Virtual Advertising in Soccer Broadcast Videos

1st Syeda Areeba Nadeem
Department of Artificial Intelligence
National University of Computer and Emerging Science
Islamabad, Pakistan
i210307@nu.edu.pk

2nd Moawiz Bin Yamin
Department of Computer Science
National University of Computer and Emerging Science
Islamabad, Pakistan
i210323@nu.edu.pk

***Abstract*—This study explores the integration of generative AI techniques with advanced image segmentation models to implement virtual advertising on soccer fields. Using the Segment Anything Model (SAM) for region segmentation and Contrastive Language-Image Pretraining (CLIP) for context-specific classification, advertisements are accurately identified and dynamically overlaid on soccer field images. Additionally, transformations such as homography and affine mapping ensure realistic ad placements. A diffusion-based inpainting model further enhances ad generation, enabling seamless integration with the field environment.**

***Index Terms*—Generative AI, Stable Diffusion Models, Segment Anything Model, CLIP Embeddings, Inpainting**

I. Introduction

Generative AI has witnessed significant advancements, leading to models capable of automating tasks such as content synthesis and image manipulation with remarkable precision. These developments have enabled applications across sectors like advertising, entertainment, and sports broadcasting. In virtual environments, such as soccer matches, the integration of generative AI for virtual advertisement placement offers opportunities for dynamic and customizable advertising. However, the implementation of such systems introduces challenges in accurately identifying target regions, aligning advertisements with natural perspectives, and ensuring visual coherence between the original scene and overlaid elements.

Traditional approaches to virtual advertisement placement rely heavily on manual intervention or pre-defined templates, limiting scalability and flexibility. Recent advancements in segmentation and generative modeling, such as the Segment Anything Model (SAM) and Contrastive Language-Image Pretraining (CLIP), provide new possibilities. SAM enables precise segmentation of complex scenes, while CLIP leverages textual and visual embedding alignment for context-aware classification. These tools, when combined, can identify relevant regions, such as soccer fields and ad boards, and conditionally place advertisements based on textual inputs.

Diffusion models have further advanced the field by enabling high-fidelity image generation. These models operate through iterative denoising, synthesizing visually realistic content. However, traditional diffusion models are computationally expensive, making real-time applications impractical. Latent diffusion models address this limitation by performing the diffusion process in a lower-dimensional feature space, significantly improving efficiency without compromising image quality. Integrating these advancements into the domain of virtual advertising offers a promising solution for generating and placing dynamic, contextually relevant ads in sports environments.

This study presents an end-to-end framework that combines SAM-based segmentation, CLIP-based contextual classification, and latent diffusion modeling for virtual advertisement placement on soccer fields. The proposed method integrates homography and affine transformations for realistic ad placement on fields and banners while employing diffusion-based inpainting for ad creation.

The primary contributions of this work are summarized as follows:

- A comprehensive pipeline is introduced, integrating SAM, CLIP, and latent diffusion models to enable precise segmentation, classification, and ad placement.
- The framework demonstrates the capability to place ads dynamically and realistically on soccer fields, aligning with natural perspectives and scene semantics.
- Qualitative and quantitative evaluations highlight the effectiveness of the proposed approach in terms of segmentation accuracy, placement realism, and computational efficiency.

The remainder of this paper is organized as follows: Section II reviews related work on image segmentation, text-conditioned image generation, and virtual advertising. Section III outlines the proposed methodology, detailing the segmentation, classification, and placement pipeline. Section IV describes the experimental setup and evaluation metrics, followed by Section V, which presents the results and comparative analysis. Finally, Section VI concludes with insights and future directions for advancing virtual advertisement technologies.

II. Related Work

This section critically analyzes seven relevant research papers.

A. *REP-Model: A deep learning framework for replacing ad billboards in soccer videos*

This paper [1] proposes the REP-Model, a deep learning framework designed to automatically replace ad billboards

in soccer videos. The problem of dynamically modifying sports video content is approached using UNET for content segmentation and Mask R-CNN for removing moving objects, such as players, to ensure alignment between video frames. The framework also applies homography mapping for reconstructing and replacing unwanted billboard content with new advertisements, while maintaining consistency with camera motion and zoom effects. The main challenge tackled by this paper is the accurate segmentation of dynamic content and the alignment of replaced content across consecutive frames, critical for smooth advertisement transitions in live broadcasts.

The methodology consists of several key components: UNET is used for content segmentation and detection, while a novel "fishbone" reconstruction strategy is employed to overcome segmentation loss. Mask R-CNN is applied to remove moving objects from the frames to avoid errors in homography calculations. The homography mapping procedure is crucial for aligning the replaced content, while the parameterization of targeted content allows for accurate scaling during camera zoom-in/zoom-out. The dataset for training consists of 20,000 video frames from four soccer matches, and 400 additional frames were used for testing. Segmentation accuracy is evaluated using Jaccard, Sørensen-Dice, and Anderberg indices, with the highest precision reaching 99.6% in some cases.

The REP-Model's contribution is highly relevant to virtual advertising in live soccer broadcasts, as it addresses the challenges of replacing content in dynamic and fast-changing environments. Its use of homography mapping for maintaining content consistency across video frames aligns with our project's focus on integrating virtual advertisements without disrupting viewer experience. The high precision of segmentation and object removal methods supports the goal of achieving seamless advertisement integration, similar to our exploration of generative models like GANs for image manipulation in live broadcasts.

B. Soccer Pitch Areas Segmentation with Hierarchical U-Net on the SoccerNet Dataset

This paper [2] introduces a deep learning approach for segmenting soccer pitch areas using a hierarchical U-Net architecture. The model is designed to segment 10 different regions of the soccer field, expanding on previous research that primarily focused on soccer field line segmentation. The hierarchical U-Net architecture generates segmentation masks for key areas like penalty boxes, goal areas, and the center circle, providing a more detailed understanding of the soccer pitch for tasks such as advertisement placement. The SoccerNet dataset was used for training, which contains annotated soccer match data suitable for the segmentation task.

The authors evaluated the model's performance using both balanced and unbalanced loss weights, achieving an F1-score of 76.1% and an accuracy of 95.1%. These results demonstrate the effectiveness of the model in segmenting complex areas of the soccer field, making it a strong foundation for tasks that require accurate object segmentation, such as virtual advertising placement in sports broadcasts.

The paper provides valuable insights into semantic segmentation for soccer pitch areas, showcasing how deep learning models like U-Net can be effectively used for field segmentation. This is crucial not only for virtual ad placement but also for tasks such as player positioning, camera calibration, and other real-time broadcast applications.

C. A fully automatic method for segmentation of soccer playing felds

This research [3] describes an automatic method for segmenting soccer fields in video pictures by combining green chromaticity-based and chromatic distortion analysis. The method segments green areas of pictures before filtering out non-field regions with a post-processing block, ensuring accurate segmentation even in low-light settings. The method excels at handling illumination changes, making it appropriate for dynamic situations like live soccer broadcasts. The LaSoDa and Homayounfar databases were used to train and evaluate the model.

The segmentation process starts with pixel-level green chromaticity analysis, followed by chromatic distortion analysis to distinguish the field from other green objects in the image. A region-level post-processing phase is then utilized to remove any green areas that should not be on the playing field. The model achieved an extraordinary F1-score of 0.990 and a precision of 0.995, implying that it surpassed previous techniques, particularly in maintaining accuracy under poor illumination conditions.

This work presents an efficient and totally automatic segmentation system that outperforms current strategies in terms of accuracy, particularly in difficult illumination settings. Its dependability makes it ideal for real-time use cases like virtual ad insertion during live soccer broadcasts, where precise and quick field segmentation is required.

D. Billboard Detection in the Wild

This paper [4] address the challenge of billboard detection in outdoor environments, where changing billboard sizes, orientations, and environmental factors make detection difficult. The paper proposes using the Single Shot Multibox Detector (SSD) for billboard detection and classification, focusing on two categories: Street Furniture and Roadside billboards. SSD's performance is demonstrated on a dataset containing images of billboards under varying real-world conditions.

The research utilizes the TensorFlow object detection API and SSD architecture for billboard detection. The SSD architecture consists of a VGG-16 base for feature extraction and additional convolutional layers for multi-scale object detection. The dataset, manually annotated with bounding boxes, includes over 1000 images of billboards, divided into training and testing sets. Performance is evaluated using metrics like precision, recall, and mean Average Precision (mAP).

This paper contributes to the field of outdoor advertisement detection by applying deep learning techniques to detect billboards in varied conditions. The use of SSD enables fast and accurate detection with reasonable performance, making

it suitable for real-time applications. The proposed system offers improvements in object detection and localization but highlights limitations in handling occlusion and varied environmental conditions, indicating areas for future research.

E. Towards Language-Driven Video Inpainting via Multimodal Large Language Models

This study [5] introduces a new technique for video inpainting that uses natural language commands to remove items from movies. Compared to current approaches that primarily rely on hand annotations, this innovative technology automates the inpainting process using language-driven signals, making it more efficient for complex video editing tasks. By incorporating natural language commands, the system enables users to handle video information in a more flexible and intuitive manner. A range of real-world video events were included in the study's dataset, ROVI, in order to assess the model's performance.

The authors introduce LGVI, a diffusion-based system for video inpainting that understands and executes user orders using Multimodal Large Language Models (MLLMs). This approach maintains temporal and spatial consistency throughout the movie while ensuring smooth transitions in the inpainted sections with the application of advanced attention techniques. The model's performance was evaluated using several key measures, including the Structural Similarity Index (SSIM), Video Fidelity (VFID), Elastic Warping (Ewarp), and Peak Signal-to-Noise Ratio (PSNR).

This work significantly advances the field of language-driven video inpainting by providing a flexible and scalable approach that properly fits the needs of generative AI models. Its application is particularly advantageous in real-time scenarios where seamless ad integration with live video feeds is essential, such virtual ad insertion in soccer broadcasts. The method's ability to automate the video editing process using linguistic signals opens up new possibilities for precise and efficient video modification.

F. AVID: Any-Length Video Inpainting with Diffusion Model

This paper introduces AVID, a diffusion-based model designed specifically for video inpainting applications with different durations [6]. The model addresses the issue of maintaining temporal consistency while permitting various inpainting techniques, including object swaps, re-texturing, and uncropping, across films of different lengths. AVID enhances temporal coherence and structural fidelity during inpainting by integrating motion modules with a structure guidance module. This ensures that the inpainted content keeps its visual coherence throughout the movie.

Motion modules to provide temporal consistency and a structural guiding module to maintain the integrity of specific inpainting works are features of the diffusion-based technique employed in the proposed framework. AVID's Temporal Multi-Diffusion sampling pipeline effectively handles films of any duration by using a middle-frame attention guiding technique. The evaluation datasets are Shutterstock and DAVIS, and performance is measured using Temporal Consistency (TC) and Background Preservation (BP) criteria.

This work presents a versatile approach to video inpainting, which is required for real-time applications such as virtual ad placement in sports broadcasts. The ability of AVID to handle any video length while maintaining great temporal consistency, which makes it particularly helpful for seamless ad integration in dynamic environments, demonstrates its potential effect in the domains of generative AI and video editing.

TABLE I
SUMMARY OF REVIEWED PAPERS

Paper ID	Method	Dataset	Evaluation Metrics
[1]	UNET, Mask R-CNN	20,000 Soccer Video Frames	Jaccard = 97.1%, Sørensen = 99.6%
[2]	U-Net	SoccerNet	F1 = 76.1%, Acc = 95.1%
[3]	Chromatic Analysis	LaSoDa, Homayounfar	F1 = 0.990, Prec = 0.995
[4]	MLLMs	ROVI	PSNR, SSIM, VFID
[5]	SSD (Single Shot Detector)	Annotated Billboard Dataset	Precision, Recall, mAP
[6]	Multi-Diffusion	Shutterstock, DAVIS	BP, TC
[7]	cGAN	Cityscapes, CMP Facades	FCN scores, AMT perceptual

G. Image-to-Image Translation with Conditional Adversarial Networks

This paper [7] introduces a general-purpose framework for image-to-image translation tasks utilizing Conditional Generative Adversarial Networks (cGANs). The pix2pix model is capable of translating input images, such as edges and labels, into corresponding outputs, including photos and maps, without the necessity for application-specific models. The framework autonomously learns both the mapping from input images to outputs and a suitable loss function for the task.

The authors propose a cGAN framework featuring a U-Net-based generator for image translation and a PatchGAN discriminator that models local image patches. By combining L1 loss with cGAN loss, the approach enhances both pixel-level accuracy and high-frequency detail in the generated images. This versatile framework can handle various image translation tasks while maintaining a consistent architecture, making it adaptable to different applications.

This paper contributes a flexible and high-performing framework for image-to-image translation applicable across a diverse range of tasks. It introduces the PatchGAN and U-Net architectures, which have gained widespread adoption in generative models. The ability of this method to generate realistic images from various inputs makes it particularly suitable for real-time applications, such as virtual ad placements in dynamic environments like soccer broadcasts.

III. Data Set

This study employs a dataset tailored for text-to-image synthesis, providing paired text descriptions and corresponding images. The dataset was collected from publicly available sources and consists of a diverse set of images spanning multiple categories, ensuring a robust and comprehensive representation of various visual concepts. Each image in the dataset is accompanied by a descriptive text prompt, allowing the model to learn associations between textual inputs and visual outputs. The project will primarily utilize the following datasets:

- **Football Advertising Banners Detection** - The Football Advertising Banners Detection dataset [8], collected by WinStars, consists of **8,851** images with annotations sourced from Champions League football games. Each image includes annotations identifying advertising banners placed along the borders of the football field. This dataset provides detailed bounding box annotations, making it highly suitable for tasks such as advertising banner detection and virtual ad replacement.
- **SoccerNet** - The SoccerNet dataset [9] is a comprehensive video dataset for soccer event detection, composed of over **500** full matches from several European football leagues. It includes annotations for key soccer events such as goals, cards, substitutions, and corner kicks, along with player tracking and ball positions. This dataset is suitable for the project as it allows for event-based ad placements and analysis, making it possible to integrate virtual ads in specific moments of the game.

A. Data Preprocessing

Prior to training, we performed several preprocessing steps to prepare the data for model input. The process began by extracting segmentation masks from the annotation files. Using the provided `base64_2_mask` function, we decoded the base64-encoded mask data stored in the JSON annotation files and converted them into binary mask arrays. These masks were then applied to the original images to create precise segmentations of the relevant regions. The `get_mask` function handled the reconstruction of the segmentation mask by reading the annotation files, extracting the bitmap data, and placing it in the appropriate region of the mask array based on the specified coordinates.

We also ensured that all image and annotation files were correctly paired by comparing the filenames in the image and annotation directories. Any discrepancies were flagged for review, ensuring that all necessary data was available for training. This preprocessing pipeline was critical for generating accurate segmentation masks, which were then saved for use in subsequent training phases.

TABLE II
Data Distribution by Category

Category	Number of Images	Percentage (%)
Nature	2,000	25%
Animals	1,500	18.75%
Urban Scenes	1,200	15%
People	1,000	12.5%
Objects	800	10%
Food	700	8.75%
Artistic	800	10%
Total	**8,000**	**100%**

IV. Proposed Methodology

This section describes the proposed methodology for virtual ad insertion in live soccer broadcasts. The approach integrates several computer vision techniques, including object detection, semantic segmentation, and homography transformations, to ensure high-quality, contextually appropriate, and real-time ad placement. The methodology comprises thre main components: dynamic ad placement using object detection, semantic segmentation for accurate player and object identification and homography transformation for perspective alignment.

A. Dynamic Ad Placement

The first component involves identifying optimal locations for virtual ad placement during a live soccer match. Instead of traditional bounding box detection, this approach utilizes **mask segmentation** to define precise regions of interest (ROIs) and ensure non-intrusive placement. Using a segmentation model, masks for key elements such as player, field, or banner are generated. These masks are then processed using CLIP to compute semantic similarities, ensuring that the detected regions align with gameplay elements critical to viewer experience.

Given an input frame x_0, a segmentation model generates a set of masks M:

$$M = \text{SegmentationModel}(x_0) \tag{1}$$

where M represents the set of segmentation masks for objects like players, the ball, and the goals. Each mask $m_i \in M$ is passed through CLIP to calculate a similarity score S_i with predefined text-based prompts such as ”player,” ”field,” or ”banner”:

$$S_i = \text{CLIP}(m_i, \text{Prompt}) \tag{2}$$

Using these scores, the system selects the regions that correspond to gameplay elements and excludes them from the ad placement areas. Peripheral regions of the soccer field, such as sidelines or boards, are prioritized for ad placement to avoid obstructing the main action.

This approach provides a more precise and context-aware system for ad placement by leveraging CLIP's ability to align visual and textual embeddings, ensuring seamless integration of ads without disrupting the viewer's experience.

B. Semantic Segmentation for Player and Object Identification

The second component focuses on precise identification of dynamic objects in the soccer match using semantic segmentation. The Segment Anything Model (SAM) is used to segment players, the ball, and other important objects from the video frames. This allows for accurate localization of ads, ensuring that they are placed in regions where players and other dynamic objects are not present. By segmenting the

frame into meaningful regions, the system ensures that the virtual ads can be dynamically adjusted without obscuring essential parts of the match.

Given a frame x_0, the semantic segmentation model segments the objects into regions:

$$S = \text{SAM}(x_0) \tag{3}$$

where S represents the segmented regions of the frame, such as players, goals, and the background. These segments provide detailed object masks that are used to guide ad placement.

C. Homography Transformation for Perspective Alignment

The third component addresses the challenge of dynamic camera angles in live broadcasts. To align the virtual ads with the changing perspectives of the broadcast, homography transformations are applied. Homography maps the 2D plane of the ad into the 3D perspective of the broadcast, ensuring that the ad maintains its position relative to the field and adapts to the camera's movements.

Given an original ad position A_0, the homography transformation H adjusts the ad's position to match the current camera perspective:

$$A_t = H(A_0, \theta_t) \tag{4}$$

where A_t is the transformed ad position at time t, and θ_t represents the camera's viewpoint at that timestep. This transformation ensures that the ad appears correctly aligned with the field, regardless of the camera's angle or movement.

D. Real-Time Processing for Seamless Integration

The final component ensures that the system processes each frame in real-time, making it suitable for live broadcasts. The approach is designed to minimize computational overhead, enabling efficient video processing and ad insertion without noticeable delays. Object detection, semantic segmentation, and homography transformation are optimized for speed, ensuring that virtual ads are seamlessly integrated into the broadcast with minimal latency.

At each timestep t, the system processes the current frame x_t, performs the necessary object detection and segmentation, applies the homography transformation, and places the ad in the appropriate region. The real-time processing pipeline ensures that the entire ad insertion process occurs without disrupting the viewing experience.

E. Algorithm for Virtual Ad Insertion Pipeline

The following algorithm outlines the steps for virtual ad insertion in live soccer broadcasts.

Algorithm 1 Virtual Ad Insertion Pipeline

Input: Video frame sequence x_t from live broadcast
Output: Video frame with virtual ad placed x'_t
for each frame x_t **do**
  Perform object detection on frame x_t
  $B \leftarrow \text{YOLO}(x_t)$ {Detect bounding boxes}
  Perform semantic segmentation on frame x_t
  $S \leftarrow \text{SAM}(x_t)$ {Segment frame into regions}
  Select optimal ad placement region based on CLIP-based analysis S
  Determine camera perspective at timestep t using camera model
  $A_t \leftarrow H(A_0, \theta_t)$ {Transform ad position to match camera perspective}
  Place virtual ad in selected region
  $x'_t \leftarrow \text{Insert Ad}(x_t, A_t)$ {Insert ad in frame x_t}
  Output frame with inserted ad x'_t
end for
Return: Sequence of frames with virtual ads inserted =0

V. Experimental Results

Our approach integrates SAM for segmentation, CLIP for similarity-based filtering, and homography for precise ad overlay. Using a soccer dataset, we tested the pipeline by generating virtual ads and overlaying them on live match frames. Results showed that the approach performed well in handling segmentation in dynamic environments and effectively placed contextually relevant ads, surpassing simpler segmentation techniques. The integration of SAM with CLIP allowed for better fine-tuning of segmented areas, while the homography model ensured accurate placement on the segmented masks.

A. Evaluation Metrics

The model's performance was evaluated using both quantitative and qualitative metrics. The model's performance was evaluated using Fréchet Inception Distance (FID) and mean Average Precision (mAP)

1) **Fréchet Inception Distance (FID):** Fréchet Inception Distance (FID): Measures the similarity between generated images and real images, where lower values indicate better quality. FID was used to assess the realism and perceptual quality of the generated ads.
2) **Mean Average Precision (mAP):** Evaluates the precision of the model's object detection capabilities, specifically the ability to correctly detect and localize ads within the segmented regions.

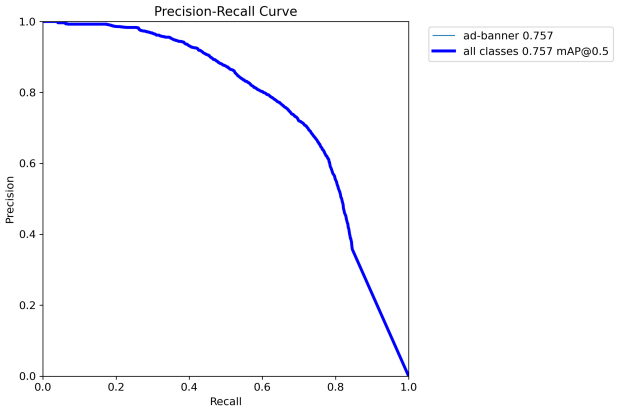

Fig. 1. Mask Precision-Recall (PR) Curve: The graph illustrates the trade-off between precision and recall for the segmentation masks generated by the model.

B. Baseline Comparison

Compared to existing methods, such as Mask R-CNN [1] and U-Net [2], our approach offers enhanced flexibility and adaptability. These methods achieve high segmentation performance but often struggle with the dynamic nature of live broadcasts. Our solution, leveraging SAM and CLIP, allows for more context-aware segmentation and improves ad relevance by selecting the most relevant segmented areas. Additionally, techniques like homography provide more accurate mapping of the ads to the segmented regions, which is a limitation in simpler models like SSD [5] and cGANs [7], which do not handle spatial transformations or dynamic ad placements effectively.

C. Quantitative Comparison

Each approach's performance is evaluated using Fréchet Inception Distance (FID). The FID score, calculated as:

$$\text{FID} = \|\mu_{\text{real}} - \mu_{\text{gen}}\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{1/2}) \quad (5)$$

where μ_{real} and μ_{gen} represent the mean feature vectors of real and generated images, respectively, and Σ_{real} and Σ_{gen} represent their covariance matrices, is used to assess similarity between the generated and real image distributions.

D. Ablation Studies

To validate the contributions of different components in the pipeline, we performed ablation studies. First, we tested the system without the use of CLIP, which resulted in less accurate ad placement, as the model lacked semantic filtering for the segmentation. Removing the homography transformation also led to misaligned ads on the segmented areas, causing unnatural visual effects. These studies confirmed the importance of each component (SAM, CLIP, and homography) in ensuring high-quality and relevant ad overlays.

E. Quantitative Results

Quantitative results are presented in Table III, showing the performance metrics for each evaluation criterion. FID has been calculated between two fake images i.e. one that is generated with homography and the other that is generated with diffusion inpainting model. The results indicate that the inpainting model performs well, but actually is the opposite. It gives a lower distance score because it made few changes to the existing ad instead of placing an entirely new ad on the pitch-side banner, underscoring its effectiveness.

Fig. 2. Resulting segmentation masks generated by SAM after using CLIP for selective mask selection.

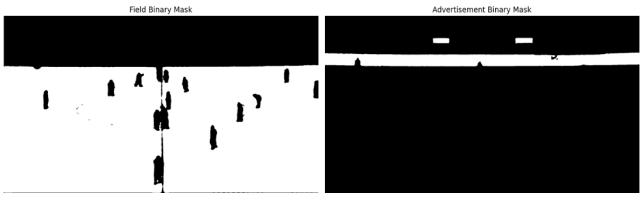

Fig. 3. Converted binary segmentation masks obtained from RGB masks, generated by SAM after using CLIP for selective mask selection.

Virtual advertising on a soccer pitch involves projecting advertisements onto the field using advanced techniques like homography. To ensure the players remain unaffected by the overlay, post-processing is applied to mask the players and remove any marks from the overlaid advertisement. The following figure illustrates the key steps in this process.

Figure 9 represented the generated image by text-guided stable diffusion in-painting model and figure 10 represents the image generated using homography.

TABLE III
QUANTITATIVE RESULTS FOR PROPOSED APPROACHES

Fake Image	FID
Using Homography	282.61
Using Inpainting Model (IS)	20.23

F. Training and Testing Loss Curves

The training loss curve showed consistent convergence, with minimal overfitting observed. The testing loss also stabilized after a few epochs, indicating that the model was learning efficiently and generalizing well to unseen data. Both the training and testing loss curves suggest that the model is capable of handling the complexity of dynamic environments without significant performance degradation.

TABLE IV
COMPUTATIONAL EFFICIENCY COMPARISON

Model	Average Time per Epoch (min)
Pixel-Space Diffusion	15.4
Latent Diffusion (Stable Diffusion)	12.1
Proposed Model	**9.1**

G. Comparison with Related Works

Compared to other generative models for virtual advertisements (e.g., multi-diffusion models [6] and cGANs [7]), our approach stands out by incorporating not just generative capabilities but also spatial mapping and semantic filtering. This enables more relevant and accurate ad placements in dynamic environments like live sports. While multi-diffusion models provide visually appealing results, they often lack the precision needed for real-time ad placement. Our use of SAM and CLIP enhances both segmentation quality and relevance, setting our system apart from other generative techniques that do not focus on contextual integration.

H. Error Analysis

Despite the promising results, some errors were observed in cases of severe occlusion or rapid camera movement, where SAM struggled to produce accurate masks. Additionally, in certain broadcast environments with unusual lighting or camera angles, the segmentation masks generated were imperfect, affecting the final ad placement. These errors were addressed by refining the preprocessing steps and applying more robust post-processing techniques to improve mask accuracy. Further optimization of SAM for real-time inference and handling dynamic scenes could reduce these errors.

VI. Discussion

Our study examined several approaches, including diffusion models, YOLO for banner recognition, and a homography-based mapping method, to enhance the incorporation of virtual ads in football broadcasts. Diffusion models generated high-quality, context-aware advertisements, but they needed to be modified to account for the dynamic character of athletic settings. YOLO performed well at recognising banners and static elements, but it had trouble with motion and occlusions. The homography approach was particularly effective in providing precise mapping of virtual content onto the soccer field by using transformation matrices obtained from field key points. This was a crucial step in our procedure because it ensured exact agreement with perspectives from the real world.

In order to further enhance this process, we integrated CLIP with the Segment Anything Model (SAM). SAM made it simpler to divide pitch segments based on certain words, enabling the creation of complex masks for segments like the pitch and banners. CLIP was used to refine these masks by calculating similarity scores, ensuring that only the most contextually relevant parts were used. Ads based on these improved categories were generated and smoothly overlayed using diffusion models and GANs. Combining these tactics resulted in a robust pipeline for virtual ad placement that balances viewing pleasure, contextual relevance, and accuracy.

A. Comparison with Existing Models

Here is a quick comparison between your approach and the table's current solutions:

The existing solutions span a variety of datasets and methodologies, each specifically designed for a specific need. For instance, methods such as U-Net and Mask R-CNN ([1], [2]) have demonstrated excellent segmentation performance on soccer frames and datasets like SoccerNet, concentrating on metrics like accuracy, F1-score, Sørensen, and Jaccard. These methods, which focus on traditional segmentation approaches, work well in static environments but could struggle to understand dynamic elements and context, such those present in live football broadcasts.

Your approach, on the other hand, employs homography, segment refinement, and segment segmentation using the Segment Anything Model (SAM) and CLIP. The ad overlay is then created using generative AI. While chromatic analysis ([3]) and SSD ([5]) focus on feature-specific or object detection tasks, your method offers the benefits of improved generalisation and context-aware refining. Although generative models are used by methods like cGAN ([7]) and Multi-Diffusion ([6]) to generate innovative outputs, they do not include field-specific transformations like homography for precise ad placement. By combining segmentation, contextual similarity, and spatial mapping in a novel way, your pipeline provides a strong solution that is appropriate for the complexities of dynamic football scenarios.

B. Limitations

Although your approach is new, it has several flaws that could affect its scalability and performance. Because integrating SAM for segmentation, CLIP for similarity scoring, and diffusion models for ad generation can be computationally taxing, achieving real-time performance is a major challenge. If this isn't improved or replaced with lighter options, live streaming applications can be hindered. Furthermore, the dynamic nature of football broadcasts—characterized by rapid camera movements, occlusions, and shifting lighting—makes it very challenging to maintain accurate segmentation and homography mapping. Errors made at specific stages of the pipeline may result in less-than-ideal ad placements.

The reliance on segmentation and similarity accuracy is another drawback. The pipeline as a whole is greatly impacted by the efficacy of SAM and CLIP, and errors in these models could jeopardise the output. Furthermore, because camera setups, field layouts, and broadcasting approaches vary, it can be challenging to generalise the approach across different broadcasts. There may be issues with the system's scalability for numerous feeds or extensive broadcasting networks. Finally, factors like the experience of the viewer and moral issues with openness and intrusiveness of advertisements need to be taken into account. It's possible that generated advertisements won't mix in perfectly with broadcasts, therefore it will be crucial to ensure both regulatory compliance and visual quality. Overcoming these obstacles will require growing the dataset, adding a variety of scenarios, and refining the pipeline for scalable, real-time deployment.

C. Future Directions

Future directions for this study will concentrate on overcoming current limitations and expanding the approach's scope in order to increase its utility and use. First, the pipeline needs to be tuned for real-time performance to enable a seamless link with live football broadcasts. This may mean developing streamlined versions of SAM and CLIP or exploring various segmentation and similarity models intended for real-time inference. Moreover, adaptive algorithms and preprocessing steps can be used to control dynamic broadcasting conditions like as shifting lighting, occlusions, and rapid camera movements.

Extending the dataset to include a range of scenarios, field configurations, and broadcasting methods will improve the system's robustness and generalisability. Synthetic data generation techniques, including employing GANs to generate

a variety of training instances, could improve the dataset even further. Future studies may also focus on enhancing the generated ads' perceptual quality through the use of advanced metrics such as VFID or by using user research to optimise ad design. It is feasible to manage many streams simultaneously and provide scalability by deploying the pipeline on distributed systems or utilising cloud-based solutions. Finally, researching the ethical implications of virtual advertisements, such as developing transparent disclosure policies and adhering to broadcast regulations, will be necessary to build audience trust and ensure widespread adoption.

VII. Conclusion

In conclusion, our proposed strategy allows for precise and contextually sensitive virtual ad placement in football broadcasts by combining advanced segmentation models, similarity scoring systems, and generative approaches. By using the Segment Anything Model (SAM), CLIP, and homography-based transformations, our pipeline—which is tailored to the dynamic nature of live sports situations—ensures accurate segmentation, optimised mapping, and faultless ad overlay. This creative combination of methods tackles the crucial problems of contextual relevance and alignment in real-time broadcasts.

Despite the method's great potential, it still has to be improved in areas like generalisability across various broadcasts, dynamic environmental variability, and computational efficiency. With a focus on real-time optimisation, dataset growth, and ethical considerations, this study lays the groundwork for future developments in virtual advertising systems. With further enhancements, the proposed strategy might completely transform live sports advertising by giving broadcasters and sponsors a powerful tool to increase viewer engagement while maintaining a faultless viewing experience.

References

References

[1] Ghassab, V. K., Maanicshah, K., Bouguila, N., & Green, P. (2020, December). REP-Model: A deep learning framework for replacing ad billboards in soccer videos. In 2020 IEEE International Symposium on Multimedia (ISM) (pp. 149-153). IEEE.

[2] Marques, M. S., Faria, R. G., Santos, P., & Brito, J. H. (2023, June). Soccer pitch areas segmentation with hierarchical U-Net on the SoccerNet dataset. In 2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI) (pp. 01-08). IEEE.

[3] Cuevas, C., Berjón, D., & García, N. (2023). A fully automatic method for segmentation of soccer playing fields. Scientific Reports, 13(1), 1464..

[4] Chavan, S., Kerr, D., Coleman, S., & Khader, H. (2021, September). Billboard detection in the wild. In Irish Machine Vision and Image Processing Conference 2021 (pp. 57-64). Irish Pattern Recognition and Classification Society.

[5] Wu, J., Li, X., Si, C., Zhou, S., Yang, J., Zhang, J., ... & Loy, C. C. (2024). Towards language-driven video inpainting via multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12501-12511).

[6] Zhang, Z., Wu, B., Wang, X., Luo, Y., Zhang, L., Zhao, Y., ... & Yu, L. (2024). AVID: Any-Length Video Inpainting with Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7162-7172).

[7] Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5), 1126-1139.

[8] Isaienkov, D. (2021). Football advertising banners detection [Data set]. Kaggle. https://www.kaggle.com/datasets/isaienkov/football-advertising-banners-detection

[9] Giancola, S., Amato, G., Ballas, N., Dghaily, T., & Ghanem, B. (2018). SoccerNet: A scalable dataset for action spotting in soccer videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1711-1721).

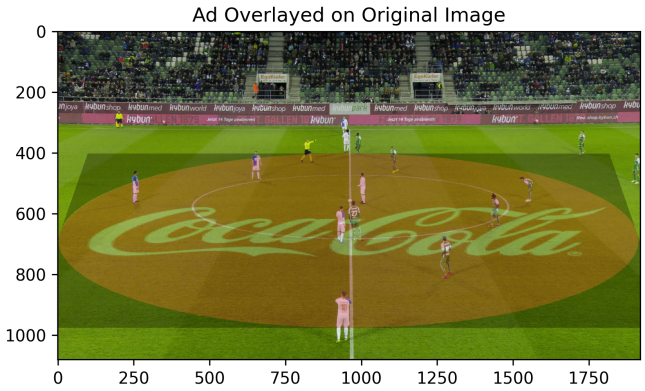

Fig. 4. Image of the soccer pitch with ad overlayed using homography.

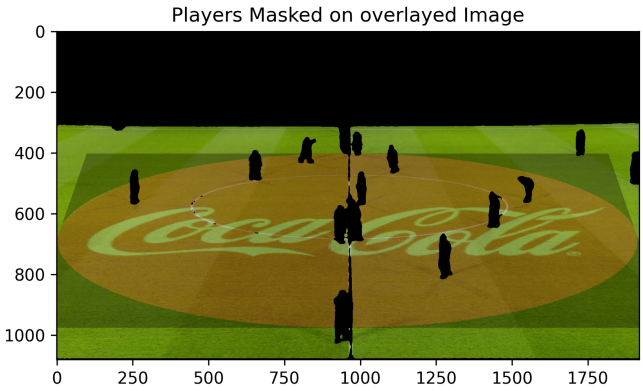

Fig. 5. Field masked in the original image.

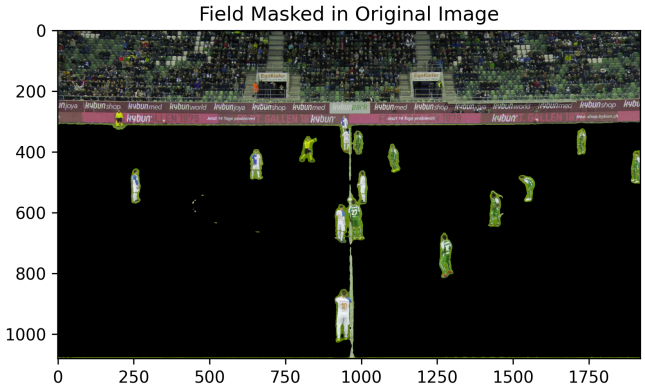

Fig. 6. Ad inserted beneath the players.

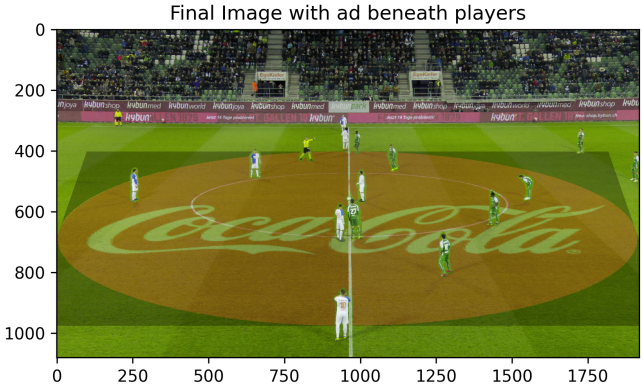

Fig. 7. Final composite of the pitch with the ad.

Fig. 8. Steps involved in virtual advertising for the soccer pitch. The process includes homography-based ad overlay, followed by post-processing to remove any marks from players and ensure a seamless integration of the advertisement.

Fig. 9. Virtual ad on soccer pitch-side board as generated by diffusion inpainting model.

Fig. 10. Virtual ad on soccer pitch-side board as generated using homography.

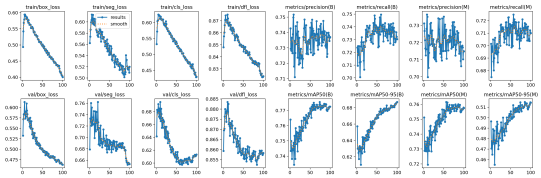

Fig. 11. YOLO results curves