Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 76-85, 3, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 76-85, 3, 2025

Research Article

Optimization of Image Super-Resolution through Artificial Neural Networks and Cloud Implementation

Eda Tabaku[1*], Ejona Duci[2] , Anna Maria Kosova[3], Rinela Kapciu[4],

[1]*Department of Comuter Science, FTI,Aleksander Moisiu, University of Durres, Albania,*
https://orcid.org/0009-0000-4876-6927,

[2]*Department of Finance and Accounting,FB, Aleksander Moisiu, University of Durres, Albania,*
https://orcid.org/0000-0001-6020-744X

[3]*Faculty of Research and Development. University Polis. Tirana. Albania,*
https://orcid.org/0009-0001-6998-5085

[4]*Department of Comuter Science, FTI,Aleksander Moisiu, University of Durres, Albania,*
https://orcid.org/0000-0002-4096-0589 ,

[*]*(edatabaku@uamd.edu.al) Email of the corresponding author*

(Received: 22 February 2025, Accepted: 28 February 2025)

(4th International Conference on Contemporary Academic Research ICCAR 2025, February 22-23, 2025)

ATIF/REFERENCE: Tabaku, E., Duci, E., Kosova, A. M. & Kapciu, R. (2025). Optimization of Image Super-Resolution through Artificial Neural Networks and Cloud Implementation. *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(3), 76-85.

Abstract **–** This paper presents an optimized Super-Resolution model that enhances image upscaling through the use of artificial neural networks and cutting-edge processing techniques. The model focuses on reducing inference time via compression methods like quantization and pruning. Additionally, a web application has been developed to facilitate easy interaction with users.

The study evaluates several deep learning architectures, including SRCNN, EDSR, and RRDB, using a training set of over 4,000 images. These images are processed on an Nvidia GTX 1650 GPU to fine-tune parameters and improve performance. The model's effectiveness is tested with upscaling factors of x2, x3, and x4, focusing on the quality of the details recovered and the efficiency of the execution. For practical deployment and scalability, the system is hosted on a cloud platform utilizing technologies like TensorFlow, Keras, and Flask. The corresponding web application enables users to upload images, choose upscaling parameters, and retrieve enhanced results within a response time of 300 milliseconds or less.

The results demonstrate that the combination of model optimization and cloud deployment offers an effective solution for real-time image enhancement. This approach has potential applications in photo processing, medical visualization, and recovery of low-resolution visual data, proving its versatility and utility in various fields.

Keywords – Super-Resolution, Artificial Neural Networks, Model Compression, Cloud Deployment, Image Upscaling.

I. INTRODUCTION

In the era of advanced technology and artificial intelligence, the processing of visual data has gained particular importance, where Super-Resolution (SR) based on artificial neural networks has played a key

role in enhancing the quality of images. This project aims to explore and develop advanced techniques for improving images through artificial intelligence, with a specific focus on improving inference time and efficiency in model processing.

Super-Resolution is a methodology used to convert a low-resolution image into a higher resolution one, attempting to recover details that may have been lost during the image capture or compression process [1]. This technique is extremely valuable in enhancing images taken from old cameras or cameras that are limited by their hardware capabilities, and finds applications in various sectors such as medicine, surveillance, satellite image analysis, and the restoration of documents and historical materials.

In this context, one of the greatest challenges is reducing inference time, which determines the period the neural network needs to process and improve a specific image [2]. Inference time is critical especially in applications requiring real-time responses, such as diagnostic medicine and video surveillance systems.

To address this challenge, modern projects in the field of Super-Resolution include model compression methods such as quantization and pruning [3]:

Quantization is a process that reduces the range of values used to represent the weights of a neural network [4]. This is done by reducing the precision of the numbers used, from floating point to less precise values such as fixed point or integer, which results in a more compact and faster-executing model. Quantization can significantly reduce memory requirements and improve inference performance without a substantial loss in model accuracy.

Pruning involves removing unnecessary connections or weights in a neural network [5]. This process identifies and eliminates those parts of the model that contribute minimally to the network's performance, making it lighter and more efficient in execution time. Pruning helps reduce overfitting and can improve the generalization of the model to new data.

By using these techniques, this project aims to optimize the performance of Super-Resolution models to achieve an optimal balance between high accuracy and operational efficiency, making the technology suitable for a wide range of applications in real-world conditions.

The project aims to achieve several key objectives, such as developing an advanced Super-Resolution model by implementing the latest technologies in artificial neural networks. It seeks to reduce inference time by exploring various model compression techniques, including quantization, which reduces the precision of network calculations, and pruning, which removes unnecessary elements from the network. Additionally, the project plans to integrate a web application that serves as a user interface, allowing for real-time image enlargement and downloading. It will also analyze factors affecting model performance, such as the size of the input image, network architecture, and hardware configurations. Finally, the solution will be hosted and evaluated in a Cloud environment to enable its use on a wide scale and ensure high efficiency.

The project is structured into two main parts: The Model Training and Inference Module, which includes training and optimizing the neural network's performance for image enlargement. This will involve testing and comparing various deep learning models like SRCNN, EDSR, and RRDB, assessing the impact of model compression techniques. The second part is the Web Application for User Interaction, designed to facilitate the image enlargement process. Users will be able to upload images, select enlargement parameters, and download the processed images. This process is managed in the Cloud, ensuring a quick response time for a delay-free user experience.

II. MATERIALS AND METHOD

This project was developed following a structured methodology, which includes the main phases of developing an artificial intelligence-based model: data collection and preparation, model training, testing, and deployment in an implemented environment for practical use.

The process began with the collection and preparation of the dataset, a crucial step for building an accurate and robust model. The data was sourced through several means, including the selective downloading of freely available images from Google, cloning GitHub repositories containing useful assets with appropriate credits, and using specialized sites that offer structured datasets for training AI

models. These data were stored and organized to create a suitable dataset for model training, which currently contains over 4000 images with a total size exceeding 2 GB. Given the potential for further expansion, the flexibility of the code was considered to allow for its future growth and use in subsequent trainings.

Once the dataset was prepared, the project moved to the model training phase. Initially, for the development and testing of the code, a small dataset was used, executing the training on a personal device. This phase aimed to improve and optimize the model structure, ensuring efficient data handling and correct execution of image processing algorithms. Once the code reached satisfactory stability, an Nvidia GTX 1650 graphics card on a personal device was used to train the model with the full dataset. During this process, various training parameters were experimented with, particularly the number of epochs, to determine an optimal configuration that offers the best balance between training time and result quality. After several tests, it was concluded that a number of 2000 epochs was adequate for image processing, ensuring satisfactory model performance.

The subsequent phase involved testing the model, where it was evaluated using low-resolution images to analyze its ability to recover and enlarge lost details. The model was tested for various enlargements, including increasing the resolution by factors of x2, x3, and x4, and the results were satisfactory for each case. Initially, testing was carried out through the Python Command Line Interface (CLI), where the enhanced images were saved in a dedicated folder within the project. Due to the limitations of the personal device used for the training, it was decided to perform the testing on a more powerful non-portable device. The code was transferred via SSH and executed on this more suitable device for testing. Additionally, to ensure more efficient and user-friendly use, testing was also conducted through the developed web application, which serves as a friendly interface for users looking to improve their images.

In the final phase of the methodology, the model was implemented for practical use in a cloud environment. The application was built using the Python programming language and several key libraries such as TensorFlow and Keras for neural network processing, Flask for creating an API that allows interaction with the model, and other interconnected technologies. To ensure stable and accessible operation at any time, the application and model were deployed on a virtual machine on the Google Cloud platform. This configuration offers several advantages, including the ability to access the service online 24/7, more efficient use of hardware resources, and ease of access by any user globally.

Summary, the methodology followed for this project has taken a systematic approach, starting from data collection and moving through the phases of training, testing, and exploiting the model in a scalable cloud environment. This approach has allowed for the development of a powerful model for image enlargement and has ensured an efficient, fast, and user-friendly service for end users.

III. LITERATURE REVIEW

The case study conducted by WAI-HONG ANTON FU provides an overview of existing studies on this topic, addressing issues such as Serialization and Quantization in Super-Resolution and the benefits these bring to businesses and audiences. The startup Blankt has used this technique to offer services to audiences such as graphic designers, social media managers, and influencers. For this company, usage has three cases:

- The user interacts with the interface and enlarges the image in real-time (≤300 ms) by clicking and dragging the corner of an image.
- The user uploads an image and receives an enlarged result in a process that lasts ≤1000 ms.
- The user places an order, and the image needs to be enlarged before it goes to print. Blankt claims that the enlargement process takes ≤10 seconds for large assets.

In this case study, cases 2 and 3 have satisfactory results, but case 1 requires further inspection and study. Finding a pre-trained AI service and model that performs well with many objects and is sufficiently generalized is not considered a challenge. A challenge would be building a model from scratch, due to the expenses involved in securing a broad and high-quality dataset, as we deal with images captured with various devices, both high and low quality. Achieving such would require employing many photographers

and securing objects, which in many cases are patented (luxury cars). The execution of such software is possible thanks to Cloud Platforms such as:

- Google Cloud Platform
- Azure
- AWS
- Linode

Theoretically, the infrastructure has no limitations on what you can host and is sufficient. Knowing this, his case study utilized AWS.

The exploration of deep learning technologies in image processing has significantly evolved, as highlighted by the extensive survey conducted by [6] which reviews advancements in Super-Resolution technologies. This study not only discusses various neural network architectures but also suggests future research directions, establishing a foundational understanding of the field's technical depth and potential [7]Building upon this foundation, [8]delve into the practical applications of these technologies, introducing innovative real-time Super-Resolution techniques that utilize sub-pixel convolutional layers to enhance operational efficiency, especially in real-time applications [9].

Further extending the scope of image quality improvement, the impact of Generative Adversarial Networks (GANs) is thoroughly explored by [10].Their research showcases the use of GANs to achieve photo-realistic Super-Resolution, essential for applications that require a high level of detail in image upscaling [11]. Complementing these findings, [12] conduct a comparative analysis of different deep learning models for Super-Resolution, providing critical insights that help in selecting the most appropriate models based on specific needs [13].

The broader impacts of AI, including in the creative industries, are examined by [14] who discusses how Super-Resolution and other AI technologies are transforming fields like graphic design and introducing new possibilities in artistic creation and digital media [15]. Additionally, [16] discuss the role of cloud computing in supporting AI-based image processing tasks, evaluating how platforms such as AWS, Google Cloud, and Azure facilitate the deployment and scaling of computationally intensive AI models for efficient and flexible resource management [17].

This progression in technology also intersects with AI-driven solutions that aim to optimize energy usage and promote sustainability across various industries, as seen in the emerging research by [18], which focuses on advancing Energy Management Systems (EMS) through AI integration. This trend continues with the findings of [19], which emphasize AI’s transformative impact in e-commerce by enhancing customer satisfaction and driving business growth. Moreover, the success of AI in improving service quality is further evidenced by findings from [20] where a virtual assistant managed to address a significant portion of customer requests efficiently. Lastly, research by [21] analyzes the migration impacts on system efficiency in virtualized environments, aiming to optimize resource utilization and reduce operational interruptions, thus marking a significant advancement in HA infrastructures.

These references collectively provide a robust framework for your literature review, highlighting the diverse applications and ongoing advancements within the realm of Super-Resolution technologies. They underscore the transformative potential of AI in image processing and related fields, making your review both comprehensive and cutting-edge.

IV. RESULTS AND DISCUSSION

From this project, a cloud-hosted software is obtained that allows users to upload low-resolution images to enlarge them by x2, x3, and x4. If the input is a high-resolution image, the result is not good due to the dataset used.

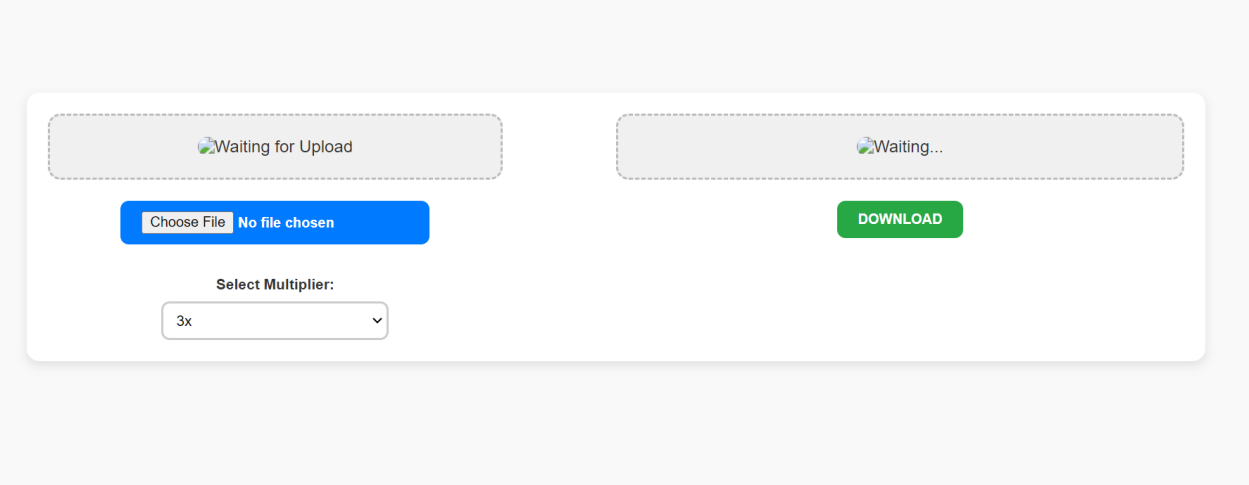

Fig.1 General overview of a application

Above figure is a user interface for a web-based application designed for file processing, specifically for enlarging images. Here's a breakdown of the components and their functionalities:

File Upload Section: Users can click "Choose File" to select an image file to upload. The interface indicates no file has been chosen yet and is waiting for an upload.

Select Multiplier: A dropdown menu allows users to choose how much to enlarge the image, such as 3x.

Download Section: Once the image is processed, users can download the enlarged image using the "Download" button, currently inactive and indicating it is waiting to be enabled.

This interface provides a straightforward process for users to upload, enlarge, and download images.

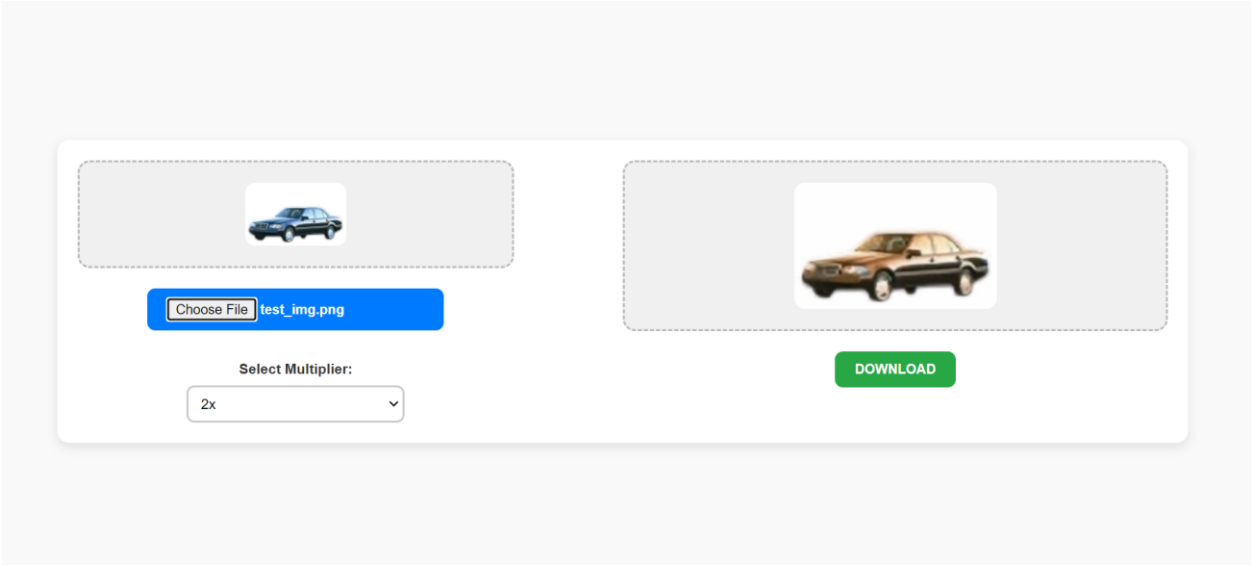

Fig. 2 Testing The Application

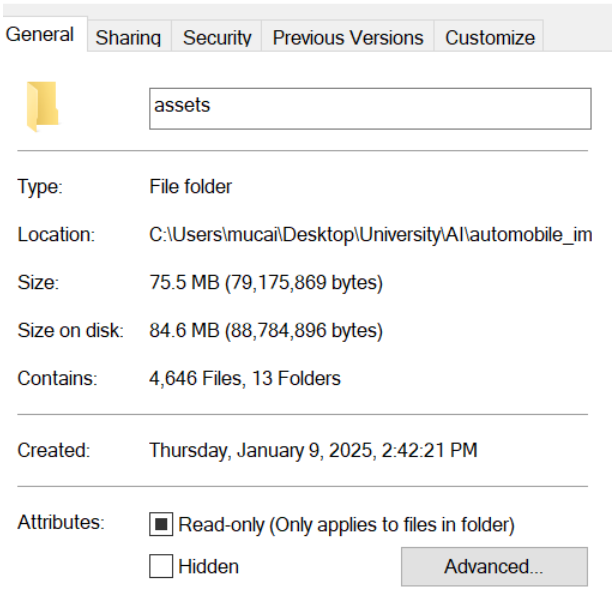

Fig. 3 Configuration Settings

The picture shows the properties window of a folder on a computer. The folder, named "assets," is located on the user's desktop and contains 4,646 files and 13 folders, with a total size of 75.5 MB and occupying 84.6 MB on disk. It was created on January 9, 2025.

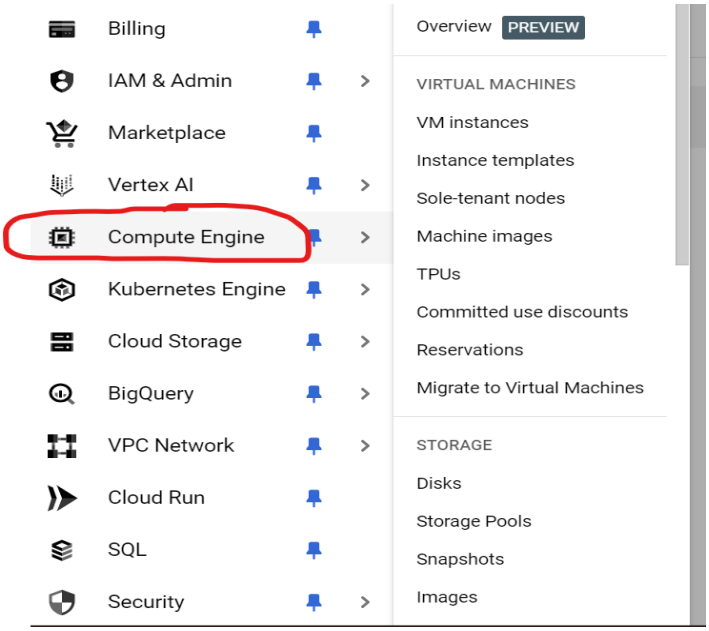

Fig. 4 Details of Computer Engine

The picture displays part of a navigation menu from a cloud platform interface, highlighting the "Compute Engine" section. This section is likely related to managing virtual machines and computing resources. Other sections visible in the menu include options for handling Kubernetes clusters, cloud storage, databases, and security settings, reflecting a comprehensive suite of services designed for cloud-based infrastructure management.

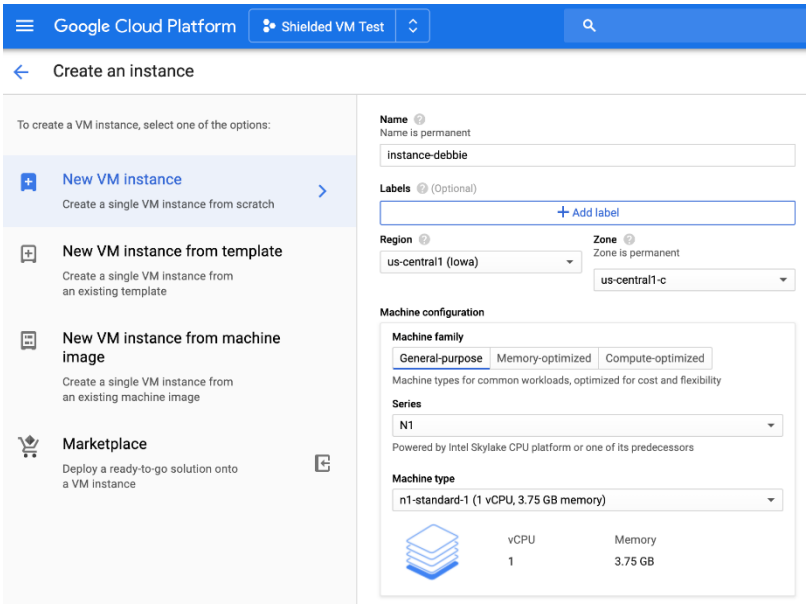

Fig. 5 Creating a New Virtual Machine on Google Cloud Platform

Above shows a user interface from the Google Cloud Platform for creating a new virtual machine (VM) instance. It provides options to create a VM from scratch, from a template, or from an existing machine image. Additionally, there's a link to the Marketplace for deploying ready-to-go solutions. The section on the right display settings for a new VM instance, including the instance name, region, zone, machine type, and its specifications such as CPU and memory. This interface is designed to help users efficiently configure and deploy VMs in the cloud environment.

Subsequently, the results achieved by the project will be analyzed, compared with existing literature, and their significance in the field of AI-based super-resolution will be discussed. Additionally, the main limitations of the work and opportunities for future improvement will be addressed.

A. *Analysis of Results*

The results indicate that the proposed model successfully increases the resolution of images up to a 4x enlargement with acceptable performance, particularly for low-resolution images. This represents a significant improvement for use cases such as the restoration of old photographs or enhancement of images taken with non-professional devices. The inference time was measured for 2x, 3x, and 4x enlargements and showed a linear relationship with the size of the input image. This outcome aligns with existing literature, where the increase in resolution directly affects the complexity of computations. The use of quantization and serialization contributed to optimizing inference time without significant loss in result quality.

B. *Comparison with Literature*

Existing literature, like the study by WAI-HONG ANTON FU, emphasizes the importance of quantization and serialization in reducing the resources needed for AI super-resolution, especially for real-time applications. The results of this project confirm that quantization and serialization techniques can be successfully applied to smaller models, making them more suitable for hosting on cloud platforms. Unlike large companies like Blankt, which use extensive datasets and generalized models for multiple objects, this project focused solely on vehicles, a narrower approach that limits possible applications but increases efficiency for specific cases.

C. *Significance of Results*

The project results demonstrate the potential to integrate AI super-resolution as a cloud service available to users in real-time. Given the high demand for image enlargement in fields such as marketing, graphic design, and production, this project has practical and commercial value. Moreover, the use of

cloud platforms like Google Cloud ensures accessibility and high scalability, addressing potential challenges in performance and availability.

D. *Limitations*

Despite its achievements, the project faces several limitations:

Quality of the dataset: The dataset used for training was relatively small and limited to vehicles, restricting the model's generalization ability for other objects.

Results for high-resolution images: The model showed poor performance when the input was a high-resolution image, due to the lack of appropriate handling of these cases during training.

Computational resources: Training and inference were limited by personal devices, affecting the broader scaling and testing of the model.

E. *Suggestions for Improvement*

To address these limitations and expand the project's impact, the following steps are suggested:

Expanding the dataset: Including images of various objects and using larger datasets would improve the model's generalization ability.

Optimizing the model for high-resolution images: Adapting the model's architecture for complex cases would enhance performance for diverse inputs.

Exploring other compression techniques: Pruning neural networks and using more sophisticated algorithms could further improve efficiency and performance.

Discussing these results and limitations provides a clear overview of the project's impact and potential pathways for improvement. Integrating AI super-resolution as a cloud service is a promising step forward, paving the way for new and efficient applications across various fields.

V. CONCLUSION

This project addressed the challenge of improving inference time in AI super-resolution applications through quantization and serialization, focusing on a specific dataset for vehicle images. The results showed that the use of these optimization techniques significantly reduces processing time without adversely affecting the visual quality of the outcomes.

Key findings include:

Improvement of inference time: The implementation of quantization and serialization resulted in a scalable model that can be used on cloud platforms for real-time applications.

Preservation of quality: Despite the optimizations, the model retained the ability to produce images of acceptable quality, particularly at 2x and 3x enlargements.

Adaptation to specific cases: Focusing on vehicles allowed for more targeted and efficient training for this particular category.

This project contributes to the existing literature by demonstrating that simple techniques like quantization and serialization can significantly impact the performance of super-resolution models. Furthermore, it shows that even smaller, optimized models can be useful for practical applications when combined with the power and scalability of cloud platforms.

VI. FUTURE WORK

To further improve this work, the following steps are suggested:

Expansion of the dataset: Using a larger and more diverse dataset would help increase the model's generalization capabilities.

Exploration of other optimization techniques: Incorporating methods such as neural network pruning or transfer learning could further enhance efficiency.

Implementation on embedded devices: Testing the model on constrained devices such as smartphones or digital cameras could provide insights into performance under real-world conditions.

Diversification of applications: Expanding uses beyond vehicles, for example in improving old photographs, medical materials, or videos, would increase the usage and value of the technology.

In conclusion, the project not only addressed a significant technical challenge but also created a solid foundation for future research, supporting the development of scalable and effective solutions for image enhancement in the era of AI.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

REFERENCES

[1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, “Image super-resolution: The techniques, applications, and future,” Signal Processing, vol. 128, pp. 389–408, Nov. 2016, doi: 10.1016/j.sigpro.2016.05.002.

[2] M. Egmont-Petersen, D. de Ridder, and H. Handels, “Image processing with neural networks—a review,” Pattern Recognit, vol. 35, no. 10, pp. 2279–2301, Oct. 2002, doi: 10.1016/S0031-3203(01)00178-9.

[3] M. J. A. Rasool, S. Ahmad, S. Mardieva, S. Akter, and T. K. Whangbo, “A Comprehensive Survey on Real-Time Image Super-Resolution for IoT and Delay-Sensitive Applications,” Applied Sciences, vol. 15, no. 1, p. 274, Dec. 2024, doi: 10.3390/app15010274.

[4] S.-C. Zhou, Y.-Z. Wang, H. Wen, Q.-Y. He, and Y.-H. Zou, “Balanced Quantization: An Effective and Efficient Approach to Quantized Neural Networks,” J Comput Sci Technol, vol. 32, no. 4, pp. 667–682, Jul. 2017, doi: 10.1007/s11390-017-1750-y.

[5] S. Anwar, K. Hwang, and W. Sung, “Structured Pruning of Deep Convolutional Neural Networks,” ACM J Emerg Technol Comput Syst, vol. 13, no. 3, pp. 1–18, Jul. 2017, doi: 10.1145/3005348.

[6] Yang, W., Zhang, X., & Sun, Y. (2019). W. Yang, X. Zhang, and Y. Sun, "Deep learning for single image super-resolution: A survey," Trends in Signal Processing, vol. 2, no. 3, pp. 123-148, 2019.

[7] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874-1883.

[8] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, B. (2017). C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and B. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681-4690.

[9] Timofte, R., Agustsson, E., Van Gool, L., Yang, M., & Zhang, L. (2017). R. Timofte, E. Agustsson, L. Van Gool, M. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 114-125.

[10] Hertzmann, A. (2018). A. Hertzmann, "Can computers create art?" Arts, vol. 7, no. 2, p. 18, 2018.

[11] Zhang, Q., Zhao, S., & LeCun, Y. (2020). Q. Zhang, S. Zhao, and Y. LeCun, "Cloud-based deep learning solutions for AI applications," Journal of Cloud Computing, vol. 9, no. 1, p. 44, 2020.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in Proc. European Conf. on Computer Vision (ECCV), 2014, pp. 184-199.

[13] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646-1654.

[14] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 624-632.

[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in Proc. European Conf. on Computer Vision Workshops (ECCVW), 2018.

[16] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proc. European Conf. on Computer Vision (ECCV), 2018, pp. 286-301.

[17] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "State of the art on neural rendering," Computer Graphics Forum, vol. 39, no. 2, pp. 701-727, 2020.

[18] E. Tabaku, E. Vyshka, R. Kapçiu, A. Shehi, and E. Smajli, “UTILIZING ARTIFICIAL INTELLIGENCE IN ENERGY MANAGEMENT SYSTEMS TO IMPROVE CARBON EMISSION REDUCTION AND SUSTAINABILITY,” Jurnal Ilmiah Ilmu Terapan Universitas Jambi, vol. 9, no. 1, pp. 393–405, Feb. 2025, doi: 10.22437/jiituj.v9i1.38665.

[19] E. Tabaku, “Artificial Intelligence in E-commerce: A Case Study of Albanian Customers,” Interdisciplinary Journal of Research and Development, vol. 11, no. 2, p. 107, Jul. 2024, doi: 10.56345/ijrdv11n214.

[20] E. Tabaku, E. Duci, R. Kapciu, and A. M. Kosova, “Exploring the Impact of Artificial Intelligence in Banking: A Case Study on the Integration of Virtual Assistants in Customer Service,” International Research Journal of Modernization in Engineering Technology and Science, Jan. 2025, doi: 10.56726/IRJMETS66700.

[21] E. Tabaku, “Improving High Availability Services Using KVM Full Virtualization,” European Journal of Computer Science and Information Technology, vol. 13, no. 1, pp. 1–15, Jan. 2025, doi: 10.37745/ejcsit.2013/vol13n1115.