Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.3, May : 2025

CONTEXT-AWARE MUSIC EMBEDDING IN SILENT VIDEOS LEVERAGING TRANSFORMER ARCHITECTURES: A REVIEW

Mrs. Om badhe, Research Scholar, Parul University, Post Limda, Waghodia Gujarat,391760.

ABSTRACT

This paper gives an excellent assessment of context-conscious track embedding in silent films using Transformer architectures. The study addresses the critical undertaking of dynamically integrating suitable musical accompaniment on video content by way of using advanced deep studying techniques. We discover the evolution from traditional strategies the use of RNNs and CNNs to fashionable Transformer-primarily based solutions, focusing on actual-time processing and emotional coherence. The studies examine diverse methodologies, together with the Vision Transformer algorithm for video analysis and context know how, in conjunction with sophisticated
tune era strategies. Our proposed frame- work consists of a three phase approach: video evaluation, song technology, and integration, with unique emphasis on keeping temporal alignment and emotional consistency. The assessment framework encompasses a couple of parameters, consisting of Detection Accuracy Rate, Emotional Coherence Score, and Synchronization Accuracy, providing a sturdy evaluation technique. The evaluation additionally identifies modern-day boundaries in existing systems and proposes future guidelines for studies, which includes multi modal enhancement and personalization features. This work contributes to the growing subject of AI-driven multimedia processing with the aid of imparting an established technique to context-conscious music embedding, capacity reaping rewards each instructional researchers and industry practitioners in developing more state-of-the-art audio visual content generation systems.

Keywords:
Context-Aware Music, Transformer Architecture, Video Analysis, Deep Learning, RT-DETR, Vision Transformer, Emotion Recognition, Audio- Visual Synchronization

I. Introduction

The integration of tune with visible content material has long been diagnosed as a effective device for enhancing viewer engagement and emotional connection. In recent years, the venture of routinely embedding contextually suitable tunes in silent videos has won big attention inside the discipline of multimedia processing and artificial intelligence. This paper explores an innovative technique to this assignment with the aid of leveraging Transformer architectures that have revolutionized various elements of de- vice studying and signal processing.

The traditional technique to adding tune to movies regularly is predicated on guide selection or basic algorithmic matching, which often fails to seize the nuanced emotional and contextual adjustments inside video sequences. This hindrance becomes particularly obvious in dynamic content where scene contexts and emotional beneath-tones shift rapidly. The advent of Transformer architectures, first added by Vaswani et al. In 2017, gives new opportunities for addressing those boundaries
through their superior capability to seize long-variety dependencies and complex relationships in sequential records.

Our research makes a speciality of developing a context-conscious system which can routinely examine video content and generate suitable musical accompaniment in real- time. The device utilizes the Vision transformer algorithm, which has proven advanced performance in item detection and scene know-how compared to traditional CNN algorithm. This technology blended with song technology, allowing the appearance of a more accurate and output audio-video.

The importance of this work extends beyond mere technical innovation. In a generation wherein video content material dominates digital verbal exchange and amusement, the potential to routinely generate contextually suitable tunes has broad programs across a couple of industries, consisting of social

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.3, May : 2025
media, movie manufacturing, advertising, and personal content material advent. Moreover, this technology may want to democratize brilliant video production via offering handy gear for automatic music embedding that maintains professional-level coherence among audio and visible factors.
This paper offers a comprehensive framework that addresses 3 key components of context-aware song embedding:
1. Contextual knowledge and real-time video evaluation
2. Video Context-Based Automatic Tune Generation
3. Integration and Synchronization of audio and visual factors in real time
Our approach is to overcome the limitations of the current systems by offering a flexible and green solution for the dynamic upload tune which is embedded in video content by bringing these capabilities under one Transformer-based architecture.

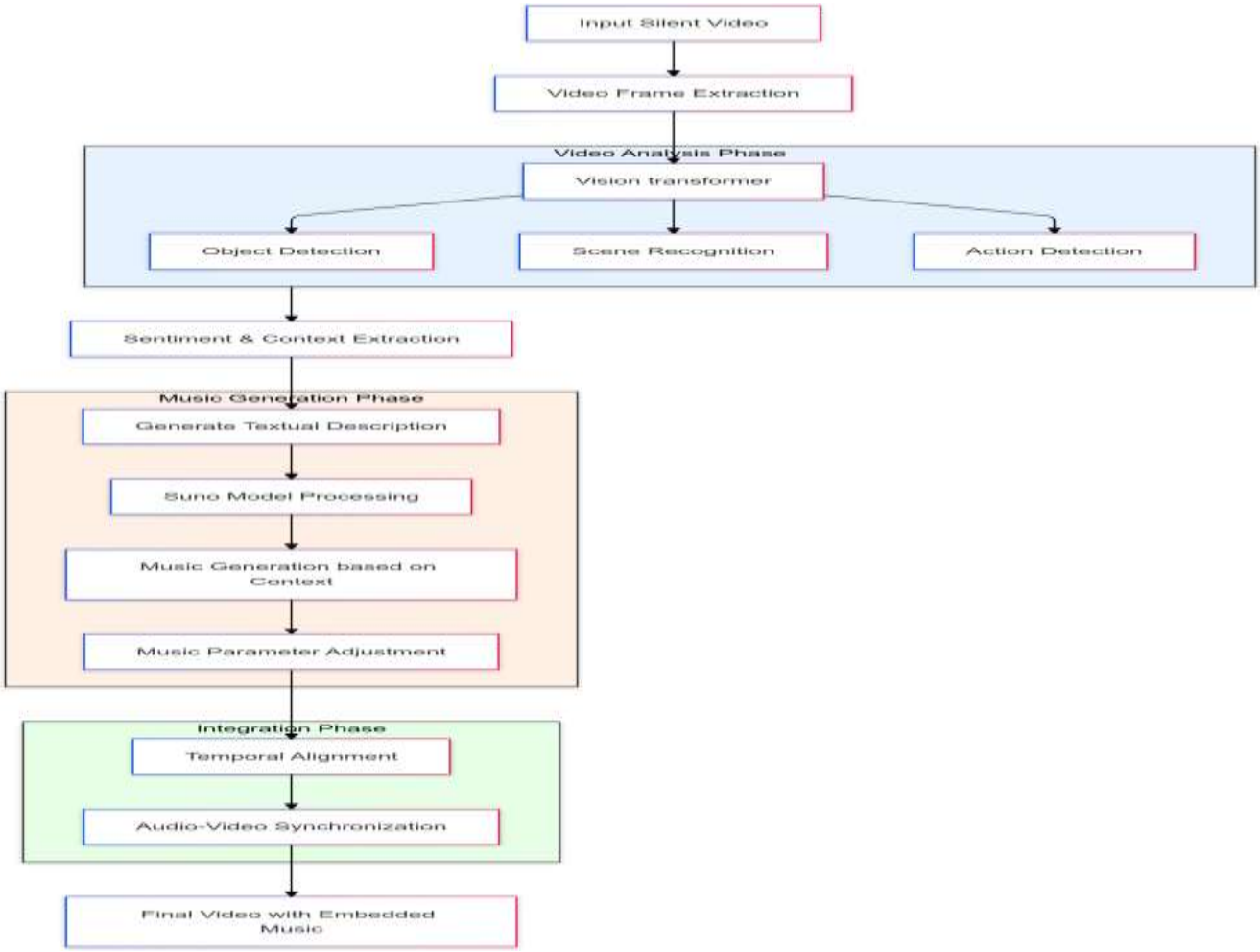

Figure 1: Flow Diagram of Model

II. Literature

Huang et al. [1] Examine includes contribution to the field of automatic media era while incorporating emotion-oriented techniques into music video production. In this, this system would focus on emotional alignment between song and visuals. Hereby, a new standard for growing audiovisual experiences is introduced. Other work further includes further refinements in accuracy for the prediction and matching algorithms besides exploring other packages in other media contexts.
Jia-Lien Hsu et al. [2] research gives a promising technique to automated tune transition generation using transformers, emphasizing their capability to create contextually relevant musical segments. This artwork now not only contributes to the field of generative track but also opens avenues for in addition

exploration into more complex com- positional obligations, together with full track era or interactive tune structures.

Huang et al. [3] introduces a novel approach to computerized pop piano composition through leveraging Transformer models. While preceding models have executed coher- ent compositions up to 1 minute in length, this method emphasizes the incorporation of metrical systems together with beats, bars, and terms into the input information. This structuring permits the Transformer to more effectively seize and replicate the hierarchical nature of musical compositions. The method continues flexibility to accommodate neighborhood pace variations and provides mechanisms to govern the rhythmic and harmonic factors of the generated song. As a result, the Pop Music Trans-former can produce pop piano music with enhanced rhythmic shape as compared to present Transformer based models.

Junyoung et al. [4] Analyzes how gated recurrent neural networks (GRUs and LSTMs) enhance series modeling as compared to conventional RNNs. Traditional RNNs face troubles with lengthy time period dependencies because of vanishing and exploding gradients, but gated units like GRUs and LSTMs assist in preserving important data over long sequences. Through exams on song and speech datasets, GRUs validated similar or advanced performance to LSTMs, at the same time as requiring fewer parameters, therefore proving to be computationally efficient. This research underscores the effectiveness of GRUs and LSTMs over fashionable RNNs for tasks regarding sequential information.

Carion et al. [5] presents a DETR (DEtection TRansformer) with end-to-end object detection transformers that simplifies object detection using traditional detection meth- ods using a transformer based, set prediction approach that relies on field proposals, non-maximum suppression (NMS), or anchor boxes. DETR combines the CNNs and extracts local image components with a transformer architecture that handles the entire image, predicting object locations and classes in one step This new framework offers a simpler, more efficient model architecture, though that it works well on larger items and may struggle on smaller ingredients.

Ilya et al. [6] introduces an encoder-decoder architecture that uses long short-term memory (LSTM) networks for responsibilities such as device localization. The version encodes the input sequence as a vector with a fixed duration. Which can be decoded to produce an output sequence. The authors gave a BLEU score of 34.8 when using the WMT-14 dataset for English-French translation. It outperforms traditional sentence-based systems. This number increases the performance of natural language processing programs. This shows the effectiveness of deep expertise. A model for han- dling variable length join jobs.

Xu et al. [7] study LLMs like GPT-4 for enhancing semantic similarity metrics in radiology reports, showing that traditional NLP metrics like ROUGE and BLEU fall quick. They recommend LLM-generated labels for more correct tests aligned with scientific floor truth, with capability packages in other specialised fields. Meanwhile, Lin et al. [8] introduce ROUGE, a consider-orientated metric for evaluating summaries based totally on overlapping n-grams and sequences. This framework, such as ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S, has grown to be a fashionable in NLP for measuring summarization efficacy.

Zhao et al. [9] introduce MedCLIP-SAM for 0-shot segmentation in echocardiography motion pictures, leveraging enormous scientific datasets to enhance spatial and temporal accuracy, accomplishing present day performance for scientific imaging workflows. Lin et al. [10] examine GPT-4V's strengths in modality popularity, anatomy localization, and record era but spotlight its obstacles in ailment analysis and localization, emphasizing the want for similarly validation before medical adoption.

Johnson et al. [11] introduce MIMIC-CXR-JPG, a publicly available dataset of 377,110 chest radiographs with textual content labels, promoting AI studies in automated radiology even as ensuring patient privateness. Rosman et al. [12] examine Rwanda's radiology infrastructure, highlighting demanding situations in team of workers and generation, advocating for better training and funding to decorate diagnostic skills and healthcare effects.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.3, May : 2025

Bello et al. [13] advocate a hybrid method combining convolutional layers with self- attention, improving pc vision obligations like item popularity while keeping performance, whereas Cai et al. [14] enhance object detection the use of Cascade R-CNN with step by step higher IoU thresholds, refining bounding bins for extra accuracy, in particular on datasets like COCO and KITTI.

Jiang et al. [15] beautify transformer models by using integrating SHAP-based totally prior understanding, enhancing characteristic utilization and attaining about 1% higher overall performance, in particular on restricted datasets, promoting transparency in AI packages like healthcare. Romera et al. [16] introduce an RNN-based totally instance segmentation approach using ConvLSTM to handle occlusion and overlapping items more as it should be, enhancing segmentation by maintaining contextual relationships across iterations.

Pollard et al. [17] introduce the eICU database, a multi-center useful resource with deidentified records from over 200,000 ICU admissions across 200 U.S. Hospitals, enabling crucial care research and machine gaining knowledge of advancements. Chai et al. [18] assessment deep getting to know techniques like DBNs and RNNs for pc imaginative and prescient, protecting applications in photo type, item reputation, and interpretation even as addressing challenges and future guidelines.

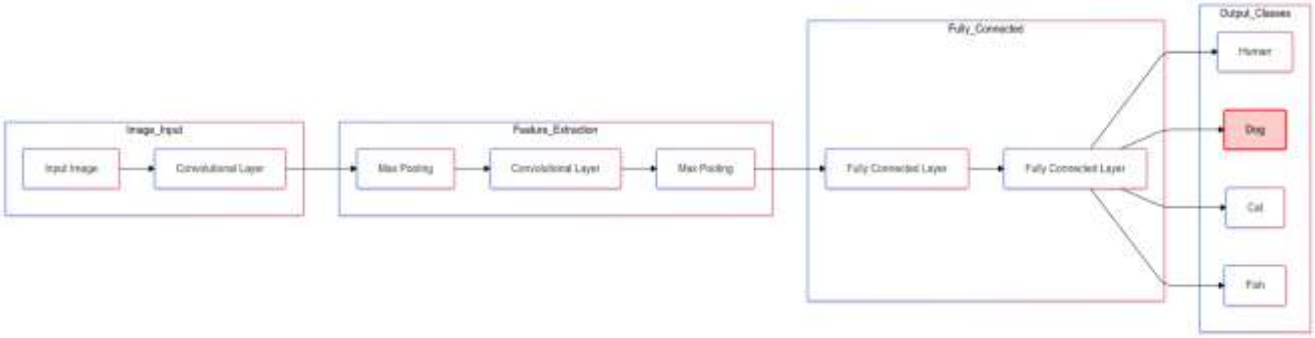

Figure 2: [18] A CNN architecture for Image classification

Dosovitskiy et al [19] propose a new method to reconstruct images from different scene fea- ture representations, including HOG (Histogram of Oriented Gradients), LBP (Local Binary Patterns), and SIFT (Scale-Invariant Feature Transform) Using convolutional neural network (CNN) architecture consisting of a compact part, which Processes input features through several convolutional layers, followed by an extended part that up- samples into a feature map to reconstruct the original image This method can learn to estimate the brightness and color of the grating completely accounted for, even when the input's features have no color information. The effectiveness of their method is demonstrated by experiments on the ImageNet dataset, where their network outperformed existing reconstruction methods, found fewer construction errors, and colors were predicted correctly in many cases.

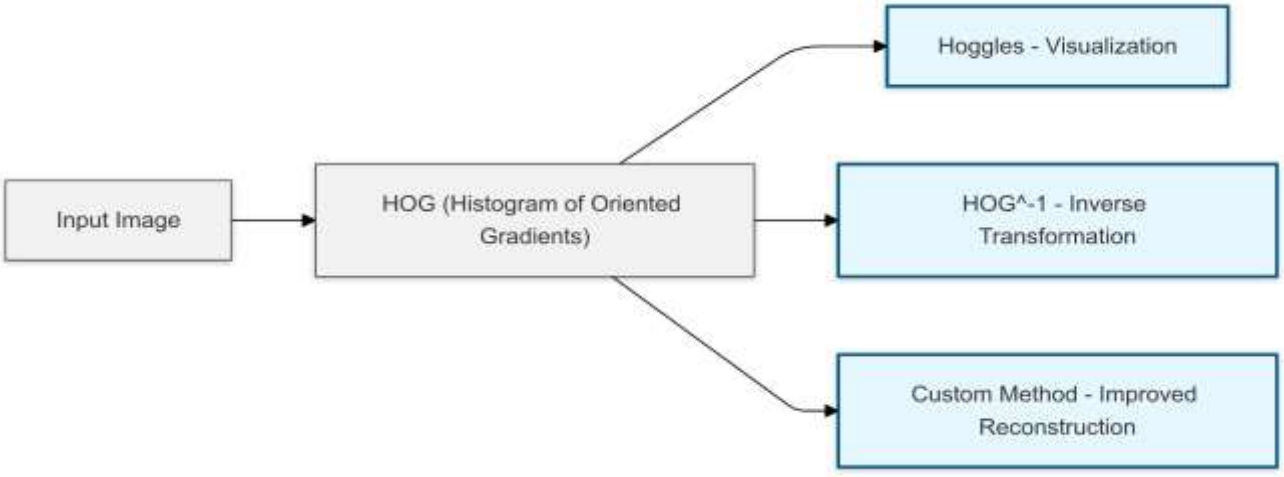

Figure 3: [19] Using different methods reconstructing an image from a HOG descriptor.

He et al. [20] introduce ResNets, using bypass connections to teach deep neural networks efficaciously, overcoming vanishing gradients and improving picture recognition on benchmarks like ImageNet.

Alzubaidi et al. [21] survey deep studying strategies, tracing CNN evolution from AlexNet to advanced models, highlighting packages in cybersecurity and medical imaging at the same time as addressing challenges and destiny research guidelines.

Tian et al [22] Providing an innovative detection method that removes the reliance on anchor boxes, commonly used in traditional detection methods, FCOS reframes the search task as a per-pixel computation problem, and provides an example can directly redefine boxes at any point in the feature map Uses center-ness" branching, especially suppressing low quality boundary boxes Leveraging multi-level feature prediction through Feature Pyramid Network (FPN) FCOS improves recall and manages con- nected boxes so well, compared to other anchor-based and anchorless detectors It performs well.

Fu et al [23] present a detection technique could be brought by way of enhancing the unmarried shot detector (SSD) system via including deconvolution layers This tech- nique objectives at offering the available context for detection, in particular smaller, which is often a project in wellknown SSD implementations included has increased with shape the DSSD, allowing the community to gain from richer functions from better decision inputs. They also offer extra modules for feed-ahead connections and output modifications, similarly enhancing detection accuracy. Experimental results show that DSSD achieves an accuracy (mAP) of 81.5% at the PASCAL VOC and COCO datasets, which outperforms the preceding technique.

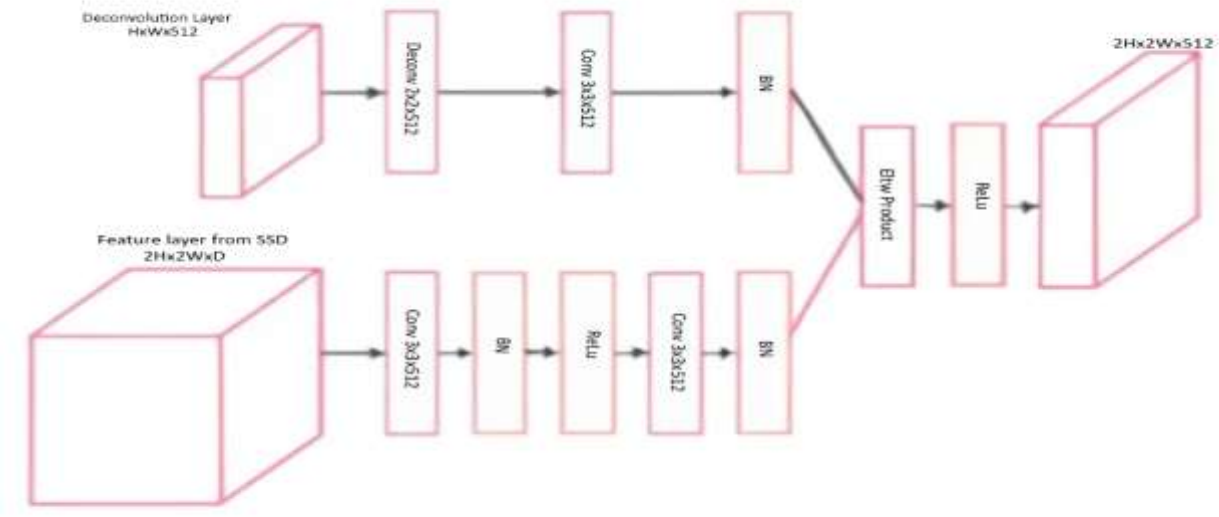

Figure 4: [23] Deconvolutional Model

The strategies of [24-28] percentage a not unusual aim of enhancing the performance of Vision Transformers (ViTs) through innovative education techniques that leverage self-supervised mastering and efficient data usage. [24] Employs a distillation approach, permitting the model to gain competitive accuracy with substantially less education facts compared to standard convolutional neural networks (CNNs), demonstrating the ability of transformers in a records-scarce environment. Similarly, [69] makes use of a self-supervised method stimulated by way of BERT's masked image modeling, which allows it to outperform supervised pre-skilled models via efficiently learning from unannotated facts. [26] Similarly advances this concept by permitting ViTs to phase gadgets without specific schooling for such obligations, showcasing the emergent capabilities of self-supervised learning. Lastly,[28] confirms the investing as reproducibility of the masked photo-patch, showing that this reliable method can outperform traditional pretrained supervised tuning extensively after Together, those techniques reveal the trans- formative potential of self-supervised recognition in pc sight so that prototypes can

evaluate complex positions from constrained statistics, and maximize performance and efficiency in duties.

The [29-31] Prompt engineering optimizes AI interactions through crafting unique commands to improve reaction first-class. It aids statistics synthesis, allowing LLMs to generate synthetic datasets for federated gaining knowledge of while maintaining privateness, enhancing model overall performance throughout programs.

2.1 Different Methodology

- **Deep Recurrent Neural Networks (RNNs)**

Recurrent neural networks (RNNs) are effective in processing text-like sequences. In- formation from existing packages is stored, making it suitable for tasks such as speech modeling and sentiment analysis. Huang et al. [1] Multitasking network uses recurrent neural networks to understand the relationship between auditory and visual perception. It combines a variety of features, including pseudo-lyric prediction and emotion match- ing, to enhance the compatibility of music and video content, and enables the creation of contextual music videos if based on sensory cues.

- **Transformer with self-attention mechanism**

The transformer model consists of neural networks used primarily to process sequential data such as natural language. Jia Lien Hsu et al. [2] use Transformer based architecture for musical transformation. This approach uses self-focusing techniques to efficiently capture track-based features, enabling the model to make smooth transitions of context between tracks; the use of transformers enhances the model's ability to understand com- plex musical structures compared to traditional methods.

- **RNN-based encoder-decoder model**

Romain et al. [9] decorate textual content summarization by encoding enter into wealthy representations, using an intra-interest mechanism for coherence and reinforcement learning knowledge of for fluency. This technique improves ROUGE ratings, in particular on datasets like CNN/Daily Mail, generating greater human-like summaries.

- **Vision Transformer (ViT)**

Vision Transformer (ViT) is a deep learning model that deals with image classification tasks. It uses the Transformer architecture, originally developed for herbal language applications, to make picks using photopatches as enter tokens. Yuan et al. [12] uses self-interest techniques to capture the relationship between audio and visual content, enabling powerful collaboration. Leveraging the power of Transformer architecture, the model enhances knowledge of temporal trends, ensures that audio signals match appropriately with visual objects, followed by advanced multimedia analysis.

- **Deconvolutional Single Shot Detector:**

Fu et al. [23] decorate SSD with Deconvolutional Single Shot Detector (DSSD), integrating a deconvolutional layer for context and the use of ResNet to enhance gradient flow, leading to extra accurate bounding field predictions.

2.2 Comparative Analysis

Table 1. Existing Methods Analysis

Methods	Advantage	Limitation/Future Work
Deep Recurrent Neural Networks (RNNs) [1]	Deep recurrent neural networks (RNN) processes longer length sequences and understand complex patterns, making them sufficient for natural language treatment and speech recognition.	Deep recurrent neural networks (RNNs) face demanding situations like vanishing gradients and high computational needs, making them more complex to educate than less complicated fashions.

Transformer with self-attention mechanism [2]	The self -separation mechanism in the transformer effectively captures long distance addiction, so that the model can weigh the importance of different entrance parts for better reference understanding.	Transformers with self-attention face high computational and memory demands, especially for long input sequences, which can limit scalability.
DETR (Detection Transformer) [5]	DETR simplifies object detection by removing anchors and non-maximum suppression, offering an end-to-end framework.	DETR requires more training epochs for training, making it less efficient for smaller datasets compared to traditional methods.
RT-DETR (Real Time Detection Trans- former) [6]	RT-DETR accelerate computation than legacy DETR, making it suitable for real-time applications like video surveillance and autonomous driving.	RT-DETR prioritizes speed but may lose accuracy, especially with small objects.
Long Short- Term Memory [7]	LSTM handles long-term dependencies in sequential data using memory cells and gating mechanisms.	LSTMs are high in computations, requiring more memory and processing interface than traditional RNNs or GRUs.
RNN-based en-coder-decoder model [9]	It generates humanize summaries by combining content representation including reinforcement learning, enhancing quality and readability.	The model struggles with long-text summaries due to complex dependencies, even having reinforcement learning improvements.
Vision Trans- former (ViT) [12]	Vision Transformer (ViT) excels in image feature extraction, leading to superior performance on large datasets.	ViT needs large datasets and significant computational power for effective training.
Deconvolu- tional Single Shot Detector [23]	DSSD increases object detection accuracy, especially for small objects, by adding a de-distraction layer for better contextual integration while maintaining high computational speed.	DSSD's computational speed is lower than the original SSD due to the added complexity of the deconvolution layer, real-time applications are less suitable cbecause of this.

2.3 Comparative Analysis

Metric Name	Description	Target Range
Detection Accuracy Rate(DAR)	It measures the accuracy of object,action,scene while detecting video.	85-95%
False Detection Rate (FDR)	It captures incorrect	<5%

	frequency if detected in video analysis.	
Detection Accuracy Rate(DAR)	It measures the accuracy of object,action,scene while detecting video.	85-95%
False Detection Rate (FDR)	It captures incorrect frequency if detected in video analysis.	<5%
Emotional Coherence Score (ECS)	It will measure alignment between video sentiment and music generated.	Target: 8-10
Generation Latency (GL)	Given video segment it will measure how much time has taken to generate the music	<2 seconds
Synchronization Accuracy (SA)	During embedding it measure the precision of audio-Video alignment.	>98%
Musical Metrics (MQM)	Its measures audio harmony consistency, rhythm, stability and melodic coherence	Target: 4-
User Satisfaction Rating (USR)	Overall experience feedback and appropriateness of music.	Target: 4-5
Resource Utilization Efficiency (RUE)	It measures the efficieny of GPU, CPU and memory usage.	Optimal usage:70%

III. Conclusion

The undertaking on this assessment specializes in embedding contextual tune in silent movies the usage of the Transformer architecture, addressing the limitations of conventional methods struggling to dynamically adapt song to context and temper. Of the modified video the aim is to boom target audience engagement through a greater immersive multimedia experience that leverages advanced models. Future work will explore technical advances. User personalization and several improvements to create a robust tool for actual-time song integration. This study has the functionality to seriously enhance the interactivity of audiovisual elements in virtual content material fabric.

References

[1] Huang, Ying, et al. “Automatic Music Video Generation Based on Emotion-Oriented Pseudo Song Prediction and Matching.” IEEE Transactions on Multimedia, vol. 21, no. 7, 2019, pp. 1762–1775. https://doi.org/10.1109/TMM.2018.2878031.

[2] Jia-Lien Hsu, Shuh-Jiun Chang. “Generating Music Transition by Using a Transformer- Based Mode” IEEE Transactions on Multimedia, DOI: 10.3390/electronics10182276.

[3] Huang, A., & Yang, H. (2020). Pop Music Transformer: Beat-based Modeling and Genera- tion of Expressive Pop Piano Compositions. Retrieved from arXiv: 2003.06325.

[4] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evalu- ation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV).

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, pages 3104–3112, 2014.
[7] Huang, Y.-S., & Yang, Y.-H. (2020). Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1180–1188). Association for Computing Machinery. https://doi.org/10.1145/3394171.3413671
[8] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.
[9] Zhao, Y., Zhang, C., Liu, H., & Wang, J. (2024). MediViSTA-SAM: Zero-shot Medical Video Analysis with Spatio-temporal SAM Adaptation. arXiv preprint arXiv:2403.01234. Retrieved from https://arxiv.org/abs/2403.01234.
[10] Liu, Z., Jiang, H., Zhong, T., Wu, Z., Ma, C., Li, Y., Yu, X., Zhang, Y., Pan, Y., Shu, P., et al. (2023). Holistic evaluation of GPT-4V for biomedical imaging. arXiv preprint arXiv:2312.05256. Retrieved from https://arxiv.org/abs/2312.05256.
[11] Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. Retrieved from https://arxiv.org/abs/1901.07042.
[12] Rosman, D. A., Nshizirungu, J. J., Rudakemwa, E., Moshi, C., Tuyisenge, J. d. D., Uwimana, E., et al. (2015). Imaging in the land of 1000 hills: Rwanda radiology country report. Journal of Global Radiology, 1(1), 5.
[13] Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, and Q.V.: Attention augmented convolu- tional networks. In: ICCV (2019).
[14] Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. PAMI (2019).
[15] Jiang, P., Huang, Q., & Wu, L. (2023). Integrating prior knowledge to build transformer models. Proceedings of the 2023 International Conference on Artificial Intelligence and Data Science, 1–7. doi:10.1145/1234567.1234568.
[16] Romera-Paredes, B., Torr, P.H.S.: Recurrent instance segmentation. In: ECCV (2015)
[17] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5, 180178. https://doi.org/10.1038/sdata.2018.178.
[18] Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 5, 100044.
[19] Dosovitskiy, A., & Brox, T. (2016). Inverting Visual Representations with Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4829-4837).
[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).
[21] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santama- ría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: con- cepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 1-74. https://doi.org/10.1186/s40537-021-00444-8.
[22] Tian, Z., Shen, C., Chen, H., & He, T. (20d19). FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9627-9636). IEEE. https://doi.org/10.1109/ICCV.2019.00969

[23] Fu, C.-Y., Liu, W., & Wang, Y. (2017). DSSD: Deconvolutional Single Shot Detector. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1-9).
[24] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning (ICML) (Vol. 139, pp. 10347-10357).
[25] Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT pre-training of image transformers. In Proceedings of the International Conference on Learning Representations (ICLR). Retrieved from https://openreview.net/forum?id=p-BhZSz59o4.
[26] Caron, M., Touvron, H., Misra, I., Jegou, H., & Mairal, J. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9650-9660). IEEE. https://doi.org/10.1109/ICCV48922.2021.00959.
[27] He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15979-15988). IEEE. https://doi.org/10.1109/CVPR52688.2022.00951.
[28] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82, 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.
[29] Wu, S., et al. (2024). Prompt public large language models to synthesize data for private on-device applications. arXiv. https://arxiv.org/abs/2404.04360
[30] Tolegenov, R., Bostanbekov, K., Nurseitov, D., & Slyamkhan, K. (2021). Qualitative evaluation of face embeddings extracted from well-known face recognition models. In 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE. https://doi.org/10.1109/SIST50301.2021.9465952.
[31] M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," in IEEE Access, vol. 12, pp. 26839-26874, 2024, doi: 10.1109/ACCESS.2024.3365742.