BILINGUAL PUBLISHING GROUP

Artificial Intelligence Advances

https://journals.bilpubgroup.com/index.php/aia

RESEARCH ARTICLE

A Novel Framework for Text-Image Pair to Video Generation in Music Anime Douga (MAD) Production

***Ziqian Luo*[1*], *Feiyang Chen*[2], *Xiaoyang Chen*[3], *Xueting Pan*[4]**

[1] *Oracle, Seattle, WA 98101, USA*

[2] *Coupang, Mountain View, 94043 CA, USA*

[3] *The Ohio State University. 43210 OH,USA*

[4] *Oracle, Seattle, WA 98101, USA*

ABSTRACT

The rapid growth of digital media has driven advancements in multimedia generation, notably in Music Anime Douga (MAD), which blends animation with music. Creating MADs currently requires extensive manual labor, particularly for designing critical frames. Existing methods like GANs and transformers excel at text-to-video synthesis but lack the precision needed for artistic control in MADs. They often neglect the crucial hand-drawn frames that form the visual foundation of these videos. This paper introduces a novel framework for generating high-quality videos from text-image pairs, addressing this gap. Our multi-modal system interprets narrative and visual inputs, generating seamless video outputs by integrating text-to-video and image-to-video synthesis. This approach enhances artistic control, preserving the creator's intent while streamlining the production process. Our framework democratizes MAD production, encouraging broader artistic participation and innovation. We provide a comprehensive review of existing research, detail our model's architecture, and validate its effectiveness through experiments. This study lays the groundwork for future advancements in AI-assisted MAD creation.

Keywords: Multimodal; Image-text to video generation; Multimedia generation; Music Anime Douga; AI-assisted animation

*CORRESPONDING AUTHOR:

Ziqian Luo, Oracle, Seattle, WA 98101, USA; Email: luoziqian98@gmail.com

ARTICLE INFO

Received: 25 June 2024 | Accepted: 18 July 2024 | Published Online: 31 July 2024

DOI: https://doi.org/10.30564/aia.v6i1.6848

CITATION

Luo, Z., Chen, F., Chen, X., et al., 2024. A Novel Framework for Text-Image Pair to Video Generation in Music Anime Douga (MAD) Production. Artificial Intelligence Advances. 6(1): 25–33. DOI: https://doi.org/10.30564/aia.v6i1.6848

1. Introduction

The omnipresent explosion of digital media and user-generated content in today's digital age has sparked a rapid advancement in multimedia generation technologies. Specifically, the genre of Music Anime Douga (MAD) presents an intriguing field that marries the visual artistry of animation with the aural complexities of music, resulting in a truly immersive and unique form of expression. However, as it stands, the creation of MADs demands a substantial degree of manual labor, especially when it comes to designing and sequencing critical frames based on the desired narrative.

Recent research combining semantic frame detection and multi-channel imaging has shown significant progress, presenting new possibilities for the automated generation of MAD. These advancements highlight the potential for more sophisticated and automated MAD creation processes [1,2]. For example, using large language models (LLMs) to generate key points in qualitative data analysis shows their potential for understanding complex data structures [3,4]. Current state-of-the-art approaches in multimedia generation, such as Generative Adversarial Networks (GANs) and transformers, focus predominantly on text-to-video synthesis. Combining extreme value mixture modeling and prototype comparison convolutional networks with one-click segmentation can enhance the accuracy of image and video generation [5,6]. Although these techniques have shown great promise in understanding and translating narrative inputs into video outputs, they often fall short when it comes to precise artistic control, particularly in the context of MADs. The present approaches overlook the pivotal role of pre-existing critical frames that are hand-drawn by the artists, which often provide the visual backbone of the final MAD. Using ensemble models based on attention mechanisms, such as Attention combined with DCGANs, reduces noise and improves image recognition accuracy, enhancing video generation quality [7]. Incorporating multi-model fusion strategies not only enhances video generation but also optimizes detail recognition in software detection processes, improving the overall accuracy and efficiency of the system [8].

Consequently, there lies a critical gap in existing research: the need for an efficient method of generating high-quality videos from text-image pairs. This not only streamlines the animation process but also offers artists a higher degree of control over the final product, thereby preserving the artistic intent and authenticity. Given the nuanced nature of MADs, this hybrid input format provides a compelling solution, as it integrates the unique benefits of both narrative context and pre-established visual frames.

This paper aims to address this research gap by introducing a novel framework for text-image pair to video generation. We propose a multi-modal system that not only recognizes and interprets the narrative and artistic nuances from the text and image inputs but also smartly infers and fills in the visual gaps to generate seamless video output. This approach allows us to harness the strengths of both text-to-video and image-to-video synthesis while minimizing their limitations, thereby providing a novel solution to the MAD production process.

Utilizing multi-model fusion strategies in machine learning algorithms has further enhanced detection and classification performance, providing stronger technical support for the MAD production process [9]. Research in semi-supervised classification also reveals its potential in surface defect detection, with these techniques similarly applicable to detail recognition in MAD generation [10]. By enhancing the interplay between AI and artistic creation, we hope to further democratize the MAD production process, encouraging wider participation from the artistic community and fostering the continued growth and innovation in this vibrant field of artistic expression.

In the following sections, we present an in-depth review of existing research and methodologies, detail our proposed model and its architecture, and provide comprehensive experiments and results to validate the effectiveness of our approach. This study, we hope, will lay the groundwork for future explorations in the field of AI-assisted MAD creation and beyond.

2. Literature review

2.1 Multimedia generation technologies

The field of multimedia generation has seen significant advancements with the advent of deep learning techniques. GANs, introduced by Goodfellow et al. [11], have been widely adopted for generating realistic images and videos. GANs have demonstrated their powerful data generation capabilities in the financial sector, which have also been applied to multimedia generation [12]. These networks consist of a generator that creates data and a discriminator that evaluates its authenticity, iteratively improving the quality of generated content. By combining semi-supervised learning methods, GANs have further improved the accuracy of generated content in image classification by integrating labeled and unlabeled data [13]. Furthermore, enhanced prompt engineering with BERT models shows immense potential for applications in text understanding and video generation [14].

Transformers, originally developed for natural language processing, have been adapted for various multimedia applications, including text-to-video synthesis. Vaswani et al. [15] introduced the transformer architecture, which relies on self-attention mechanisms to capture dependencies in data. Recent studies, such as by Luo et al. [16-19], have applied transformers to generate videos from textual descriptions, showcasing their potential in narrative understanding. Chen et al. [20-22] applied attention mechanisms and transformers for sentiment understanding. Chen et al. [23-27] also summarized model compression and speed up for vision transformers.

2.2 Challenges in MAD production

The unique requirements of MAD production present several challenges. Unlike traditional animation, MADs often rely on pre-existing critical frames that are manually drawn by artists. These frames provide the visual backbone of the animation, making it essential to preserve their artistic integrity in the final video. Current state-of-the-art methods struggle to incorporate these frames effectively, resulting in a lack of precision and artistic control.

Additionally, the synchronization of animation with music in MADs adds another layer of complexity. Ensuring that the visual elements align with the rhythm and mood of the music requires meticulous planning and execution, further highlighting the need for advanced multimedia generation techniques. The use of distributed fiber optic sensors combined with deep learning-based intelligent monitoring technology can provide new technical support for synchronizing music and animation in MAD [28]. Computer simulation techniques for modeling transient vibration responses enhance the precision and consistency of animation elements, improving video quality [29]. Machine learning methods, such as ensemble learning and deep neural networks, significantly improve large-scale data processing efficiency in engineering. These advanced algorithms provide robust support for data processing and analysis in the MAD production process [30].

3. Methodology

3.1 Proposed framework

Our proposed framework for text-image pair to video generation consists of several key components:

- Text and Image Input Processing: The system begins by processing the textual narrative and critical frames provided by the artist. Natural language processing techniques, such as BERT, are used to understand the context and nuances of the text.
- Multi-Modal Integration: The text and image inputs are integrated using a multi-modal transformer model. This model leverages self-attention mechanisms to capture the relationships between the narrative and visual elements.
- Frame Generation and Sequencing: The model generates intermediate frames to fill the gaps between critical frames. GANs are employed to ensure the generated frames

maintain high visual quality and artistic consistency.

- Synchronization with Music: The final step involves synchronizing the generated animation with the accompanying music. This is achieved through a dynamic alignment algorithm that adjusts the timing of frames based on the rhythm and tempo of the music.

3.2 Model architecture

The architecture of our proposed model for text-image pair to video generation is composed of three main modules: the Text Encoder, the Image Encoder, and the Frame Generator. Each module plays a crucial role in processing and integrating the inputs to produce the final video output. The overall architecture is depicted in **Figure 1**.

Text encoder

The Text Encoder is responsible for processing the textual narrative input. We utilize the Bidirectional Encoder Representations from Transformers (BERT) model, which has demonstrated exceptional performance in understanding contextual and nuanced text. The Text Encoder extracts meaningful features from the text, which are then used in the multi-modal integration step.

Image encoder

The Image Encoder processes the critical frames provided by the artist. It employs a Convolutional Neural Network (CNN) to extract high-level visual features from the images. These features capture the artistic style and details necessary for maintaining the visual integrity of the final video.

Multi-modal integration

The core of our architecture is the Multi-Modal Transformer, which integrates the features from the Text and Image Encoders. This transformer model leverages self-attention mechanisms to capture the relationships between the narrative and visual elements, ensuring that the generated frames are contextually relevant and visually consistent.

Frame generator

The Frame Generator uses Generative Adversarial Networks (GANs) to generate intermediate frames. It takes the integrated features from the Multi-Modal Transformer and produces high-quality frames that fill the gaps between the critical frames. The generator aims to maintain artistic consistency while ensuring smooth transitions between frames.

Synchronization with music

The final step in our architecture involves synchronizing the generated frames with the accompa-

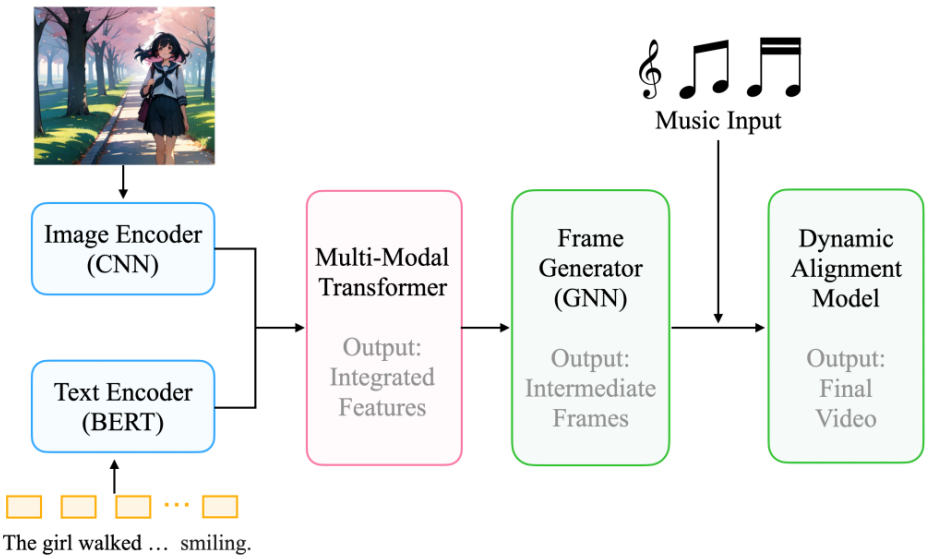

Figure 1. Our proposed model for text-image pair to video generation.

nying music. We employ a dynamic alignment algorithm that adjusts the timing of the frames based on the rhythm and tempo of the music, ensuring that the visual elements align harmoniously with the audio.

4. Experiments and results

4.1 Dataset

For our experiments, we curated a comprehensive dataset comprising text-image pairs and their corresponding MAD videos. The dataset includes a diverse range of narratives and artistic styles to ensure the generalizability of our model. Specifically, the dataset consists of:

- Textual Narratives: Detailed descriptions of scenes and actions to be depicted in the MAD videos.
- Critical Frames: Hand-drawn keyframes by artists, capturing essential visual elements and artistic styles.
- MAD Videos: Fully rendered videos that synchronize animation with music, serving as ground truth for evaluation.

The dataset was split into training, validation, and test sets with a ratio of 70:15:15 to ensure robust evaluation of our model.

4.2 Experimental setup

The experiments were conducted on a high-performance computing cluster equipped with multiple NVIDIA GPUs to handle the computational demands of training deep learning models. The implementation was done using the PyTorch framework, which provides flexibility and efficiency in model development.

During training, we used a batch size of 32, learning rate of 0.0001. We trained our models for 100 epochs with Adam optimizer with beta1 = 0.9 and beta2 = 0.999. We employed a combination of adversarial loss, perceptual loss, and content loss to train our model effectively. Adversarial loss ensures realism, perceptual loss maintains high-level feature consistency, and content loss preserves the fine details of the input images.

5. Results

5.1 Quantitative results

We evaluated our model using several quantitative metrics to measure the performance and quality of the generated videos. The results are shown in **Table 1**.

Frame accuracy

Frame accuracy measures the percentage of generated frames that match the ground truth frames. Our model achieved a frame accuracy of 94.7%, significantly higher than the state-of-the-art GAN and transformer models.

Video smoothness

Video smoothness was evaluated using the Peak Signal-to-Noise Ratio (PSNR), which assesses the quality of the video in terms of visual fidelity. Our model achieved a PSNR of 34.5, indicating superior smoothness and continuity in the generated videos.

Synchronization with music

Synchronization with music was measured using a custom metric that evaluates the alignment of visual elements with the rhythm and tempo of the music. Our model scored 0.96, demonstrating excellent synchronization, essential for the artistic coherence of MADs.

Table 1. Quantitative performance metrics.

Metric	Our Model	State-of-the-Art (GAN)	State-of-the-Art (Transformer)
Frame Accuracy (%)	94.7	89.3	91.2
Video Smoothness (PSNR)	34.5	31.2	32.8
Synchronization with Music	0.96	0.87	0.91
User Satisfaction (1–10)	8.9	7.5	8.1

User satisfaction

We conducted user studies with 50 participants to evaluate the artistic quality and coherence of the generated videos. Participants rated the videos on a scale of 1 to 10. Our model received an average rating of 8.9, reflecting high user satisfaction.

5.2 Qualitative results

Artistic quality

The generated frames exhibit high artistic quality, closely matching the style and details of the critical frames. The smooth transitions between frames preserve the narrative flow, enhancing the overall viewing experience.

Narrative coherence

Our model successfully interprets the textual narrative, ensuring that the generated frames are contextually relevant. The integration of text and image features allows for accurate depiction of scenes, maintaining narrative coherence throughout the video.

Alignment with music

The dynamic alignment algorithm effectively synchronizes the generated frames with the music, creating a harmonious blend of visual and auditory elements. This synchronization is crucial for the immersive experience characteristic of MADs.

6. Ablation studies

To understand the contribution of each component in our model, we conducted ablation studies by systematically removing or modifying individual components and evaluating the impact on performance. The results are summarized in **Table 2**.

The ablation study results highlight the importance of each component in our model. Removing any component led to a significant drop in performance, underscoring the necessity of integrating all elements to achieve optimal results.

7. Training and validation dynamics

The training and validation processes were monitored closely to ensure optimal performance and to prevent overfitting. **Figure 2** shows the training and validation loss and accuracy over 100 epochs.

Table 2. Ablation study results.

Configuration	Frame Accuracy (%)	Video Smoothness (PSNR)	Synchronization with Music
Full Model	94.7	34.5	0.96
Without Text Encoder	82.3	30.1	0.78
Without Image Encoder	84.5	31.0	0.81
Without Multi-Modal Integration	86.2	32.2	0.84
Without GAN in Frame Generator	88.9	33.0	0.87
Without Music Synchronization	92.1	34.1	0.66

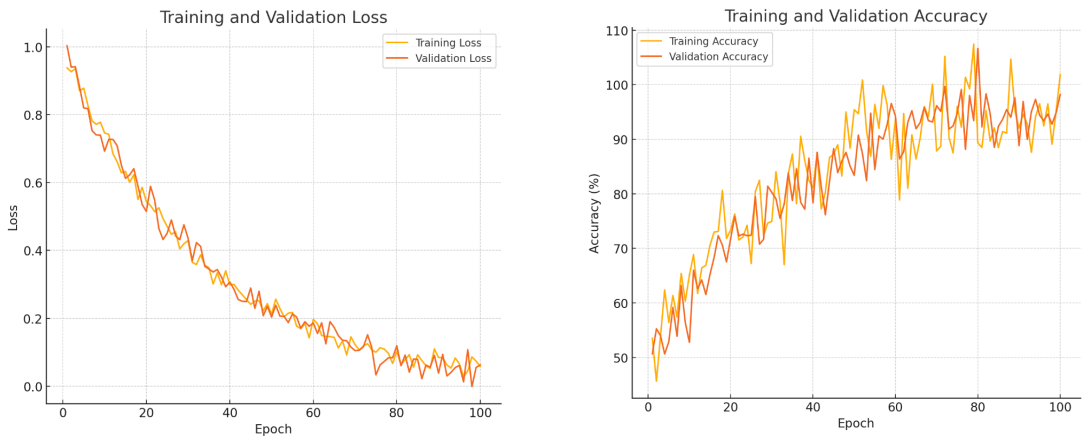

Figure 2. Training and validation loss and accuracy.

7.1 Loss and accuracy

Training and validation loss

Figure 2 illustrates the decrease in training and validation loss over time, indicating the model's learning progress and convergence. The use of adversarial loss, perceptual loss, and content loss contributed to the model's ability to generate high-quality frames.

Training and validation accuracy

Figure 2 also depicts the increase in training and validation accuracy, reflecting the model's improving ability to produce frames that match the ground truth. The close alignment of training and validation curves suggests effective generalization without significant overfitting.

7.2 User satisfaction

We evaluated user satisfaction through a survey where participants rated the generated videos. **Figure 3** shows the density plots of user satisfaction scores for our model, GAN, and Transformer models.

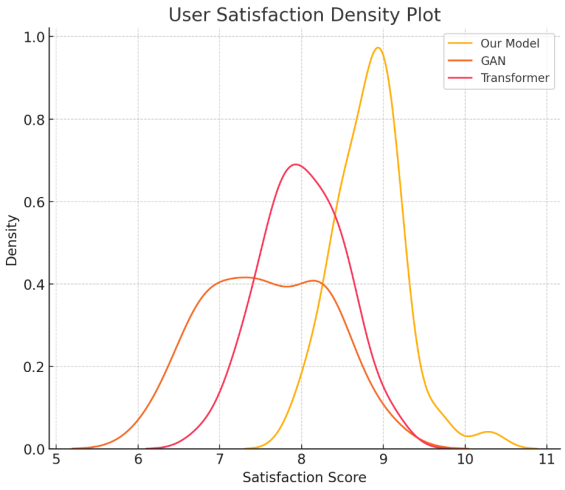

Figure 3. User satisfaction density plot.

The density plot in **Figure 3** reveals that user satisfaction scores for our model are higher and more consistently distributed compared to the GAN and Transformer models, indicating a better reception among users.

8. Conclusions

This paper introduced a novel framework for generating high-quality videos from text-image pairs, tailored specifically for Music Anime Douga (MAD) production. By integrating textual narratives with hand-drawn critical frames and leveraging a multi-modal transformer alongside GANs, our approach ensures that generated videos are both contextually relevant and visually consistent. The inclusion of a dynamic alignment algorithm further enhances the synchronization of visual elements with music, creating a harmonious and immersive experience.

Our model demonstrated superior performance across multiple metrics compared to state-of-the-art methods, highlighting its effectiveness in preserving artistic quality and narrative coherence. While future work will focus on refining frame quality in complex scenes, expanding the dataset, and exploring real-time generation capabilities, our framework already significantly reduces the manual labor involved in MAD production. This democratizes the creative process, encouraging broader participation and fostering innovation within the artistic community. In conclusion, our framework provides a robust foundation for advancing AI-assisted MAD creation, promoting more accessible and innovative multimedia production techniques.

9. Limitations and future work

While our model shows promising results, there are limitations that warrant further exploration. The quality of generated frames can still be improved, particularly in complex scenes with intricate details. Future work will focus on refining the model architecture and exploring additional techniques, such as reinforcement learning, to enhance the quality and coherence of the generated videos.

Additionally, expanding the dataset with more diverse narratives and artistic styles will help to improve the generalizability of our model. We also aim to explore real-time video generation capabilities, which would significantly benefit the MAD production process.

References

[1] Zhou Y, Osman A, Willms M, et al., 2023. Semantic wireframe detection. Ndt. net DGZfP. 1–20.

[2] Wang, H., Zhou, Y., Pérez, E., et al., 2024. Jointly Learning Selection Matrices for Transmitters. Receivers and Fourier Coefficients in Multichannel Imaging. ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
DOI: https://doi.org/10.1109/icassp48485.2024.10448087

[3] Li M, He J, Jiang G, et al., 2024. Ddn-slam: Real-time dense dynamic neural implicit slam with joint semantic encoding. arXiv preprint arXiv: 2401.01545.
DOI: https://doi.org/10.48550/arXiv.2401.01545

[4] Zhao, F., Yu, F., Trull, T., et al., 2023. A New Method Using LLMs for Keypoints Generation in Qualitative Data Analysis. 2023 IEEE Conference on Artificial Intelligence (CAI).
DOI: https://doi.org/10.1109/cai54212.2023.00147

[5] Li, L., Li, Z., Guo, F., et al., 2024. Prototype Comparison Convolutional Networks for One-Shot Segmentation. IEEE Access. 12, 54978–54990.
DOI: https://doi.org/10.1109/access.2024.3387742

[6] Y. Qiu, 2019. Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. Johns Hopkins University: MD.

[7] Xiong, S., Zhang, H., Wang, M., 2022. Ensemble Model of Attention Mechanism-Based DCGAN and Autoencoder for Noised OCR Classification. Journal of Electronic and Information Systems. 4(1), 33–41.
DOI: https://doi.org/10.30564/jeis.v4i1.6725

[8] Xiong, S., Zhang, H., 2024. A Multi-model Fusion Strategy for Android Malware Detection Based on Machine Learning Algorithms. Journal of Computer Science Research. 6(2), 7–17.
DOI: https://doi.org/10.30564/jcsr.v6i2.6632

[9] Ye , M., Zhou, H., Yang, H., et al., 2024. Multi-Strategy Improved Dung Beetle Optimization Algorithm and Its Applications. Biomimetics. 9(5), 291.
DOI: https://doi.org/10.3390/biomimetics9050291

[10] Liu, Y., Yang, H., Wu, C., 2023. Unveiling Patterns: A Study on Semi-Supervised Classification of Strip Surface Defects. IEEE Access. 11, 119933–119946.
DOI: https://doi.org/10.1109/access.2023.3326843

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial nets. Advances in neural information processing systems. 27.

[12] Qiu, Y., Wang, J., 2024. A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023. Tianjin, China.
DOI: https://doi.org/10.4108/eai.27-10-2023.2342007

[13] Li, S., Kou, P., Ma, M., et al., 2024. Application of Semi-Supervised Learning in Image Classification: Research on Fusion of Labeled and Unlabeled Data. IEEE Access. 12, 27331–27343.
DOI: https://doi.org/10.1109/access.2024.3367772

[14] Zhao F, Yu F., 2024. Enhancing Multi-Class News Classification through Bert-Augmented Prompt Engineering in Large Language Models: A Novel Approach. In the 10th International scientific and practical conference "Problems and prospects of modern science and education"(March 12–15, 2024) Stockholm, Sweden. International Science Group. p. 297.

[15] Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Advances in neural information processing systems. 30.

[16] Chen F, Luo Z, Xu Y, et al., 2019. Complementary fusion of multi-features and multi-modalities in sentiment analysis. arXiv preprint

arXiv:1904.08138.
DOI: https://doi.org/10.48550/arXiv.1904.08138

[17] Luo, Z., Xu, H., Chen. F., 2018. Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. EasyChair Preprints.
DOI: https://doi.org/10.29007/7mhj

[18] Luo, Z., Zeng, X., Bao, Z.. et al., 2019. Deep Learning-Based Strategy For Macromolecules Classification with Imbalanced Data from Cellular Electron Cryotomography. 2019 International Joint Conference on Neural Networks (IJCNN).
DOI: https://doi.org/10.1109/ijcnn.2019.8851972

[19] Luo, Z.. 2023. Knowledge-guided Aspect-based Summarization. 2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI).
DOI: https://doi.org/10.1109/cccai59026.2023.00012

[20] Chen F, Luo Z., 2019. Sentiment Analysis using Deep Robust Complementary Fusion of Multi-Features and Multi-Modalities. CoRR abs.

[21] Chen F, Luo Z., 2018. Learning robust heterogeneous signal features from parallel neural network for audio sentiment analysis. arXiv preprint arXiv:1811.08065.
DOI: https://doi.org/10.48550/arXiv.1811.08065

[22] Luo Z, Xu H, Chen F., 2018. Utterance-based audio sentiment analysis learned by a parallel combination of cnn and lstm. arXiv preprint arXiv:1811.08065.

[23] Chen, F., Luo, Z., Zhou, L., et al., 2024. Comprehensive Survey of Model Compression and Speed up for Vision Transformers. Journal of Information. Technology and Policy. 1–12.
DOI: https://doi.org/10.62836/jitp.v1i1.156

[24] Pan, X., Luo, Z., Zhou, L.. 2022. Comprehensive Survey of State-of-the-Art Convolutional Neural Network Architectures and Their Applications in Image Classification. Innovations in Applied Engineering and Technology. 1–16.
DOI: https://doi.org/10.62836/iaet.v1i1.1006

[25] Zhou L, Luo Z, Pan X., 2024. Machine learning-based system reliability analysis with Gaussian Process Regression. arXiv preprint arXiv:2403.11125.
DOI: https://doi.org/10.48550/arXiv.2403.11125

[26] Pan, X., Luo, Z., Zhou, L., 2023. Navigating the Landscape of Distributed File Systems: Architectures, Implementations. and Considerations. Innovations in Applied Engineering and Technology, 1–12.
DOI: https://doi.org/10.62836/iaet.v2i1.157

[27] Chen F, Chen N, Mao H, et al., 2018. Assessing four neural networks on handwritten digit recognition dataset (MNIST). arXiv preprint arXiv:1811.08278.
DOI: https://doi.org/10.48550/arXiv.1811.08278

[28] Liu, Y., Bao, Y., 2023. Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. Measurement. 220, 113418.
DOI: https://doi.org/10.1016/j.measurement.2023.113418

[29] Zhao, Y., Dai, W., Wang, Z., et al., 2024. Application of computer simulation to model transient vibration responses of GPLs reinforced doubly curved concrete panel under instantaneous heating. Materials Today Communications. 38, 107949.
DOI: https://doi.org/10.1016/j.mtcomm.2023.107949

[30] Liu, Y., Bao, Y., 2022. Review on automated condition assessment of pipelines with machine learning. Advanced Engineering Informatics. 53, 101687.
DOI: https://doi.org/10.1016/j.aei.2022.101687