https://doi.org/10.2352/EI.2025.37.12.HPCI-172

Write Sentence with Images: Revisit the Large Vision Model with Visual Sentence

Quan Liu, Department of Computer Science, Vanderbilt University, Nashville, TN
Can Cui, Department of Computer Science, Vanderbilt University, Nashville, TN
Ruining Deng, Department of Computer Science, Vanderbilt University, Nashville, TN
Tianyuan Yao, Department of Computer Science, Vanderbilt University, Nashville, TN
Yuechen Yang, Department of Computer Science, Vanderbilt University, Nashville, TN
Yucheng Tang, NVIDIA Cooperation, Redmond, WA
Yuankai Huo, Department of Computer Science, Vanderbilt University, Nashville, TN

Yuankai Huo is the corresponding author, e-mail: yuankai.huo@vanderbilt.edu

Abstract

This paper introduces a novel framework for generating high-quality images from "visual sentences" extracted from video sequences. By combining a lightweight autoregressive model with a Vector Quantized Generative Adversarial Network (VQGAN), our approach achieves a favorable trade-off between computational efficiency and image fidelity. Unlike conventional methods that require substantial resources, the proposed framework efficiently captures sequential patterns in partially annotated frames and synthesizes coherent, contextually accurate images. Empirical results demonstrate that our method not only attains state-of-the-art performance on various benchmarks but also reduces inference overhead, making it well-suited for real-time and resource-constrained environments. Furthermore, we explore its applicability to medical image analysis, showcasing robust denoising, brightness adjustment, and segmentation capabilities. Overall, our contributions highlight an effective balance between performance and efficiency, paving the way for scalable and adaptive image generation across diverse multimedia domains.

Introduction

Image generation from textual descriptions has emerged as a significant area of research within the field of artificial intelligence, driven by its potential applications in diverse domains such as art creation, virtual reality, and assistive technologies. This process involves transforming a given text into a visually coherent and contextually relevant image, leveraging advanced deep learning models. The ability to generate images from text not only aids in visualizing concepts but also enhances human-computer interaction by enabling machines to understand and represent visual information in a human-like manner.

Despite the progress made, generating images from visual sentences derived from video sequences presents unique challenges. Visual sentences extracted from videos encapsulate dynamic and complex scenes, requiring the generation model to comprehend and faithfully represent the temporal and spatial nuances embedded in the textual descriptions. Traditional models often struggle with this complexity, leading to suboptimal image quality and coherence.

The primary objective of this research is to develop an efficient method for generating high-quality images from textual descriptions, particularly those derived from video sequences. This will be achieved through the implementation of a light autoregression model coupled with VQGAN, aiming to balance the trade-off between computational efficiency and image quality as shown in Fig 1.

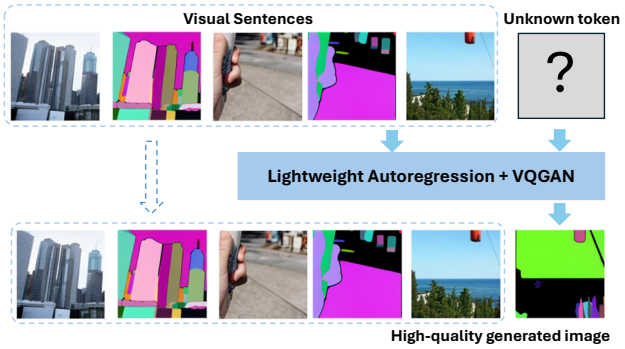

Figure 1. *Overview of the proposed pipeline: given partially annotated images (treated as "visual sentences") and an unknown token, the system uses a lightweight autoregressive model combined with VQGAN to predict and fill in missing content, producing high-quality, coherent image completions.*

This paper makes several key contributions to the field of image generation based on the visual sentence:

1. Introduction of the Light Autoregression Model: We explore a novel, lightweight autoregressive model designed to improve the efficiency of image generation. This model reduces the computational burden while maintaining the capability to generate detailed and accurate images from textual descriptions.

2. Integration with VQGAN: by integrating the light autoregression model with Vector Quantized Generative Adversarial Network (VQGAN), we leverage the strengths of both approaches. VQGAN is known for producing high-fidelity images utilizing a quantized latent space, which enhances texture and detail preservation.

3. Case study of visual sentence on medical image analysis: to leverage the knowledge of LLM to medical image, we provide insights on medical visual sentence construction.

Related work

Image generation under text guidance

In recent years, the field of image generation from text has witnessed significant advancements, driven by the development of sophisticated deep learning models [7, 8, 9, 10, 3]. Existing methods for generating images from textual descriptions primarily leverage autoregressive models and Variational Autoencoders (VAEs) [2, 11]. Autoregressive models, such as DALL-E and GPT-3, have shown remarkable ability to generate coherent and contextually relevant images by predicting each pixel or patch sequentially, thereby capturing intricate details from textual input [12, 13]. One of the prominent advancements in this domain is the introduction of VQGAN (Vector Quantized Generative Adversarial Network), which combines the strengths of GANs and VAEs to produce high-quality images [14, 15]. VQGAN excels in generating high-fidelity images by utilizing a quantized latent space, thus enabling better texture and detail preservation compared to traditional GANs [16, 17]. Alongside quality, the efficiency of image generation models has garnered considerable attention, with research focusing on optimizing model performance and reducing computational costs [18, 19, 20, 21, 22]. Techniques such as model pruning, knowledge distillation, and the integration of more efficient architectures have been explored to enhance the speed and scalability of image generation systems [23, 6]. This review aims to provide an in-depth analysis of these methods, highlighting their contributions and ongoing challenges in the quest for efficient and high-quality image generation from text.

Autoregression model in LLM

The autoregression model is a fundamental approach in large language models (LLMs), where the generation of text is conditioned on the sequential prediction of the next token based on previously generated tokens. This model has been widely adopted in prominent LLM architectures, such as GPT (Generative Pre-trained Transformers) [1], which utilize autoregressive techniques to predict the next word in a sequence by leveraging a unidirectional attention mechanism. This autoregressive nature allows for efficient text generation, as each token is conditioned on the history of the sequence, mimicking human-like text generation. However, autoregressive models also inherit challenges, such as compounding errors over long sequences and slower inference times compared to parallel decoding models, since tokens are generated one at a time. Despite these limitations, autoregression remains a cornerstone for generating coherent and contextually relevant outputs in various NLP tasks, including text completion, translation, and summarization, demonstrating its robustness in practical applications.

Method

The pipeline is shown in Fig. 2. The core idea is to treat an input sequence of partial or annotated images—referred to as a "visual sentence"—as a tokenized representation. A lightweight autoregressive model predicts missing tokens in this sequence, leveraging an embedding layer, a compact RNN (GNU), an attention mechanism, and a decoder to learn spatial and contextual relationships efficiently. These predicted tokens are subsequently transformed into the final high-quality image by a VQGAN decoder, resulting in a coherent completion that naturally blends with the given visual context.

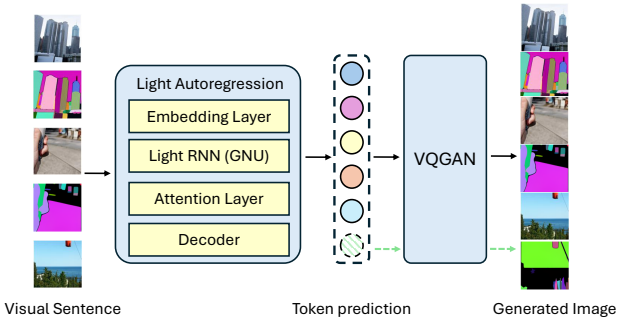

Figure 2. *Illustration of our image completion framework. A "visual sentence" (sequence of partial or annotated images) is processed by the Light Autoregression module—consisting of an embedding layer, a lightweight RNN (GNU), an attention mechanism, and a decoder—to predict missing tokens. These tokens are then fed into VQGAN, which synthesizes the final high-quality image completion.*

Visual Sentence Extraction

The first step in our proposed method involves the extraction of visual sentences from video sequences. Visual sentences are composed by analyzing the frames of a video and generating descriptive textual representations for each significant segment. This process includes detecting keyframes that capture the essence of the scene, followed by the application of natural language processing techniques to generate coherent sentences that describe the visual content. These sentences encapsulate the dynamic and spatial information present in the video, providing a comprehensive textual description that serves as input for the image generation model.

Light Autoregression Model Architecture

The light autoregression model is designed to enhance efficiency while maintaining high performance in generating images from text. The architecture consists of several key components:

Embedding Layer: Converts the input text into a dense vector representation. Recurrent Layers: Utilize lightweight recurrent neural networks (RNNs), such as GRUs (Gated Recurrent Units), to capture sequential dependencies in the text while keeping computational overhead low. Attention Mechanism: Incorporates an attention layer to focus on relevant parts of the text, improving the model's ability to generate contextually accurate images. Decoder: Transforms the processed text embeddings into pixel values, constructing the image sequentially.

VQGAN Integration

To further improve the quality of the generated images, we integrate VQGAN with the light autoregression model. VQGAN, a variant of GAN that operates in a quantized latent space, helps produce high-fidelity images with detailed textures and structures. We initialize VQGAN with pretrained weights to leverage existing knowledge (transfer learning), incorporate custom loss functions that emphasize both pixel-level accuracy and perceptual quality to ensure the generated images are both realistic and faithful to the input descriptions, and employ adversarial training techniques to refine image quality using a discriminator network to distinguish between real and generated images. The step-by-

step process for generating images from visual sentences involves the following: extracting keyframes from the video and generating textual descriptions for each segment (visual sentence extraction), converting the visual sentences into dense vector representations using the embedding layer (text embedding), using the light autoregression model to generate an initial image representation from the text embeddings (autoregressive generation), and passing the initial image representation through VQGAN to refine and enhance the image quality (VQGAN enhancement). The final high-quality image is then output, corresponding to the input visual sentence. This integrated approach ensures that the generated images are not only accurate representations of the input text but also exhibit high visual fidelity, meeting the dual objectives of efficiency and quality in image generation from textual descriptions.

Data and experimental design

Dataset

SA-1B is a large-scale dataset designed to enhance the capabilities of AI models in image generation tasks from [3]. Comprising over 1 billion image-text pairs, this dataset provides a diverse range of high-quality images paired with detailed textual descriptions. The dataset includes various categories such as objects, scenes, activities, and abstract concepts, ensuring comprehensive coverage of visual and contextual information.

Evaluation Metrics

The evaluation of our image generation model uses the Dice score to measure the similarity between generated images and ground truth images. The Dice score, defined as

$$\text{Dice Score} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

where A is the set of pixels in the generated image and B is the set of pixels in the ground truth image. Dice score assesses the overlap between two sets of data, A higher Dice score indicates greater similarity, reflecting the model's ability to produce accurate and high-quality images. This metric provides a rigorous and quantitative evaluation of the model's performance.

Baseline Comparison

The model setting follows LVM [4] and DeLVM [5], VQGAN is used as the image generator, the pretrained LLM is used for the auto-regression. The two components are most relevant with the model setting: LLM and VQGAN. The LLM is resource-intensive. To optimize efficiency of the LVM pipeline, we proceed in two ways: (1) use smaller LLM model and (2) maintain smaller codebook of VQGAN. The LLM is from LlaMA pretrained model by Meta. The size of LLM model are range from 7B, 1B and 300M. The default codebook parameter is 256 token, codebook shape is 8192×64. To test the effect of codebook shape, we change the codebook size to smaller size ranging from 8192, 4096, 2048 and 1024.

Medical image analysis case

To evaluate the SimLVM model performance on medical image analysis tasks, we create the medical image dataset: pathology image denoising dataset, MoNuSeg dataset [6] and pathology image brightness dataset. We construct the medical image visual sentences as the prompts and target images as query images. Three tasks are evaluated: image denoising, segmentation and brightness tuning.

Results

Our studies evaluate the model from two aspects: quantitative results by dice score and qualitative results by showing the generated images.

Quantitative result

The model is performed on the LLM model and VQGAN model with difference parameters. The purpose is to provide more efficiency model structure. The LVM model performance is shown in Table. 1. The model use LlaMA-300M and VQGAN when cookbook size is 8192 has the superior performance.

Qualitative result

In this section, we provide the qualitative results by showing the image generated by the SimLVM. In figure1, segmentation results provided by difference size of LLM is shown in Fig. 3. It is obvious to see the segmentation results by LlaMA-1B shows more details. The qualitative results are evaluated by showing the denoise and brightness adaptation on the pathology image data with pretrained model in Fig. 4,

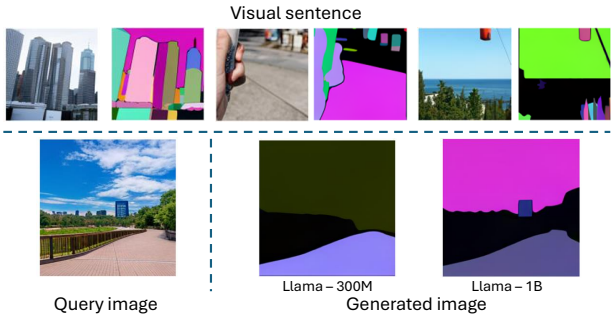

Figure 3. *Comparison of image completion results under different model capacities. The top row shows the "visual sentence" (contextual partial images), while the bottom row presents a new query image (left) and two generated completions (middle and right) produced by our Llama-300M and Llama-1B models, respectively.*

(2) One failed task on pathology image segmentation is also shown in Fig. 5. While the model correctly segments some regions, it overlooks subtle boundaries and misclassifies critical tissue structures (indicated by the highlighted regions). As a result, the predicted segmentation deviates significantly from the ground truth, emphasizing the challenges of capturing fine-grained features in complex histopathological data.

SimLVM application on medical image sentence

We perform the visual sentence on medical image fields from three aspects: image denoising, image brightness tuning and instance segmentation. In Fig. 4 first row, we generated results on the medical image denosing task. In Fig. 4 second row, changing the brightness of medical image is performing.

Method	LLM	Token	Codebook size	Codebook dim	Dice (%)
Original	7B	256	8192	64	31.01
SimLVM	1B	256	8192	64	50.79
	300M	256	8192	64	**61.25**
	300M	256	4096	64	23.79
	300M	256	2048	64	24.80
	300M	256	1024	64	21.70

For the LLM size, 'B' represent billion, 'M' represent million.

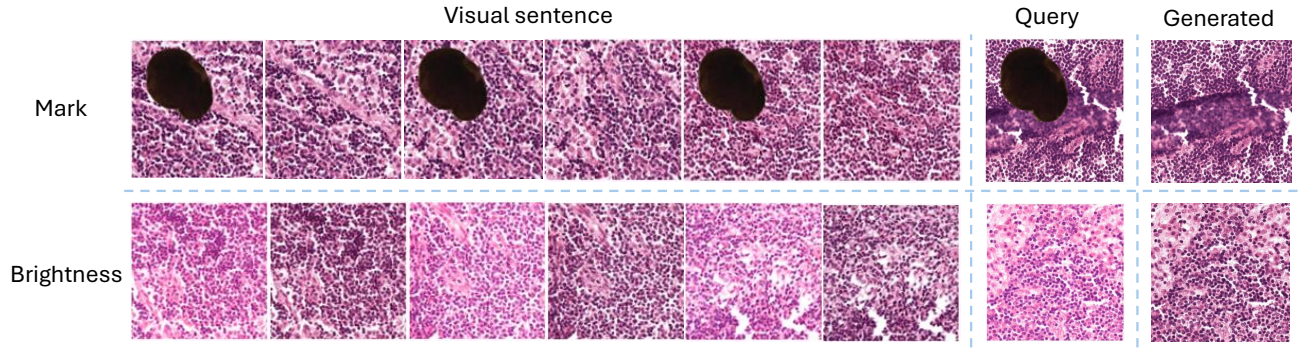

Figure 4. *Demonstration of domain transformations in histopathology images. The top row ("Mark") shows examples where a dark circular mark appears in different positions, while the bottom row ("Brightness") illustrates various intensity levels. For both transformations, the "visual sentence" (middle columns) provides partial contexts, followed by a new query image (second-to-last column) and our model's completed result (last column).*

Failed generation in medical image analysis task

Due the field gap between the medical image and natural image, the SimLVM fails on certain medical image analysis tasks. Fig. 5 shows the failure on pathology image segmentation task.

Discussion

The light autoregression model, with its streamlined architecture, reduces computational overhead without compromising image quality. The integration with VQGAN further enhances the visual fidelity of the generated images, resulting in outputs that are both detailed and realistic. These findings highlight the potential of our approach to generate high-quality images efficiently.

However, there are limitations to our current approach. One limitation is the dependency on the quality of the visual sentence extraction process; inaccuracies in textual descriptions can lead to suboptimal image generation. Additionally, while our model performs well on standard datasets, its robustness in handling highly complex or abstract visual scenes requires further validation. The computational cost, although reduced, remains significant for very large-scale applications.

Future work will focus on addressing these limitations and further improving the model. One potential direction is to enhance the visual sentence extraction process using more advanced natural language processing techniques to ensure more accurate and comprehensive textual descriptions. Another avenue for research is to optimize the light autoregression model and VQGAN integration for even greater efficiency, possibly through techniques like model distillation or the development of more compact network architectures. Moreover, expanding the training dataset to include a wider variety of scenes and contexts could improve the model's robustness and generalization capabilities. Exploring the use of multimodal inputs, such as combining text with additional metadata like audio or sensor data, could also enhance the richness and accuracy of the generated images.

Conclusion

In this work, we presented a lightweight autoregressive model integrated with VQGAN to address the task of image generation from visual sentences derived from videos. By tokenizing partially annotated frames and leveraging autoregressive prediction, our approach captures fine-grained sequential context while maintaining computational efficiency. The experimental evaluations, conducted on both general and medical imaging datasets, validated the ability of the proposed framework to generate high-fidelity images at lower computational cost compared to existing methods. Notably, our system demonstrated versatility in medical image tasks such as denoising, brightness tuning, and segmentation, underscoring the practical potential of adopting lightweight yet powerful generative models in specialized domains.

Despite these promising outcomes, several avenues remain open for future work. Refining the visual sentence extraction process with more advanced natural language and domain-specific techniques could yield richer contextual information, ultimately improving generation quality. Investigating strategies to further reduce the codebook size in VQGAN without compromising fidelity may benefit real-time applications even more. Additionally, applying the framework to higher-dimensional data (e.g., 3D volumetric images) and exploring multimodal cues (e.g., audio, text captions) represent natural extensions to broaden the impact of this research. By addressing these challenges, the proposed pipeline holds the potential to advance efficient, high-quality image generation for a range of academic, industrial, and clinical

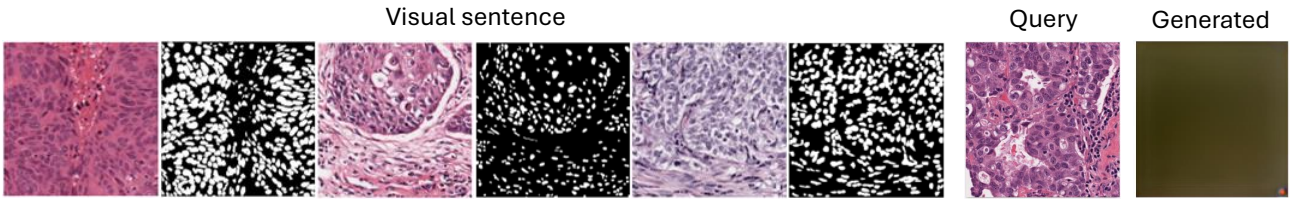

Figure 5. *Failure case example in medical image segmentation. The generated segmentation mask shows the limitation of SimLVm on segmentation tasks.*

applications.

Acknowledgment

YH and QL acknowledge support from NIH under contract R01DK135597. YH is the corresponding author.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, (2023).

[2] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. "Masked Autoencoders are Scalable Vision Learners," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, (2022).

[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. "Segment Anything," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, (2023).

[4] Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A. L., Darrell, T., Malik, J., and Efros, A. A. "Sequential Modeling Enables Scalable Learning for Large Vision Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22861–22872, (2024).

[5] Guo, J., Hao, Z., Wang, C., Tang, Y., Wu, H., Hu, H., Han, K., and Xu, C. "Data-Efficient Large Vision Models Through Sequential Autoregression," *arXiv preprint arXiv:2402.04841*, (2024).

[6] Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, Ö F., Tsougenis, E., Chen, H., Heng, P.-A., Li, J., Hu, Z., et al. "A Multi-Organ Nucleus Segmentation Challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, (2019).

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., et al. "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, (2020).

[8] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. "Zero-Shot Text-to-Image Generation," *International Conference on Machine Learning (ICML)*, (2021).

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, (2017).

[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al. "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, (2023).

[11] Kingma, D. P. and Welling, M. "Auto-Encoding Variational Bayes," *2nd International Conference on Learning Representations (ICLR)*, (2014).

[12] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017).

[13] Chen, M., Radford, A., Child, R., et al. "Generative Pretraining from Pixels," *International Conference on Machine Learning (ICML)*, (2020).

[14] Dhariwal, P. and Nichol, A. "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, (2021).

[15] Gafni, O., Polyak, A., Ashual, O., et al. "Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors," *arXiv preprint arXiv:2203.13131*, (2022).

[16] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. "Image-to-Image Translation with Conditional Adversarial Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, (2017).

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NeurIPS)*, (2014).

[18] Ronneberger, O., Fischer, P., and Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, (2015).

[19] Zhou, H., Ramesh, S., Kannan, H., et al. "LAFITE: Towards Language-Free Training for Text-to-Image Generation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022).

[20] Yin, X. and Liu, H., and Shao, L. "A Survey of Deep Learning for Image Synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2022).

[21] Xie, S., Wang, L., and Moens, M.-F. "A Survey on Temporal and Multimodal Deep Learning for Video Classification," *IEEE Transactions on Multimedia*, vol. 24, pp. 1076–1093, (2022).

[22] Hou, L., Chen, X., Hu, Y., et al. "MinCutNet: A Minimum-Cut Network for Tissue Segmentation in Histopathology," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 653–661, (2022).

[23] Radford, A., Wu, J., Child, R., et al. "Language Models are Unsupervised Multitask Learners," *OpenAI Technical Report*, (2019).

Author Biography

Quan Liu is a PhD in Computer Science at Vanderbilt University, working with Dr. Yuankai Huo. He received his Bachelor's degree from Huazhong University of Science and Technology. His work has focused on the medical image analysis with computer vision techniques.