**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

**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

# 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

Black-and-white photographs represent an invaluable part of historical archives, yet they inherently lack the rich color information that conveys full visual and emotional context. Image colorization seeks to bridge this gap by predicting plausible colors, thereby enhancing the cultural and historical significance of these images. This thesis focuses on developing a deep learning approach to colorize black-and-white photographs related to Vietnamese history before 1950.

A U-Net-based model was proposed to infer the chrominance channels from the luminance input in the CIELAB color space. Three loss functions were explored: Mean Squared Error (MSE), Perceptual Loss with VGG16, and Perceptual Loss with ResNet50. Extensive experiments on a dataset of over 250,000 images demonstrated that while MSE achieved the lowest pixel-wise errors, perceptual losses produced more realistic and visually pleasing colorizations. Specifically, the perceptual loss using VGG16 resulted in smoother textures and better semantic coherence, whereas the version with ResNet50 captured finer structural details. When evaluated on a separate curated test set of 23,000 images and two real historical photographs, the perceptual-based models consistently outperformed MSE in qualitative assessments, generating colors that aligned more naturally with human perception. And then, a web application , that allows  users to upload and colorize grayscale images directly, was also deployed using Streamlit Cloud.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my academic advisor, Dr. Tran Van Ha, for the invaluable guidance, encouragement, and insightful feedback throughout the course of this research. His expertise and mentorship have been instrumental in shaping the direction and rigor of this thesis.

I also extend my sincere thanks to my colleagues, friends, and family, whose emotional support and thoughtful suggestions provided me with the strength and motivation to persist through challenges. Their presence has been a constant source of encouragement during this journey.

I am grateful for the availability of open-source tools and communities – particularly TensorFlow, Keras, Streamlit, and other related libraries – which enabled the development, experimentation, and deployment of this project in a practical and accessible way. These resources have played a crucial role in transforming this thesis from a theoretical concept into a functioning application usable by end-users.

To all who have accompanied me on this journey, thank you.

# Table of Contents

# List of Tables

# List of Figures

# 1. Introduction

## 1.1. 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.

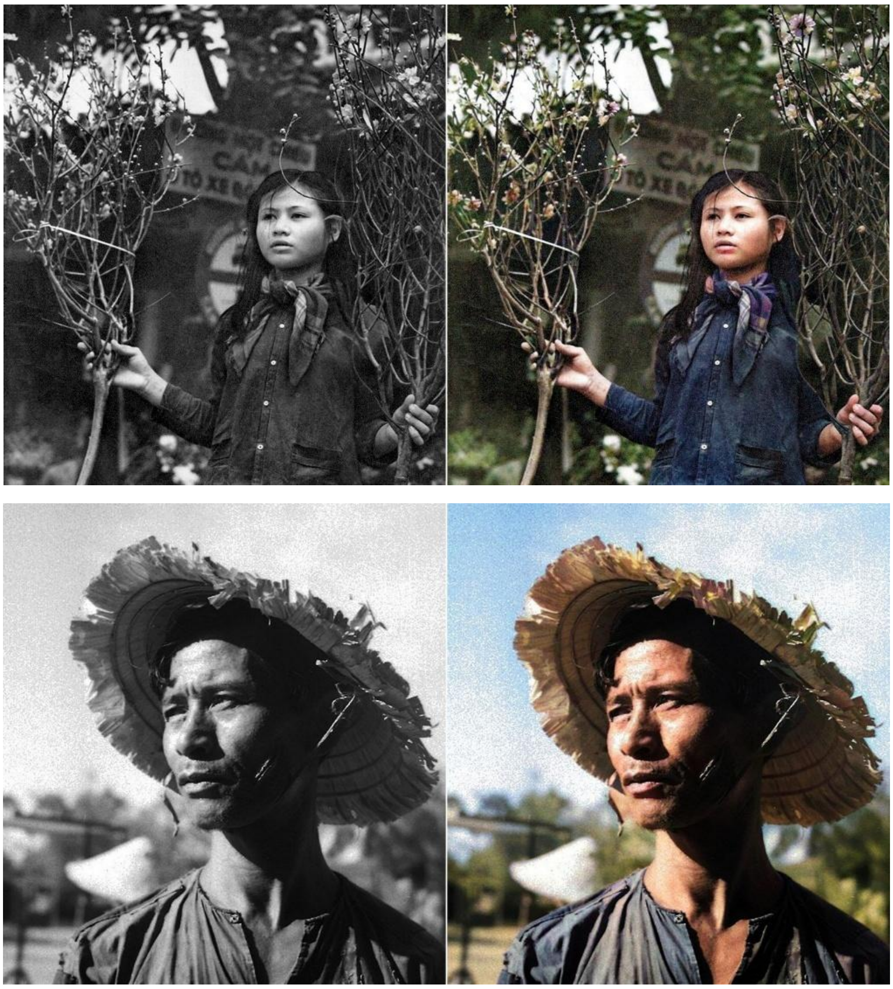

**Figure 1.** Example colorization of a grayscale Vietnamese historical portrait.
Source: DeOldify model output [1]

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 [2] [3]. Figure 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 in [4] and [5], 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, Zhang et al. proposed using class-rebalanced CNNs trained on large datasets to produce semantically meaningful colors [6]. 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.

## 1.2. Research Objectives

This thesis aims to develop a deep learning-based system capable of colorizing historical black-and-white photographs from Vietnam before 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:

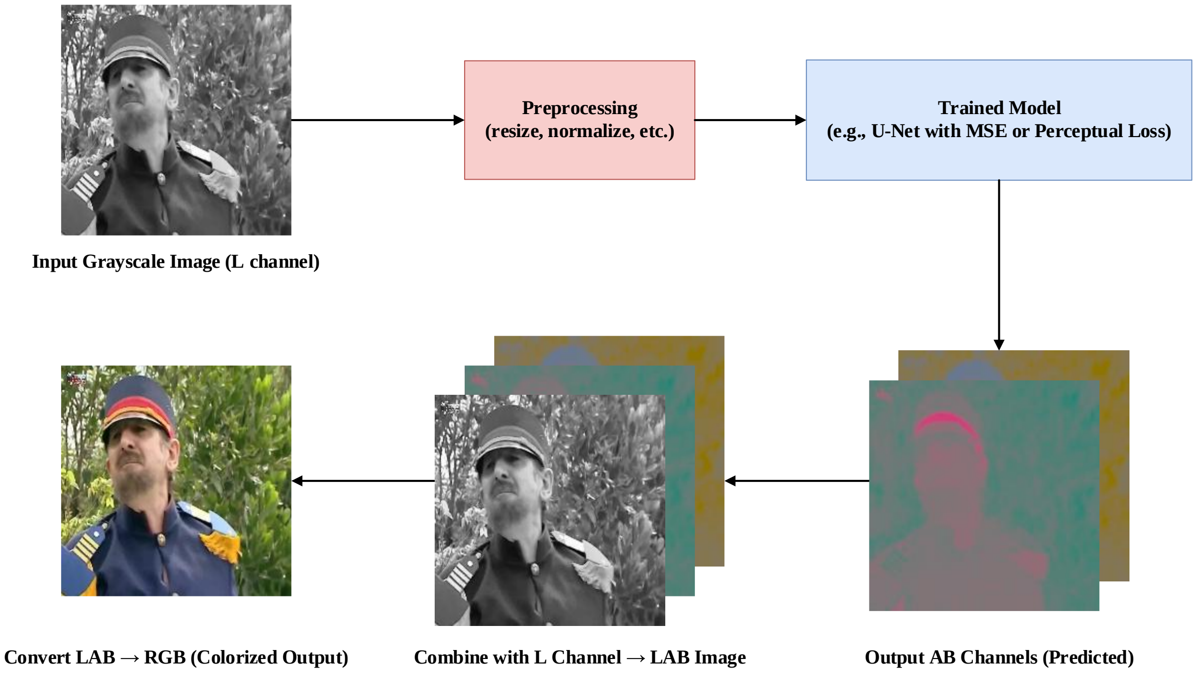
Preprocessing
(resize, normalize, etc.)
Trained Model
(e.g., U-Net with MSE or Perceptual Loss)
Input Grayscale Image (L channel)
Convert LAB → RGB (Colorized Output)
Combine with L Channel → LAB Image
Output AB Channels (Predicted)

**Figure 2.** The pipeline of grayscale input and its colorized output

1. **Develop a robust image colorization model using deep learning techniques:** A U-Net-based architecture is employed as the core model due to its proven effectiveness in image-to-image translation tasks [7]. 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 2 shows the pipeline of grayscale input and its colorized output using one of the trained models.

2. **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:

   a. Mean Squared Error (MSE): A traditional pixel-wise loss that minimizes average squared differences.

   b. Perceptual Loss using VGG16: Computes loss in feature space based on intermediate activations of a pre-trained VGG16 network [8], encouraging perceptual similarity.

   c. Perceptual Loss using ResNet50: Similar to VGG16-based loss but uses a ResNet50 [9] backbone to leverage residual learning and deeper hierarchical features.

3. **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.

## 1.3. 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.

# 2. Literature Review

## 2.1. 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, which applied predefined heuristics to guide color propagation or transfer from reference images. For instance, Welsh et al. proposed a method that transfers color from a reference color image to a grayscale image based on pixel similarity in a luminance-weighted neighborhood [4]. Similarly, Levin et al. framed colorization as an optimization problem that encourages neighboring pixels with similar intensities to have similar chrominance values [5]. However, such systems often failed when encountering unfamiliar content or complex textures due to their reliance on low-level features and lack of semantic understanding.

With the rise of machine learning, especially deep learning, data-driven approaches began to dominate the field [6] [10] [11]. 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 [5] [12], CNNs trained on large-scale datasets [6] [11], and more recently, approaches based on Generative Adversarial Networks (GANs) [1] [13] [14] and transformer architectures [15] [16]. These new paradigms leverage semantic understanding, global context, and attention mechanisms to improve the realism of colorization.

Other innovative efforts have incorporated semantic segmentation [17], user-guided scribble input [18], and reference-based image retrieval [19] to enhance control and plausibility. Additionally, some research has focused on applying these methods to historic or domain-specific datasets such as anime [20], medical imaging [21], or cultural heritage restoration [3] [22].

## 2.2. Deep Learning-based Approaches

**CNN-based methods:** CNNs are a class of deep learning models well-suited for image-related tasks. These methods are typically fast, well-understood, and can be trained with standard loss functions like cross-entropy or mean squared error. In image colorization, CNNs learn to predict the chrominance (A and B channels in the LAB color space) based on grayscale luminance (L channel).

Zhang et al. proposed a pioneering CNN framework that treats colorization as a classification problem by quantizing the ab color space into bins [6]. 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.

Iizuka et al. introduced a more sophisticated model that incorporates both global image features (to infer scene semantics) and local textures (to preserve fine details) [10]. This hybrid design enhances the contextual understanding of the model, allowing it to apply more appropriate colors based on object identity and position.

Cheng et al. designed a fully convolutional network that produces smoother and more globally consistent colorizations [12]. U-Net architecture, initially developed for biomedical segmentation, has also been adapted for colorization tasks due to its encoder-decoder structure with skip connections, enabling both low-level detail retention and high-level abstraction [7].

**GAN-based methods:** GANs, introduced by Goodfellow et al., have revolutionized image generation tasks [13]. In the context of colorization, it consist of a generator that predicts color and a discriminator that distinguishes between real and fake color images.

DeOldify is a well-known GAN-based model built on top of a ResNet generator and perceptual loss, demonstrating exceptional results in restoring and colorizing old photographs [1]. Other works like Pix2Pix [14] and its variant Pix2PixHD [23] offer conditional GAN-based solutions for image-to-image translation tasks including colorization.

However, GANs face challenges such as training instability, mode collapse, and hallucinated artifacts—especially problematic when dealing with low-resolution or degraded historical images. Recent efforts focus on stabilizing GAN training with techniques like spectral normalization, progressive growing, and Wasserstein loss [24].

**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 [15], 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 and hybrid approaches [16].

## 2.3. 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 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 by Johnson et al., this loss compares the similarity of high-level feature representations between the predicted and ground truth images using a pretrained CNN [25]. Perceptual loss aims to produce colorizations that are more perceptually convincing and structurally coherent.

3. **Adversarial Loss:** Derived from GANs [13], this loss encourages the generator to produce outputs that fool the discriminator. It helps generate vivid and photorealistic colorizations, particularly when combined with perceptual loss.

4. **Style and Content Losses:** Some methods borrow from neural style transfer to encourage preservation of style patterns or global color palettes [26].

More details about the implementation of these losses will be discussed in the next chapter.

## 2.4. Evaluation Metrics for Colorization

Evaluating the quality of colorized images is a non-trivial task, as there can be multiple plausible colorizations for a single grayscale input. Therefore, objective evaluation metrics are essential to fairly compare different colorization methods. This section presents three widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

1. PSNR is a traditional metric that measures the pixel-wise fidelity between the generated and ground truth color image [27]. It is calculated based on the Mean Squared Error (MSE), with higher PSNR values indicating better similarity. However, PSNR often fails to correlate with human perception, especially in tasks like colorization where minor pixel differences may not affect visual quality significantly.

2. SSIM improves upon PSNR by considering luminance, contrast, and structural information between images, also better aligns with human visual perception by evaluating structural similarity rather than raw pixel differences [28]. As such, it has become a popular choice for evaluating image restoration tasks, including colorization.

3. LPIPS is a more recent perceptual metric that uses deep features extracted from pretrained neural networks to compare images [29]. This has been shown to correlate strongly with human judgments of perceptual similarity. It is especially useful in generative tasks where visual plausibility matters more than exact replication of ground truth.

# 3. Methodology

## 3.1. 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 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.

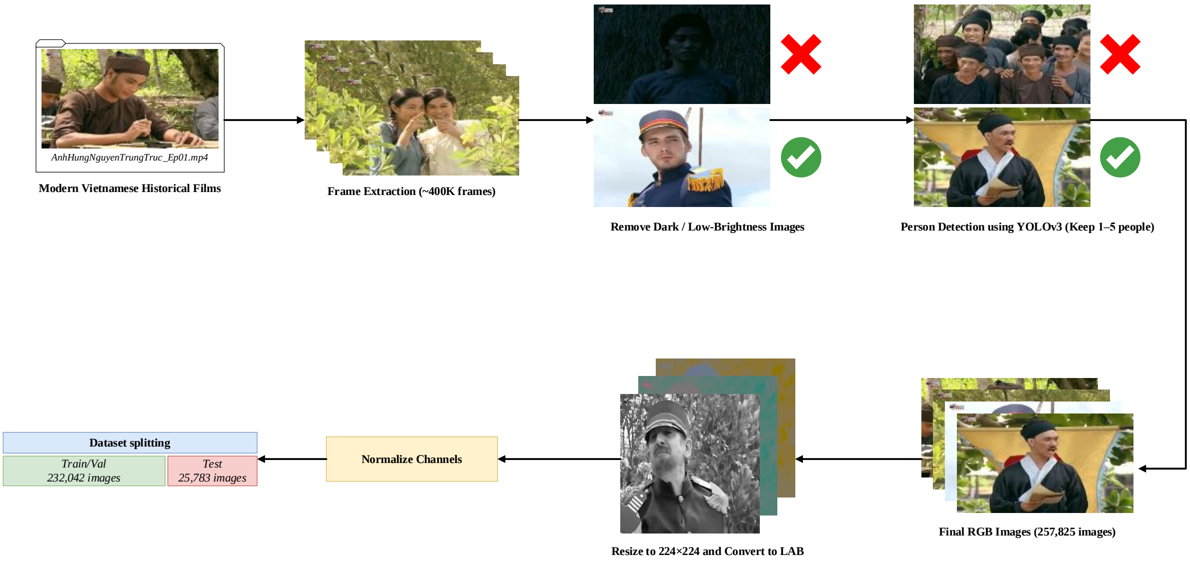
AnhHungNguyenTrungTruc_Ep01.mp4
Modern Vietnamese Historical Films
Frame Extraction (~400K frames)
Remove Dark / Low-Brightness Images
Person Detection using YOLOv3 (Keep 1–5 people)
Final RGB Images (257,825 images)
Resize to 224×224 and Convert to LAB
Normalize Channels
Dataset splitting
Train/Val
232,042 images
Test
25,783 images

**Figure 3.** The pipeline of dataset construction.

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 [30]. 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, ensuring 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 $[0,1]$ by dividing by 100:

$$L_{\text{norm}} = \frac{L}{100} \tag{1}$$

For the A and B (viết hoa vì bên dưới e cũng viết hoa) channels, which typically range from approximately $[-128,127]$, values were normalized to the range $[-1,1]$ using the following transformations:

$$a_{\text{norm}} = \frac{a-128}{128}, \qquad b_{\text{norm}} = \frac{b-128}{128} \tag{2}$$

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.

## 3.2. Proposed Model Architecture

This thesis utilizes a U-Net-based deep neural network for the task of image colorization, focusing on black-and-white images from Vietnamese historical contexts prior to 1950. U-net architecture is particularly well-suited for this task due to its encoder-decoder structure and skip connections, which enable effective feature reuse across different levels of abstraction [7]. The objective of the model is to infer plausible color information (A and B channels) from a single-channel grayscale input (L channel), which carries all luminance-related information.

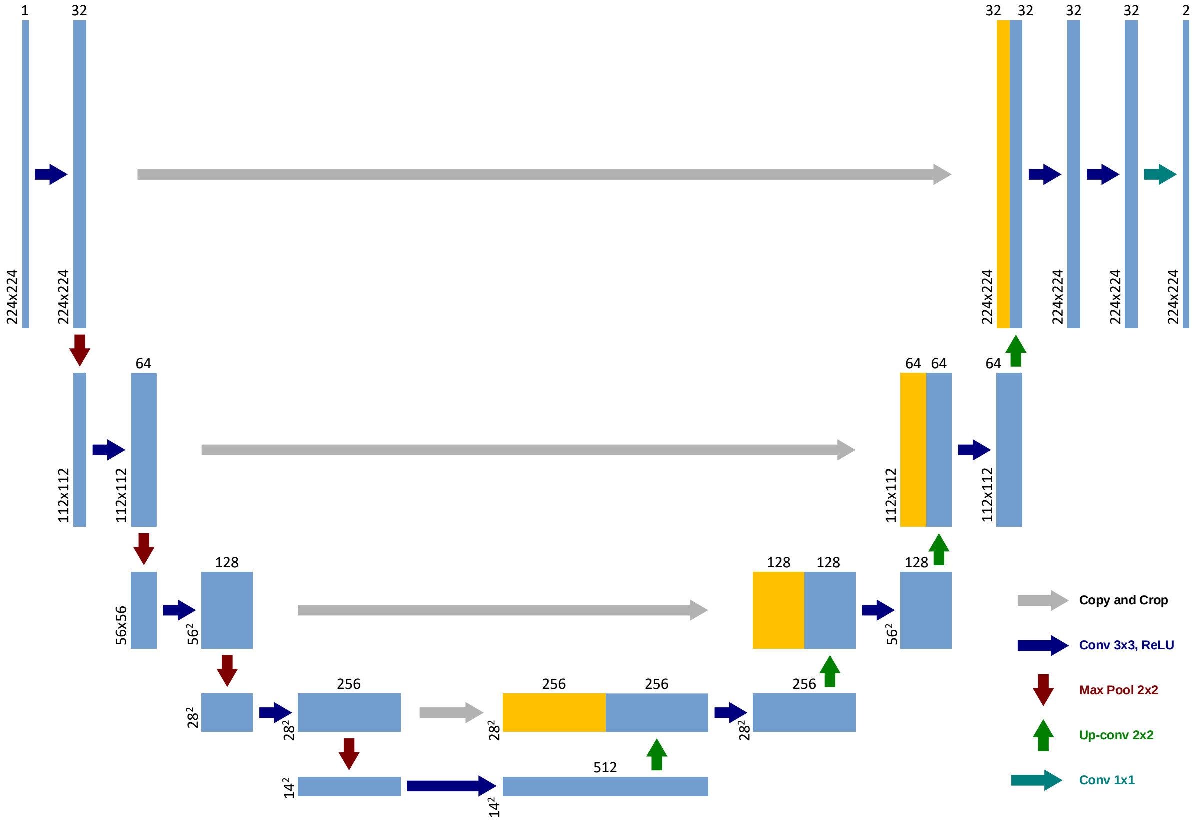

**Figure 4.** U-Net-based architecture for grayscale-to-color image colorization

Figure 4 shows the pipeline of the colorization process. The model receives an input image in the L channel with a shape of (224, 224, 1). This L channel is first passed through a series of convolutional and pooling layers in the encoder, followed by a symmetric decoder with transposed convolutions. The network then predicts the missing A and B chrominance channels. These outputs are combined with the original L input to reconstruct a full LAB image, which is then converted into RGB for visualization or storage.

The encoder portion of the network consists of several blocks of two `Conv2D` layers (with ReLU activation) followed by `MaxPooling2D`. With each downsampling step, the number of filters doubles, ranging from 32 to 512. These layers progressively capture increasingly abstract features while reducing spatial resolution. This hierarchical encoding allows the model to learn both low-level texture features and high-level semantic context.

At the center of the U-Net is a bottleneck layer, which serves as a compact latent representation of the entire image. This layer encodes the most important semantic information necessary for accurate colorization. The decoder then upsamples this latent space back to the original resolution using `Conv2DTranspose` layers. Each decoder block is complemented with skip connections that concatenate the corresponding encoder block's feature maps, ensuring that spatial details lost during downsampling are restored during reconstruction.

These skip connections are vital to the performance of the U-Net, especially for tasks like colorization, where edge detail and object boundaries must be accurately preserved. Without such connections, the model would rely solely on high-level latent features, often resulting in blurry or less realistic outputs. With the skip connections, the decoder can leverage both low-level and high-level features, thus producing sharper and more consistent colorized images.

The final output layer of the network is a `Conv2D` layer with two filters (corresponding to the A and B channels) and a `tanh` activation function. This ensures that the output values fall in the normalized range of [-1, 1], which can later be rescaled back to the original LAB value range for reconstruction. The output shape of the model is (224, 224, 2), consistent with the expected AB channels.

Overall, the model contains approximately 4.7 million parameters, making it lightweight yet expressive enough to handle a dataset of over 230,000 images. The architecture is summarized in Table 1, which presents the layers, output dimensions, and parameter counts in a structured form. This helps in understanding the computational complexity and depth of the network.

The training process of the model involves experimenting with three different loss functions, each designed to optimize the colorization quality from different perspectives. These include Mean Squared Error (MSE), Perceptual Loss using VGG16, and Perceptual Loss using ResNet50. The implementation and rationale behind these losses will be discussed in detail in the following section (Section 3.3).

**Table 1.** U-Net Architecture Summary

| Layer Name | Type | Output Shape | Params | Notes |
|---|---|---|---|---|
| input_1 | InputLayer | (224, 224, 1) | 0 | Grayscale L channel |
| conv2d | Conv2D | (224, 224, 32) | 320 | First encoder block |
| max_pooling2d | MaxPooling2D | (112, 112, 32) | 0 | Downsampling |
| conv2d_1 | Conv2D | (112, 112, 64) | 18,496 | |
| max_pooling2d_1 | MaxPooling2D | (56, 56, 64) | 0 | |
| conv2d_2 | Conv2D | (56, 56, 128) | 73,856 | |
| max_pooling2d_2 | MaxPooling2D | (28, 28, 128) | 0 | |
| conv2d_3 | Conv2D | (28, 28, 256) | 295,168 | |
| max_pooling2d_3 | MaxPooling2D | (14, 14, 256) | 0 | Bottleneck starts here |
| conv2d_4 | Conv2D | (14, 14, 512) | 1,180,160 | Bottleneck |
| conv2d_transpose | Conv2DTranspose | (28, 28, 256) | 1,179,904 | Upsampling |
| concatenate | Concatenate | (28, 28, 512) | 0 | Skip connection from conv2d_3 |
| conv2d_5 | Conv2D | (28, 28, 256) | 1,179,904 | |
| conv2d_transpose_1 | Conv2DTranspose | (56, 56, 128) | 295,040 | |
| concatenate_1 | Concatenate | (56, 56, 256) | 0 | Skip connection from conv2d_2 |
| conv2d_6 | Conv2D | (56, 56, 128) | 295,040 | |
| conv2d_transpose_2 | Conv2DTranspose | (112, 112, 64) | 73,792 | |
| concatenate_2 | Concatenate | (112, 112, 128) | 0 | Skip connection from conv2d_1 |
| conv2d_7 | Conv2D | (112, 112, 64) | 73,792 | |
| conv2d_transpose_3 | Conv2DTranspose | (224, 224, 32) | 18,464 | |
| concatenate_3 | Concatenate | (224, 224, 64) | 0 | Skip connection from conv2d |
| conv2d_8 | Conv2D | (224, 224, 32) | 18,464 | Final decoding block |
| conv2d_9 | Conv2D | (224, 224, 32) | 9,248 | |
| conv2d_10 | Conv2D | (224, 224, 2) | 66 | Output: AB channels |
| **Total Parameters** | **-** | **-** | **4,711,714** | **All layers are trainable** |

## 3.3. Loss Function Selection

In the task of automatic image colorization, the choice of loss function directly affects the quality of the generated images. A good loss function should not only minimize pixel-level differences but also help preserve semantic coherence and perceptual realism, especially when colorizing historical images where ground truth may have limited consistency. This study evaluates three loss functions: Mean Squared Error (MSE), Perceptual Loss using VGG16, and Perceptual Loss using ResNet50.

The MSE loss, one of the most commonly used loss functions for regression tasks, measures the average squared differences between the predicted values and the ground truth. In image colorization, MSE computes the pixel-wise error between the predicted AB chrominance channels and the ground truth AB channels in the CIELAB color space. The formulation is:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{\text{true}} - y_i^{\text{pred}})^2 \tag{3}$$

where $y^{true}$ and $y^{pred}$ represent the ground-truth and predicted AB values respectively, and $N$ is the number of pixels. While MSE is simple and stable during training, it often leads to blurry outputs when multiple plausible colorizations exist for a single grayscale input [6]. This limitation is particularly evident in ambiguous regions such as sky, clothing, or background textures.

To overcome this, Perceptual Loss has been introduced in recent literature [25], focusing on high-level feature similarity instead of raw pixel values. The perceptual loss is calculated by comparing feature activations of the predicted and ground truth color images extracted from a pre-trained convolutional neural network. Its general form is:

$$\mathcal{L}_{\text{perceptual}} = \sum_{\text{l}=1}^{\text{L}} \frac{1}{N_l} \sum_{i=1}^{N_l} \left( F_l^{\text{true}}(i) - F_l^{\text{pred}}(i) \right)^2 \tag{4}$$

Here, $F_l^{\text{true}}$ and $F_l^{\text{pred}}$ denote the feature maps of the ground truth and predicted LAB images at layer $l$, and $N_l$ is the number of elements in the feature map. This loss encourages the model to generate outputs that are perceptually closer to the reference, even if pixel-wise deviations exist.

In this thesis, two popular backbone networks are used for computing perceptual loss: VGG16 and ResNet50. For VGG16, the selected layers are `block1_conv2`, `block2_conv2`, and `block3_conv3`. These layers are chosen to capture both low-level textures and mid-level semantics, which are crucial for producing realistic color transitions while maintaining spatial coherence. This selection follows prior works in neural style transfer and super-resolution [8], where these specific layers showed strong correlation with perceptual quality.

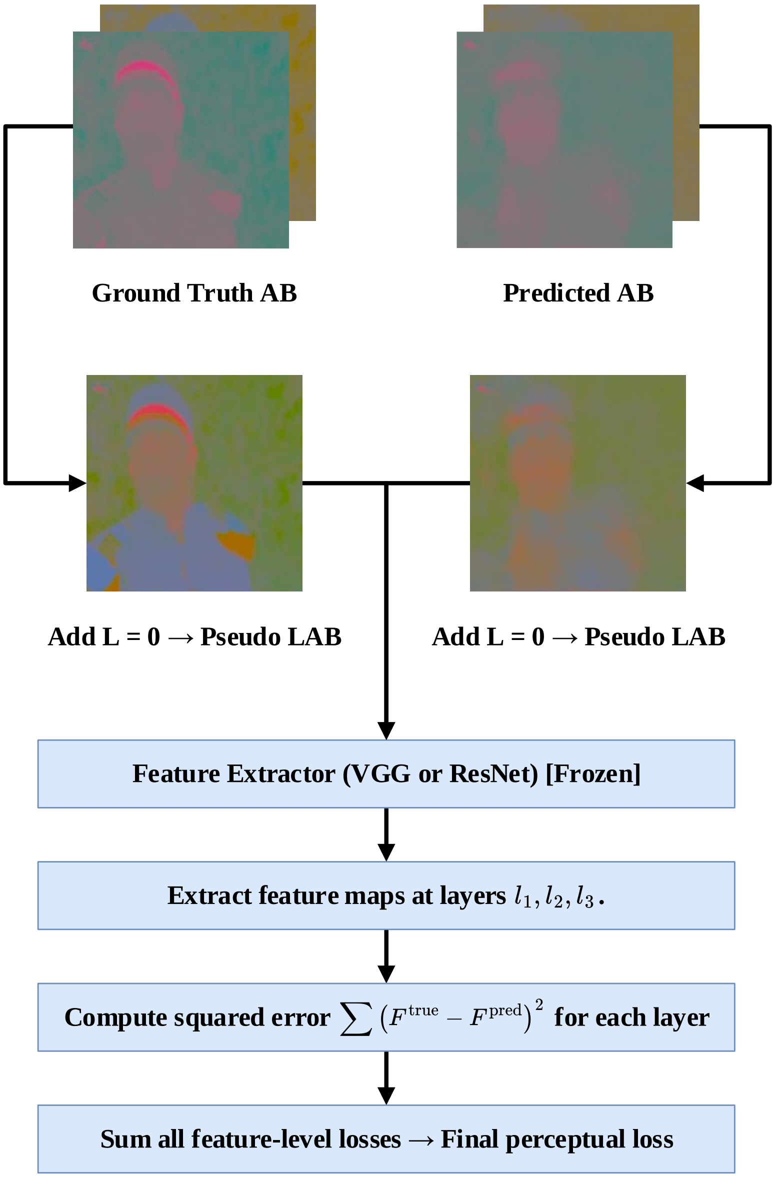
Ground Truth AB
Predicted AB
Add L = 0 → Pseudo LAB
Add L = 0 → Pseudo LAB
Feature Extractor (VGG or ResNet) [Frozen]
Extract feature maps at layers $l_1, l_2, l_3$.
Compute squared error $\sum\left(F^{\text{true}} - F^{\text{pred}}\right)^2$ for each layer
Sum all feature-level losses → Final perceptual loss

**Figure 5.** Workflow of Perceptual Loss Calculation

For ResNet50, the layers used are `conv1_relu`, `conv2_block3_out`, and `conv3_block4_out`. These layers span across different residual blocks to capture progressively more abstract features. The skip connections in ResNet50 help preserve low-level details while enabling deeper feature extraction, which is particularly helpful in learning structure from historical images that may include noise, blur, or occlusions. Compared to VGG16, ResNet50 offers a more compact architecture and improved training stability due to residual learning [9].

Figure 5 illustrates the computation pipeline of perceptual loss. The predicted and ground truth AB channels are first merged with a dummy L channel to form pseudo-LAB images. These LAB images are then passed through a frozen feature extractor (VGG16 or ResNet50), and feature maps are collected from the selected layers. The loss is computed as the sum of squared differences between the corresponding feature maps at each layer. This setup allows the model to focus on structural alignment and visual consistency rather than raw pixel accuracy. Also Table 2 illustrates a simplified comparison of the three loss functions used in this thesis:

**Table 2.** Comparison of Loss Functions

| **Loss Function** | **Computation Level** | **Advantages** | **Limitations** |
|---|---|---|---|
| MSE | Pixel-wise | Simple, stable, widely used | Produces blurry outputs |
| Perceptual (VGG16) | Feature-level (shallow-mid) | Captures texture and perceptual similarity, enhances realism | Computationally expensive, large model |
| Perceptual (ResNet50) | Feature-level (deep + skip) | Maintains fine textures, better abstraction, faster training | May generalize poorly to out-of-domain examples |

As will be shown in the next chapter, the perceptual losses yield more visually pleasing and realistic colorizations than MSE, especially in scenes involving complex human features, clothing, or artistic background elements. However, the increased computation and potential for overfitting must also be considered in practice.

# 4. Experiments and Results

## 4.1. Loss Function Comparison

To assess the impact of different loss functions on model performance, three U-Net-based models were trained using distinct loss configurations: MSE, Perceptual Loss using VGG16, and Perceptual Loss using ResNet50. Each model was trained for 30 epochs with a batch size of 16, leveraging a combined training and validation dataset of approximately 232,042 preprocessed images. The training was conducted on a cloud-based NVIDIA RTX 3090 GPU, ensuring stable performance even for computationally intensive perceptual loss models.

During training, the evolution of both the loss and mean absolute error (MAE) metrics was monitored over each epoch. Figure 6, 7, 8 illustrate the learning curves for the MSE, VGG16 perceptual, and ResNet50 perceptual models, respectively. These figures help visualize the convergence patterns, generalization capabilities, and potential overfitting behaviors associated with each approach.

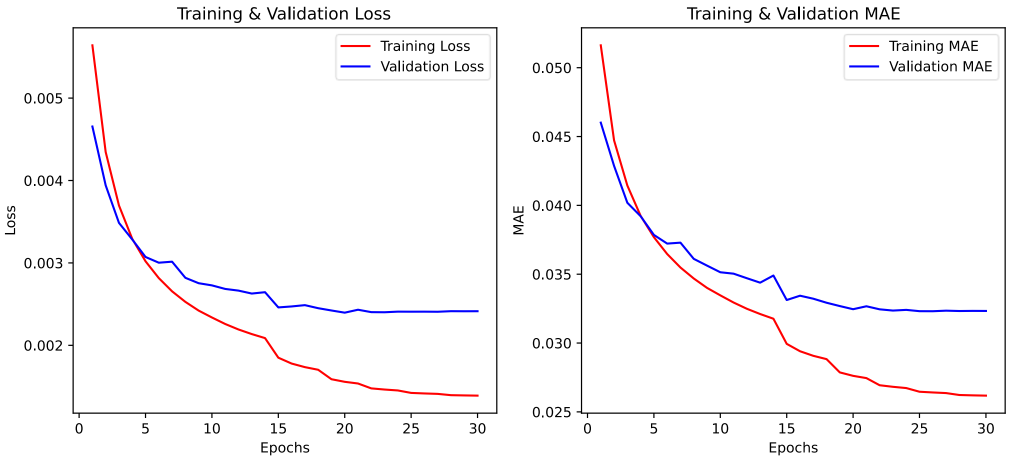

**Figure 6.** Training and validation loss (left) and MAE (right) curves for the model trained with MSE loss.

As shown in Figure 6, the MSE-based model demonstrates a smooth and consistent decline in both training and validation loss. The MAE curve follows a similar pattern, indicating stable convergence. The model reaches a validation MAE of approximately 0.033, suggesting good pixel-level accuracy. Notably, the gap between the training and validation curves is minimal, implying that the model generalizes well without overfitting. Moreover, this configuration has the shortest training time, approximately 5 hours, due to the simplicity of the pixel-wise loss calculation.

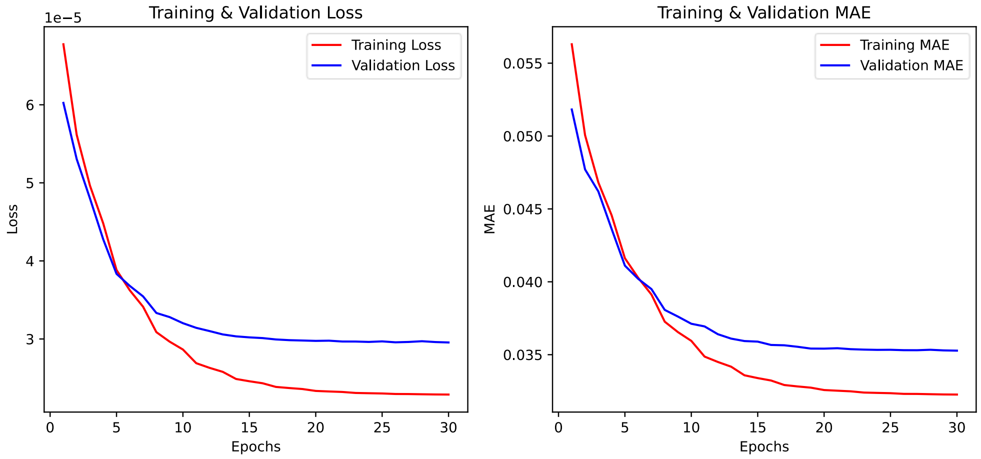

**Figure 7.** Training and validation loss (left) and MAE (right) curves for the model trained with Perceptual Loss using VGG16.

In contrast, the model trained with Perceptual Loss using VGG16 exhibits different characteristics, as seen in Figure 7. The training and validation losses are much smaller in scale due to being computed in feature space, and the convergence is slightly slower. Although the final validation MAE is slightly higher (≈ 0.0348), this does not necessarily imply inferior performance, since perceptual loss emphasizes the preservation of semantic features and visual realism rather than mere pixel accuracy. This model took the longest training time, approximately 10 hours, largely due to the overhead of feature extraction using the VGG16 backbone.

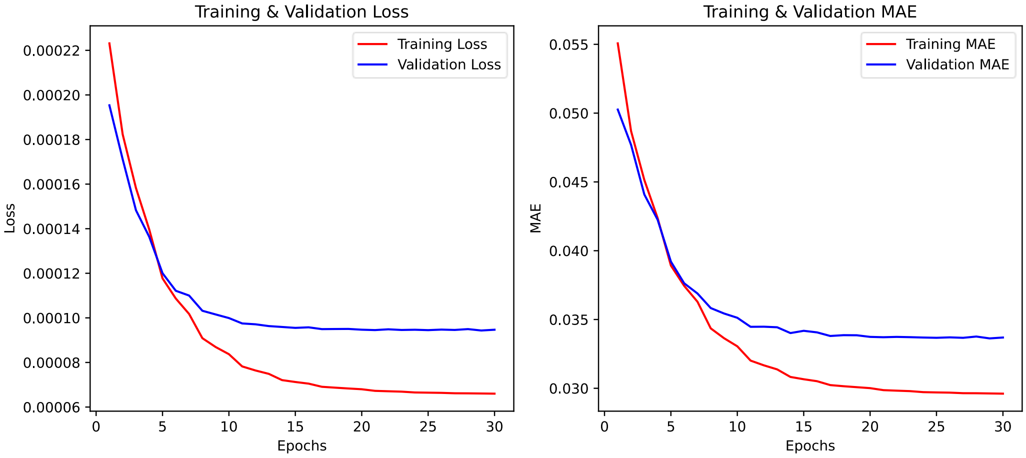

**Figure 8.** Training and validation loss (left) and MAE (right) curves for the model trained with Perceptual Loss using ResNet50.

Figure 8 shows the results for the model trained with Perceptual Loss based on ResNet50. This model strikes a balance between accuracy and training efficiency. Its validation MAE converges around 0.0335, slightly better than the VGG-based model, while the training time is reduced to around 8.5 hours. The training and validation curves are smoother and closer together, indicating consistent learning. ResNet's residual connections likely contribute to better feature reuse and improved convergence.

Across all three models, the validation metrics begin to stabilize after approximately 20 epochs, with diminishing returns beyond that point. This suggests that early stopping or fine-tuning learning rate schedules could further optimize training efficiency in future work. While the MSE model achieves the lowest validation MAE, it is important to emphasize that MAE does not fully capture the perceptual quality of colorized images. Perceptual models, although scoring slightly higher MAE, are expected to produce more visually realistic and semantically consistent results – an aspect that will be explored through qualitative analysis in Section 4.3.

## 4.2. Evaluation Metrics

To quantitatively assess the performance of the proposed colorization models, three standard image quality metrics were utilized: Mean Absolute Error (MAE), PSNR, and SSIM. These metrics were computed between the colorized outputs generated by each model and their corresponding ground-truth color images, over a test set consisting of 25,783 grayscale images.

MAE is a basic yet widely used evaluation metric in image restoration tasks. It computes the average absolute difference between predicted and true pixel values. While MAE is simple to implement and interpret, it lacks sensitivity to human visual perception, often resulting in blurred outputs when used as the sole loss function.

PSNR is another pixel-wise fidelity metric, calculated using the ratio between the peak possible pixel value and the mean squared error between the predicted and ground-truth images. A higher PSNR implies better image quality and lower distortion. However, like MAE, PSNR is less effective in evaluating the perceptual plausibility of colors and textures.

SSIM is more aligned with the human visual system and has become a standard metric in image processing tasks. A value closer to 1 indicates that the predicted image is structurally similar to the reference.

**Table 3.** Quantitative Evaluation on Test Set (25,783 images).

| **Model** | **MAE↓** | **PSNR↑** | **SSIM↑** |
|---|---|---|---|
| U-Net + MSE Loss | 5.5189 | 30.3006 | 0.9555 |
| U-Net + Perceptual VGG16 | 6.0318 | 29.5798 | 0.9533 |
| U-Net + Perceptual ResNet50 | 5.7462 | 29.9902 | 0.9513 |

The results of all three models on the test set are summarized in Table 3. The MSE-trained model achieved the lowest MAE (5.5189) and the highest PSNR (30.3006) and SSIM (0.9555), indicating that it produced outputs closest to the ground-truth in terms of raw numerical accuracy. These outcomes are consistent with previous studies showing that MSE encourages pixel-wise precision [6], but may result in less visually compelling outputs.

In contrast, the model trained with perceptual loss based on VGG16 performed less optimally in terms of MAE (6.0318), PSNR (29.5798), and SSIM (0.9533). This result, however, is expected due to the perceptual loss’s emphasis on high-level semantic features rather than pixel alignment. Prior research has shown that perceptual loss tends to generate outputs with better textures and perceptual coherence [25].

The ResNet50-based perceptual model sits in between the two, yielding a MAE of 5.7462, PSNR of 29.9902, and SSIM of 0.9513. By leveraging residual connections and deeper feature representations, ResNet-based perceptual losses can capture more complex textures while maintaining competitive quantitative results [9].

These results emphasize the trade-off between numeric accuracy and perceptual realism. While MSE provides precise but often over-smoothed results, perceptual losses with pretrained networks offer more visually engaging outputs. As such, the effectiveness of each approach will be further analyzed through qualitative visual comparisons in the next section.

## 4.3. Qualitative Results

In addition to quantitative metrics, qualitative evaluation was conducted to provide a more intuitive understanding of model performance. The visual comparisons were made on two types of data: (1) selected images from the large test set of 25,783 images extracted from historical Vietnamese films, and (2) authentic real-world black-and-white photographs of Vietnam before 1950.

Figure 9 displays qualitative results on images sampled from the test set. These images represent synthetic grayscale inputs derived from color frames of modern historical films. Each row shows, from left to right, the Ground Truth (original color image), the Grayscale version (L-channel input), and the colorized outputs from the three models: MSE-based, Perceptual VGG-based, and Perceptual ResNet-based. Overall, all three models successfully recover the primary color structures of the scenes. However, clear differences emerge upon closer inspection. The MSE-trained model tends to produce slightly muted colors and softer transitions, leading to less vivid and less sharp results, particularly noticeable in clothing and background textures.

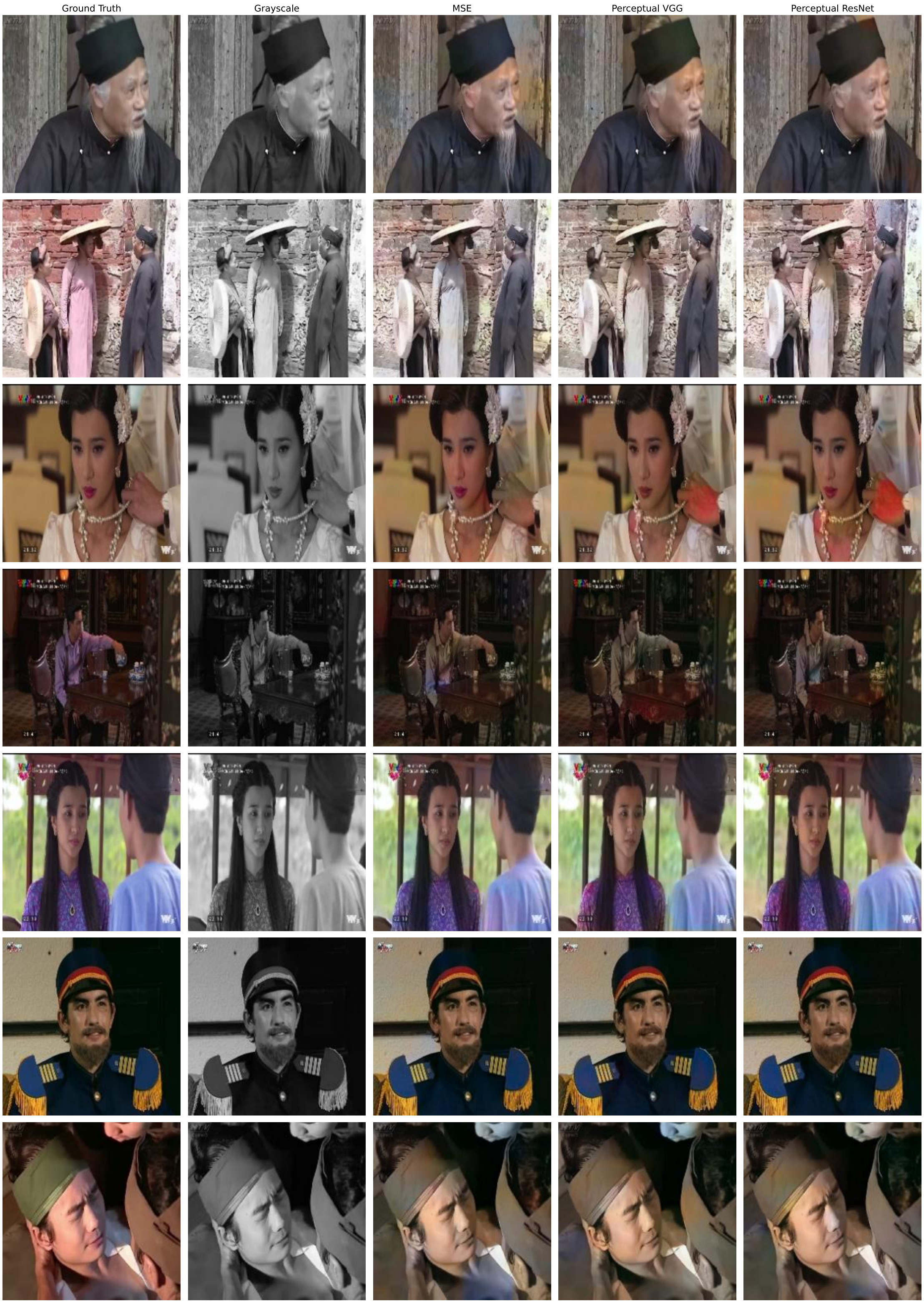

**Figure 9.** Visual comparison on the test dataset images (synthetic grayscale from modern historical films)

The Perceptual VGG-based model generates richer and more vibrant colorizations. It better recovers fine-grained textures such as fabric details and background elements, resulting in a more lively and historically plausible appearance. However, minor artifacts and color inconsistencies, especially in complex regions, are sometimes observed.

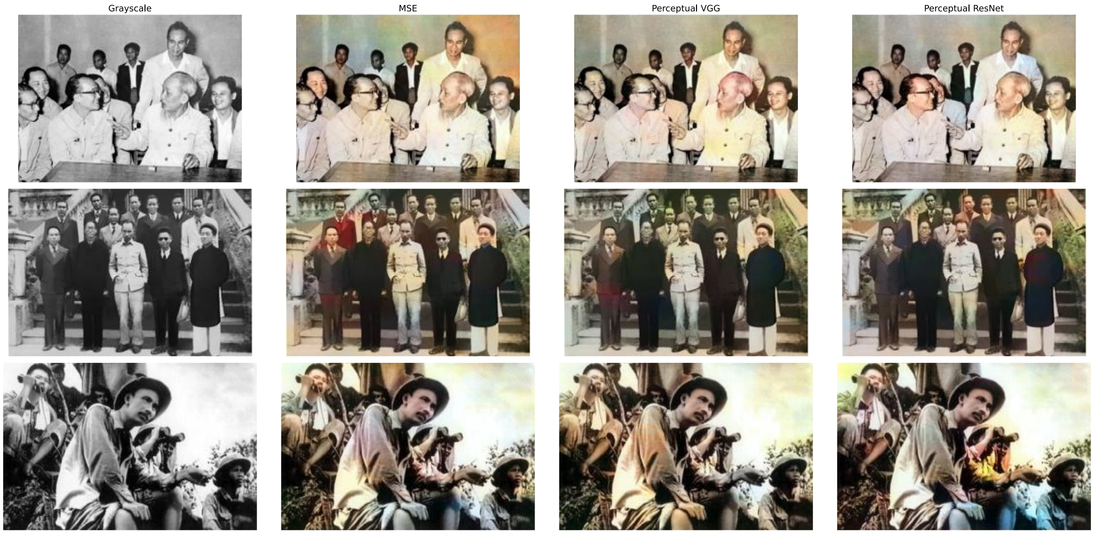

**Figure 10.** Visual comparison on real historical photographs (set 1)

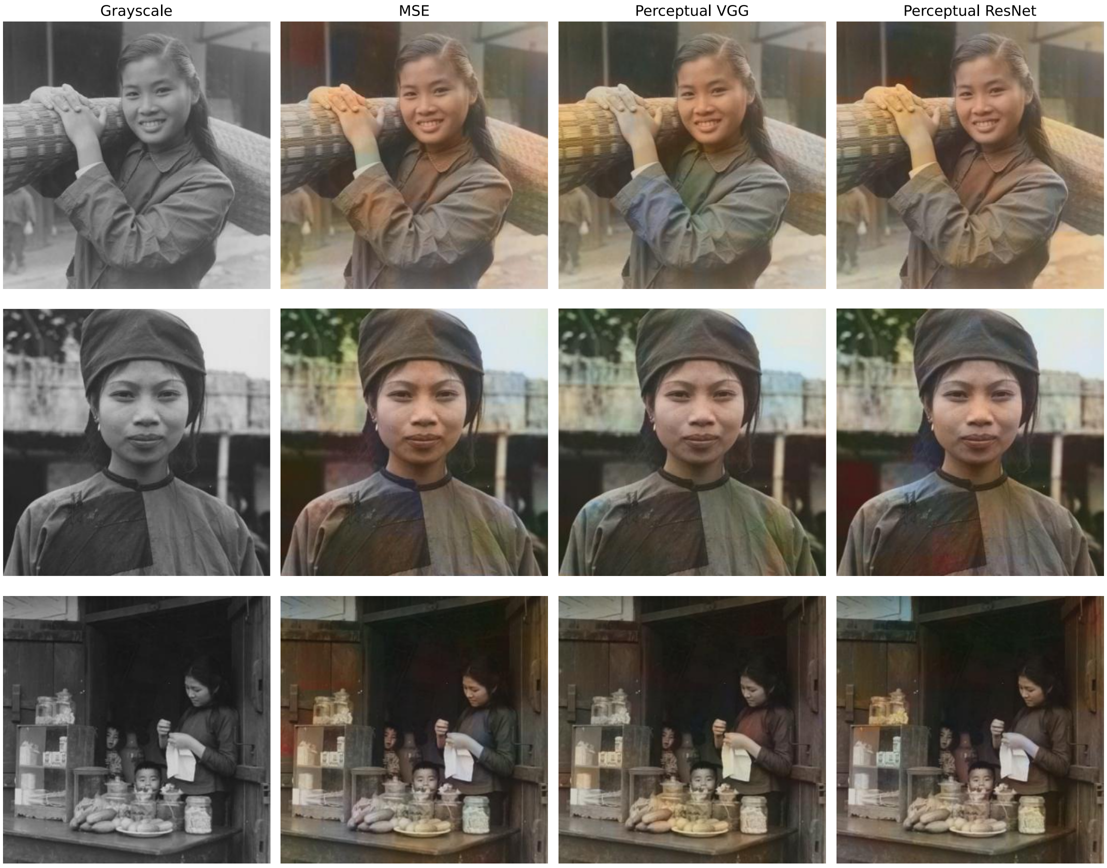

**Figure 11.** Visual comparison on real historical photographs (set 2)

Meanwhile, the Perceptual ResNet-based model strikes a balance between the two: its outputs are vivid but slightly less saturated than VGG, while maintaining superior texture smoothness and fewer visible artifacts. Facial features, object contours, and shading appear more natural and coherent in the ResNet outputs. These observations reinforce the effectiveness of perceptual losses, particularly in enhancing realism for visual restoration tasks.

Moving beyond synthetic examples, Figure 10 and Figure 11 present the results on authentic real-world historical photographs. Figure 10 shows comparisons on images from real historical archives. These black-and-white images contain portraits and social scenes captured during important periods of Vietnamese history. Since no ground truth color information is available, evaluation is purely subjective. The Grayscale input columns show significant information loss in terms of material differentiation and emotional perception. All three models succeed in reintroducing plausible chromaticity into the scenes.

The MSE model again produces relatively conservative colorizations, maintaining structural coherence but lacking vibrancy. The Perceptual VGG model introduces more expressive tones, especially in clothing and facial areas, making the images feel more "alive." However, it sometimes exaggerates contrasts or introduces mild color misplacements. The Perceptual ResNet model produces slightly more stable and smoother outputs, providing a good balance between vividness and authenticity. It is particularly effective at reconstructing consistent backgrounds and preserving the naturalness of human faces.

Figure 11 provides an additional set of real historical examples. In these examples, perceptual models again demonstrate superior ability to infer realistic skin tones, clothing colors, and environmental backgrounds. Fine details such as folds of fabric, facial shadows, and architectural textures are better retained in the Perceptual ResNet outputs, while the Perceptual VGG outputs provide more saturated but occasionally less consistent results. MSE outputs are noticeably duller, especially when recovering complex color distributions.

Across all qualitative evaluations, perceptual loss models clearly outperform traditional MSE-based training in delivering more visually plausible and emotionally resonant colorizations. Between perceptual methods, ResNet50 generally offers slightly better structural stability, while VGG16 provides higher color vibrancy at the cost of occasional artifacts. These qualitative findings are highly consistent with the quantitative results presented earlier and demonstrate the value of feature-level loss functions in the colorization of historical Vietnamese imagery. Table 4 summarizes the key strengths and weaknesses observed during the qualitative evaluation across the three trained models:

**Table 4.** Summary of qualitative strengths and weaknesses of each loss function-based model

| Loss Function | Computation Level | Advantages |
|---|---|---|
| MSE | Stable training, preserves general structures well, avoids extreme hallucinations. | Produces dull colors, lacks vividness and emotional richness, tends to blur fine details. |
| Perceptual (VGG16) | Generates more vibrant and lively colorizations, better restores textures and high-level features. | May introduce color inconsistencies or artifacts in complex regions, slightly unstable colors. |
| Perceptual (ResNet50) | Balances realism and consistency, reconstructs fine textures well, smoother color transitions. | Colors may be slightly less saturated compared to VGG16, generalization to unseen data is moderate. |

# 5. Discussion

## 5.1. Impact of Different Loss Functions

The choice of loss function plays a crucial role in determining the final quality of the colorized outputs. Through both quantitative evaluation (Section 4.2) and qualitative analysis (Section 4.3), it is evident that different losses introduce distinctive patterns of strengths and weaknesses. Models trained with the Mean Squared Error (MSE) loss tend to produce stable and structurally accurate results. They successfully minimize pixel-wise differences between predicted and ground truth images, leading to relatively consistent but often desaturated or blurry colorizations. As observed in the test results, MSE-trained models achieved the lowest Mean Absolute Error (5.5189) and the highest PSNR (30.3006), but the resulting images lacked vibrancy and emotional resonance, especially when compared to perceptually trained models.

In contrast, models trained with Perceptual Loss using VGG16 demonstrated an ability to generate more vivid and semantically meaningful colors. By computing loss in feature space instead of pixel space, these models better captured textures and scene semantics. The qualitative results showed that VGG16-based models produced richer skin tones, more natural backgrounds, and restored fine details such as clothing patterns. However, due to relying on mid-level feature activations (`block1_conv2`, `block2_conv2`, `block3_conv3`), some color bleeding and instability were occasionally observed in complex backgrounds. Quantitatively, the VGG16-based model had slightly higher MAE (6.0318) and lower SSIM (0.9533) than the MSE model, reflecting the trade-off between perceptual realism and pixel-wise fidelity.

Similarly, the model trained with Perceptual Loss using ResNet50 achieved a compromise between stability and perceptual enhancement. ResNet50, with its deeper architecture and residual connections, allowed the model to leverage both low-level and higher-level features (`conv1_relu`, `conv2_block3_out`, `conv3_block4_out`). As a result, the ResNet-based model was able to maintain fine texture and structure more effectively than VGG16-based models, while introducing fewer artifacts. Although its MAE (5.7462) and PSNR (29.9902) fell between the MSE and VGG16 models, visual inspection suggested that ResNet produced more balanced colorizations without severe distortions. Therefore, based on both metric performance and human evaluation, perceptual losses, especially when combined with a strong feature extractor like ResNet50, provide meaningful improvements for historical image colorization.

Overall, while MSE Loss guarantees pixel-level accuracy, it often sacrifices visual appeal. Perceptual Losses, despite slightly higher numerical errors, enhance human interpretability and the emotional connection to the colorized images – an essential factor for applications in historical restoration and education. This confirms the importance of considering both objective and subjective criteria when evaluating colorization systems.

## 5.2. Limitations

Despite achieving promising results, the current system still exhibits several limitations that should be addressed in future work. One notable issue is that the colorization quality remains imperfect. Even with perceptual enhancements, the colors predicted by the models are sometimes inaccurate compared to historical reality. This stems from the inherent ambiguity of the colorization task, where multiple plausible color mappings exist for a single grayscale input, especially in the absence of contextual metadata.

The training dataset, although extensive with over 232,042 images, lacks sufficient diversity. The data was primarily extracted from modern Vietnamese historical films, which, while thematically appropriate, might not fully replicate the textures, clothing, environmental conditions, or photographic artifacts of genuine historical images before 1950. As a result, when applying the models to real-world archival photographs, there may be discrepancies in texture fidelity, background reconstruction, or costume accuracy.

Another limitation relates to model generalization. Since the models were trained solely on a specific style of modern film frames, their ability to handle genuinely old photographs, which often contain noise, scratches, fading, and uneven lighting, may be limited. Although perceptual loss improved robustness to some distortions, further techniques such as data augmentation with artificial degradation, or domain adaptation methods, could enhance real-world performance.

The computational cost of training perceptual models, particularly with VGG16 and ResNet50 backbones, is relatively high. Training the perceptual VGG16-based model took approximately 10 hours and required significant GPU memory resources (RTX 3090), compared to only 5 hours for the MSE-based model. This limitation may hinder the deployment or retraining of these models in resource-constrained environments.

# 6. Model Deployment Using Streamlit

To make the developed historical image colorization models accessible to a broader audience, a web-based application was deployed using Streamlit Cloud. The goal was to create a simple and intuitive interface that allows users to upload grayscale images and receive immediate colorized results without requiring any technical expertise. The deployment process focused on maintaining ease of use, minimal latency, and high-quality output visualization.

The overall processing pipeline of the application involves several sequential steps. First, users can upload grayscale images in common formats such as JPG, JPEG, or PNG, with a file size limit of up to 200MB. Upon successful upload, the image undergoes preprocessing: it is resized to a standardized resolution of 224×224 pixels, converted from RGB to LAB color space, and normalized following the same conventions used during model training (i.e., scaling the L channel to $[0,1]$ and the AB channels to $[-1,1]$). This ensures compatibility with the input expectations of the trained models.

Next, the preprocessed image is passed through three different colorization models: the baseline MSE-trained model, the perceptual loss model using VGG16, and the perceptual loss model using ResNet50. Each model predicts the AB channels corresponding to the input L channel. The predicted outputs are then merged back with the original L channel and converted from LAB to RGB space to reconstruct full-color images. These results are displayed side by side for easy comparison.

Furthermore, the application provides download options for each colorized result. Users can selectively download the output generated by the MSE model, the VGG16-based model, or the ResNet50-based model, depending on their preference. This feature enhances user interactivity and flexibility, allowing the application to serve a wide range of purposes such as educational use, personal archiving, or cultural heritage restoration.

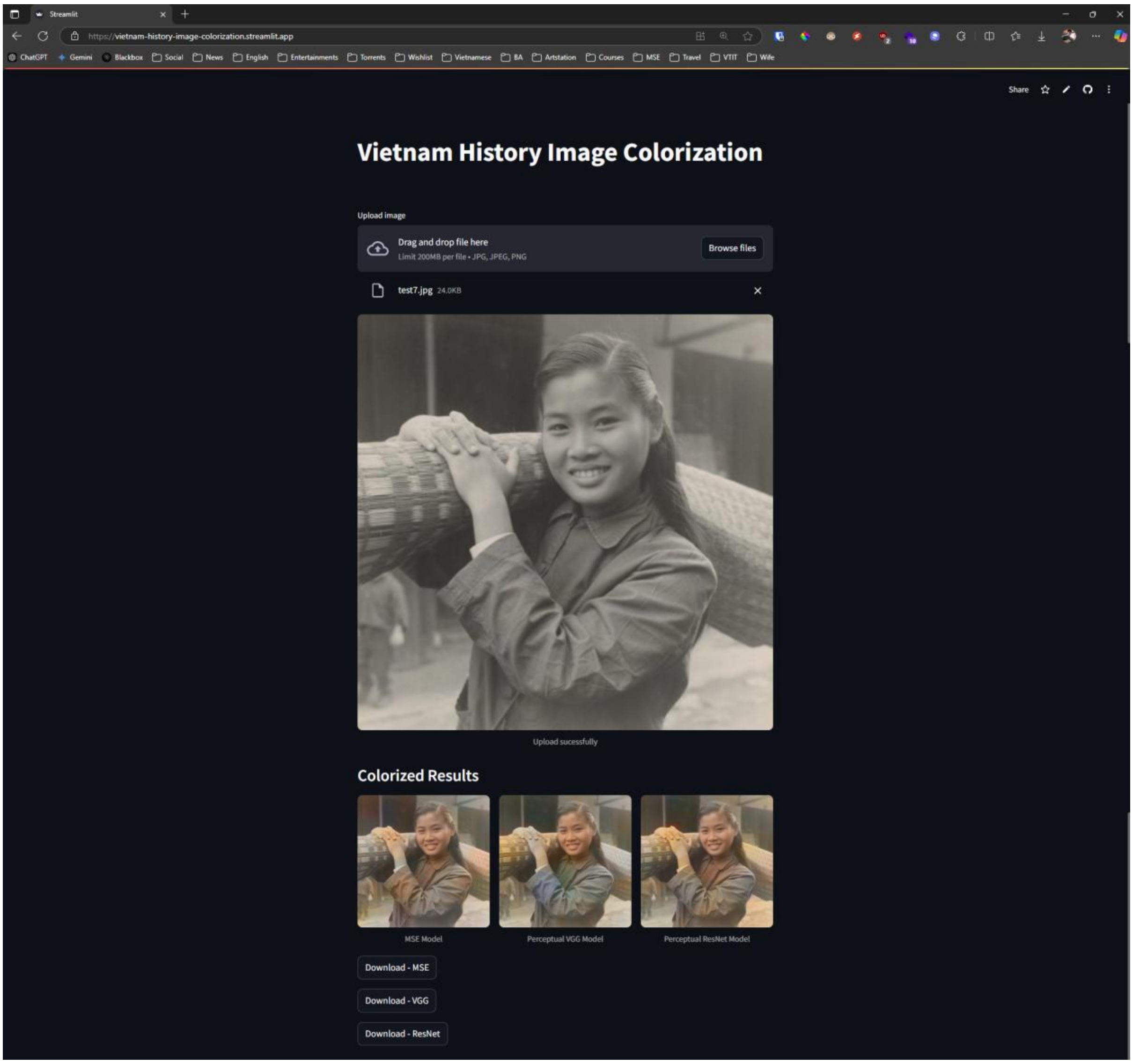

**Figure 12.** User interface of the Vietnam History Image Colorization application deployed using Streamlit.

Figure 12 illustrates the user interface of the deployed application. As shown, the layout is designed to be clean and minimalistic. Users first select and upload a grayscale image. The colorized results are then automatically generated and displayed beneath, categorized according to the respective models. The download buttons provide seamless interaction, enabling users to retrieve the results directly. The deployment was executed via Streamlit Cloud, ensuring that the system is freely accessible online without the need for local installation. The application is hosted at the following URL: https://vietnam-history-image-colorization.streamlit.app/

# 7. Conclusion and Future Work

## 7.1. Summary of Contributions

This thesis has proposed a deep learning-based approach for colorizing historical Vietnamese black-and-white photographs using a U-Net architecture. Unlike conventional manual or rule-based methods, the model automatically learns to infer plausible color information from the grayscale input, focusing on images depicting Vietnam before 1950. One key contribution is the comparison of different loss functions—namely, Mean Squared Error (MSE), Perceptual Loss using VGG16, and Perceptual Loss using ResNet50. The results demonstrate that while MSE yields lower pixel-wise error metrics, perceptual losses, especially those leveraging deep feature maps, produce more visually realistic and detailed colorizations.

To translate the research into a practical tool, a web-based application was developed and deployed using Streamlit Cloud. This application allows users to upload grayscale images and instantly receive colorized outputs generated by all three models. The project thus not only advances academic understanding of colorization techniques but also provides an accessible tool for educators, historians, and the public to interact with historical imagery. The thesis bridges the gap between theoretical deep learning research and real-world application in cultural preservation and storytelling.

## 7.2. Future Work

While the current results are promising, several directions can be pursued to further enhance the system. First, the model architecture can be upgraded by incorporating GANs [13], which have shown strong capabilities in generating more realistic and vibrant outputs. Additionally, integrating semantic segmentation models could guide the colorization process by providing object-level understanding, thus improving context-aware color assignment. Second, the training dataset can be expanded to include authentic historical photographs from Vietnamese archives. This would address domain gaps between modern reenactment scenes and actual vintage photos, improving the model’s generalization to true historical contexts. Finally, the system's performance and scalability could be enhanced by reengineering the deployment stack. Specifically, using FastAPI for backend model inference combined with Streamlit for the frontend interface would significantly optimize response time and allow more concurrent users to access the service. These enhancements aim to make the historical colorization tool more accurate, faster, and impactful for both academic research and cultural heritage preservation.

# References

[1] J. Antic, "DeOldify: A deep learning based project for colorizing and restoring old images," 2019. [Online]. Available: https://github.com/jantic/DeOldify.

[2] 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.

[3] 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.

[4] T. Welsh, M. Ashikhmin and K. Mueller, "Transferring color to greyscale images," in *SIGGRAPH*, San Antonio, Texas, 2002.

[5] A. Levin, D. Lischinski and Y. Weiss, "Colorization using Optimization," in *SIGGRAPH*, Los Angeles, CA, 2004.

[6] R. Zhang, P. Isola and A. A. Efros, "Colorful image colorization," in *ECCV*, Amsterdam, 2016.

[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] 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.

[11] G. Larsson, M. Maire and G. Shakhnarovich, "Learning representations for automatic colorization," in *ECCV*, Amsterdam, 2016.

[12] Z. Cheng, Q. Yang and B. Sheng, "Deep Colorization," in *ICCV*, Santiago, 2015.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," 2014.

[14] C. Vondrick, H. Pirsiavash and A. Torralba, "Generating Videos with Scene Dynamics," in *NIPS*, Barcelona, 2016.

[15] M. Kumar, D. Weissenborn and N. Kalchbrenner, "Colorization Transformer," in *ICLR*, 2021.

[16] P. Sangkloy, N. Burnell, C. Ham and J. Hays, "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies," in *SIGGRAPH*, Los Angeles, CA, 2017.

[17] D. Martin, C. Fowlkes, D. Tal and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *ICCV*, Vancouver, 2001.

[18] E. Reinhard, M. Ashikhmin, B. Gooch and P. Shirley, "Color Transfer between Images," in *IEEE Computer Graphics and Applications*, 2001.

[19] R. Irony, D. Cohen-Or and D. Lischinski, "Colorization by Example," in *EGSR*, Konstanz, 2005.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," in *NIPS*, Long Beach, CA, 2017.

[21] A. Radford, L. Metz and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *ICLR*, Caribe Hilton, San Juan, 2016.

[22] K. Nazeri, E. Ng and M. Ebrahimi, "Image Colorization using Generative Adversarial Networks," Springer, 2018.

[23] T. Nguyen, K. Mori and R. Thawonmas, "Image Colorization Using a Deep Convolutional Neural Network," in *ASIAGRAPH*, Tokyo, 2016.

[24] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak and D. Chen, "Deep Exemplar-based Video Colorization," in *CVPR*, Long Beach, CA, 2019.

[25] J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *ECCV*, Amsterdam, 2016.

[26] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan and A. W. Muzaffar, "Image Colorization: A Survey and Dataset," Elsevier, 2024.

[27] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *20th International Conference on Pattern Recognition*, Istanbul, 2020.

[28] Z. Wang and H. R. Sheikh, "Image quality assessment: From error visibility to structural similarity," in *IEEE Transactions on Image Processing*, 2004.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *CVPR*, Salt Lake City, Utah, 2018.

[30] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018.