2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10647534

BMT-BENCH: A BENCHMARK SPORTS DATASET FOR VIDEO GENERATION

Ziang Shi[⋆] *Yang Xiao*[†] *Da Yan*[§] *Min-Te-Sun*[‖] *Wei-Shinn Ku*[¶] *Bo Hui*[‡]

[⋆] Columbia University [†]Nankai University, [§]Indiana University Bloomington
[¶]Auburn University, [‖]National Central University, [‡]University of Tulsa

ABSTRACT

In recent years, there has been a growing interest among researchers and scholars in the analysis of sports activities, driven by the advancements of machine learning and the increased availability of public data. However, there remains a scarcity of comprehensive sports video datasets that possess the necessary attributes to address various research tasks effectively. We present the "Badminton Benchmark" (BMT-BENCH) to facilitate reproducible machine learning research in the sports domain. This dataset comprises high-quality, high-speed video clips collected from official badminton tournaments involving two team players. The dataset is labeled and unlabeled, catering to different research problems such as video generation and real-time object detection. we feature a baseline system mainly for video generation tasks and provide a thorough evaluation of the challenges posed by the dataset's unique nature. The dataset is publicly accessible at https://drive.google.com/drive/folders/1moYDb8tp5K-VDxPJU3sTorfYE7NnwVpf?usp=sharing and the baseline system is available at https://github.com/ziangshi/BMT_BENCH_baseline_repo.

Index Terms— Benchmark Dataset, Badminton, Sport Dataset, Video Generation, Object Detection

1. INTRODUCTION

Sports analytics has received a lot of attention in the sports domain with advancements in video recording technology and analytics methods. It has been widely used in complex applications of sports such as video summarization [6], highlight generation [9], aid in coaching [10], player's fitness [11], weaknesses and strengths assessment [12], sports robots [13], etc. While analyzing a sequence of motion is crucial for these tasks, sports videos, recorded for live viewing, are commonly collected and leveraged for large-scale data mining. At the same time, machine learning (ML), especially deep learning, has shown impressive performance in various computer vision tasks including object detection and video generation. Deep learning has also been applied to improve the performance of various tasks in the sports domain such as action recognition [20], and player movement prediction [22, 23, 24].

Historically, high-quality and large datasets have played significant roles in advancing research. However, large-scale sports videos with high resolution remain inaccessible to the ML research community. Moreover, manually tagging with labels (e.g., player's action) in the video data is time-consuming, which poses a challenge to supervised learning. We remark that existing sports datasets present issues that may negatively impact future research.

First, due to the substantial costs associated with collecting and fine-grained labeling by domain experts, the availability of public badminton datasets is limited. To the best of the author's knowledge, there are only two sports datasets related to badminton matches: BadmintonDB and ShuttleSet22. BadmintonDB is extremely small and there are only 9 matches recorded. Therefore, it is inadequate for statistical studies on different players. The ShuttleSet22 dataset only provides the locations of players in each stroke. In real matches, a lot of factors (e.g., posture and timing) besides the player's location have an influence on the next play. Moreover, both two datasets are based on stroke level. We argue that fine-grained data is crucial to model playing patterns in sports with machine learning methods. Here, we introduce a video dataset to address the shortcomings of existing datasets. Table 1 shows the difference between our Badminton Benchmark (BMT-BENCH) and previous datasets. Different from BadmintonDB and ShuttleSet22, BMT-BENCH provides videos of badminton matches. The video is manually labeled in fine granularity and shows the interactions between players, shuttlecocks, and racket movements in badminton tournaments. Last but not least, our benchmark is comprehensive in terms of both shot type and players. All play types or strategies are collected to avoid imbalanced data in machine learning. Both male and female players are involved in the dataset. Our dataset can empower the rigorous advancements of machine learning and artificial intelligence in sports.

In addition to building the datasets, we also perform extensive benchmark experiments for the dataset. Recently, significant methodological advances have been made in generative models [15, 16, 8, 7] which have produced promising results in diverse

Dataset	Feature	Level	Size
BadmintonDB	Shot type	Stroke-level	9,671 strokes
ShuttleSet22	Player location, shot type, score	Stroke-level	33,612 strokes
BMT-BENCH	Match Video	Clip-level	2,005 clips

Table 1: Comparison with previous badminton datasets.

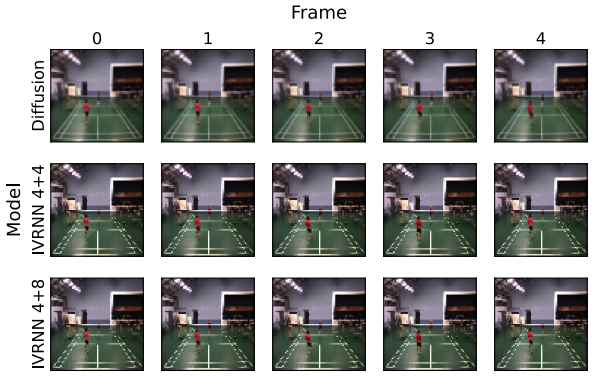
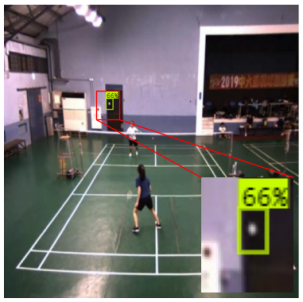


Fig. 1: Video generation

(a) R-CNN



(b) YOLO

Fig. 2: Real-time object detection

domains. For example, a diffusion model has been used to generate high-resolution videos, and the superiority is verified on the prediction of robot pushing motions [17].

We mainly focus on future frame generation tasks based on the previous frames containing information of the player's movement. Figure 3 displays frames in the generated video. Figure 2 showcases the suitable usage of datasets on object detection tasks. Through the experiments and ablation studies, we show the value of the benchmark to compare different models in a fair manner. Also, we highlight research challenges and opportunities provided by our dataset, especially on (1) movements of multiple agents (i.e., players) and (2) improving performance on videos with a dark background.

BMT-BENCH is an open-source and ongoing benchmark. Over time, we plan to label more data and include new deep-learning models in various tasks. The BMT-BENCH project welcomes input from the sports and research community. The contributions of this work are summarized as follows:

- We present the "Badminton Benchmark" (BMT-BENCH) to facilitate reproducible research and support fairly comparing different model designs in the sports domain.
- Different from previous datasets, BMT-BENCH includes videos and fine-grained labels that can empower advancements in machine learning and computer vision.
- We explore the performance of well-known baselines and highlight research challenges which we hope to stimulate the machine learning progress in the sports domain.

The rest of this paper is organized as follows. Section 2 reviews the related works and points out the shortcomings of previous datasets. In Section 3, we describe the data collection and labeling process. In Section 4, we present the experimental result. Section 5 discusses the potential challenges and future works. Finally, we conclude our paper and discuss the open challenges in Section 6.

2. SHORTCOMINGS OF CURRENT BENCHMARKS

Sports datasets play a crucial role in advancing research and analytics in the field of sports science, computer vision, and artificial intelligence. These datasets provide valuable insights into player movements, actions, and sports-specific challenges, facilitating the development of models and technologies to enhance sports performance, coaching, and analysis.

2.1. Related Works

The availability of public badminton datasets has been limited due to the substantial costs associated with collecting and fine-grained labeling by domain experts [4]. In recent times, researchers have taken the initiative to release badminton datasets with the aim of supporting the sports community.

For example, the Shuttlecock dataset by Cartron [1] comprises 8,000 images of shuttlecocks, each resized to 640x640 pixels. This dataset includes annotations for the position of shuttlecocks, making it a valuable resource for training object detection models.

Additionally, the BadmintonDB dataset by Ban et al. [2] offers annotations for rallies, strokes, and match outcomes between two players. It serves as a useful dataset for player-specific match analysis and prediction tasks, particularly at the stroke level. It's worth noting that BadmintonDB primarily features a single matchup, that of Kento Momota and Anthony Sinisuka Ginting, which limits its diversity. Furthermore, BadmintonDB covers matches from 2018 to 2020, which may not capture the latest developments and tactics.

In contrast, ShuttleSet22 [3] collects data from high-ranking matches in 2022, aiming to reflect the state-of-the-art tactic records. It consists of 30,172 strokes within 2,888 rallies in the training set, 1,400 strokes within 450 rallies in the validation set, and 2,040 strokes within 654 rallies in the testing set, complete with detailed stroke-level metadata. To provide a comparative view of the differences between ShuttleSet22 and BadmintonDB, we have summarized the discrepancies in Table 1.

2.2. Sport Dataset

In this context, the badminton dataset stands out as a valuable resource for researchers and enthusiasts interested in the sport of badminton. This dataset, tailored to the specific dynamics of badminton tournaments, offers a unique perspective on the interactions between players, shuttlecocks, and racket movements. It provides a detailed and comprehensive view of the sport's gameplay and captures the nuances that are essential for in-depth analysis.

The DeepSport dataset, primarily designed for ball detection in sports, can offer valuable insights into object detection and tracking in a single-viewpoint scenario. However, its focus on ball interaction with players and poor contrast against the background may not directly align with *badminton*, which primarily involves player actions and racket-shuttle interactions. Nonetheless, the efficient CNN architecture for real-time ball detection in DeepSport could serve as a source of inspiration for object detection in *badminton*, although model adaptation would be necessary to suit the nuances of the *badminton* dataset.

On the other hand, GolfDB, created with the objective of golf swing sequencing and detecting key events in the golf swing, shares similarities with *badminton* in terms of biomechanical analyses and the detection of key events in sports movements. The provision of labeled event frames, bounding box information, player-related data, and club type and view type details in GolfDB sets it apart from other datasets. The approach taken in SwingNet, a lightweight deep neural network introduced in GolfDB, may offer insights into object detection and action recognition in a sports setting, though adaptation would be necessary to cater to the unique dynamics of *badminton*.

ShuttleSet22, a dataset specific to badminton, presents a valuable resource for studying player actions, shuttlecock interactions, and stroke-level metadata within rallies. It is highly relevant to *badminton* analytics and offers a direct parallel to the sport's gameplay. The dataset's stroke-level information and rally-based approach make it particularly suitable for research objectives. Additionally,

the dataset facilitates stroke forecasting, aligning with future frame prediction tasks in *badminton*.

In contrast, *badminton* is player-centric and specifically tailored to the interactions between players, shuttlecocks, and racket movements in the context of *badminton* tournaments. It aligns directly with research objectives for object detection (multi or single) and future frame prediction in the unique setting of *badminton*. *Badminton* offers a specialized perspective on the sport's dynamics and is well-suited for research goals, making it a valuable asset for research in *badminton* analytics.

3. DATA OVERVIEW

We have a total of 2,005 un-annotated clips featuring badminton player rounds, and we've employed Shot-By-Shot (S2) Labeling as our primary labeling method. These clips capture the player's actions from the initiation of ball serving until a player scores. All the images from these clips are stored in PNG file format. Furthermore, these clips are categorized based on the scoring for each player. The player on the farther side of the court is denoted as "red," while the player on the closer side is denoted as "blue."

3.1. Row data collecting

Based on the accuracy of the movements and to ensure a comprehensive dataset, we carefully selected practice game videos of 19 skilled players from National Central University's badminton school team. These players, comprising 15 male and 4 female players, demonstrated a skill level just below that of professional athletes, making their movements and tactical ball skills valuable references for our analysis. The chosen players allowed us to capture a diverse range of badminton actions, including distinct stroke types.

Our cooperative relationship with the school team further facilitated data collection. Not only did it enable us to record practice matches featuring players with proficiency in various badminton actions, but it also provided the opportunity to collect data on less frequent movements or other relevant aspects. It enables the diversity of tactics and movements used during the matches. This cooperative approach ensured the comprehensiveness and depth of our dataset.

3.2. Camera Settings

During data collection, we utilized a high-quality camera from The Imaging Source, specifically the model DFK 37AUX273. This camera was equipped with specific parameters to capture detailed video footage of badminton actions. The camera's resolution was set at 1280*960 pixels to ensure a clear visual representation of the badminton movements. Operating at a frame rate of 60fps, the DFK 37AUX273 captured the fast-paced movements inherent in badminton, resulting in smooth and high-definition video recordings. The XRGB video format was selected to preserve essential color information during data collection.

Strategic camera placement was considered essential. The camera was positioned approximately 2 meters behind the baseline of the badminton court and elevated to a height of approximately 4.5 meters. Tilted at a 30-degree angle, the DFK 37AUX273's lens offered an optimal field of view, reducing distortions and facilitating precise documentation of the various badminton actions. Utilizing this specific model from The Imaging Source allowed for the accurate capture of the intricate movements and techniques of the players, contributing to the comprehensive and detailed dataset.

3.3. Data Preprocessing

We employed a wide-angle lens and positioned the camera high up to simulate the broadcast camera used in formal games. However, the distortion caused by the wide-angle lens can impact the training of the model, leading to inaccuracies in court coordinates, deviations in player and shuttlecock positions, and consequently, errors in the training results. To address this, we utilized OpenCV to calculate the necessary calibration parameters for removing lens distortion. This involved obtaining the 2D coordinates of a chessboard in the 3D world and integrating the information captured from multiple angles of the chessboard image into the calculation.

Before video correction, the originally straight boundary lines of the court appeared significantly curved under the wide-angle lens. This curvature could introduce errors in judging court boundary lines, affecting the accuracy of player and shuttlecock positions. After correction, the degree of curvature was significantly reduced, improving the accuracy of spatial representation compared to the pre-correction state. However, due to the correction, the screen was compressed. Therefore, it was necessary to cut out the necessary parts of the video and reprocess them into the required size after correction.

3.4. Human/Expert labeling

Human labeling was a critical aspect of data processing. Five students, including a member of Central University's badminton team and two from different badminton clubs, actively participated in the labeling process. Their expertise and experience in the sport ensured accurate identification and labeling of the diverse badminton actions captured in the video data.

The labeled data underwent meticulous review by the head coach of Central University's badminton team. The coach diligently examined the annotations, addressing any discrepancies or errors encountered during the labeling process. This rigorous review process ensured the reliability and validity of the labeled data, minimizing potential errors.

3.5. Labeling process

We use Shot-By-Shot (S2) Labeling as our main labeling tool. This tool consists of four parts: recording basic game data, cropping the video for each round, recording the score, and labeling micro-level game data.

The process of labeling a badminton match video involves four stages. The first stage involves obtaining the match video and recording fundamental details about the competition. In the second stage, the beginning, end, and score of each rally are marked. Stage three focuses on pre-processing the video, which includes distortion correction and cropping. In stage four, each shot is analyzed in detail including the type of stroke.

3.6. Data Post-processing

Upon completing the data labeling, our research embarked on the task of segmenting the competition's content, which consisted of two primary areas. First, we addressed video segmenting: the full-length match videos were strategically divided into individual clips corresponding to specific stroke types. This segmentation was executed post data annotation, enhancing its applicability in video categorization or action recognition tasks. The slicing was performed based on ffmpeg, following the use of the Shot-By-Shot (S2) Labeling tool, with segmentation guided by the labeled CSV file generated

by this tool. File naming followed a convention encapsulating essential details like date, time, ID, start and end timestamps, and types of stroke. Second, our data augmentation process filled gaps where some actions were less commonly found in actual gameplay. We enriched the dataset by employing controlled ball-feeding techniques and utilized the previously described camera setup to capture the diversity and complexity of players' actions. These combined efforts serve to create a robust and nuanced dataset, vital for our analytical and recognition tasks within the context of the sport.

4. BADMINTON FUTURE FRAME GENERATION

We provide two sets of data for distinct task to perform. To generate future frames, our primary approach involves evaluating various deep neural network models with the "Unlabeled Badminton" partial dataset, focusing on tasks related to video generation and predicting future frames. The future frames would predict the players' stroke movement based on previous frames of actions. It's important to note that the "Unlabeled Badminton" dataset typically contains various frames lengths for the video clip depicting rounds of badminton players. These frames are carefully chosen to minimize the occurrence of duplicate frames that depict similar player movements. Notably, within this specific dataset, the shuttlecock in the game is intentionally left without labels to maintain the clarity of image frames and prevent the model from receiving misleading information.

The "Labeled Badminton" partial dataset is specifically designed for real-time object detection tasks. Within this dataset, each image frame undergoes a pre-processing stage where the shuttlecock is labeled with bounding box; the storage of label information in the chosen data format depends on the specific needs of the models. Importantly, for this dataset, we exclusively select rounds in which the shuttlecock remains within the camera's field of view throughout the entire match, ensuring continuity in tracking its trajectory.

These two datasets serve different purposes, each tailored to specific tasks. Significantly, the datasets demonstrate their suitability for future frame prediction tasks, owing to their high-speed frame rates. We consider the following representative models below as our baseline systems.

- **SVG** [26]: Stochastic Video Generation with a Learned Prior. The model use recurrent inference networks to estimate the latent distribution for each time step and product future frame prediction in pixel level.
- **IVRNN** [27]: Improved Conditional VRNNs for Video Prediction. This approach involves with an improved VAE model for video predictions. It uses hierarchical latent and a higher capacity likelihood network to improve upon previous VAE approaches on longer temporal space.
- **FutureGAN** [28]: Encoder-decoder GAN model that predictes future frames of a video sequence condition on a sequence of past frames.
- **RetroGAN** [29]: Retrospective Cycle GAN. A unified generative adversarial network for predicting accurate and temporally consistent future frames over time through retrospective cycle constraints
- **RVD** [30]: Residual Video Diffusion. A Denoising diffusion probabilistic model generates future frames by correcting a deterministic next-frame prediction using a stochastic residual acquired by an inverse diffusion process

Model	PSNR	SSIM	LPIPS	FVD
RVD	15.4889	0.6080	0.0659	1129.1934
IVRNN	35.8143	0.9894	0.0062	967.3911
SVG-LP	33.3668	0.9796	0.0167	1663.7945
RetroGAN	25.7551	0.8904	0.0574	1344.5718
FutureGAN	27.8907	0.7305	0.0421	1601.2293

Table 2: Badminton Future Frame Generation Performance

Model	PSNR	SSIM	LPIPS	FVD
RVD(4,8)	29.8227	0.9531	0.0158	1437.3270
RVD(2,10)	29.3458	0.9440	0.06374	1533.4711
IVRNN(4,8)	21.6440	0.8176	0.1736	1934.3771
IVRNN(2,10)	22.0285	0.8312	0.1575	2325.2832

Table 3: Ablation study on training with different context frame sequence length, where(c,p) denotes c context frames and p prediction frames

The models yield average prediction accuracy results assessed through various numerical metrics, including the commonly employed Frechet Video Distance, which compares predicted future frames to actual frames. Additionally, other performance metrics such as Peak Signal-to-Noise Ratio, Structural Similarity Index Measure, and Learned Perceptual Image Path Similarity are employed to evaluate the model's performance in the context of future frame prediction datasets.

4.1. Badminton Future Frame Generation

The results presented in Table 2 provide a comprehensive overview of the performance metrics for the task of future frame prediction, where the models are tasked with predicting 12 future frames based on 8 input frames. The metrics under consideration encompass Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Fréchet Video Distance (FVD).

Among the models, RVD showcases the lowest PSNR and SSIM values, indicative of relatively lower pixel-level fidelity and structural similarity with the ground truth frames. However, it is worth noting that RVD attains a notable FVD score, which suggests that it maintains a certain level of temporal consistency in the generated video sequences. These results highlight a trade-off between pixel-level accuracy and temporal coherence in RVD's predictions.

IVRNN, in contrast, excels in terms of SSIM, achieving significantly higher values. This performance indicates that IVRNN generates future frames with superior pixel-level fidelity and structural similarity when compared to the ground truth frames. Moreover, IVRNN's exceptionally low LPIPS score implies that it attains a remarkable level of perceptual similarity to the reference frames. The model's FVD score denotes a reasonable level of video consistency. IVRNN's strong performance suggests its potential for high-quality future frame prediction tasks.

The results presented in Table 2 showcase the varying strengths and weaknesses of the models in the context of future frame prediction. IVRNN stands out with exceptional PSNR, SSIM, and LPIPS scores, indicating its ability to generate high-quality and perceptually similar future frames.

In Figure 3, we visualize the generated frames from input frames (context). Specifically, 4 frames (Frame 0 to Frame 3) are fed into

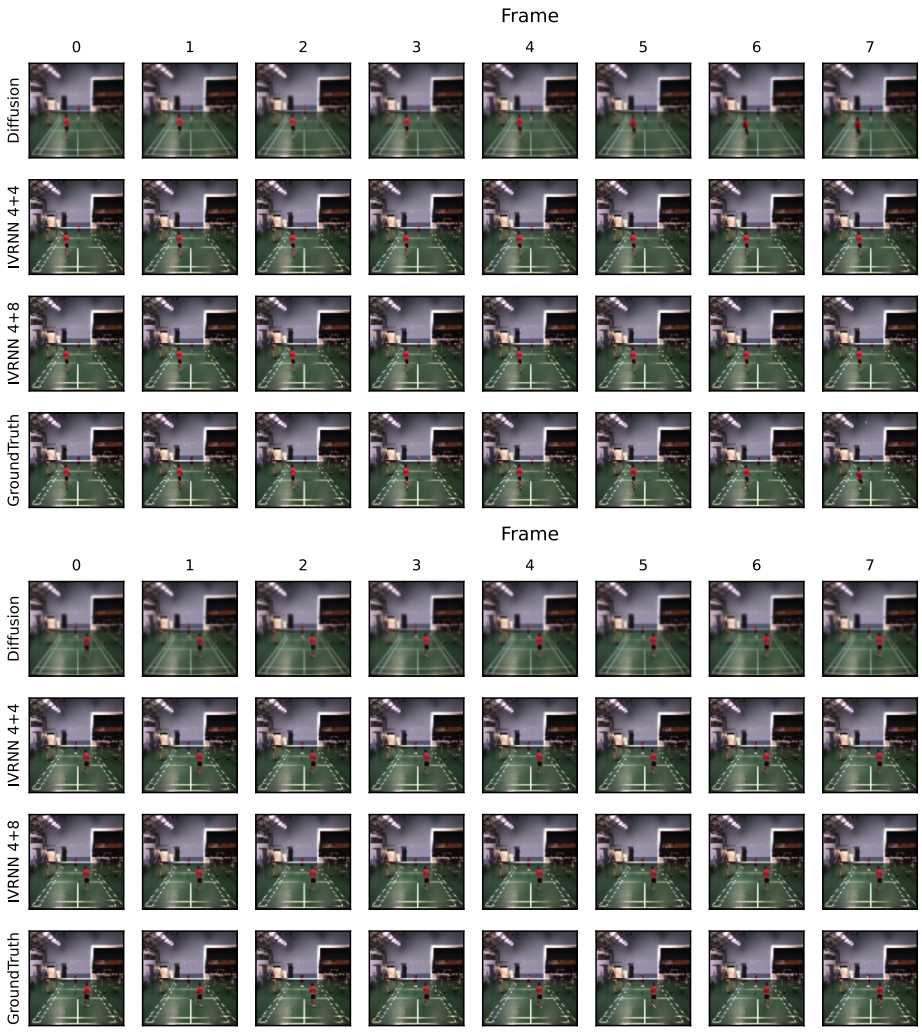


Fig. 3: Predict future 4 frames with 4 input frames

two video generation models (RVD and IVRNN) to generate future 4 frames (Frame 4 to Frame 7). We can observe the generated frames are similar to the ground truth. Compared with IVRNN and ground truth, RVD based on a diffusion model can further improve the quality of images. For example, we can see that the white line in the site is much smoother. We also investigate the performance of IVRNN when the number of frames to be generated increases. Compared with IVRNN 4+4 which predicts 4 future frames with 4 input frames, IVRNN 4+8 predicts 8 future frames and the quality of generated images is slightly lower. It indicates the number of frames to be generated has a negative effect on the training process.

In Figure 4, we show the performance of RVD while the setting is varying. The label "3-5" indicates that we use the previous 4 frames to generate future 4 frames with RVD. Similar to IVRNN, when the number of input frames and the number of frames to be generated increase, the image quality will decrease slightly.

4.2. Ablation Study

We conducted ablation studies to assess the influence of the number of context frames applied during the model training stage. The performance of predictions, based on an alignment of 12 frames, varied depending on the chosen number of context frames and predicted frames. As depicted in Table 3, the RVD model consistently exhibits relatively higher performance, indicating robustness, particularly in short context frame sequences. On the other hand, the performance of the IVRNN model appears to be affected by the number of context frames used during training. While the specific metrics show variability, a consistent observation is that a shorter context frame sequence tends to negatively impact the model's predictive performance

5. OPEN CHALLENGES

We remark that there are several challenges to generate videos in the sports domain, especially for badminton:

1. There are multiple agents (players and shuttlecock) in the videos. It is different from existing benchmarks with only one object in the video. While one player is closer to the camera, it poses a challenge to balance the quality of animation between different agents.
2. Intuitively, the context (i.e., the layout of the field) is important for video generation. It is unknown how to leverage white

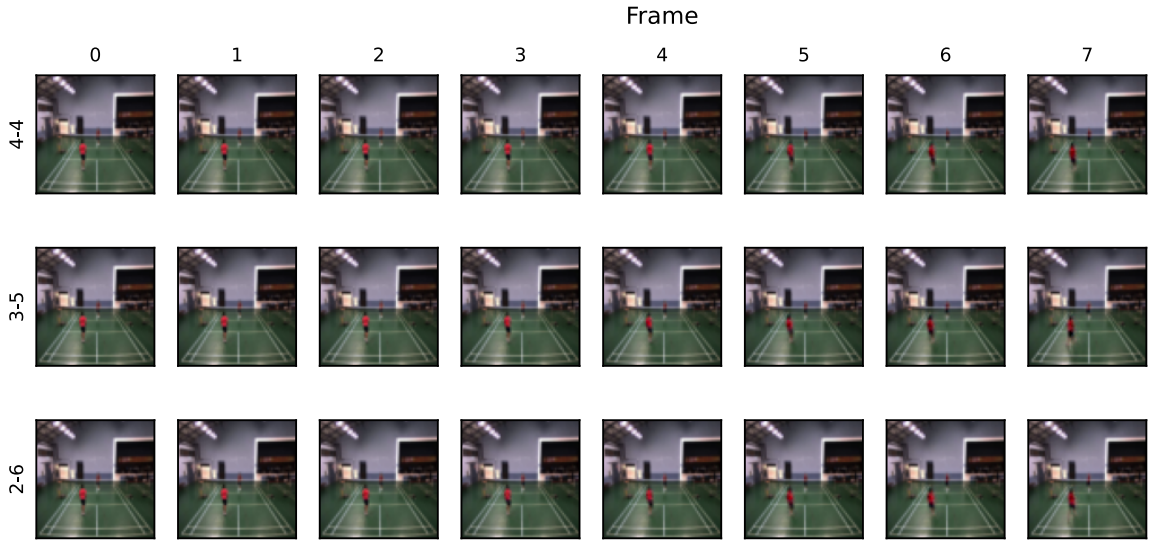


Fig. 4: Varying # of frames in context for RVD

line marks to further improve the quality of generated videos.

3. The dark background and reflection further pose challenges in video generation.

As we look forward, several areas of future work present themselves for further exploration and improvement:

1. Refine the object detection models, leveraging the labeled shuttlecock information in "Labeled Badminton," and explore techniques to improve object tracking.
2. Enhance the quality of future frame predictions by investigating advanced neural network architectures, data augmentation methods, and fine-tuning strategies.
3. Extend the evaluation of the object detection algorithms and frame prediction models, incorporating additional performance metrics, such as PSNR, SSIM, and LPIPS.
4. Investigate the potential for real-time object detection within the context of live badminton matches, considering the constraints and challenges of real-world applications.

6. CONCLUSION

In conclusion, the datasets utilized in this study, the "Unlabeled Badminton" and "Labeled Badminton," have proven to be highly suitable for both object detection and future frame prediction tasks. The "Unlabeled Badminton" dataset, with its high-speed frame rates, lends itself well to the challenges of video generation and predicting future frames. The absence of labeled shuttlecock information maintains the clarity of image frames, making it a valuable resource for future frame prediction. On the other hand, the "Labeled Badminton" dataset is purpose-built for object detection and shuttlecock trajectory prediction, with shuttlecock labeling and continuity in tracking ensuring its suitability for such tasks.

The combination of these datasets enables researchers to explore a wide range of tasks related to badminton, from real-time object detection to video generation. The variety and quality of data available within these datasets offer exciting opportunities for future research.

We discuss the open challenges of generating video in extreme settings such as dark backgrounds and multiple objects. These future directions will not only advance the field of computer vision but also contribute to a deeper understanding of badminton dynamics and the development of intelligent systems for the sport.

Acknowledgment

This research has been funded in part by the U.S. National Science Foundation grants CRII 2348177.

7. REFERENCES

[1] Cartron, J. (2022). Shuttlecock datasets. arXiv:2306.15664. https://arxiv.org/abs/2306.15664

[2] Ban, S., et al. (2022). BadmintonDB: A dataset for player-specific match analysis. arXiv:2306.15664. https://arxiv.org/abs/2306.15664

[3] ShuttleSet22. (2022). Dataset. arXiv:2306.15664. https://arxiv.org/abs/2306.15664

[4] Wang, A. (2022). Challenges in collecting badminton datasets. Journal of Sports Data and Analytics, 1(1), 45-58.

[5] Kwon, H., Shim, W. & Cho, M. Temporal U-Nets for Video Summarization with Scene and Action Recognition. *2019 IEEE/CVF International Conference On Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. pp. 1541-1544 (2019), https://doi.org/10.1109/ICCVW.2019.00192

[6] Vasconcelos, N. & Lippman, A. A Spatiotemporal Motion Model for Video Summarization. *1998 Conference On Computer Vision And Pattern Recognition (CVPR '98), June 23-25, 1998, Santa Barbara, CA, USA*. pp. 361-366 (1998), https://doi.org/10.1109/CVPR.1998.698631

[7] Jiang, C., Hui, B., Liu, B. & Yan, D. Successfully Applying Lottery Ticket Hypothesis to Diffusion Model. *ArXiv Preprint ArXiv:2310.18823*. (2023)

[8] Gao, S., Hui, B. & Li, W. Image Generation of Egyptian Hieroglyphs. *Proceedings Of The 2024 16th International Conference On Machine Learning And Computing*. pp. 389-397 (2024)

[9] Ghanem, B., Kreidieh, M., Farra, M. & Zhang, T. Context-aware learning for automatic sports highlight recognition. *Proceedings Of The 21st International Conference On Pattern Recognition (ICPR2012)*. pp. 1977-1980 (2012)

[10] Mlakar, M. & Lustrek, M. Analyzing tennis game through sensor data with machine learning and multi-objective optimization. *Adjunct Proceedings Of The 2017 ACM International Joint Conference On Pervasive And Ubiquitous Computing And Proceedings Of The 2017 ACM International Symposium On Wearable Computers, UbiComp/ISWC 2017, Maui, HI, USA, September 11-15, 2017*. pp. 153-156 (2017), https://doi.org/10.1145/3123024.3123163

[11] Fieraru, M., Zanfir, M., Pirlea, S., Olaru, V. & Sminchisescu, C. Aifit: Automatic 3d human-interpretable feedback models for fitness training. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 9919-9928 (2021)

[12] Behera, S., Agrawal, P., Awekar, A. & Vedula, V. Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary. *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*. pp. 673-679 (2019), https://doi.org/10.1109/ICMLA.2019.00122

[13] Xu, L. Application analysis of sports robots based on pose recognition and action feature analysis. *Int. J. Syst. Assur. Eng. Manag.*. **14**, 519-528 (2023), https://doi.org/10.1007/s13198-021-01245-1

[14] Rauter, G., Zitzewitz, J., Duschau-Wicke, A., Vallery, H. & Riener, R. A tendon-based parallel robot applied to motor learning in sports. *2010 3rd IEEE RAS & EMBS International Conference On Biomedical Robotics And Biomechatronics*. pp. 82-87 (2010)

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial networks. *Commun. ACM*. **63**, 139-144 (2020), https://doi.org/10.1145/3422622

[16] Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances In Neural Information Processing Systems 33: Annual Conference On Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*. (2020)

[17] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M. & Fleet, D. Video Diffusion Models. *NeurIPS*. (2022)

[18] Ebert, F., Finn, C., Lee, A. & Levine, S. Self-Supervised Visual Planning with Temporal Skip Connections. *1st Annual Conference On Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*. **78** pp. 344-356 (2017), http://proceedings.mlr.press/v78/frederik-ebert17a.html

[19] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference On Computer Vision And Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp. 779-788 (2016), https://doi.org/10.1109/CVPR.2016.91

[20] Hong, J., Fisher, M., Gharbi, M. & Fatahalian, K. Video Pose Distillation for Few-Shot, Fine-Grained Sports Action Recognition. *2021 IEEE/CVF International Conference On Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. pp. 9234-9243 (2021), https://doi.org/10.1109/ICCV48922.2021.00912

[21] Zhu, G., Xu, C., Huang, Q., Gao, W. & Xing, L. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. *Proceedings Of The 14th ACM International Conference On Multimedia, Santa Barbara, CA, USA, October 23-27, 2006*. pp. 431-440 (2006), https://doi.org/10.1145/1180639.1180728

[22] Felsen, P., Agrawal, P. & Malik, J. What will Happen Next? Forecasting Player Moves in Sports Videos. *IEEE International Conference On Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. pp. 3362-3371 (2017), https://doi.org/10.1109/ICCV.2017.362

[23] Kong, Y., Gao, S., Sun, B. & Fu, Y. Action Prediction From Videos via Memorizing Hard-to-Predict Samples. *Proceedings Of The Thirty-Second AAAI Conference On Artificial Intelligence, (AAAI-18), The 30th Innovative Applications Of Artificial Intelligence (IAAI-18), And The 8th AAAI Symposium On Educational Advances In Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. pp. 7000-7007 (2018), https://doi.org/10.1609/aaai.v32i1.12324

[24] Su, S., Hong, J., Shi, J. & Park, H. Predicting Behaviors of Basketball Players from First Person Videos. *2017 IEEE Conference On Computer Vision And Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. pp. 1206-1215 (2017), https://doi.org/10.1109/CVPR.2017.133

[25] Vázquez, J., Liniger, A., Schwarting, W., Rus, D. & Gool, L. Deep Interactive Motion Prediction and Planning: Playing Games with Motion Prediction Models. *CoRR*. **abs/2204.02392** (2022), https://doi.org/10.48550/arXiv.2204.02392

[26] Denton, E. & Fergus, R. Stochastic Video Generation with a Learned Prior. *Proceedings Of The 35th International Conference On Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. **80** pp. 1182-1191 (2018), http://proceedings.mlr.press/v80/denton18a.html

[27] Castrejón, L., Ballas, N. & Courville, A. Improved Conditional VRNNs for Video Prediction. *2019 IEEE/CVF International Conference On Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. pp. 7607-7616 (2019), https://doi.org/10.1109/ICCV.2019.00770

[28] Aigner, S. & Körner, M. FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing Autoencoder GANs. *CoRR*. **abs/1810.01325** (2018), http://arxiv.org/abs/1810.01325

[29] Kwon, Y. & Park, M. Predicting Future Frames Using Retrospective Cycle GAN. *IEEE Conference On Computer Vision And Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. pp. 1811-1820 (2019)

[30] Yang, R., Srivastava, P. & Mandt, S. Diffusion Probabilistic Modeling for Video Generation. *CoRR*. **abs/2203.09481** (2022), https://doi.org/10.48550/arXiv.2203.09481