**MINISTRY OF EDUCATION AND TRAINING**

**FPT UNIVERSITY**

A DEEP LEARNING APPROACH FOR COLORIZING BLACK-AND-WHITE PHOTOGRAPHS IN VIETNAM BEFORE 1950

by

Le Tuan Anh

A thesis submitted in conformity with the requirements  
for the degree of Master of Software Engineering

© Copyright by Le Tuan Anh – 2025

**MINISTRY OF EDUCATION AND TRAINING**

**FPT UNIVERSITY**

A DEEP LEARNING APPROACH FOR COLORIZING BLACK-AND-WHITE PHOTOGRAPHS IN VIETNAM BEFORE 1950

by

Le Tuan Anh

A thesis submitted in conformity with the requirements  
for the degree of Master of Software Engineering

Supervisor:

Dr. Tran Van Ha

© Copyright by Le Tuan Anh – 2025

A Deep Learning Approach for Colorizing Black-and-White Photographs in Vietnam before 1950

Le Tuan Anh

Degree Master of Software Engineering

FPT University

2025

# Abstract

<Write here>

Giới thiệu ngắn gọn về bài toán colorization:

* Ảnh đen trắng là một phần quan trọng trong lịch sử, nhưng do lúc đó khoa học kỹ thuật chưa phát triển, nên chỉ có thể tạo những bức ảnh đen trắng, thiếu mất thông tin màu sắc.
* Colorization có thể giúp khôi phục (đánh màu, từ khôi phục không chuẩn lắm, vì vốn dĩ nó ko có màu, giờ mình đánh màu) màu sắc, tạo ra giá trị văn hóa và nghiên cứu.

Phương pháp đề xuất:

* Áp dụng mô hình U-Net để tô màu ảnh lịch sử.
* Kết hợp MSE Loss và Perceptual Loss để cân bằng giữa độ chính xác màu và chi tiết.
* Dữ liệu huấn luyện: 250,000 ảnh từ phim hiện đại nhưng có bối cảnh lịch sử Việt Nam trong tước những năm 1950.

Kết quả đạt được:

* So sánh giữa các loss function khác nhau
* Cần đưa ra các con số cụ thể của các kết quả đạt được bên trên

Phát triển ứng dụng Streamlit giúp người dùng thử nghiệm mô hình trực tiếp.

Hướng phát triển tương lai:

* Cải thiện kiến trúc mạng bằng Transformer-based models.
* Mở rộng tập dữ liệu với ảnh lịch sử thực tế.
* Tối ưu tốc độ inference để ứng dụng thực tế.

# Acknowledgements

<Write here>

Lời cảm ơn đến giảng viên hướng dẫn, đã cung cấp kiến thức và định hướng trong nghiên cứu.

Cảm ơn đồng nghiệp, bạn bè, gia đình đã hỗ trợ về mặt tinh thần và góp ý trong quá trình làm luận án.

Cảm ơn các tài nguyên mở như TensorFlow, Streamlit giúp phát triển và triển khai mô hình thực tiễn cho người dùng cuối.

# Table of Contents

[Abstract I](#_Toc196232116)

[Acknowledgements II](#_Toc196232117)

[Table of Contents III](#_Toc196232118)

[List of Tables IV](#_Toc196232119)

[List of Figures V](#_Toc196232120)

[1. Introduction 1](#_Toc196232121)

[1.1. Problem Statement 1](#_Toc196232122)

[1.2. Research Objectives 2](#_Toc196232123)

[1.3. Thesis Contribution 4](#_Toc196232124)

[2. Literature Review 4](#_Toc196232125)

[2.1. Overview of Image Colorization 4](#_Toc196232126)

[2.2. Deep Learning-based Approaches 5](#_Toc196232127)

[2.3. Loss Functions for Colorization 6](#_Toc196232128)

[3. Methodology 6](#_Toc196232129)

[3.1. Dataset Preparation 6](#_Toc196232130)

[3.2. Proposed Model Architecture 6](#_Toc196232131)

[3.3. Loss Function Selection 7](#_Toc196232132)

[4. Experiments and Results 7](#_Toc196232133)

[4.1. Training Process 7](#_Toc196232134)

[4.2. Evaluation Metrics 8](#_Toc196232135)

[4.3. Qualitative Results 9](#_Toc196232136)

[5. Discussion 9](#_Toc196232137)

[5.1. Impact of Different Loss Functions 9](#_Toc196232138)

[5.2. Limitations 10](#_Toc196232139)

[6. Model Deployment Using Streamlit 10](#_Toc196232140)

[7. Conclusion and Future Work 10](#_Toc196232141)

[7.1. Summary of Contributions 10](#_Toc196232142)

[7.2. Future Work 10](#_Toc196232143)

[References 12](#_Toc196232144)

# List of Tables

**No table of figures entries found.**

# List of Figures

[Figure 1‑1. Example colorization of a grayscale Vietnamese historical portrait. Source: DeOldify model output [6] 2](#_Toc196232145)

[Figure 1‑2. The pipeline of grayscale input and its colorized output. 3](#_Toc196232146)

# Introduction

## Problem Statement

Black-and-white photographs have long served as crucial visual records of the past. However, the absence of color in these images presents a major limitation – while grayscale photos retain spatial and structural information such as shapes and contrast, they fail to convey the full spectrum of visual cues that color naturally provides. Color plays a vital role in helping viewers perceive materials, distinguish objects, and emotionally connect with scenes. As a result, black-and-white images are often perceived as distant or abstract, particularly by modern audiences accustomed to color-rich media.

In the context of Vietnamese history, especially the period before 1950, most visual archives are preserved in black and white – including photographs of people, landscapes, and socio-political events. These images, despite their historical value, are difficult to relate to for younger generations. Restoring color to them is not merely an aesthetic enhancement but a culturally meaningful act. It bridges generations, supports educational storytelling, and helps preserve national heritage by reviving the vibrancy of lost moments in time. Recent Vietnamese media articles also highlight the efforts of local youth in reviving black-and-white photos of historical figures and events using digital tools, emphasizing the emotional and educational importance of this task [1] [2]. Figure 1‑1 demonstrates how a black-and-white image of a Vietnamese historical figure is transformed through automatic colorization. The addition of realistic skin tones, background hues, and clothing colors significantly enhances the interpretability and emotional connection of the visual.

Despite its promise, image colorization remains a technically complex and underdetermined problem. A single grayscale pixel can correspond to many plausible colors, making the task highly ambiguous. The “one-to-many” mapping issue in colorization was first emphasized by early works such as [3] and [4], which highlighted the inherent ambiguity of inferring chrominance from luminance alone. Moreover, historical photographs often suffer from additional degradation such as blurriness, noise, or uneven lighting – all of which further reduce the reliability of pixel-wise inference.

To tackle these challenges, recent deep learning-based methods aim to predict chrominance information (A and B channels in LAB color space) from the luminance channel alone. Notably, [5] proposed using class-rebalanced CNNs trained on large datasets to produce semantically meaningful colors. However, even these methods face difficulties with generalization, especially when applied to historical Vietnamese data, which lacks large-scale annotated datasets. Ensuring the realism and cultural accuracy of the colors thus remains a key research challenge.



Figure ‑. Example colorization of a grayscale Vietnamese historical portrait.  
Source: DeOldify model output [6]

## Research Objectives

This thesis aims to develop a deep learning-based system capable of colorizing historical black-and-white photographs from Vietnam prior to 1950. These photographs are primarily portraits extracted from modern films set in historical Vietnamese contexts. The ultimate goal is to enhance the realism and emotional resonance of such images while preserving cultural authenticity. To achieve this goal, the following objectives are pursued:

1. **Develop a robust image colorization model using deep learning techniques:** A U-Net-based architecture [7] is employed as the core model due to its proven effectiveness in image-to-image translation tasks. The model is trained to infer the chrominance (ab channels) from the luminance (L channel) in the CIELAB color space. Input images are preprocessed to (224×224) resolution, normalized, and separated into channels. Figure 1‑2 shows the pipeline of grayscale input and its colorized output using one of the trained models.



Figure ‑. The pipeline of grayscale input and its colorized output.

1. **Evaluate and compare different loss functions for image colorization:** The thesis experiments with three loss configurations to determine which best preserves realism and structural accuracy:
   1. Mean Squared Error (MSE): A traditional pixel-wise loss that minimizes average squared differences.
   2. Perceptual Loss using VGG16: Computes loss in feature space based on intermediate activations of a pre-trained VGG16 network [8], encouraging perceptual similarity.
   3. Perceptual Loss using ResNet50: Similar to VGG16-based loss but uses a ResNet50 [9] backbone to leverage residual learning and deeper hierarchical features.
2. **Deploy the best-performing model as a user-facing application:** A web-based demo is built using Streamlit, allowing users to upload grayscale photos and receive instant colorized results. The application aims to make AI-based colorization more accessible to non-technical audiences, especially educators, historians, and the general public. The interface is simple and optimized for usability.

## Thesis Contribution

This thesis makes the following key contributions to the field of automatic image colorization, especially within the context of preserving Vietnamese cultural heritage:

1. **Domain-specific adaptation for Vietnamese historical imagery**: While most existing image colorization research focuses on generic datasets (e.g., ImageNet, Places), this study is one of the first to build and experiment on a large-scale dataset (~250,000 images) extracted from modern Vietnamese films with historical settings. This ensures that the model is trained on culturally relevant visual data, enhancing its ability to restore authentic colors for Vietnamese portraits and traditional clothing styles.
2. **Comparative study of perceptual loss with both VGG16 and ResNet50:** Previous works commonly adopt perceptual loss with VGG16 [8] for style transfer and image enhancement. In contrast, this thesis extends the perceptual loss paradigm by incorporating ResNet50 [9] – a deeper architecture with residual connections – to investigate whether deeper semantic features offer better realism or detail preservation in historical contexts. The direct comparison between MSE, VGG16-based, and ResNet50-based perceptual losses over a real-world dataset contributes new insight to the choice of loss functions in image restoration tasks.
3. **Practical deployment through an interactive web application:** Beyond model performance, this work places emphasis on usability. A lightweight colorization tool is deployed using Streamlit, allowing non-experts – including educators, historians, archivists, and general users – to directly engage with AI-powered color restoration. This prototype bridges the gap between academic research and real-world cultural applications, offering a foundation for public-facing historical storytelling platforms.

# Literature Review

## Overview of Image Colorization

Image colorization refers to the process of converting grayscale images into visually plausible color images. In its earliest forms, the task was largely manual: artists or editors would hand-paint black-and-white photographs or use color templates. These manual techniques, although precise, were labor-intensive and not scalable for large archives.

To alleviate the manual burden, rule-based systems were developed. These methods operated by defining sets of hand-crafted rules or heuristics to propagate color from known regions to unknown areas. However, such systems often failed when encountering unfamiliar content or complex textures.

With the rise of machine learning, especially deep learning, data-driven approaches began to dominate the field. By learning mappings directly from data, these methods avoid the rigidity of hand-crafted rules and can generalize across a wide variety of inputs. This shift has made colorization not only faster but also more robust and adaptable to diverse image types, including degraded or historic photographs.

Representative milestones include early exemplar-based learning methods [3] [4], followed by convolutional neural networks (CNNs) trained on large-scale datasets [5], and more recently, approaches based on Generative Adversarial Networks (GANs) [6] and transformer architectures [10] [11]. These new paradigms leverage semantic understanding, global context, and attention mechanisms to improve the realism of colorization.

## Deep Learning-based Approaches

**CNN-based methods:**

Convolutional Neural Networks (CNNs) are a class of deep learning models well-suited for image-related tasks. In image colorization, CNNs learn to predict the chrominance (A and B channels in the LAB color space) based on grayscale luminance (L channel).

[5] proposed a pioneering CNN framework that treats colorization as a classification problem by quantizing the ab color space into bins. The model is trained to output a probability distribution over these bins for each pixel. This approach significantly improved the diversity and plausibility of color outputs.

[12] introduced a more sophisticated model that incorporates both global image features (to infer scene semantics) and local textures (to preserve fine details). This hybrid design enhances the contextual understanding of the model, allowing it to apply more appropriate colors based on object identity and position.

Such CNN-based methods are typically fast, well-understood, and can be trained with standard loss functions like cross-entropy or mean squared error.

**GAN-based methods:**

Generative Adversarial Networks (GANs), introduced in [13], have revolutionized image generation tasks. In the context of colorization, GANs consist of a generator that predicts color and a discriminator that distinguishes between real and fake color images.

DeOldify [6] is one of the most well-known GAN-based colorization projects. It leverages a ResNet-based generator and perceptual loss to achieve impressive visual results. GANs allow for more vivid and realistic colors, often generating outputs that are closer to human artistic interpretations. However, they come with challenges such as training instability, mode collapse, and artifacts, particularly when applied to noisy or low-quality historical images.

**Transformer-based Methods:**

More recently, transformer architectures – originally developed for natural language processing – have been adapted to computer vision tasks. Transformers use self-attention mechanisms to model long-range dependencies, allowing for global contextual reasoning across an image.

In colorization, this has led to models like the Colorization Transformer [11], which outperform CNNs in capturing semantic coherence and maintaining consistent color across similar regions. Although still in early stages, these methods show potential for more human-like, context-aware colorization, especially when combined with large-scale pretraining.

## Loss Functions for Colorization

Loss functions play a critical role in guiding the learning process of deep models for colorization. Two primary types are commonly used:

1. **Pixel-wise Loss**: The **Mean Squared Error (MSE)** measures the average squared difference between predicted and ground truth pixel values. While it ensures numerical accuracy, MSE tends to produce blurry results when the model is uncertain about color choices.
2. **Perceptual Loss**: Introduced in [13], this loss compares the similarity of high-level feature representations between the predicted and ground truth images using a pretrained CNN. Perceptual loss aims to produce colorizations that are more perceptually convincing and structurally coherent.

More details about the implementation of these losses will be discussed in **Chapter 3: Methodology**.

# Methodology

## Dataset Preparation

The dataset plays a central role in enabling the deep learning model to learn the mapping between grayscale input and plausible color output. However, acquiring colorized photographs of Vietnam before 1950 is nearly impossible due to technological and archival limitations. Most available images from that era are in black and white, and their quality is often poor due to age, scanning artifacts, and lack of annotations. To address this challenge, a novel yet practical strategy is adopted: building a large dataset by extracting frames from modern Vietnamese films that are set in historical contexts. These films, while produced in recent decades, are carefully crafted to replicate the appearance and atmosphere of earlier periods through costumes, makeup, sets, and lighting. Consequently, they offer a valuable proxy for historical colorization training, as they simulate authentic visual styles from past centuries.

**Figure 3.1** illustrates the full process, starting from raw video files to a clean, structured dataset ready for training. The figure shows the stages of frame extraction, image filtering (brightness and person detection), LAB conversion, normalization, and dataset splitting. This comprehensive and scalable pipeline lays the groundwork for developing a data-driven image colorization system that is both efficient and historically meaningful.

From these films, a **frame extraction process** was applied using OpenCV. Frames were extracted at regular intervals or scene transitions, depending on the source material. This step generated **over 400,000 initial RGB frames**. However, this raw pool contained many unsuitable frames, such as those that were overly dark, scenes without people, or frames with extreme blur or distortion. To ensure data quality, a multi-stage cleaning process was implemented.

First, **dark images were automatically removed**. This was done by converting each frame to grayscale and calculating its mean brightness. Frames below a brightness threshold were discarded, as they often corresponded to transitions, black screens, or nighttime scenes where little structural information was visible. This helped reduce noise in the training data and improved consistency in luminance-based learning.

Next, frames were passed through a **person detection filter** using the YOLOv3 object detection model [1]. Since the goal of the thesis focuses on portrait-centric colorization, scenes without any humans or those featuring large crowds were excluded. Frames with between **1 to 5 detected persons** were retained, striking a balance between having enough foreground detail and avoiding cluttered compositions. This filter also helped maintain semantic consistency in the dataset, ensuring that the network learns to colorize facial features, clothing, and body regions more accurately.

After these filtering steps, a final dataset of **257,825 clean RGB images** was obtained. These images are thematically relevant and visually consistent with Vietnamese history before 1950. The dataset covers diverse scenes from various time periods including **the Nguyễn Dynasty, French colonial rule, and post-revolutionary Vietnam**, as depicted in the modern film sources. While not truly historical, the artistic realism and attention to detail in these films make them valuable substitutes for authentic imagery, especially when large-scale real-world data is lacking.

Once curated, the RGB images were **preprocessed and converted into the LAB color space**, which separates lightness (L channel) from color information (a and b channels). This separation is critical for colorization tasks, as the model learns to infer only the chrominance (a, b) given the luminance (L) input. Each image was **resized to a resolution of 224×224** pixels to match the input shape required by the U-Net model. Following the resizing, the RGB images were converted to LAB.

The **normalization step** was then applied to each channel, ensures that the model receives inputs and outputs within consistent, bounded numerical ranges, which improves training stability and convergence. The **L channel**, which ranges from 0 to 100 in LAB space, was scaled to [0,1][0, 1][0,1] by dividing by 100:

For the **a and b channels**, which typically range from approximately , values ere normalized to the range using the following transformations:

The final step was **splitting the dataset** into three subsets: training, validation, and test. A random split was applied, where **90% (232,042 images)** were allocated for training and validation, and the remaining **10% (25,783 images)** were reserved for testing. Within the 90%, a secondary split assigned approximately 10% to validation, resulting in a training set of around **208,800 images** and a validation set of **23,200 images**. This division supports model generalization and prevents overfitting, while the hold-out test set is used for final performance evaluation.

## Proposed Model Architecture

<Write here>

chỗ này e nên vẽ một sơ đồ pipeline về cấu trúc mô hình đưa ra xong sau đó miêu tả từng phần trong cấu trúc đưa ra, như thế người đọc sẽ có cái nhìn tổng thể về nội dung e làm

**U-Net architecture:**

1. Encoder với Conv2D và MaxPooling.
2. Decoder với Conv2D Transpose.
3. Đầu ra là 2 kênh màu (AB) trong không gian Lab

## Loss Function Selection

<Write here>

**MSE Loss**

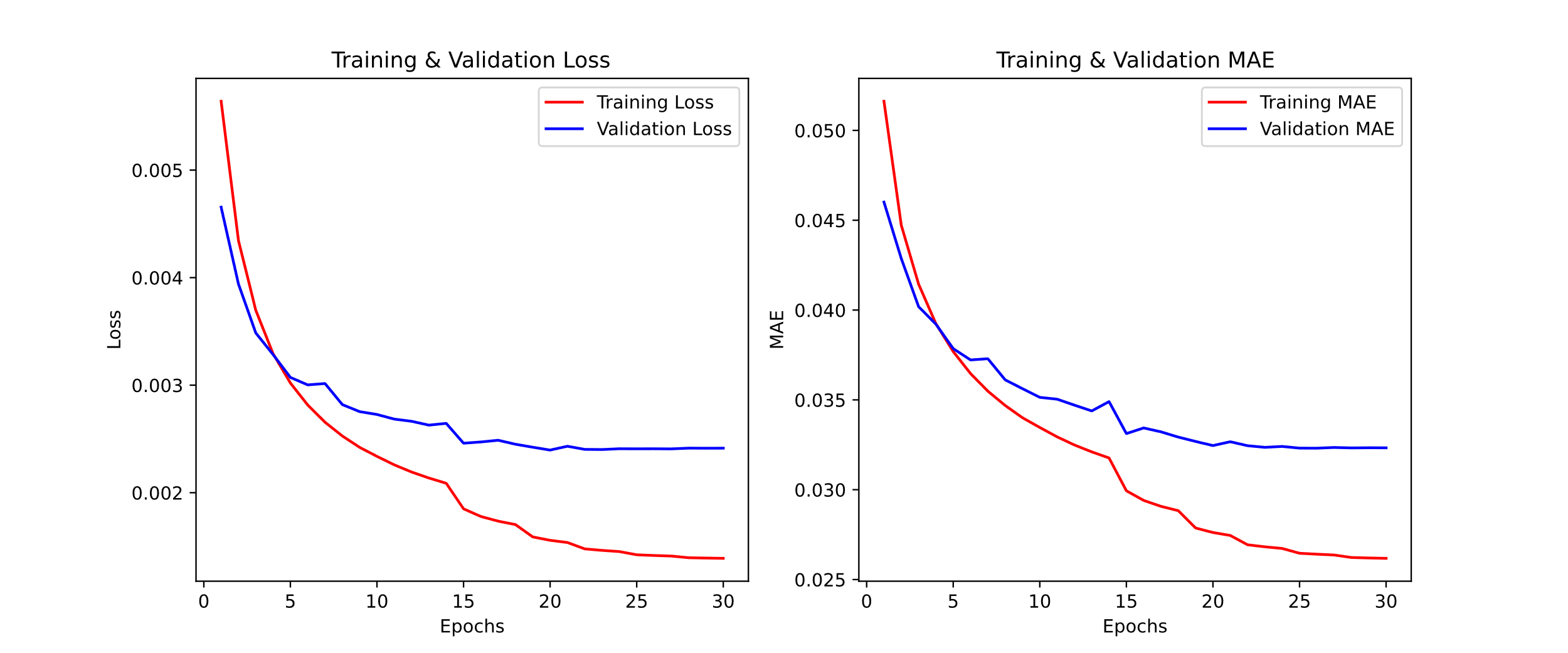
**Perceptual Loss** with VGG16

**Perceptual Loss** with ResNet50

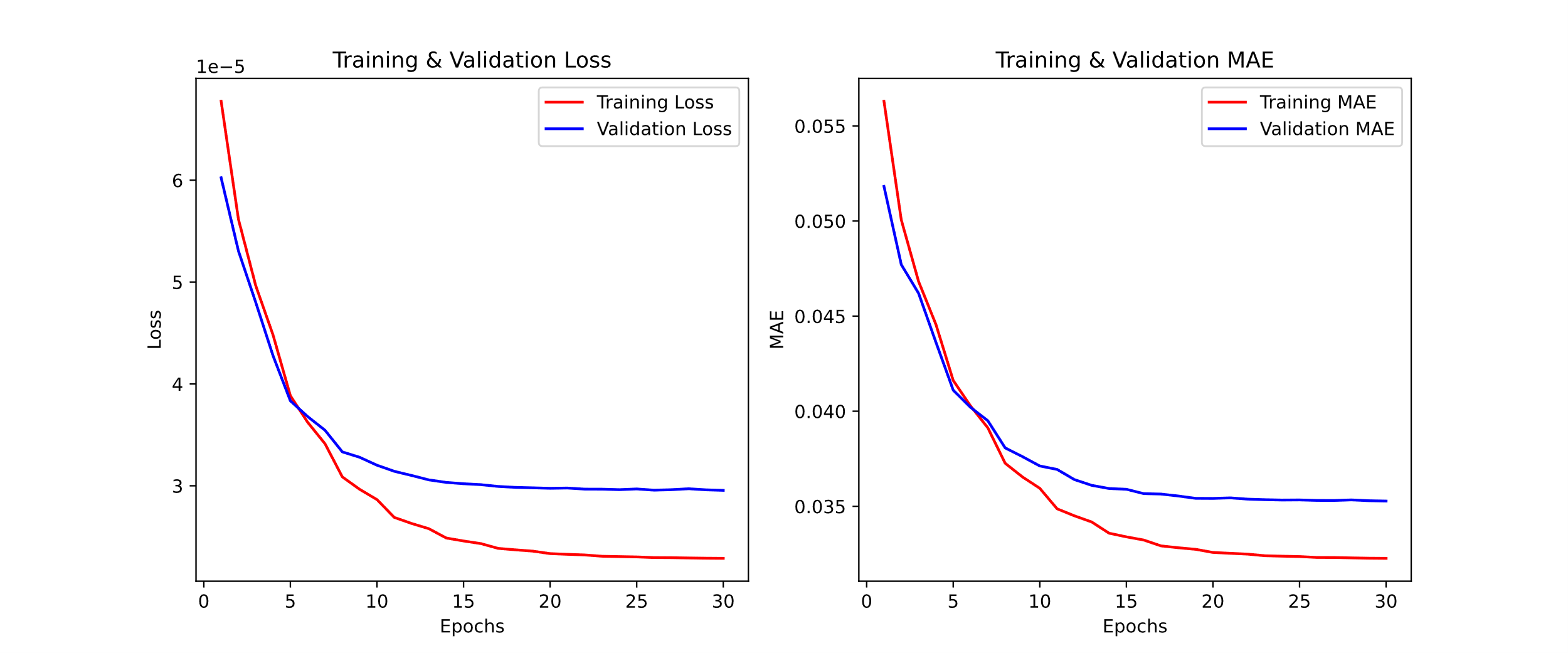
# Experiments and Results

## Training Process

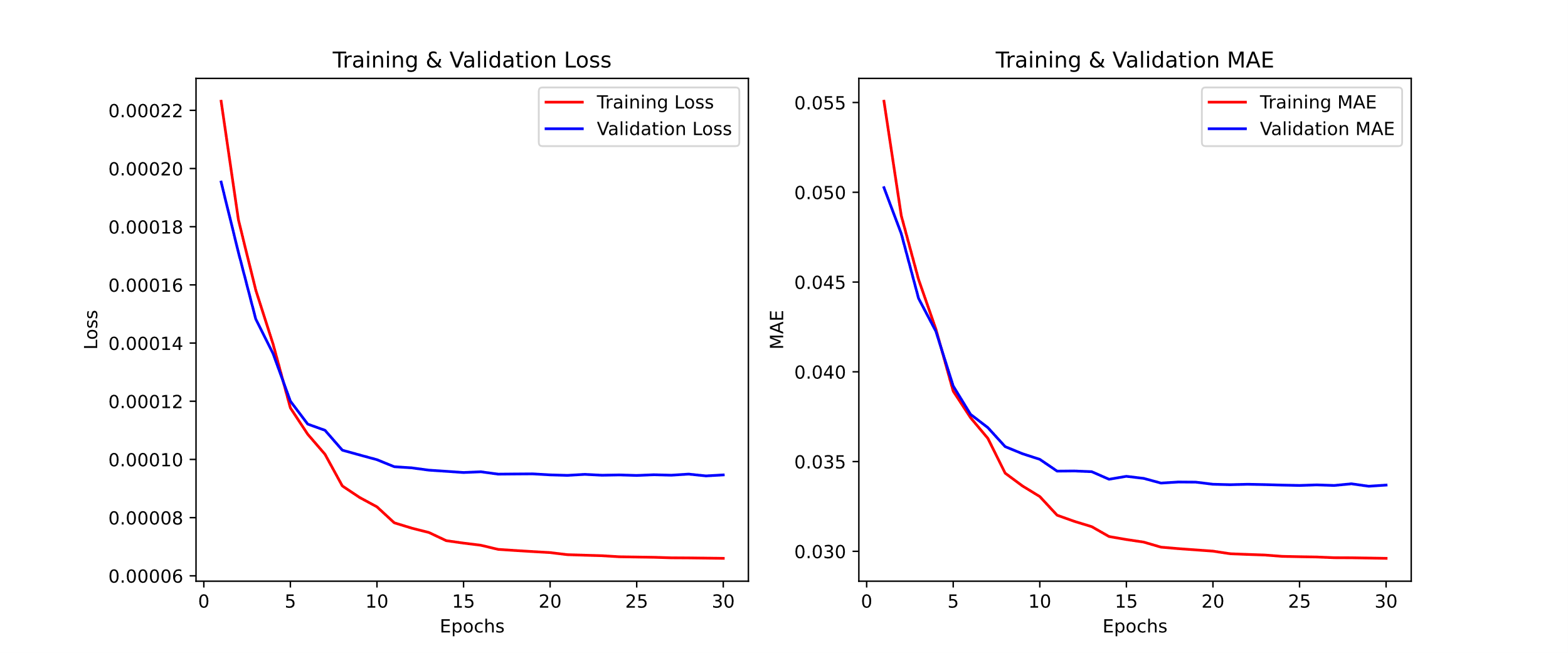
<Write here>



MSE



Perceptual VGG



Perceptual ResNet

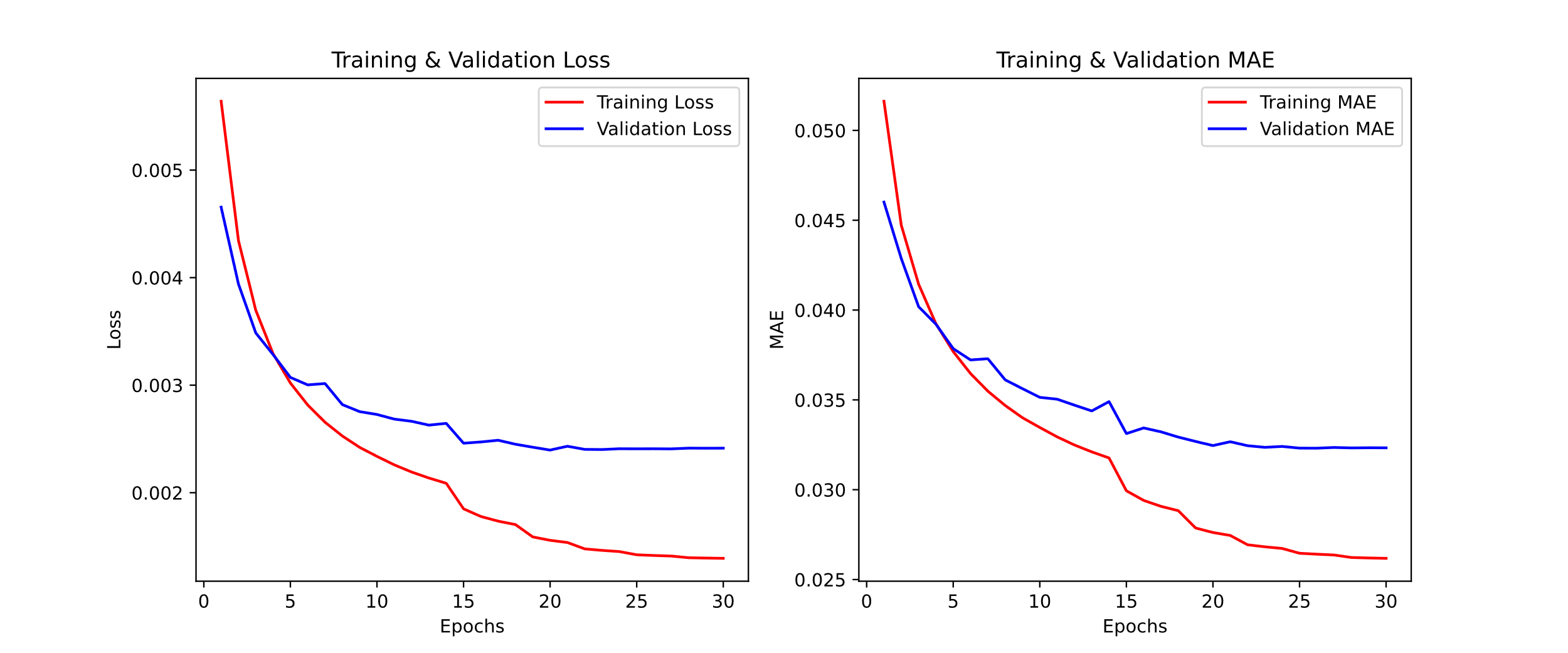
Huấn luyện mô hình trong **30 epochs,** batch size = 16 trên tập dataset 230k ảnh.

Theo dõi quá trình giảm loss trong từng epoch.

## Evaluation Metrics

<Write here>

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | PSNR | SSIM |
| MSE | 5.5189 | 30.3006 | 0.9555 |
| Perceptual VGG | 6.0318 | 29.5798 | 0.9533 |
| Perceptual ResNet | 5.7462 | 29.9902 | 0.9513 |

Kết quả của test 23k ảnh tập test

**Mean Absolute Error (MAE)**: Đánh giá sự khác biệt pixel giữa ảnh gốc và ảnh tô màu.

**PSNR (Peak Signal-to-Noise Ratio), SSMI (Structural Similarity Index Measure)**: Đánh giá chất lượng ảnh đầu ra so với ảnh gốc.

## Qualitative Results

<Write here>



ảnh mẫu tạm thời

So sánh kết quả trực quan giữa các loss function:

1. **MSE Loss:** Màu sắc có thể bị nhòe.
2. **Perceptual Loss with VGG:** Cải thiện chi tiết nhưng có thể thiếu màu chính xác.
3. **Perceptual Loss with ResNet:**

Trình bày kết quả bằng hình ảnh từ tập test.

# Discussion

## Impact of Different Loss Functions

<Write here>

So sánh MSE vs. Perceptual Loss with VGG vs. Perceptual Loss with ResNet.

Nhận xét về ảnh hưởng của từng loại loss lên chất lượng ảnh.

1. **MSE** hoạt động tốt nhưng dễ gây nhòe màu.
2. **Perceptual Loss with VGG** giúp màu sắc tự nhiên hơn nhưng có thể bị nhiễu.
3. **Perceptual Loss with ResNet**

## Limitations

<Write here>

Chất lượng màu sắc chưa hoàn hảo: Màu có thể chưa khớp 100% với thực tế.

Dataset chưa đủ đa dạng: Hầu hết dữ liệu huấn luyện đến từ phim hiện đại bối cảnh lịch sử.

Tốc độ xử lý còn chậm: Khi chạy trên Streamlit, tốc độ xử lý cần được tối ưu hơn.

# Model Deployment Using Streamlit

<Write here>

**Mục tiêu:** Xây dựng giao diện đơn giản cho phép người dùng upload ảnh đen trắng và nhận ảnh màu.

**Pipeline xử lý:**

1. Upload ảnh.
2. Tiền xử lý (chuyển đổi sang Lab, resize).
3. Dự đoán bằng mô hình.
4. Chuyển đổi lại RGB và hiển thị ảnh.

**Công cụ triển khai:** Streamlit Cloud.

# Conclusion and Future Work

## Summary of Contributions

<Write here>

Xây dựng mô hình colorization ảnh lịch sử Việt Nam bằng U-Net.

So sánh hiệu quả của các loss function khác nhau.

Phát triển ứng dụng web sử dụng Streamlit để thử nghiệm mô hình.

## Future Work

<Write here>

Cải thiện kiến trúc model: Chuyển đổi sang mô hình GAN và bổ sung mô hình nhận dạng ngữ nghĩa

Bổ sung tập huấn luyện ảnh lịch sử thực tế để cải thiện độ chính xác.

Tối ưu hóa tốc độ ứng dụng bằng cách triển khai web bằng FastAPI + Streamlit thay vì chỉ dùng Streamlit.

# References

|  |  |
| --- | --- |
| [1] | Tuổi Trẻ Cuối Tuần, “Tô màu cho quá khứ: Mò kim giữa dòng chảy lịch sử,” 02 06 2020. [Online]. Available: https://cuoituan.tuoitre.vn/to-mau-cho-qua-khu-mo-kim-giua-dong-chay-lich-su-1558595.htm. |
| [2] | Dân Trí, "Bạn trẻ 9X làm sống lại lịch sử nhờ đam mê phục chế màu tư liệu cũ," 10 09 2021. [Online]. Available: https://dansinh.dantri.com.vn/dien-dan-dan-sinh/ban-tre-9x-lam-song-lai-lich-su-nho-dam-me-phuc-che-mau-tu-lieu-cu-20210910095736038.htm. |
| [3] | T. Welsh, M. Ashikhmin and K. Mueller, "Transferring color to greyscale images," in *SIGGRAPH*, San Antonio, Texas, 2002. |
| [4] | A. Levin, D. Lischinski and Y. Weiss, "Colorization using Optimization," in *SIGGRAPH*, Los Angeles, CA, 2004. |
| [5] | R. Zhang, P. Isola and A. A. Efros, "Colorful image colorization," in *ECCV*, Amsterdam, 2016. |
| [6] | J. Antic, "DeOldify: A deep learning based project for colorizing and restoring old images," 2019. [Online]. Available: https://github.com/jantic/DeOldify. |
| [7] | O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, Munich, 2015. |
| [8] | K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, San Diego, CA, 2015. |
| [9] | K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, Las Vegas, Nevada, 2016. |
| [10] | Z. Cheng, Q. Yang and B. Sheng, "Deep Colorization," in *ICCV*, Santiago, 2015. |
| [11] | M. Kumar, D. Weissenborn and N. Kalchbrenner, "Colorization Transformer," in *ICLR*, 2021. |
| [12] | I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," 2014. |
| [13] | J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *ECCV*, Amsterdam, 2016. |
| [14] | S. Iizuka, E. Simo-Serra and H. Ishikawa, "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," in *SIGGRAPH*, Anaheim, CA, 2016. |