videos were generated by combining 52 unique video sequences with 35 distinct masks in a non-regular manner. All videos share the same resolution of 854x480 pixels, were captured at 24 frames per second, and have duration's ranging from 3 to 7 seconds, with an average duration of 6.8 seconds. The video sequences were carefully selected to encompass a wide range of spatial and temporal information as shown by figure 1, and based on SITI values recommended in ITU Recommendation 910 [14]. These video sequences were collected from a variety of sources that offer free-use licenses. Additionally, some videos were sourced from the dataset referenced in [2], [9].

The dataset is categorized into three sections, each representing a specific video inpainting application: object removal, hole video completion, and post-stabilization video inpainting. Each category utilizes a different type of mask, which will be explained in more detail in subsequent sections.

The technique of video inpainting finds wide-ranging applications in real-world scenarios, where it is used to reconstruct damaged or missing parts of videos. These targeted regions may not necessarily correspond to objects within the video; instead, they can take on various forms such as text, lines, or random shapes. The versatility of video inpainting allows for the seamless restoration of these diverse elements, ensuring the visual coherence and continuity of the video content.

1) Video hole completion: In this specific category, we offer a collection of 20 video sequences that encompass a range of characteristics, including diverse camera movements, varying levels of texture complexity, dynamic backgrounds, and scale variations. To facilitate the inpainting process for this category, we provide 7 distinct masks, which have been carefully crafted using Adobe tools. These masks are designed to cover the most commonly encountered shapes in real-world video inpainting applications, ensuring a comprehensive and representative coverage of inpainting scenarios. Figure 2 depicts a number of used masks.

2) Object removal: This particular category consists of 25 video sequences, each paired with a corresponding mask. Each mask precisely annotates a distinct object present within its corresponding video sequence. The selection of these videos

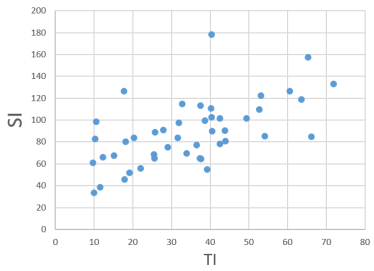


Fig. 1: Spatial (SI) and temporal (TI) information of all videos.

was conducted meticulously, aiming to encompass a wide range of scenarios encountered in real-world inpainting applications. Several crucial factors were taken into consideration during the video selection process, including object size, object movement (both static and dynamic), object occlusion, object speed, scale variations, non-appeared areas, and camera motion. By considering these diverse factors, the dataset aims to provide a comprehensive representation of the challenges and complexities encountered in real video inpainting applications.

3) Post-stabilization video inpainting: In general, digital video stabilization methods employ 2D geometric transformations on video sequences. However, this approach often results in the loss of certain parts along the frame borders. To mitigate this issue, some researchers have resorted to reducing the frame size. Nevertheless, several studies, such as [15], [16], have proposed an alternative solution by utilizing inter-frame inpainting to fill in the missing areas. In other words, these inpainting techniques can serve as a post-processing step in the video stabilization process. It is worth noting that this form of video inpainting differs from previous approaches, as it primarily focuses on addressing the large missing areas located at the borders.

In our dataset, we have addressed this type of inpainting by generating custom masks that resemble the scenarios encountered in video stabilization. The post-stabilization inpainting category encloses a total of seven distinct sequences, each of which accurately reflects the challenges encountered in stabilization scenarios involving walking, driving, or biking. Additionally, three different types of masks have been included: low shake, large shake (with significant missing areas), and quick shake. These masks are essential for conducting comprehensive video inpainting in this context.

B. Test sequences

The proposed dataset was created by integrating three distinct video inpainting methods representative of the major approaches in the literature. The first method adopts a patch-based optimization approach, as described in [17]. The second method employs a flow-based approach, which is detailed in [8]. Lastly, the third method utilizes a deep learning-based approach, as outlined in [5].

We applied the three aforementioned methods to all videos in the dataset, resulting in a total of 558 inpainted videos

(a) (b) (c) (d) (e)

Fig. 2: Frames extracted from spatio-temporal masks used in video completion inpainting. (a) static square in the center, with 8.5% of the frame area, (b) line moving left and right (4%) , (c) static text (subtitling) and moving up text (d) spread-out mask (scribble),(e) splash with random moving and changing in size with speed motion splash.

(186×3). To facilitate result presentation, we assigned a shorthand notation to each method based on the first letter of the first author's last name. Specifically, "H" represents the method proposed by Huang in [8], "N" corresponds to the method introduced by Newson in [17], and "X" denotes the method developed by Xu in [5].

C. Subjective experiment design

In the subjective experiment, a single stimulus paradigm was employed to gather subjective quality ratings. In this approach, participants were presented with the inpainted video sequences without being shown any reference sequences. Their task was to evaluate and rate the subjective quality of the inpainted videos based solely on their visual perception and experience. The rationale behind choosing the single stimulus paradigm for our experiment was rooted in the understanding that reference videos are typically unavailable in real-world inpainting applications. Consider scenarios like object removal or post-stabilization, where the original video containing the object or the unstabilized footage cannot serve as a reference for the inpainted or stabilized sequence, respectively. Hence, by omitting the reference video in our subjective experiment, we aimed to align our evaluation process with the practical constraints and challenges faced in real inpainting scenarios.

In evaluating the quality of the inpainted videos, we utilized the degradation category rating (DCR) method with certain modifications. In our approach, we focus on evaluating the perceived impairment caused by the inpainting process. To facilitate this evaluation, a five-level scale for rating the observed degradation is selected using the following categories: (5) imperceptible, (4)perceptible but not annoying, (3) slightly annoying, (2) annoying, or (1) very annoying.

To ensure a reasonable duration for the subjective evaluation process, we employed a selective approach by focusing on the most informative cases within each inpainting category. Specifically, for the video completion category, we handpicked the first fourteen videos and combined them with masks 1, 2, 3, 5, and 6. We deliberately omitted mask 4 (moving text) due to its similarity to masks 3 and 2 (static text and line moving from left to right). Additionally, to avoid redundancy, we excluded mask 7, which is a faster version of mask 6.
In the case of object removal, we made the decision to exclude the bus, man1, surf, and woman3 videos. These exclusions were motivated either by their usage during training or to prevent repetitive content during evaluation. By employing these selection criteria, we aimed to streamline the subjective experience while still capturing the essential aspects of each inpainting category.
Following the guidelines provided by ITU-T recommendations [14] for subjective tests, we have meticulously prepared four distinct playlists for our evaluation including object removal, video stabilization, and two instances for video hole completion. The object removal playlist consists of 63 videos, resulting from the combination of 21 videos and three different inpainting methods. Similarly, the video stabilization playlist comprises 42 videos obtained from 14 sequences. The video hole completion playlists consist of 101 and 103 videos, respectively, created by combining 68 videos with the three inpainting methods.

To conduct our subjective study, we utilized the "Tobii studio" software, which allows the recording of users' gaze and eases interactions. The former data is not described nor analyzed in this paper. Following the playback of each video, participants were presented with a form containing multiple choices to rate the perceived quality. They were instructed to select the score that best represented their assessment and proceed to the next video by clicking the "next" button using the mouse. All videos were played at a frame rate of 24 fps, except for specific shorter object removal videos, which were played at 15 fps to streamline the subjective evaluation process. By implementing these procedures, we aimed to uphold recognized standards for conducting subjective evaluations while creating a user-friendly and efficient evaluation environment.

D. Conducting the subjective experiment

The subjective experiment was conducted in two different places. The first part was performed in the psycho-physical test room of the XLIM laboratory, set up according to international recommendations [18], with controlled lighting and dark walls to avoid any light reflection on the screen. The used display is the 30" EIZO CG303W. Participants were seated on a fixed chair at a distance of 0.9 m from the screen and were asked to avoid excessive movement so as not to disturb the eye tracking performed by the Tobii eye tracker. The second part was performed at the LTIR lab using a 15,6" laptop and a standard room environment with ambient lighting. The quality assessment of the inpainted videos was conducted individually, with a total of 29 participants (15 at XLIM and 14 at LTIR) of various ages and genders. Prior to the psychophysical

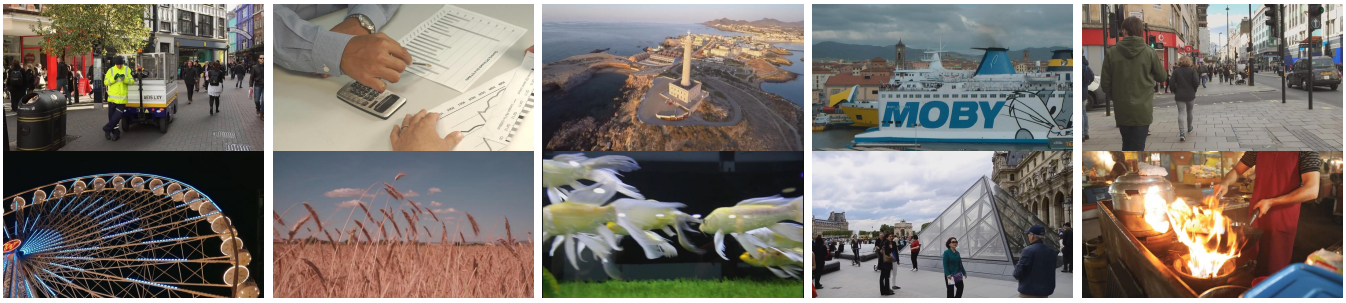


Fig. 3: Some video frames from the proposed dataset

.

evaluation, all participants were checked in terms of visual acuity using Snellen charts and color blindness using Ishihara test.

Participants with correct acuity and color vision, follow a training session to familiarize themselves with the test. This training session helps them become acquainted with the specific tasks and procedures involved in evaluating the quality of inpainted videos. By providing this training, we aim to ensure that participants are well-prepared and have a clear understanding of their role in the evaluation process.

To ensure a comprehensive evaluation, each participant was tasked with viewing and assessing all the videos and playlists. On average, the evaluation process took approximately 52 minutes per participant. To mitigate viewer fatigue, participants were advised to take two breaks, each lasting 10 minutes, during the evaluation session. These scheduled breaks provided participants with an opportunity to rest and refresh their focus.

To enhance the meaningfulness of the study and minimize potential contextual and memory biases, five different orders of video presentation were prepared. The content of successive video presentations was carefully selected to ensure variation and avoid any influence from previous or subsequent videos. This approach aimed to create a balanced and unbiased assessment environment for all participants.

III. Subjective Data Processing and Results Discussion

In this section, we begin by examining the validity of the subjective scores collected, focusing on two crucial aspects: the alignment of individual subject scores with the group mean (outlier detection and rejection) and the overall consistency of the obtained scores. We then proceed to conduct an in-depth analysis of inpainting quality, exploring the impact of key inpainting attributes.

A. Processing of subjective data

1) Screening of observers: After collecting the ratings from all participants, we proceeded to assess the consistency of their judgments by employing the outliers rejection procedure recommended by ITU [18]. To perform this analysis, we calculated the mean score $\bar{\mu}_{j,k}$, the standard deviation $\sigma_{j,k}$, and the Kurtosis coefficient $\beta_{2,j,k}$ for each presentation. The used algorithm enables us to identify participants whose ratings significantly deviate from the mean and standard deviation of the overall ratings. By implementing this procedure, we aim to ensure the reliability and consistency of the collected subjective judgments.

The outlier rejection procedure was separately conducted for each test group, and subsequently, the Mean Opinion Score (MOS) was computed for each stimulus. The correlation between the MOS scores obtained from the first and second tests exhibited a notable agreement of 96%. Therefore, we were able to merge the results from both tests.

In our study, only one subject was identified as an outlier and subsequently excluded. As a result, the MOS value for each presentation was determined based on the remaining 28 scores. Figure 4 illustrates the distribution of MOS scores obtained for all inpainted videos in our dataset. The plot reveals a range of scores spanning from good to poor, indicating the validity of the video selection process for the proposed dataset.

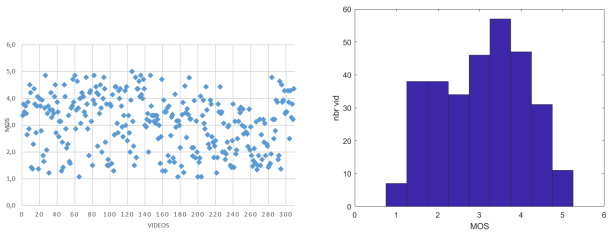


Fig. 4: Histogram of all MOS in the proposed dataset.

2) Global inter-subject consistency: We randomly divided the scores of the 28 participants into two sets. Then, for each set, we calculated the MOS of each video. Then, we calculated the correlation between the MOS of the two sets using Spearman correlation. We repeated this operation 100 times and the overall average correlation is equal to 95%, which validates that the obtained scores are well inter-consistent.

B. Inpainting quality study

As the result of the inpainting process depends on two inputs, the mask, and the input video, we study the influence

of the attributes of each input on the quality of the results, namely the mask study and the video characteristics study. Finally, we discuss the results of post-stabilization inpainting apart from the specificity of these masks and videos.

1) Mask study: After conducting our investigation, we identified three key attributes of masks that can significantly impact the inpainting results: size, motion, and shape. To validate this observation, we employed the Analysis of Variance (ANOVA) statistical technique, commonly used in experimental research to examine the influence of multiple factors on a dependent variable. In our case, the MOS scores served as the dependent variable, while the three mask attributes were treated as independent variables (or factors). We considered both object removal masks and video completion masks in our analysis.

Using the R-studio tool, we computed the results of the ANOVA test. The p-values for the three variables (size, motion, and shape) were found to be $3.39\text{e}{-14}$, $4.65\text{e}{-14}$, $5.55\text{e}{-4}$, respectively. All three p-values were below the significance threshold of $p < 0.05$, indicating that each of the mask attributes has a statistically significant effect on the quality of inpainting. Specifically, the size and motion of the mask exerted a stronger influence on the inpainting results compared to the shape of the mask, as evidenced by their considerably smaller p-values.

To visually represent the relationship between mask size and inpainting quality, we present in fig 5 a scatter plot of MOS scores as a function of mask size. Our analysis revealed that there is a negative correlation between mask size and inpainting quality. Specifically, as the mask size increases, the MOS scores decrease. This can be attributed to the fact that inpainting large masks often leads to more noticeable and distracting artifacts. These artifacts become more prominent because they occupy larger areas of the video. Moreover, reconstructing large areas is inherently more challenging for inpainting algorithms, and many algorithms struggle to generate satisfactory content, particularly when the mask covers a non-isotropic background. Mask movement also has an effect

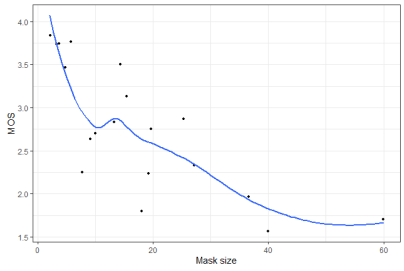


Fig. 5: Scatter plot of MOS versus mask size, the blue line is a local regression for comfortable viewing.

on the inpainting result. Indeed, when the mask moves, the inpainting algorithms can easily fill in the target area in frame t, with the information available in the neighboring frames at the same spatial position that becomes unmasked. Figure 6 confirms that the average score of videos with dynamic masks is higher than that of static masks. The study of the relationship

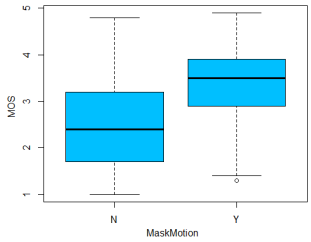


Fig. 6: Impact of mask motion on inpainting quality.

between mask shape and OR inpainting quality, reveals that compacted masks have lower MOS.

2) Video characteristics study: We conducted an investigation to explore the relationship between subjective scores of inpainted videos and five characteristics associated with the video content. We employed the statistical technique ANOVA to analyze the data. The characteristics under consideration were spatial complexity, temporal complexity, camera movement, non-appeared areas, and occlusion. To quantify spatial and temporal complexity, we utilized the spatial and temporal information (SI, TI) metrics as defined in [18]. These metrics provide measures of complexity based on the spatial and temporal properties of the video content. For the remaining three characteristics (camera movement, non-appeared areas, and occlusion), we represented them as Boolean values, where "true" indicates the presence of the characteristic and "false" indicates its absence. The results of the ANOVA test, which assesses the significance of these characteristics on the subjective scores, are presented in Table I.

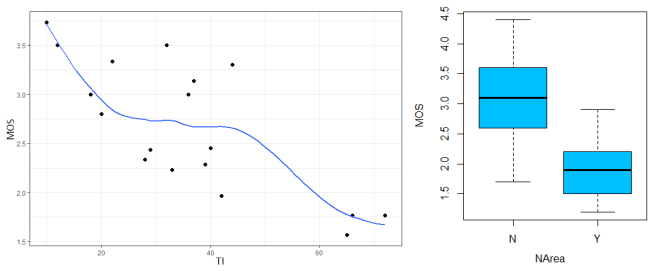
TABLE I: P-value of ANOVA test.

	OR	VC (repeated masks)	ALL (one mask for each video)
TI	**0,00001**	0,88	**0.034**
SI	0,80	0,39	0.185
Narea	**0,00003**	0,07	**0.019**
CameraMot	0,48	0,35	0.569
OCC	0,46	-	0.084

We can see that only the spatial complexity (TI) and non-appeared area (Narea) have a significant impact on inpainting quality since their p-values are less than 0.05. The other attributes SI, camera movement, and occlusion do not affect inpainting quality. This finding is stronger for the OR category and missing for the VC category. This can be explained by the fact that the results for the VC category are heavily skewed by the multiplicity of masks applied. Indeed, each video in the OR category is inpainted with a single mask, while in the VC category, the same video is repeated five times with different masks. The third column confirms this when only one mask is considered for each video to rule out mask interference. and its results have the same significance as those in column 1.

The figures 7a and 7b show the influence of TI,and the non-appeared area on the inpainting quality, respectively.

3) Inpainting for post-stabilization : In the context of post-stabilization, we conducted a study to examine the influence of two factors: the degree of shaking and the speed of shaking.



(a) (b)

Fig. 7: (a) Scatter plot of MOS versus TI, the blue line is a local regression for comfortable, (b) The impact of non-appeared areas on inpainting quality.

Figure 8 illustrates the relationship between these factors and the inpainting results. Interestingly, the results indicate that the speed of shaking has minimal impact on the quality of inpainting. This observation can be explained by the nature of the post-stabilization process.

As the degree of shaking increases, the area lost due to the shaking also increases. Consequently, the difficulty of inpainting the lost area becomes more challenging. In contrast, the shaking speed primarily affects the duration for which the lost area remains uncaptured until it is captured again. Since the shaking speed does not directly influence the size or complexity of the lost area, its impact on the inpainting results is relatively negligible.

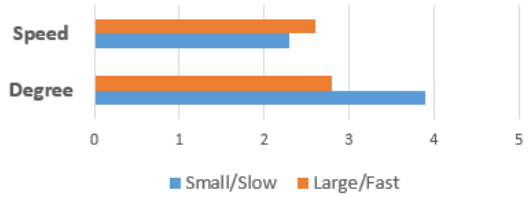


Fig. 8: Average MOS versus degree and speed of shaking.

IV. Conclusion

This article presents a novel contribution in the field of video inpainting by introducing the first dedicated dataset accompanied by a comprehensive subjective experiment. The dataset encompasses three distinct types of real-world inpainting scenarios: video hole completion, object removal, and post-stabilization inpainting. This study identified significant factors that contribute to the overall quality of the inpainted videos. Our analysis revealed that attributes such as mask size, motion, temporal complexity, and non-appeared areas significantly affect the inpainting quality, whereas attributes like mask shape, spatial complexity, camera movement, and occlusion have minimal impact. The natural extension of this work involves analyzing eye-tracking data in relation to inpainting approaches and result quality. By examining the visual attention of observers while interacting with inpainted videos, we can gain insights into how viewers perceive the inpainted content. Finally, the proposed dataset could be seen as an important step that opens the floor for the development of objective assessment models dedicated to inpainting quality.

References

[1] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5792–5801.

[2] M. Granados, K. I. Kim, J. Tompkin, J. Kautz, and C. Theobalt, "Background inpainting for videos with dynamic objects and a free-moving camera," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*. Springer, 2012, pp. 682–695.

[3] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 562–17 571.

[4] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li, "Fuseformer: Fusing fine-grained information in transformers for video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 040–14 049.

[5] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3723–3732.

[6] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf, "Flow-edge guided video completion," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 713–729.

[7] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, and H. Jin, "An internal learning approach to video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2720–2729.

[8] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Temporally coherent completion of dynamic video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016.

[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 724–732.

[10] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 603–619.

[11] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3d gated convolution and temporal patchgan," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9066–9075.

[12] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5296–5305.

[13] R. Szeto and J. J. Corso, "The devil is in the details: A diagnostic evaluation benchmark for video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 054–21 063.

[14] I. Union, "Itu-t recommendation p. 910: Subjective video quality assessmentmethods for multimedia applications," 2008.

[15] V. Voronin, V. Frantc, V. Marchuk, I. Shrayfel, N. Gapon, S. Agaian, and S. Stradanchenko, "Video stabilization using space-time video completion," in *Mobile Multimedia/Image Processing, Security, and Applications 2016*, vol. 9869. SPIE, 2016, pp. 44–52.

[16] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE Transactions on pattern analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1150–1163, 2006.

[17] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *Siam journal on imaging sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[18] I. R. BT, "500-14, methodologies for the subjective assessment of the quality of television images," *Geneva: International Telecommunication Union*, 2019.