FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Interactive Deep Learning Model for Color Restoration of Artwork Images

Rosana García

Postgraduate program en Ciencias de Datos y Aprendizaje Automático
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
December of 2025

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Interactive Deep Learning Model for Color Restoration of Artwork Images

Rosana García

Master's Thesis submitted to the Postgraduate Program en Ciencias de Datos y Aprendizaje Automático, Facultad de Ingeniería of the Universidad de la República, as part of the necessary requirements for obtaining the title of Master in Ciencias de Datos y Aprendizaje Automático.

Directors:
Ph.D Lara Raad
Ph.D Gregory Randall

Academic director:
Ph.D Marcelo Fiori

Montevideo – Uruguay
December of 2025

García, Rosana

Interactive Deep Learning Model for Color Restoration of Artwork Images / Rosana García. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2025.

XXI, 141 p. $29,7$cm.

Directors:

Lara Raad

Gregory Randall

Academic director:

Marcelo Fiori

Tesis de Maestría – Universidad de la República, Program en Ciencias de Datos y Aprendizaje Automático, 2025.

Bibliography: p. 135 – 141.

1. Restauración de imágenes, 2. Restauración Interactiva, 3. Restauración Cromática, 4. Aprendizaje Profundo, 5. LoRA, 6. Patrimonio Cultural, 7. Joaquín Torres García. I. Raad, Lara, Randall, Gregory, . II. Universidad de la República, Postgraduate Program en Ciencias de Datos y Aprendizaje Automático. III. Title

MEMBERS OF THE THESIS DEFENSE COURT

Ph.D Coloma Ballester

Ph.D Marcelo Bertalmío

Ph.D Pablo Musé

Montevideo – Uruguay
December of 2025

Acknowledgements

I would like to express my gratitude to all those who contributed to this work, whether through their time, guidance, support, or companionship throughout this process.

To Gregory and Lara, my advisors, for all the time they devoted to me, for their patient and rigorous guidance, and for accompanying every stage of this work with generosity and trust.

To Gabriele, for giving me the opportunity to visit ENS Paris–Saclay, and participate in an event like MLBrief, which broadened my horizons and taught me to look at research papers in a different way.

To Carlos and Alejandro from the Museo Torres García, for opening the doors of the archive, sharing their knowledge with such dedication, and supporting this project with enthusiasm.

To Ignacio, Julio, and Jaime from LaPA, for their support and their generosity in sharing technical expertise.

To Natalia and Santiago, who so often offered me moral support through WhatsApp.

To Marcelo, who planted the idea of pursuing a Master's degree (though I suspect he may not know it).

To my colleagues at work, whom I turned to countless times with questions and doubts.

To my friends, and especially to my closest women friends.

To my family (those who are here and those who are no longer with us).

To my extended family.

To Titina, who is always ready to help me.

And most especially, to Guille, for everything.

Somewhere in time, beyond time, the world was gray. Thanks to the Ishir people, who stole color from the gods, the world today is resplendent with colors that dazzle the eyes of all who look at them.

Ticio Escobar lent a hand to a film crew that came to the Chaco to shoot scenes of daily life among the Ishir.

An Indian girl pursued the director, a silent shadow glued to his side, staring into his face as if she wanted to jump into his strange blue eyes.

The director turned to Ticio, who knew the girl and understood her language. Through him, she confessed, "I want to know what colors you see."

The director smiled, "The same as you".

"And how do you know what colors I see?"

Eduardo Galeano, Points of view, Voices of Time.

RESUMEN

Este trabajo aborda el desafío de restaurar digitalmente la apariencia cromática de imágenes de las obras del artista uruguayo Joaquín Torres García perdidas en el incendio del Museo de Arte Moderno de Río de Janeiro en 1978. Basándose en documentación monocromática heterogénea, el trabajo evalúa las limitaciones de los métodos de colorización totalmente automáticos y utiliza un enfoque híbrido con intervención humana-experta basado en la arquitectura iColoriT. Para adaptar el modelo al dominio del artista, se construyeron conjuntos de datos con diversas características, que abarcan desde un núcleo curado por expertos a partir de obras de Torres García hasta un corpus más amplio que integra diversas fuentes de archivo. Se implementó una estrategia de ajuste fino mediante Adaptación de Bajo Rango (LoRA) para optimizar el Vision Transformer (iColorit), capturando la paleta del artista y preservando las capacidades interactivas del modelo. Los resultados experimentales demuestran que los modelos adaptados con LoRA superan consistentemente al modelo base de iColoriT tanto en métricas cuantitativas (PSNR, LPIPS) como en la fidelidad cualitativa, especialmente en escenarios con escasa guía del usuario. Por último, para los curadores, este trabajo adapta una herramienta de software interactiva que permite a los usuarios tomar muestras de color de imágenes de referencia cromáticas y condicionadas por la luminancia. Esto garantiza restauraciones históricamente y estilísticamente plausibles.

Palabras clave:
Restauración de imágenes, Restauración Interactiva, Restauración Cromática, Aprendizaje Profundo, LoRA, Patrimonio Cultural, Joaquín Torres García.

ABSTRACT

This thesis addresses the challenge of digitally restoring the chromatic appearance of artworks by the Uruguayan artist Joaquín Torres García, lost in the 1978 fire at the Museum of Modern Art in Rio de Janeiro. Relying on heterogeneous monochromatic documentation, the work evaluates the limitations of fully automatic colorization methods in the artistic domain and uses a hybrid approach based on the iColoriT architecture. To adapt the model to the artist's domain, a hierarchy of progressively inclusive datasets was constructed, expanding from a curated core of Torres García to a broader corpus incorporating diverse archival sources. A parameter-efficient fine-tuning strategy using Low-Rank Adaptation (LoRA) was implemented to optimize the Vision Transformer, capturing the artist's palette while preserving the model's interactive capabilities. Experimental results demonstrate that the LoRA-adapted models consistently outperform the iColoriT baseline in both quantitative metrics (PSNR, LPIPS) and qualitative fidelity, particularly in regimes with sparse user guidance. Finally, for curators, this work adapts an interactive software tool that allows users to take color samples from chromatic and luminance-conditioned reference images. This ensures historically and stylistically plausible restorations.

Keywords:
Image Restoration, Interactive Restoration, Color Restoration, Deep Learning, LoRA, Cultural Heritage, Joaquín Torres García.

List of Figures

List of Tables

‘

Table of Contents

Chapter 1

Introduction

1.1. Motivation

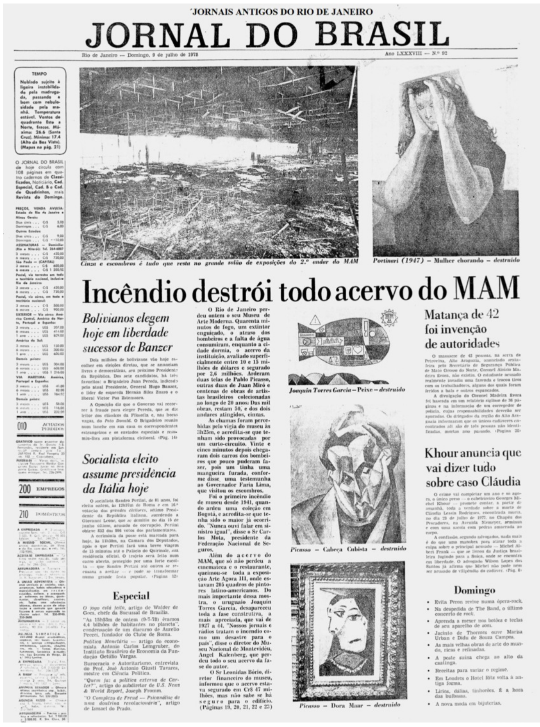
In the early hours of July 8, 1978, a devastating fire destroyed the Museum of Modern Art of Rio de Janeiro, reducing its valuable collection to ashes.

Since World War II, no comparable artistic and cultural catastrophe had been recorded. The disaster caused astonishment and grief worldwide. The press reported it with dramatic headlines: "*The greatest disaster in modern art*", "*World shock at the irreparable loss*", "*The greatest catastrophe for Latin America*". (. . .)

On that tragic day, works by Van Gogh, Picasso, Dalí, Léger, Miró, Max Ernst, Kandinsky, Matisse, and others disappeared; But the most affected artist was the Uruguayan Joaquín Torres García. (. . .) The works had been carefully selected from his Constructive period—perhaps the most representative of his talent. They had been chosen for the major exhibition organized in his honor by the Museum of Modern Art of the City of Paris in June 1975.

Those 73 works are now lost.

Jacques Lassaigne, 1981.
Curator of the Museum of Modern Art of Paris.
Excerpted from *Torres García. Works Destroyed in the Fire at the Museum of Modern Art of Rio de Janeiro*.
Fundación Torres García, Montevideo, 1981.



JORNAIS ANTIGOS DO RIO DE JANEIRO

JORNAL DO BRASIL

Rio de Janeiro — Domingo, 9 de julho de 1978 Ano LXXXVIII — N.º 92

Cinza e escombros é tudo que resta no grande salão de exposições do 2.º andar do MAM

Portinari (1947) – Mulher chorando – destruído

Incêndio destrói todo acervo do MAM

Bolivianos elegem hoje em liberdade sucessor de Banzer

Socialista eleito assume presidência da Itália hoje

Joaquín Torres Garcia – Peixe – destruído

Matança de 42 foi invenção de autoridades

Khour anuncia que vai dizer tudo sobre caso Cláudia

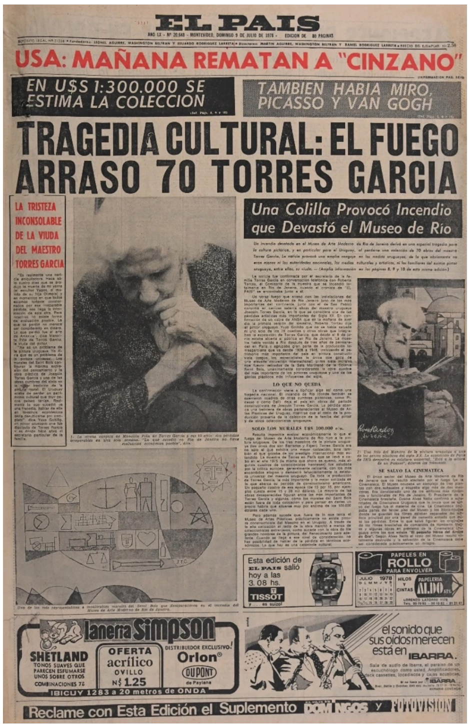
Picasso – Cabeça Cubista – destruído

Picasso – Dora Maar – destruído

Especial

Domingo

(a) *Jornal do Brasil*, July 9, 1978.



EL PAIS

AÑO LX - Nº 20.648 - MONTEVIDEO, DOMINGO 9 DE JULIO DE 1978 - EDICION DE 80 PAGINAS

USA: MAÑANA REMATAN A "CINZANO"

EN U$S 1:300.000 SE ESTIMA LA COLECCION

TAMBIEN HABIA MIRO, PICASSO Y VAN GOGH

TRAGEDIA CULTURAL: EL FUEGO ARRASO 70 TORRES GARCIA

LA TRISTEZA INCONSOLABLE DE LA VIUDA DEL MAESTRO TORRES GARCIA

Una Colilla Provocó Incendio que Devastó el Museo de Río

LO QUE NO QUEDA

Reclame con Esta Edición el Suplemento DOMINGOS y FOTOVISION

(b) *El País*, July 9, 1978.

Figure 1.1: Front pages of Brazilian and Uruguayan journals reporting the Rio de Janeiro Museum of Modern Art (MAM) fire.

The destruction of 73 artworks by Joaquín Torres García in the 1978 fire at the Museum of Modern Art (MAM) of Rio de Janeiro remains one of the most significant cultural losses in Uruguay and Latin America. Although the original works were irretrievably destroyed, various forms of documentation survived: monochromatic photographs, printed reproductions, audiovisual fragments, and portions of murals. These heterogeneous materials constitute the only remaining visual record of the lost paintings.

Forty years later, the Museo Torres García revisited this loss through the 2018 exhibition *Tiempo de Mirar* (Museo Torres García, 2018). This exhibition combined surviving fragments, archival documents, and an augmented-reality experience that allowed visitors to visualize some of the destroyed works using the available images. The initiative underscored both the fragility of cultural heritage and the potential of digital media to mediate access to artworks that no longer exist in physical form.

1.1.1. Digital preservation and computational approaches

In recent years, computer vision and deep learning have become increasingly influential in heritage documentation, analysis, and reconstruction (Mitric et al., 2024). Their integration into museums and conservation labs enables levels of inspection, comparison, and material analysis that would be unattainable through manual or purely analog methods (Sarkar & Singh, 2025).

A prominent example is the *Operation Night Watch* project (Gabrieli et al., 2021) at the Rijksmuseum, where hyperspectral imaging and a computer-controlled scanning system produced a detailed material map of Rembrandt's *The Night Watch*. The resulting multimodal dataset—later aligned using computer vision—revealed pigment distributions, underdrawings, and compositional adjustments invisible to the naked eye.

Beyond high-fidelity documentation, research has increasingly explored how deep learning can support chromatic reconstruction. The study *Artificial Intelligence-Based Color Reconstruction of Mogao Grottoes Murals Using Computer Vision Techniques* (Y. Zhang & Bunyasakseri, 2025) exemplifies this direction, combining high-resolution imaging, curated pigment databases, and spectral analyses to infer original chromatic values of the Mogao Caves murals. Its data-driven framework demonstrates how computational methods can help recover complex chromatic structures in the absence of fully preserved material.

These examples highlight a common requirement across digital restoration workflows: the sensitivity of models to the quality and consistency of their input data. Marrocchesi and Erdmann (2024) demonstrates that even small variations in illumination, focus, or camera alignment can significantly affect mosaicking, calibration, or material detection, emphasizing the importance of controlled acquisition and calibrated equipment.

In contrast to these ideal conditions, the present thesis relies on a heterogeneous and historically constrained corpus, particularly regarding the lost works. As noted, the surviving records consist of aged, low-resolution photographs, printed reproductions of uneven quality, and images captured under diverse lighting and unknown acquisition parameters. Although a small subset of new photographs was acquired following controlled guidelines, the majority of the dataset inevitably reflects the fragmentary and non-standard nature of archival material. This heterogeneity is intrinsic to the corpus's historical circumstances and fundamentally informs the methodological choices of this thesis.

1.2. General Objective

The objective of this work is to propose a digital color restoration pipeline for Joaquín Torres García's lost artwork images, based on interactive image colorization techniques that integrate algorithmic inference with expert artistic knowledge.

Image colorization methods aim to recover plausible chromatic information from grayscale images by leveraging learned color priors and, in interactive settings, incorporating user-provided color hints to resolve ambiguities inherent to the problem. Such interaction is particularly relevant in artistic restoration contexts, where multiple color interpretations may be valid, and expert guidance is essential to ensure historical and stylistic coherence.

Several state-of-the-art colorization approaches are analyzed within this framework, and a hybrid interactive method is selected to address the specific challenges of restoring Torres García's works. Given that the primary end users of the proposed system are museum curators and art experts, part of this work also focuses on adapting and extending an interactive demo to facilitate expert-driven experimentation and qualitative evaluation. Although the development of such an interface was not initially planned, it became necessary for the medium-term validation of interactive restoration workflows in collaboration with the museum.

1.2.1. Ethical considerations

Digital color restoration inevitably involves estimating information that is no longer present, particularly when the physical object has been lost. The resulting color proposals should therefore be interpreted as plausible approximations rather than definitive recoveries. Furthermore, model predictions must be scrutinized for potential biases inherited from the training datasets.

The primary aim of this work is to assist experts in the field. It must be emphasized that this algorithmic approach does not claim to reproduce the "true" colors of the lost pieces, as the ground truth is irretrievably lost. Instead, it offers hypothetical reconstructions constrained by the artist's known chromatic vocabulary.

This thesis also follows a commitment to transparency and reproducibility: all processing stages are documented, and the interactive interface records the sequence of actions performed during each session to ensure full traceability. However, the datasets used here are not publicly available. Access was granted through institutional agreements with Fundación Gurvich, Cecilia de Torres, and the Museo Torres García, which authorized their use for research but do not permit redistribution or open publication of the full corpus. Within these constraints, the methodology and implementation are described in detail to support verifiable and extendable research practices. In addition, the source code will be made available through a public GitHub repository to further support reproducibility.

1.2.1.1. Use of AI-assisted tools

AI-based tools were utilized for writing styling, translation and linguistic refinement during the preparation of this thesis.[1]

While these tools supported the writing process, all conceptual development, methodological reasoning, data analysis, and scientific conclusions remain the sole responsibility of the author.

[1] Tools used: ChatGPT (OpenAI, https://chatgpt.com), Gemini (Google, https://gemini.google.com) and DeepL (https://www.deepl.com).)

1.3. Summary of Contributions

This thesis builds upon and consolidates research developed during the author's participation in two peer-reviewed publications in the MLBrief IPOL series, both of which include a detailed description of the methods, for reproducibility, and the corresponding online demos. Its main contributions are:

A Brief Analysis of iColoriT for Interactive Image Colorization (García et al., 2024a): This work describes and analyzes iColoriT (Yun et al., 2023), a hybrid colorization method based on a Vision Transformer. The model propagates user hints to relevant regions of a grayscale image while utilizing color priors learned from large datasets. This approach provides users with enhanced control over color inference and demonstrates an efficient workflow for achieving high-quality results

A Short Analysis of BigColor for Image Colorization (García et al., 2024b): This article analyzes the BigColor method (Kim et al., 2022), a fully automatic approach designed to generate realistic and vivid colorizations for complex images. Based on a BigGAN-inspired encoder-generator network, the method utilizes a spatial feature map to enable single-forward-pass colorization and supports arbitrary input resolutions. The contribution includes an analysis of the method's performance, highlighting both its achievements and limitations.

Adaptation of an interactive colorization framework for expert-guided restoration: An extension of an existing interactive colorization interface with additional tools designed to support curator-driven experimentation and qualitative assessment, presented in Chapter 6.

Fine-tuning of a colorization model adapted to Torres García's chromatic palette: The development and evaluation of LoRA-based adaptations trained on curated datasets of the artist's production, assessed under varying hint densities using quantitative metrics and visual analysis.

Construction of curated image datasets: Four non-public image datasets of constructive artworks were assembled as part of this work, providing a reusable visual corpus for future research within the Image Group at the Facultad de Ingeniería (Udelar). Additionally, a small dataset of 4 paired images combining traditional analog grayscale film techniques with high-definition digital imaging was constructed as part of the broader research activities associated with this thesis, enabling future investigations in image.

1.4. Organization of this thesis

This thesis is organized into seven chapters:

Chapter 1 introduces the motivation, cultural context, objectives, and contributions.
Chapter 2 establishes the theoretical framework, covering color representation, classical and modern colorization methods, and key deep learning concepts necessary for understanding the subsequent chapters.
Chapter 3 describes the datasets constructed for this thesis, detailing the surviving material of the lost works, preprocessing steps, and the challenges posed by heterogeneous archival sources.
Chapter 4 presents the methodological framework, explaining the iColoriT model and its adaptations, as well as the training and fine-tuning.
Chapter 5 reports the evaluation procedures, experimental results, including performance across hint regimes, comparisons of LoRA configurations, and both quantitative (PSNR/LPIPS) and qualitative analyses.
Chapter 6 details the improved interactive demo and discusses its current application at the Museo Torres García.
Chapter 7 summarizes the findings, discusses limitations, and outlines directions for future work.

Chapter 2

Theoretical Foundations

This chapter presents the theoretical foundations of the research, organized into three main areas: color theory, image colorization, and deep learning methods.

Section 2.1 reviews color perception and the color spaces used for digital representation. Section 2.2 introduces the colorization challenge and outlines both classical and learning-based solutions. Finally, Section 2.3 explores the relevant deep learning architectures and training strategies, including optimization, hyperparameter tuning, and parameter-efficient fine-tuning (LoRA). These foundations underpin the proposed strategy for adapting pretrained models to artistic image restoration.

2.1. Color

The representation of color through standardized spaces provides a quantitative bridge between human vision and digital systems, allowing chromatic information to be measured, compared, and reproduced consistently across devices.

The theoretical foundations and mathematical definitions are summarized in this section, which are primarily based on the work of Ford and Roberts (1998) and the document HunterLab (2015), complemented by the formal standards established by the Commission Internationale de l'Éclairage (CIE, 1932, 2022a, 2022b).

2.1.1. Color perception and representation

As presented by Ford and Roberts (1998), color is a subjective and personal phenomenon. Although measuring how the brain reacts to what we see is complex, it is essential for describing and sharing colors consistently between people and across devices. Biologically, color is not a property of objects themselves, but a percept created by our visual system in response to light. Although a color can be described physically by its spectral power distribution, the human visual system reduces this continuous information through three types of cone cells, each roughly sensitive to long-, medium-, or short-wavelength light. Signals from the cones, together with those from rods that sense brightness, are integrated by the brain to produce the perceptual experience of color. The CIE formally defined a set of perceptual attributes to describe this experience, which are summarized by Hunt and Pointer (2011).

Brightness: how much light a stimulus appears to emit or reflect. For example, a phone screen showing the same image at minimum and maximum backlight has identical colors but very different brightness.

Hue: the attribute that places a stimulus within a basic color family (red, yellow, green, blue). A dark green leaf and a light green leaf share the same hue, even if they differ in lightness.

Colorfulness: the extent to which a color appears different from Gray, i.e., the degree to which a stimulus exhibits its hue.

Lightness: how light or dark a surface appears relative to a similarly illuminated white. The same gray patch may look light when surrounded by darker areas and dark when surrounded by lighter ones.

Chroma: the strength or vividness of a color compared to the brightness of an equivalent white. A red paint swatch has the same chroma indoors and outdoors, even if it appears more colorful in daylight.

Saturation: the strength or purity of a color relative to its own brightness. A very bright pink can have lower saturation than a deep red because its high brightness makes the hue appear less pure.

While these perceptual attributes characterize how humans see color, they do not by themselves provide a way to measure or reproduce it. Perception depends on a physical situation that must be specified before it can be quantified. As described by HunterLab (2015), seeing color requires the interplay of three elements: a light source, an object, and an observer. The light provides the spectral energy that makes color visible; the object selectively absorbs and reflects different wavelengths; and the visual system interprets the resulting stimulus.

To measure color, this physical situation is replaced by standardized numerical descriptions. The light source becomes a CIE illuminant, a reference spectral power distribution such as D65 for daylight (CIE, 2022b). The object is described by its spectral reflectance curve, which gives the proportion of incident light reflected at each wavelength. The observer is represented by the CIE Standard Observer (CIE, 1932), defined by color-matching functions experimentally established in 1931 and refined in 1964 to represent average human visual sensitivity (HunterLab, 2015). Together, these three standardized components make it possible to assign reproducible numerical coordinates to a color stimulus.

Mathematically, the illuminant, the object's spectral reflectance, and the CIE Standard Observer functions are multiplied, wavelength by wavelength, and then integrated across the visible spectrum to produce the CIE tri-stimulus values (X, Y, Z) (HunterLab, 2015). This computation underpins the CIE system for color measurement and specification. The resulting tri-stimulus values uniquely characterize a stimulus under defined viewing conditions and form the basis of all subsequent CIE color spaces. A color specified by its CIE coordinates will be reproduced consistently when the same illuminant and observer are used (Ford & Roberts, 1998).

2.1.1.1. Color spaces

Once color is defined in standardized physical and perceptual terms, color is encoded using numerical coordinate systems known as *color spaces*. Each space dictates how colors are represented and related: every color maps to a point in a three-dimensional coordinate system whose interpretation relies on the chosen space (Ford & Roberts, 1998).

Color spaces serve distinct purposes based on their application. Device-dependent spaces, such as RGB or CMYK, specify color reproduction using particular primaries (lights or inks); thus, numerical values depend on the specific device characteristics, whether a printer or a display (Ford & Roberts, 1998). Consequently, identical triplets may yield different perceptual colors across hardware.

A key concept is the *color gamut*, the subset of colors a device or space can reproduce (Ford & Roberts, 1998). In displays, the gamut is determined mainly by the chromaticities of the primaries and the *white point*—the reference chromaticity coordinates that define "neutral white" at maximum intensity. In printers, it depends on the spectral properties of inks and substrates. Because gamuts differ, some colors may be representable on one device but not another.

Conversely, *device-independent* spaces are designed to be invariant. Built upon the CIE framework, they define tri-stimulus values (X, Y, Z) that uniquely characterize stimuli under standardized conditions. Derived spaces—such as CIELAB and CIELUV—reorganize XYZ coordinates to better reflect perceptual attributes while preserving device independence (Ford & Roberts, 1998).

Color calibration Color calibration quantifies a device's colorimetric behavior. For displays, this entails determining the chromaticities of the primaries, the white point, and the transfer function (or tone response). The latter describes the non-linear relation between input channel values and emitted luminance, commonly modeled with a gamma curve to compensate for device properties and the non-linear sensitivity of human vision.

For printers, calibration requires measuring the spectral characteristics of inks and substrates. The resulting characterization is stored in a *device profile*, which maps the native color space to a device-independent reference such as CIE XYZ or CIELAB (Ford & Roberts, 1998).

These differences become evident in real settings: multiple uncalibrated displays may show noticeably different colors even when driven by identical image data (see Figure 2.1).

Color management system (CMS) A CMS ensures consistency when images move between devices. Using device profiles, it converts colors from a device-dependent space to another via a device-independent reference. This process compensates for differences in primaries, white points, and gamuts, maintaining a consistent appearance across calibrated systems (Ford & Roberts, 1998).

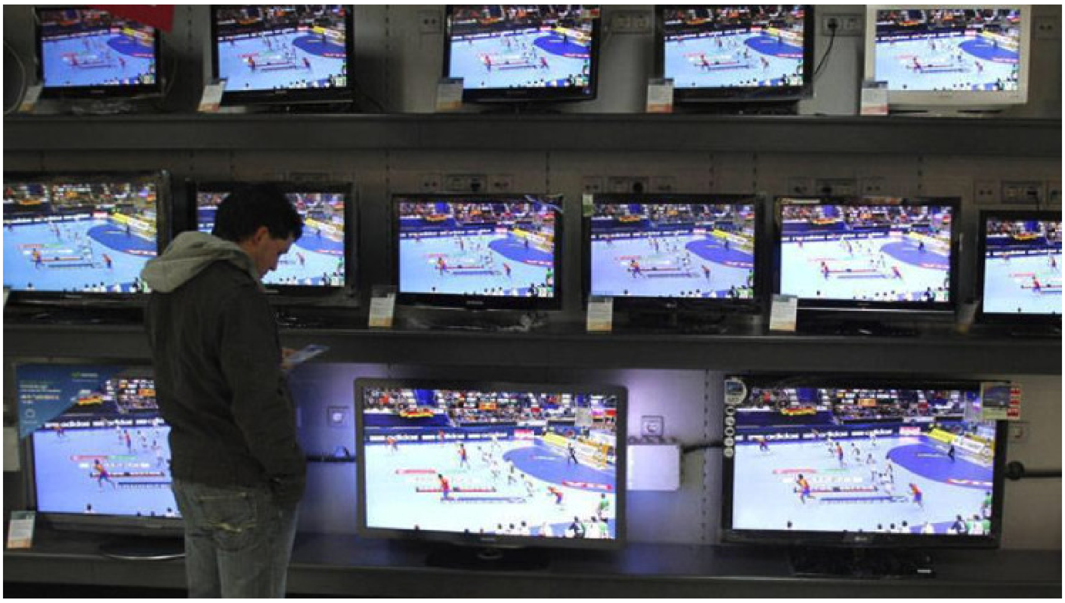


Figure 2.1: Illustration of how identical RGB color content appears differently on multiple displays, Image from Bergasa (2014).

2.1.2. Mathematical formulations of color spaces

The following equations summarize the standard definitions established by the CIE for the 1931 XYZ system and the 1976 CIELAB color space (CIE, 1932; ISO/-CIE, 2019). These formulations provide the numerical framework for describing, comparing, and transforming colors. The presentation below adheres the conventions established by Ford and Roberts (1998).

The CIE 1931 XYZ system. The CIE 1931 XYZ space (CIEXYZ) represents a color stimulus through three stimulus values (X, Y, Z), computed from the illuminant, the object's spectral reflectance, and the CIE Standard Observer. By construction, Y corresponds to luminance, and the coordinates are non-negative.

Chromaticity is expressed using the normalized coordinates:

$$x = \frac{X}{X+Y+Z}, \qquad y = \frac{Y}{X+Y+Z}, \tag{2.1}$$

with the third coordinate given by $z = 1 - x - y$. This representation, often denoted as CIE Yxy, separates luminance (Y) from chromaticity (x, y). However, Euclidean distances in this chromaticity diagram are not perceptually uniform.

The CIE 1976 L*a*b* color space. The CIE 1976 L*a*b* (CIELAB) space is a nonlinear transformation of CIEXYZ that makes Euclidean distances more closely correspond to perceived color differences. The lightness component L^* is defined as:

$$L^* = \begin{cases} 116 \left(\frac{Y}{Y_n}\right)^{1/3} - 16 & \text{if } \frac{Y}{Y_n} > \epsilon, \\ 903.3 \left(\frac{Y}{Y_n}\right) & \text{if } \frac{Y}{Y_n} \leq \epsilon, \end{cases} \tag{2.2}$$

where (X, Y, Z) are the tri-stimulus values of the stimulus, (X_n, Y_n, Z_n) are those of the reference white, and $\epsilon \approx 0.008856$.

Defining the nonlinear transformation function $f(t)$:

$$f(t) = \begin{cases} t^{1/3} & \text{if } t > \epsilon, \\ 7.787\, t + \frac{16}{116} & \text{if } t \leq \epsilon, \end{cases} \tag{2.3}$$

The chromatic components are calculated as:

$$a^* = 500 \left[f(X/X_n) - f(Y/Y_n)\right], \qquad b^* = 200 \left[f(Y/Y_n) - f(Z/Z_n)\right]. \tag{2.4}$$

A polar representation, useful for relating CIELAB to perceptual attributes, is given by:

$$C^* = \sqrt{a^{*2} + b^{*2}}, \qquad h_{ab} = \arctan\left(\frac{b^*}{a^*}\right), \tag{2.5}$$

where C^* denotes chroma and h_{ab} represents the hue angle. Note that CIELAB does not explicitly define a saturation coordinate.

2.1.2.1. Conversion from CIELAB to RGB

To reproduce a CIELAB color on a display, it must first be converted to a device-independent space (CIEXYZ) and then to a device-dependent RGB space. The conversion below adheres to the sRGB standard IEC 61966-2-1 (International Electrotechnical Commission, 1999), characterized by fixed primaries, a D65 white point, and a nonlinear gamma curve (Ford & Roberts, 1998).

From CIELAB to CIEXYZ. Let (L^*, a^*, b^*) denote the CIELAB coordinates and (X_n, Y_n, Z_n) the reference white. We define the intermediate variables:

$$f_Y = \frac{L^* + 16}{116}, \qquad f_X = f_Y + \frac{a^*}{500}, \qquad f_Z = f_Y - \frac{b^*}{200}. \tag{2.6}$$

Using the constant $\delta = 6/29$, the inverse nonlinear function $f^{-1}(t)$ is:

$$f^{-1}(t) = \begin{cases} t^3 & \text{if } t > \delta, \\ 3\delta^2 \left(t - \frac{4}{29}\right) & \text{if } t \leq \delta. \end{cases} \tag{2.7}$$

The tri-stimulus values are obtained as:

$$X = X_n \, f^{-1}(f_X), \qquad Y = Y_n \, f^{-1}(f_Y), \qquad Z = Z_n \, f^{-1}(f_Z). \tag{2.8}$$

From CIEXYZ to linear RGB. For the sRGB space, the linear RGB components are computed via the following matrix multiplication:

$$\begin{pmatrix} R_{\text{lin}} \\ G_{\text{lin}} \\ B_{\text{lin}} \end{pmatrix} = \begin{pmatrix} 3.2406 & -1.5372 & -0.4986 \\ -0.9689 & 1.8758 & 0.0415 \\ 0.0557 & -0.2040 & 1.0570 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}. \tag{2.9}$$

Gamma correction (linear RGB to sRGB). Each linear component $V_{\text{lin}} \in \{R_{\text{lin}}, G_{\text{lin}}, B_{\text{lin}}\}$ is transformed by the sRGB gamma function to obtain the final component V:

$$V = \begin{cases} 12.92\, V_{\text{lin}} & \text{if } V_{\text{lin}} \leq 0.0031308, \\ 1.055\, V_{\text{lin}}^{1/2.4} - 0.055 & \text{if } V_{\text{lin}} > 0.0031308. \end{cases} \tag{2.10}$$

The resulting values are clipped to the range $[0, 1]$ to produce the sRGB triplet (R, G, B).

Inverse transformation. The conversion from RGB back to CIELAB follows the inverse sequence of these transformations.

2.2. Image Colorization

2.2.1. Inverse problems

Inverse problems arise when one seeks to recover an unknown quantity x from indirect or incomplete observations y, obtained through a forward operator $\mathcal{A}$:

$$y = \mathcal{A}(x) + \eta, \tag{2.11}$$

where η models noise or discrepancies between the model and the physical process.

An inverse problem is considered *well-posed* if a solution exists, is unique, and depends continuously on the data. When any of these conditions fail, the problem becomes *ill-posed*, and meaningful solutions require the introduction of prior knowledge or regularization (Bertero & Boccacci, 1998). A classical stabilized formulation is given by:

$$\hat{x} = \min_{x} \left\{ \|\mathcal{A}(x) - y\|^2 + \lambda\, R(x) \right\}, \tag{2.12}$$

where $R(x)$ encodes prior assumptions and $\lambda > 0$ controls their influence.

2.2.1.1. Colorization as an inverse problem

Image colorization fits squarely into this framework. Here, the unknown x corresponds to the full color image u, while the observation y corresponds to its luminance component l. Consequently, the forward operator $\mathcal{A}$ reduces to a deterministic luminance-extraction operator, denoted as K.

While converting a color image to grayscale ($l = K(u)$) is a straightforward problem, the inverse task—recovering plausible chromatic information from l—is highly underdetermined. Since multiple distinct colors map to the same luminance value, infinite solutions exist.

To address this ambiguity, colorization is formulated as a regularized minimization problem that balances two complementary goals:

$$\min_{u} \left\{ \underbrace{\| \, l - K(u) \, \|^2}_{\text{Data fidelity}} + \underbrace{\lambda \, R(u)}_{\text{Regularization / Prior}} \right\}. \tag{2.13}$$

Data fidelity: Ensures the estimated color image $\hat{u}$ is consistent with the input; i.e., applying K to $\hat{u}$ must yield the observed grayscale l.

Regularization (Prior): Constrains the search space by enforcing assumptions about natural images. These priors may range from classical spatial smoothness to complex statistical regularities learned by deep neural networks.

2.2.2. Classification of colorization approaches

To solve this ill-posed problem, literature proposes various methods distinguished primarily by the nature of the prior $R(u)$ and the guidance mechanism.

Classical methods (optimization-based). hese approaches formulate colorization as a variational problem using explicit mathematical priors, typically assuming spatial smoothness in the chrominance channels. A representative example is the Total Variation (TV) based colorization proposed by Kang and March (2007), which adapts the classical restoration model of Rudin et al. (1992):

$$\min_{u} \; \|K(u) - l\|^2 + \lambda \int |\nabla u| \, dx, \tag{2.14}$$

which encourages piecewise-smooth chromatic fields but often fails to capture high-level semantic textures.

Classical guided methods (scribble-based). To reduce ambiguity, these methods utilize sparse user-provided color strokes ("scribbles") as constraints to guide chrominance propagation. Levin et al. (2004) proposed minimizing a weighted quadratic energy:

$$J(u) = \sum_{p,q} w_{pq} \, \|u(p) - u(q)\|^2, \tag{2.15}$$

subject to hard constraints on the user-labeled set $\mathcal{S}$. Here, w_{pq} represents the affinity between neighboring pixels p and q based on luminance similarity.

Classical guided methods (exemplar-based). Instead of relying on manual strokes, these approaches transfer chromatic information from a reference color image (the exemplar) that is semantically similar to the target grayscale image. The pioneer work by Welsh et al. (2002) assumes that pixels with similar luminance and texture neighborhoods share similar colors.

The process typically involves matching each pixel p in the target grayscale image l to a pixel q in the reference image r by minimizing a distance metric in a feature space $\mathcal{F}$:

$$q^* = \arg\min_{q \in r} \left\| \mathcal{F}\big(l(p)\big) - \mathcal{F}\big(r(q)\big) \right\|, \tag{2.16}$$

where $\mathcal{F}$ usually consists of luminance intensity and local variance statistics. Once the best match is found, the chromatic coordinates (a, b) of q are transferred to p. While effective for simple textures, these methods often fail when the target and reference possess distinct semantic structures.

Deep learning methods (automatic). Modern approaches replace explicit, handcrafted priors with statistical priors learned from large-scale datasets in a supervised manner. A neural network G_θ learns a mapping from the luminance input l to a predicted color image u by minimizing a loss function over a training set of pairs (l, u_{real}). Typically, this involves minimizing the expected error:

$$\min_{\theta} \; \mathbb{E}_{(l,u_{\text{real}})}[\mathcal{L}\left(G_\theta(l),\, u_{\text{real}}\right)]\,, \tag{2.17}$$

where u_{real} denotes the ground-truth chrominance associated with l. Seminal work in this category involves end-to-end Convolutional Neural Networks (CNNs), such as the model proposed by Iizuka et al. (2016), which originally utilized standard regression losses (e.g., L_2 distance).

However, methods relying on unimodal regression losses tend to generate desaturated (sepia-toned) results because the network minimizes error by predicting the statistical average of all plausible colors. To address this, different strategies have emerged. GAN-based approaches frame colorization as an image-to-image translation problem, encouraging realism through adversarial losses (Isola et al., 2017; Kim et al., 2022; Vitoria et al., 2020). Methods that predict distributions model the inherent ambiguity of colorization by predicting color histograms or classes instead of a single regression target (Larsson et al., 2016; R. Zhang et al., 2016). Multimodal colorization produces multiple plausible outputs (Deshpande et al., 2017; Kumar et al., 2021; Royer et al., 2017). And, object/instance-aware methods incorporate object-level structure to reduce color bleeding and improve semantics (Pucci et al., 2021; Su et al., 2020).

Hybrid methods (deep learning with guidance). Some deep learning-based approaches combine two types of color priors: those learned from a large dataset and those provided by external users as color hints (Yun et al., 2023; R. Zhang et al., 2017) or as a color reference image (He et al., 2018). Thus, these methods combine the semantic understanding of deep networks with user guidance.

Exemplar-based deep learning. A reference image provides the chromatic style and the network utilize feature matching or attention mechanisms to transfer color from the exemplar r to the target l:

$$u = G_\theta(l, r). \tag{2.18}$$

He et al. (2018) leverage this paradigm by aligning deep semantic features to transfer colors even when the reference and target have different structures Bai et al. (2021) utilizes a sparse attention mechanism to select only the most semantically relevant regions from the reference image, ignoring background noise and improving the accuracy of color transfer in complex scenes.

Scribble-guided deep learning. The model accepts user color scribbles as an additional input channel. To enforce fidelity to the user inputs while maintaining robustness, methods like R. Zhang et al. (2017) minimize a Huber loss (Smooth L_1) over the predicted image:

$$\mathcal{L}(\theta) = \sum_{p \in \Omega} \mathcal{L}_\delta \left(G_\theta(l, u_{\mathcal{S}})(p) - u_{\text{real}}(p) \right), \tag{2.19}$$

where $\mathcal{L}_\delta$ is the Huber loss, which behaves quadratically for small errors and linearly for large errors.

More recently, diffusion models are being used more and more for high-fidelity, controllable colorization within a general image-to-image generative framework (Liang et al., 2025; Saharia et al., 2022).

2.3. Deep Learning Models

Deep learning has reshaped computer vision by enabling the acquisition of hierarchical representations directly from data. Several architectural families have played influential roles in this evolution. Convolutional Neural Networks (CNNs) have long underpinned recognition tasks; Generative Adversarial Networks (GANs) introduced adversarial strategies for synthesis and restoration; transformer-based models introduced global attention mechanisms; and diffusion-based models have recently achieved state-of-the-art performance in generation (Goodfellow et al., 2014; Ho et al., 2020; LeCun et al., 2015; Vaswani et al., 2017).

This section provides an overview of the architectures most relevant to this thesis, with a particular emphasis on transformer-based vision models.

2.3.1. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) (LeCun et al., 1998) constitute a foundational architecture in computer vision. Their defining operation is the convolution: a learnable filter applied locally across spatial neighborhoods to extract hierarchical features. Early layers typically detect low-level cues, such as edges or textures, while deeper layers capture increasingly abstract semantic structures (Zeiler & Fergus, 2014).

CNNs incorporate key inductive biases—locality, weight sharing, and translation equivariance—that render them computationally efficient and data-effective. For years, they represented the dominant paradigm for image classification, retrieval, and segmentation (Krizhevsky et al., 2012).

2.3.1.1. VGG16

Among CNN architectures, VGG16 (Simonyan & Zisserman, 2015) serves as an example of a deep network constructed from uniform building blocks. As can be seen in Figure 2.2, the architecture features a stack of 13 convolutional layers organized into five blocks, using exclusively 3×3 filters, followed by max-pooling operations and fully connected layers. This design strategy—stacking small kernels to increase depth—proved highly effective, offering strong representational capacity while preserving architectural simplicity.

Due to its high-quality feature hierarchy, VGG16 is widely deployed as a feature extractor (Sharif Razavian et al., 2014; Yosinski et al., 2014). When the fully connected classification head is removed, the remaining convolutional backbone produces high-level embeddings that encode semantic structure. These feature vectors can be compared using metrics such as cosine similarity and Euclidean distance, enabling tasks such as image retrieval and clustering.

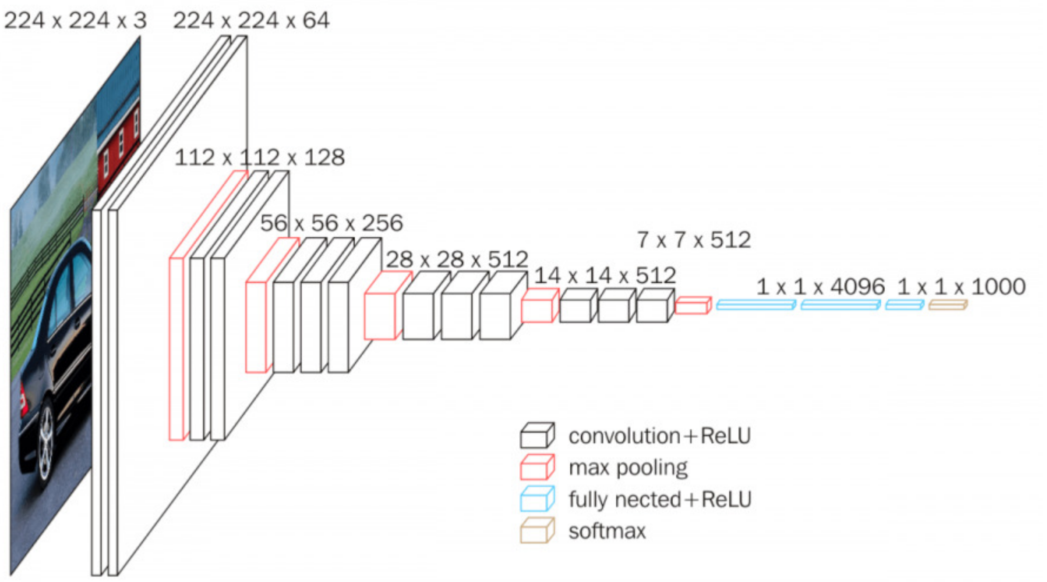


Figure 2.2: General schematic of the VGG16 architecture. Image from (Kamal & Ez-Zahraouy, 2023)

2.3.2. Transformer Architecture

Transformers were originally introduced for sequence modeling in natural language processing by Vaswani et al. (2017). Their core innovation is the *self-attention* mechanism, which enables each token to aggregate information from the entire input sequence simultaneously, independent of proximity. The following equations describe the standard Transformer architecture. The conceptual synthesis of these components draws on Thickstun (2022).

Self-Attention Mechanism Given an input sequence of token embeddings $X \in \mathbb{R}^{N \times d}$, the model computes Queries (Q), Keys (K), and Values (V) via learned linear projections:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{2.20}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are the weight matrices.

The scaled dot-product attention, $\boldsymbol{\alpha}$, is defined as:

$$\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{2.21}$$

This operation allows each token to attend to all others, weighting them according to learned relevance scores. The output of the attention layer (Figure 2.3, left) is then obtained by applying these attention weights to the value vectors:

$$\mathrm{Attention}(Q, K, V) = \boldsymbol{\alpha}V. \tag{2.22}$$

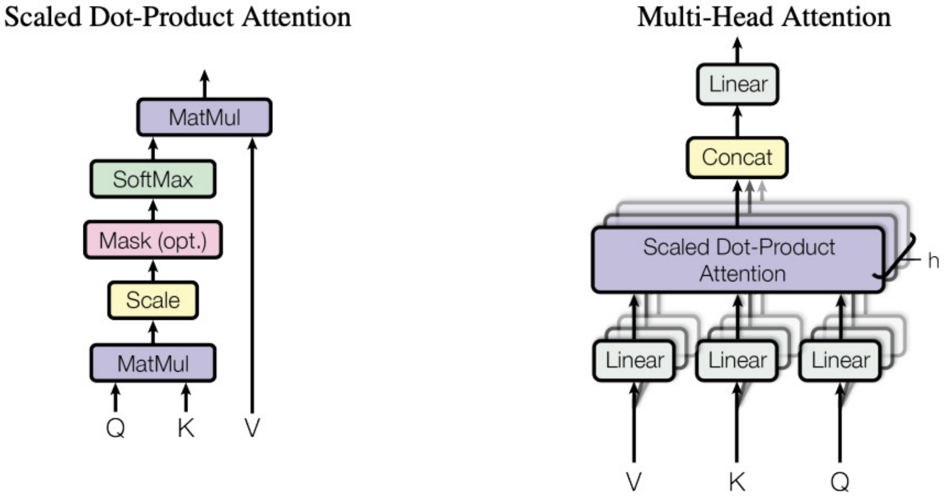


Figure 2.3: Schematic illustration of the scaled dot-product attention mechanism and Multi Head attention. Image from (Vaswani et al., 2017)

Multi-Head Attention Instead of computing a single attention map, Transformers employ multiple attention heads that operate in parallel. Each head attends to the input sequence through an independent linear projection, allowing the model to capture different types of relationship simultaneously.

Formally, given an input representation $X \in \mathbb{R}^{N\times d}$, multi-head attention is defined as:

$$\mathrm{MSA}(X) = \mathrm{Concat}(H_1, \ldots, H_h)W_O, \tag{2.23}$$

where h denotes the number of attention heads. Each head output $H_i \in \mathbb{R}^{N \times d_k}$ is computed independently using scaled dot-product attention, with $d_k = d/h$.

The $\mathrm{Concat}(\cdot)$ operation concatenates the head outputs along the feature dimension, producing a matrix of size $\mathbb{R}^{N \times (h \cdot d_k)} = \mathbb{R}^{N \times d}$. Finally, the learned output projection $W_O \in \mathbb{R}^{d \times d}$ mixes information across heads and restores the original embedding dimensionality. This can be seen in Figure 2.3, on the right.

Positional Encodings Because standard self-attention is permutation-invariant, the model requires explicit positional information to interpret sequence order. This is achieved by injecting positional encodings into the input embeddings. These encodings may be fixed sinusoidal functions (Vaswani et al., 2017) or learnable parameters, depending on the implementation.

Transformer Encoder Block A Transformer encoder block integrates the multi-head attention mechanism with normalization layers and Multi-Layer Perceptron (MLP) blocks.Following the original architecture proposed by Vaswani et al. (2017), the block employs a residual connection around each of the two sub-layers, followed by layer normalization (NL).

Given an input X, the output of the attention sub-layer, denoted as Z_{attn}, is computed as:

$$Z_{attn} = \mathrm{NL}(X + \mathrm{MSA}(X)). \tag{2.24}$$

Subsequently, the MLP processes this representation. The MLP consists of two linear transformations with a non-linear activation function in between. Formally:

$$\mathrm{MLP}(Z_{attn}) = \sigma(Z_{attn} W_1 + b_1) W_2 + b_2, \tag{2.25}$$

where $W_1 \in \mathbb{R}^{d \times d_{ff}}$ and $W_2 \in \mathbb{R}^{d_{ff} \times d}$ are the learned weight matrices, b_1 and b_2 are the bias vectors, and $\sigma(\cdot)$ is a non-linear activation function (typically ReLU or GELU). The inner dimension d_{ff} is usually larger than the model dimension d (e.g., $d_{ff} = 4d$).

The final output of the encoder block, Z_{out}, incorporates the residual connection for the MLP:

$$Z_{out} = \mathrm{NL}(Z_{attn} + \mathrm{MLP}(Z_{attn})). \tag{2.26}$$

Trainable Parameters The set of learnable parameters for a single Transformer encoder block, denoted by Θ, consists of the projection matrices of the attention mechanism and the parameters of the feed-forward network (MLP) and NL. Specifically:

$$\Theta = \left\{ \begin{array}{ll} \{W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}\}_{h=1}^{H}, W_O, & \text{(Attention weights)} \\ W_1, b_1, W_2, b_2, & \text{(MLP weights and biases)} \\ \gamma_1, \beta_1, \gamma_2, \beta_2 & \text{(NL scale and shift)} \end{array} \right\} \tag{2.27}$$

where γ and β represent the learnable gain and bias parameters of the NL (Thickstun, 2022).

2.4. Vision Transformer (ViT)

The Vision Transformer (ViT) (Dosovitskiy et al., 2021) represents a paradigm shift in computer vision, successfully adapting the pure Transformer architecture—originally designed for sequence modeling in NLP—to image analysis. Unlike Convolutional Neural Networks (CNNs), which process visual information via local receptive fields, ViT treats an image as a sequence of discrete patches, enabling global context modeling from the earliest processing stages.

As shown in Figure 2.4, the Transformer encoder follows a Pre-Norm design in which Layer Normalization (NL) is applied before both the multi-head self-attention (MSA) and the MLP (feed-forward) sub-blocks, and residual connections are added afterward. This configuration improves training stability and contrasts with the original Transformer architecture proposed by Vaswani et al. (2017), which employed a Post-Norm scheme (applying normalization after residual addition). Finally, regarding the learnable parameters, the ViT encoder retains the same internal parameter set Θ for each of its L blocks as defined in Section 2.3.2 (Attention

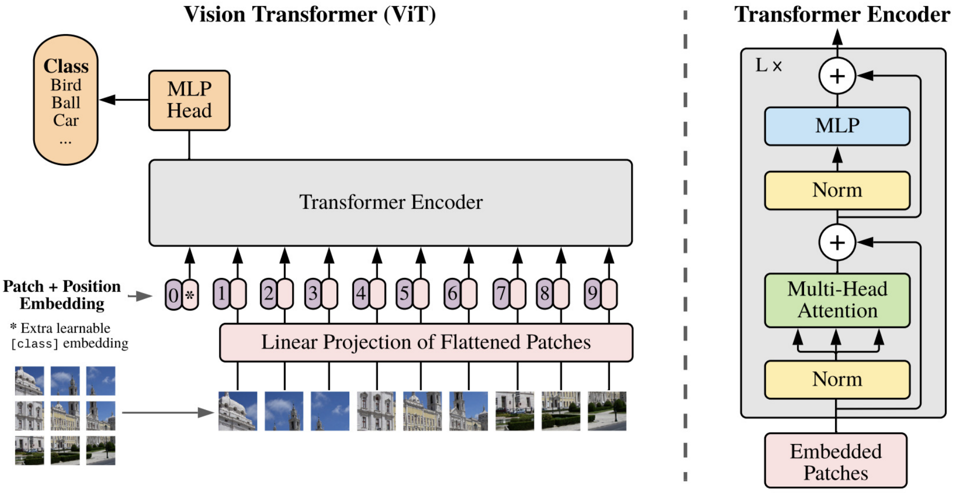


Figure 2.4: General structure of the Vision Transformer (ViT). The image is divided into patches, embedded into tokens, and processed through self-attention layers to capture global dependencies. Image from (Dosovitskiy et al., 2021).

weights W_Q, W_K, W_V, W_O, MLP weights, and NL parameters). However, adapting the architecture to visual data introduces two additional sets of parameters at the input stage: the linear projection matrix $E \in \mathbb{R}^{(P^2 \cdot C) \times d}$ used to embed the flattened patches, and the position embeddings $E_{pos} \in \mathbb{R}^{N \times d}$, which are learned during training to preserve spatial information (Dosovitskiy et al., 2021).

2.4.1. Architecture and Input Processing

To apply the Transformer architecture to 2D images, ViT reshapes the input image $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 C)}$, where $P \times P$ is the patch size and $N = (HW)/P^2$ is the resulting sequence length. Each patch is then linearly projected to a fixed latent dimension d via a trainable embedding matrix. Since the standard self-attention mechanism is permutation-invariant, learnable *positional embeddings* are added to these patch embeddings to preserve spatial information regarding the relative positions of the patches (Dosovitskiy et al., 2021).

A critical component of this input formulation, adopted from BERT (Devlin et al., 2019), is the inclusion of a specialized learnable classification token (`[class]`) prepended to the patch sequence. Intuitively, this token serves as a global information aggregator. While standard patch tokens update their representations based on their relationship with other patches, the `[class]` token interacts with the entire sequence across all layers without corresponding to a specific image region. By the final Transformer layer, its state serves as a compact, global feature vector representation of the image, which is then used by the Multi-Layer Perceptron (MLP) head for downstream tasks such as classification or colorization guidance.

2.4.2. Inductive Bias and Model Scaling

The operational principle of ViT introduces a fundamental difference in *inductive bias*—the structural assumptions a model makes about data—compared to CNNs. CNNs are designed with strong biases towards *locality* (pixels are locally correlated) and *translation equivariance* (features are recognizable regardless of position). In contrast, ViT has minimal image-specific bias; only its MLP layers are local, while the self-attention mechanism is inherently global.

This lack of hard-coded spatial assumptions makes ViT significantly more flexible but also more "data-hungry," as it must learn spatial relationships entirely from raw data. Empirical studies indicate that ViT models typically underperform on mid-sized datasets such as ImageNet (approx. 1.3 million images) due to overfitting. When pre-trained on massive datasets like JFT-300M (300M images), ViT learns robust representations that outperform state-of-the-art convolutional architectures on downstream tasks while requiring fewer computational resources to train (Dosovitskiy et al., 2021). Analysis suggests that for smaller datasets, the convolutional bias is crucial, but for sufficiently large datasets, learning relevant patterns directly from raw data is superior, yielding robust, scalable representations (Dosovitskiy et al., 2021; Steiner et al., 2022).

2.5. Training neural networks

While deep learning architectures define a model's structural capacity, its performance is ultimately determined by the optimization trajectory through the loss landscape. This section outlines the training protocols, optimization algorithms, and hyperparameter strategies employed to govern this process.

2.5.1. Data Partitioning and Model Selection

To ensure rigorous evaluation and prevent data leakage, the dataset is stratified into three distinct subsets:

- **Training Set:** Used to compute gradients and update model parameters (θ).
- **Validation Set:** Serves as a proxy for generalization performance during training. Crucially, this split guides *model selection* and hyperparameter tuning—such as the learning rate—serving as a compass to prevent overfitting to the training data.
- **Test Set:** Reserved exclusively for the final evaluation, providing an unbiased estimate of the model's performance on unseen data.

2.5.2. Optimization Algorithm

The fundamental goal of the training process is to find the optimal set of parameters θ^* that minimizes a scalar objective function $\mathcal{L}(\theta)$. This is achieved via iterative gradient-based optimization.

For Transformer-based architectures, standard Stochastic Gradient Descent (SGD) is often replaced by adaptive methods. Notable among these is the Adam optimizer, which attempts to approximate the diagonal of the inverse Hessian to achieve adaptive learning rates for each parameter (Loshchilov & Hutter, 2017). In this thesis, however, the AdamW optimizer (Loshchilov & Hutter, 2019) is employed. This modification ensures that regularization is applied more effectively, resulting in superior generalization and training stability for deep models like ViT compared to traditional adaptive optimizers.

2.5.3. Learning Rate Scheduling

Within this optimization framework, the learning rate (lr) is the pivotal hyperparameter that governs the step size of parameter updates. It directly dictates convergence behavior: an excessively high lr risks instability or divergence, while a strictly low lr may result in slow convergence or entrapment in suboptimal minima. To address this trade-off, *learning rate schedules* dynamically anneal the lr, typically initializing with a larger value to accelerate exploration and gradually decreasing it to refine convergence.

2.5.3.1. Cosine Learning Rate Schedule

The cosine schedule (Loshchilov & Hutter, 2017) is a prevalent strategy that decays the learning rate smoothly following a cosine function rather than abrupt steps. This gradual reduction typically supports stable convergence and can improve generalization. The learning rate η_t at training step t is defined as:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\frac{T_{cur}}{T_{cycle}}\pi\right)\right],$$

where $\eta_{\max}$ and $\eta_{\min}$ are the maximum and minimum learning rates in a cosine cycle, T_{cur} is the number of optimization steps elapsed since the most recent restart, and T_{cycle} is the total number of steps in that cycle. Consequently, at the beginning of each cycle ($T_{cur} = 0$), the learning rate resets to $\eta_t = \eta_{\max}$, promoting exploration of the loss landscape. As $T_{cur} \to T_{cycle}$, η_t approaches $\eta_{\min}$, ensuring precise updates. In the warm-restart configuration, a restart is triggered when $T_{cur} = T_{cycle}$, after which T_{cur} resets to 0 and a new cosine cycle begins. However, this approach depends on a pre-defined cycle length (T_{cycle}); if restarts are delayed, disabled, or set too infrequently, causing the learning rate to remain near $\eta_{\min}$ for an extended period, gradient norms can become very small, and learning may stagnate (Defazio et al., 2024).

2.5.3.2. Schedule-Free Optimization

Hand-designed learning-rate schedules can be effective, but many require specifying a target training horizon T (or cycle length) in advance. Defazio et al. (2024) proposed *Schedule-Free Optimization* to remove this dependency by replacing explicit learning-rate decay with an iterate-averaging mechanism used together with a constant step size. The method maintains a *base* parameter sequence z_t and its running (Polyak–Ruppert) average x_t. A schedule-free momentum parameter β interpolates between these two sequences to define the point where gradients are evaluated:

$$y_t = (1-\beta)z_t + \beta x_t.$$

The base iterate is then updated using gradients at y_t,

$$z_{t+1} = z_t - \gamma \nabla f(y_t),$$

and the averaged sequence is updated online,

$$x_{t+1} = \left(1 - \frac{1}{t+1}\right) x_t + \frac{1}{t+1} z_{t+1}.$$

This construction yields behavior comparable to a linear decay schedule whose effective horizon is the current step t, rather than a fixed, pre-specified T, enabling *any-time* training without committing to a total number of steps in advance. Empirically, the authors show that schedule-free variants of SGD and AdamW match or exceed cosine decay on vision and language benchmarks (including Transformer models), and avoid the stagnation that can arise when schedule-based methods are poorly matched to the training horizon (Defazio et al., 2024).

2.6. Transfer Learning

Transfer learning is a machine learning paradigm that leverages representations learned from a source task to improve generalization on a related target task. Formally, given a source domain and task for which abundant data is available, the goal is to improve the learning of a predictive function in a target domain where data may be scarce (Bishop & Bishop, 2023; Pan & Yang, 2009).

This approach relies on the hierarchical nature of deep neural networks. As noted by Bishop and Bishop (2023), the early layers of a network trained on large-scale datasets (e.g., ImageNet) learn low-level features—such as edges, textures, and color gradients—that are general and transferable across different visual tasks. Conversely, deeper layers encode increasingly abstract and semantic representations specific to the original training classes. By initializing a model with these pretrained weights, the optimization process in the target domain starts from a robust internal representation, significantly reducing the computational cost and the risk of overfitting compared to training from scratch (Bishop & Bishop, 2023; Yosinski et al., 2014).

In the context of artistic image restoration, transfer learning allows the model to inherit "priors" about natural image statistics from massive datasets, adapting them to the specific stylistic nuances of the target artistic domain.

Three primary adaptation strategies are typically employed:

Feature Extraction (Frozen Backbone): The parameters of the pretrained network are treated as a fixed feature extractor. As described in Bishop and Bishop (2023), the input data is passed through the frozen layers to generate embeddings, and only the final task-specific layers (the "head") are trained. This is computationally efficient but limits the model's ability to adapt to domains that differ significantly from the source.

Full Fine-Tuning: The entire network is used as an initialization, and all parameters are updated during training on the target dataset. This corresponds to the standard definition of *fine-tuning* in Bishop and Bishop (2023), typically performed with a very small learning rate to preserve the pretrained knowledge while adapting the representations to the new task. While this offers maximum flexibility, it is computationally expensive and prone to overfitting if the target dataset is small.

Parameter-Efficient Fine-Tuning (PEFT) (Xu et al., 2023): This family of methods represents a middle ground, enabling the adaptation of large models without updating all parameters. Techniques such as adapter layers (Houlsby et al., 2019) or Low-Rank Adaptation (LoRA) (Hu et al., 2022) inject a small number of trainable parameters into the frozen architecture. This strategy substantially reduces memory requirements while preserving most of the model's original representational capacity.

2.6.1. Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) (Hu et al., 2022) is a popular Parameter-Efficient Fine-Tuning (PEFT) technique for adapting large pretrained models to specific downstream tasks. Rather than updating all the parameters of the dense layers, LoRA freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, typically targeting the Query W_Q and Value W_V projection matrices in the self-attention mechanism. See Figure 2.5.

Following the notation in Hu et al. (2022), let $W_0 \in \mathbb{R}^{d \times k}$ denote a frozen weight matrix from the pretrained backbone. During full fine-tuning, the model learns a weight update ΔW such that the final weight is $W = W_0 + \Delta W$.

LoRA constrains this update by hypothesizing that the change in weights possesses a low "intrinsic rank." Consequently, the update ΔW is decomposed into the product of two low-rank matrices, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where the rank r satisfies $r \ll \min(d, k)$. The forward pass is thus modified as:

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{2.28}$$

During training, W_0 is frozen, and only A and B are optimized. To ensure training stability, A is typically initialized with random Gaussian noise, while B is initialized to zero, ensuring that $\Delta W = 0$ at the beginning of adaptation.

Why the Low-Rank Hypothesis Holds. The effectiveness of LoRA is grounded in the concept of *intrinsic dimension*. Inspired by the work of Aghajanyan et al. (2021), who demonstrated that over-parameterized models reside on a low intrinsic dimension, Hu et al. (2022) hypothesized that the change in weights during adaptation also possesses a low "intrinsic rank". Consequently, LoRA operationalizes this insight by optimizing the update ΔW within a low-rank manifold. Empirical evidence supports this hypothesis: Hu et al. (2022) shows that a very low rank (e.g., (r=1) or (r=2)) is often sufficient to recover the performance of full fine-tuning, as the update matrix amplifies specific features already present in the pretrained model rather than learning entirely new ones

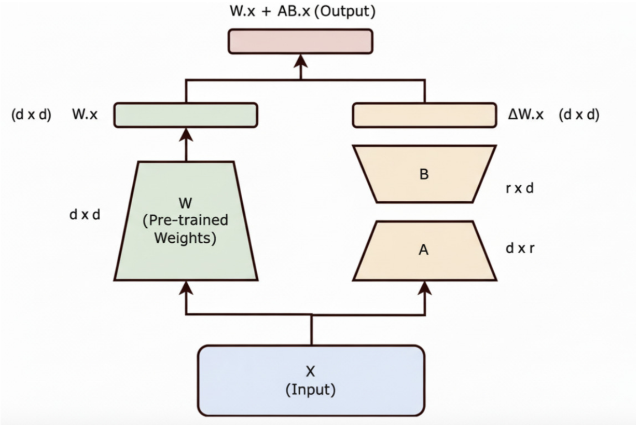


Figure 2.5: Illustration of the Low-Rank Adaptation (LoRA) mechanism (Jawade, 2023). The pretrained weights W_0 remain frozen, while a low-rank decomposition BA (representing the update ΔW) is trained.

2.6.2. Efficiency and Comparison with Adapters

To understand the efficiency gains of LoRA, it is necessary to first contextualize it against previous approaches. Prior to LoRA, adapter-based methods (Houlsby et al., 2019; Pfeiffer et al., 2020) were the strategy for parameter-efficient fine-tuning. These methods operate by inserting small "bottleneck" feed-forward layers sequentially between the Transformer's self-attention and MLP sub-layers.

While adapters successfully reduce the number of trainable parameters, they fundamentally alter the model architecture by extending the computational graph. As noted by Hu et al. (2022), this sequential addition introduces measurable inference latency, as the data must pass through these extra layers during the forward pass. This overhead can be problematic for real-time applications where every millisecond counts.

LoRA addresses this limitation by applying the low-rank updates in parallel rather than sequentially. This design choice leads to several critical advantages:

No Inference Latency: Unlike adapters, LoRA introduces zero overhead during inference. After training, the learned low-rank matrices A and B can be explicitly multiplied and added to the frozen weights W_0 (i.e., $W_{new} = W_0 + B \cdot A$). This effectively "merges" the adaptation into the original backbone, restoring the original architectural structure and ensuring that inference speed remains identical to the pretrained model.

Parameter Efficiency: By constraining the update rank r, LoRA reduces the number of trainable parameters by up to 10,000 times compared to full fine-tuning (Hu et al., 2022). This significantly reduces GPU memory requirements, enabling fine-tuning of massive models on consumer-grade hardware.

Storage Efficiency and Modularity: Since only the small matrices A and B need to be saved (often just a few megabytes), it is possible to store multiple task-specific adaptations for a single frozen backbone, switching between them easily without reloading the massive base model.

Chapter 3

Data

This chapter describes the data used in this work, including the identification and organization of the different data types and their sources.

The core corpus consists of images of artworks by Joaquín Torres García and by the artists of the Taller Torres García. These images were gathered from multiple archives, institutions, and digitization processes. Due to this heterogeneity, considerable effort was required to detect duplicates, normalize formats, and establish consistent naming conventions across the various sources.

To ensure consistency and comparability of results, the chapter also describes the hierarchical design of the datasets used for model training, validation, and testing, along with the methodology for both automated and manual image curation.

To build those datasets, collaboration agreements were established with several institutions and individuals connected to the artistic legacy of Joaquín Torres García, including Cecilia de Torres, Fundación Gurvich, and the Museo Torres García. These images are not publicly available, and their use was authorized exclusively for the research purposes of this work. For this reason, the datasets cannot be publicly released.

3.1. Data Sources

3.1.1. The lost artworks

The devastating fire that occurred in 1978 at the Museum of Modern Art of Rio de Janeiro (hereafter, MAM Rio) destroyed an exhibition that had originated four years earlier in Montevideo. This exhibition was a major retrospective of Joaquín Torres García (hereafter JTG), organized in 1974 at the Museo Nacional de Artes Visuales (hereafter MNAV) to commemorate the centenary of the artist's birth (Giner, 2003). It brought together a large and representative selection of his works, including the seven monumental murals JTG painted in 1944 at the Hospital Saint Bois in Montevideo. In total, the Taller Torres García (hereafter TTG) created thirty-five murals for the hospital: seven executed by JTG himself and the remaining twenty-eight by his students. The seven murals were carefully detached from the hospital walls in 1973 and later exhibited at the MNAV as part of the centennial retrospective (Giner, 2003). After its presentation in Montevideo, the exhibition traveled to the Musée d'Art Moderne de la Ville de Paris (hereafter MAM Paris) in 1975, before being sent to the MAM Río, where it was completely destroyed in the 1978 fire (Giner, 2003). The remaining Hospital Saint Bois murals produced by the TTG were transferred in 1997 to the Torre de las Telecomunicaciones in Montevideo, where they are preserved today.

A fundamental part of this work involved identifying the lost artworks, along with their existing records and any reference images, in order to generate potential color references. Special thanks are due to Alejandro Díaz and Carlos Serra, from the Museo Torres García (hereafter MTG), for their essential contribution in compiling and providing the inventory of the artworks destroyed in the 1978 fire. This inventory, which includes images and metadata such as title, year, technique, dimensions, and catalog reference—based on the *Catalogue Raisonné* («Joaquín Torres García Catalogue Raisonné», 2003)—forms the starting point of this work. Table 3.1 summarizes the seventy-three lost works documented by the MTG.

Title	Year	Technique / Material	Dimensions (cm)	Catalog
The Sun	1944	Enamel paint on wall varnish	192.5 × 662	1944.26
The Fish	1944	Enamel paint on wall varnish	189 × 285.5	1944.28
White Locomotive	1944	Enamel paint on wall varnish	103 × 129.5	1944.31
Pacha Mama	1944	Enamel paint on wall varnish	87 × 280	1944.12
The Tram	1944	Enamel paint on wall varnish	189.5 × 657	1944.30
Pax in Lucem	1944	Enamel paint on wall varnish	110 × 427	1944.27
Form	1944	Enamel paint on wall varnish	122 × 193	1944.29
Head of a Man Reading a Newspaper	1924	Polychrome wood	18 × 18	1921.14
Seated Man	1927	Painted wood	14 × 7 × 7	1927.86
Abstract Man	1929	Polychrome wood	27 × 10 × 5	1929.77
Sculptural Structure	1929	Polychrome wood	17 × 8 × 7	1929.84
Red Man	1929	Polychrome wood	16 × 7 × 4	1929.98
Superposed Aerial Structure	1929	Polychrome wood	32 × 25 × 5	1929.02
Composition in Pink	1929	Oil on canvas	65 × 54	1929.55
Constructive Painting	1929	Oil on canvas	60 × 73	1929.18
Mask with Teeth	1928	Polychrome wood	34 × 14 × 4	1928.206
Constructive	1930	Oil on cardboard	48 × 31	1930.48
Universal Composition	1930	Oil on wood	43.5 × 41	1930.57
Constructive	1931	Oil on canvas	45 × 38	1931.70

Title	Year	Technique / Material	Dimensions (cm)	Catalog
Constructive Composition	1931	Oil on cardboard	100 × 81	1931.69
Primitive Constructive in Red	1931	Oil on wood	95 × 43	1931.98
Homage to Van Rees	1931	Oil on cardboard	45 × 35	1931.34
Constructive in White and Pink	1932	Oil on wood	62 × 31	1932.80
Constructive Painting	1931	Oil on canvas	128 × 89	1931.43
Constructive in White	1931	Oil on canvas	73 × 60	1931.59
TN5	1931	Oil on canvas	45 × 35	1931.09
Relief Structure	1932	Incised and polychrome wood (multi-plane)	50 × 30 × 7	1932.91
Constructive with Anchor	1932	Painted wood relief	41 × 26 × 6	1932.82
Constructive	1932	Oil on canvas	73 × 60	1932.25
Constructive Painting	1932	Oil on canvas	53 × 39	1932.41
Constructive Composition	1932	Oil on wood	58 × 41	1932.55
Constructive Painting	1932	Oil on canvas	54 × 45	1932.43
Graphic on Wood	1932	Oil on wood	65 × 41	1932.47
Constructive Composition	1932	Oil on cardboard	45 × 35	1932.46
Model of Constructive Monument	1932	Incised wood	20 × 14 × 10	1932.81
Small Monument with Universal Man	1932	Incised wood	15 × 10 × 10	1932.77

Title	Year	Technique / Material	Dimensions (cm)	Catalog
Constructive with Mask and Triangle	1932	Oil on canvas	65 × 53	1932.48
Symbolic Constructive	1932	Oil on canvas	70 × 53	1932.61
Constructive Graphic with Brushstroke Background	1932	Oil on canvas	100 × 81	1932.26
Graphic in Black and White	1932	Oil on canvas	54 × 72	1932.42
Colored Tubular Structure	1937	Tempera on cardboard	85 × 52	1937.16
Structure	1933	Oil on cardboard	45 × 32	1933.12
Constructive Painting	1937	Oil on cardboard	80 × 100	1937.33
Abstract Man	1938	Painted wood	51 × 15	1938.41
Structure	1937	Oil on cardboard	45 × 32	1937.34
Constructive Painting	1937	Oil on cardboard	60 × 100	1937.39
Painting	1937	Oil on cardboard	106 × 86	1937.40
Seven-Pointed Star	1933	Incised wood	20 × 10	1933.36
Constructive Composition	1938	Tempera on cardboard	101 × 82	1938.34
Primitive Composition	1939	Oil on canvas	38 × 40	1939.25
Figurative Relief Structure	1938	Oil on canvas	39 × 46	1938.35
Small Constructive Monument	1942	Painted wood	14 × 15 × 5	1942.39
Constructive Painting	1942	Oil on canvas	104 × 76	1942.45
Constructive Art	1942	Oil on cardboard	58 × 63	1942.31

Title	Year	Technique / Material	Dimensions (cm)	Catalog
Constructive Art with Double Colored Line	1942	Oil on cardboard	51 × 59	1942.27
Structure in Gray and Ochre	1942	Oil on cardboard	91 × 55	1942.22
Infinity	1942	Oil on canvas	76 × 55	1942.19
Constructive Uruguay	1943	Oil on cardboard	50 × 70	1943.61
Composition with Writing	1942	Oil on lined cardboard	80 × 100	1942.25
Constructive Graphic	1943	Oil on lined cardboard	55 × 39	1943.74
Constructive Art	1943	Oil on cardboard	72 × 54	1943.94
Constructive Painting	1943	Oil on cardboard	66 × 52	1943.89
Constructive Painting	1943	Oil on cardboard	66 × 52	1943.96
Painting	1943	Oil on cardboard	52 × 68	1943.65
Constructive Port	1943	Oil on cardboard	55 × 65	1943.100
Two Birds	—	Painted wood	4 pieces	T3.122
A Woman, a Man and a Dog	—	Painted wood	8 pieces	T2.416
A Dog	—	Painted wood	4 pieces	T3.206
Number Game	—	Painted wood	55 pieces	T4.616
Alphabet	—	Painted wood	28 pieces	T4.617
City	—	Painted wood	11 pieces	T4.622
Golden Compass	—	Painted wood	220	1946.56
Triangle	—	Painted wood	62 × 54	1946.55

Table 3.1: Artworks by Joaquín Torres García destroyed in the 1978 fire of the Rio de Janeiro's Museum of Modern Art: title, year, technique, dimensions, and catalog reference (Catalogue Raisonné).

Among the destroyed works, we identified that 22 of 73 were three-dimensional objects, including toys, sculptures, and other relief-based pieces, usually made of painted and/or carved wood. Their volumetric and lighting characteristics differ substantially from those of two-dimensional paintings. The term “3D” will henceforth refer to this category of objects.

In 2007, three fragments of one of the 7 murals, *Pax in Lucem*, were rediscovered in Montevideo. These pieces were sent from the MAM Rio de Janeiro to Uruguay shortly after the 1978 fire and remained stored at the MNAV for nearly 30 years. Their recovery provided a rare opportunity to study surviving material traces of one of the lost works, as shown in Figure 3.1.

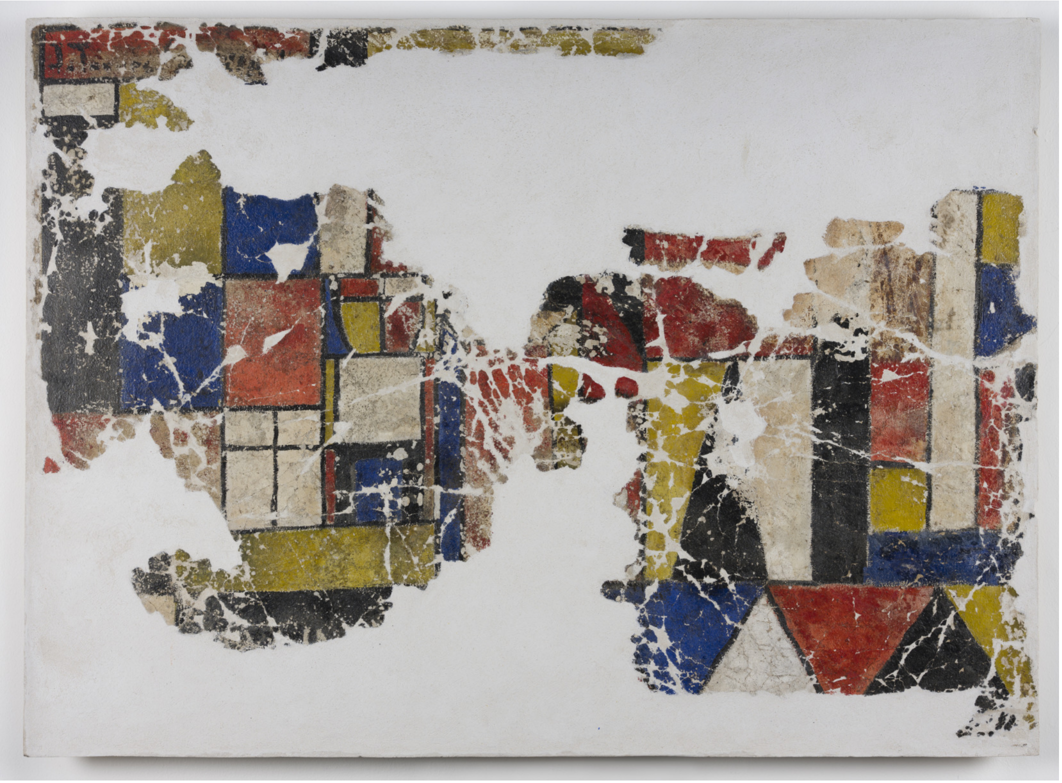


Figure 3.1: One of the surviving fragments of *Pax in Lucem*. Courtesy of the Museo Torres García. Photograph by Luis Sosa.

Photographic documentation was produced when the murals were removed from the walls of the Hospital Saint Bois. This visual record was taken by the artist and restorer Carlos Giaudrone (Giner, 2003) and preserved by the MTG. Color images of several lost artworks were also found in exhibition catalogs from Paris, and in a set of slides derived from the original footage of the documentary film *Joaquín Torres García: su vida y su obra* (1979), directed by Adolfo Fabregat and Walter

Acosta (Fabregat & Acosta, 1979). The film documents the 1974 retrospective exhibition held at the MNAV in Montevideo, which later became the one lost in the 1978 fire at the MAM Rio. The original work, a 16 mm color film with sound, and a duration of 14 minutes and 17 seconds, belongs to the Instituto de Cinematografía de la Universidad de la República (ICUR-UdelaR) collection and is preserved in the General Archive of the University (AGU-UdelaR). Its digitization was carried out by the Laboratorio de Preservación Audiovisual (LAPA) using a telecine transfer system. However, the film (and hence the digitized version) has lost its original color information, as it had been stored for nearly 30 years in a closed cabinet before its recovery. Fortunately, a small number of slides made from the original film still preserve the chromatic information of certain scenes, providing color references. Special thanks are extended to Julio Cabrio for locating these slides, and to Jaime Vázquez and Ignacio Seimanas from LAPA for their collaboration in the various image-acquisition processes that contributed to this work.

While these audiovisual materials offer valuable contextual evidence, their chromatic data cannot be considered fully reliable due to film degradation and analog reproduction limitations, as well as the lack of color calibration in the different steps of the process, as shown in Figure 3.2.

In 2022, Gregory Randall and Lara Raad visited the MAM Paris, where the seventy-three works had been exhibited in 1975 before being sent to Rio de Janeiro. They found in the museum's archives a set of color slide photographs of that exhibition, some of them include calibration information recorded at the time. With the authorization of MAM Paris, these slides were scanned by Rafael Grompone and Lara Raad. Special thanks are extended to Gregory, Lara, and Rafael for their collaboration in this process, and to the MAM Paris for granting access to these materials and authorizing their digitization.

3.1.2. Reference image sources

As noted earlier, the image datasets used in this thesis were compiled from multiple archives and institutional collections, ~~summarized in Table 3.2~~. Each source presented specific characteristics in terms of image format, resolution, color fidelity,

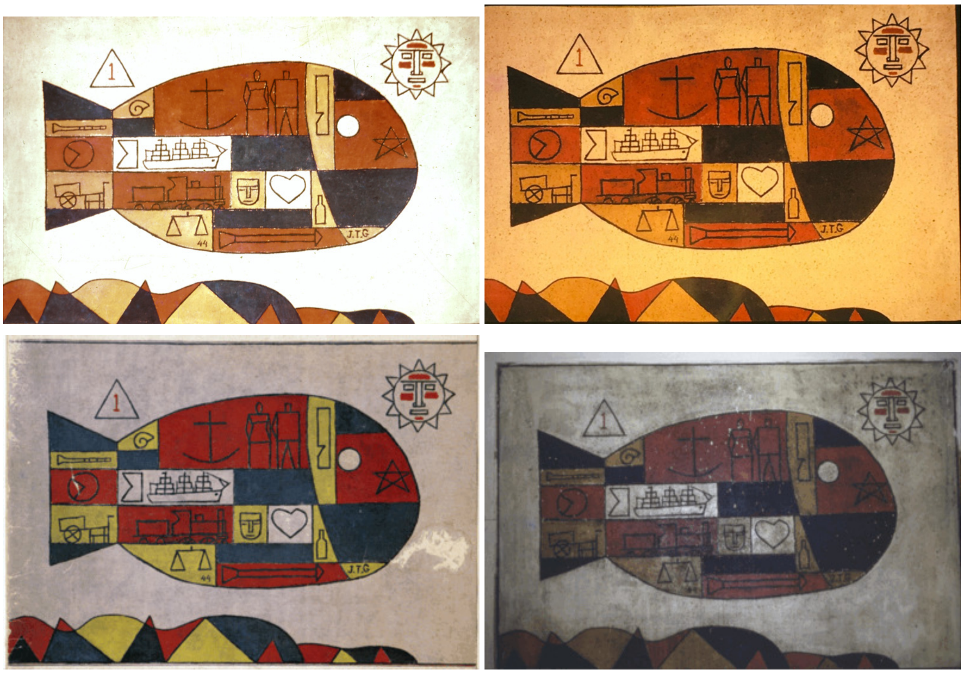


Figure 3.2: Different images for the same artwork: "The fish" (1944-28). Top, left to right: (1) LAPA version, (2) MAM Paris version. Bottom, left to right: (3) Giaudrone version and (4) Cecilia de Torres HD version. Note the differences in color between them.

and naming conventions (JTG did not give a name for most of his artworks), often resulting in multiple digital versions of the same artwork, as illustrated in Figure 3.2. This diversity required a careful curation process to detect duplicates, harmonize metadata, and ensure consistency between datasets.

The specific sources comprising the corpus are named and detailed below and summarized in Table 3.2.

JTG Catalog, named as IMGC. The *Catalogue Raisonné* («Joaquín Torres García Catalogue Raisonné», 2003) serves as the base reference for JTG's artworks. Developed under the direction of Cecilia de Torres, this online scholarly resource documents all known paintings, sculptures, and toys by JTG. It provides comprehensive metadata for each artwork, covering provenance, exhibition history, and bibliographic references, and assigns each piece a unique catalog number based on its year of production. As previously noted, JTG did not assign titles to most of his artworks; therefore, the normalized titles and numbers follow those used in the estate's posthumous inventory and in the catalog itself. While the images available online are relatively small (typically around 204×300 pixels, 29 KB, JPEG) and include both color and grayscale reproductions, this catalog remains an essential resource for establishing standardized references and metadata consistency across all datasets. This source has only one image per artwork.Notably, this source includes a single image per artwork, but collectively represents the full corpus of JTG's known production.

Cecilia de Torres Gallery Archive named as IMGCec. Additional high–resolution images (400–12,000 KB, JPEG/PNG, full color) were kindly provided by Cecilia de Torres, enhancing the visual quality of a subset of artworks represented in IMGC. These images are not available for download from the public catalog. This source has more than one image per represented artwork.

Museo Torres García named as IMGM. The MTG provided several complementary datasets expanding upon the materials from the *Catalogue Raisonné*. Together, these datasets represent the full corpus of Joaquín Torres García's known artworks, as in IMGC, while providing additional visual documentation for a subset of works. They consist of high–resolution TIFF images (average size: 100 MB). Within this collection, a curated selection of color images was organized chronologically to provide clearer chromatic references for comparative analysis. While not all artworks are represented by multiple images, some works include more than one visual record, allowing for richer comparative and contextual interpretation.

Gonzalo Fonseca Archive from MTG, named as IMGF. Images related to the work of Gonzalo Fonseca, a TTG member, were generously provided by MTG. Files are in TIFF format (average size: 30 MB). This source has only one image per represented artwork.

MAM Paris, named as IMGP. This subset contains the color photographs produced by MAM Paris for the 1975 exhibition, conserved in their archives and digitized by Rafael Grompone and Lara Raad. Files are provided in TIFF format, ranging from 14 MB to 67 MB, and include calibration data for a portion of the slides, making them particularly valuable for assessing chromatic consistency. This source has only one image per artwork from a subset of artworks represented in IMGC.

LaPA Audiovisual Preservation Laboratory from UdelaR, named as IMGL. This set comprises digitized stills and slide scans derived from historical audiovisual materials preserved by the General Archive of the University (AGU). Image quality and color fidelity vary due to film aging and analog reproduction processes. Files are in JPEG format, ranging between 200 KB and 1 MB. This source has more than one image per represented artwork from a subset of artworks represented in IMGC.

Fundación José Gurvich, named as IMGG. This collection was made available through a cooperation agreement between the Universidad de la República (UdelaR) and the Fundación José Gurvich. It includes high–quality color reproductions (TIFF/PNG, 100 MB each) of works by members of the Taller Torres García, particularly José Gurvich. Special thanks are extended to Eugenia Méndez, from Fundación José Gurvich, for her coordination and support in granting access and permissions. This source has only one image per artwork.

Studio Captures from MTG and Torre de las Telecomunicaciones, named as IMGS. This small but technically rigorous dataset consists of high–resolution studio captures (TIFFs, up to 300 MB) and grayscale film images, all produced under controlled lighting conditions and subsequently calibrated. Two separate photographic sessions were conducted: the first at MTG, and the second at the Torre de las Telecomunicaciones (ANTEL), where the remaining murals painted by the TTG at Hsopital Saint Bois are currently located. Both sessions were conducted under standardized lighting and color conditions (see Figure 3.3). This source has more than one image per represented artwork.

At the MTG, paired analog and digital captures were produced for selected works by Joaquín Torres García and Gonzalo Fonseca. Photographer Luis Sosa, together with Ignacio Seimanas, carried out the session, combining traditional analog film techniques with high–definition digital imaging. The analog process employed

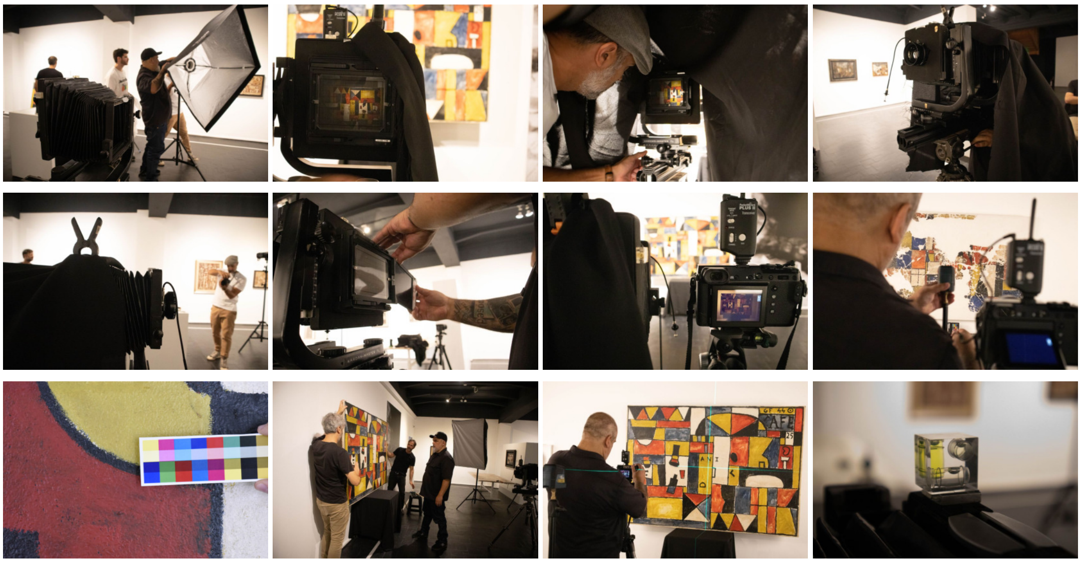


Figure 3.3: Documentation of the photographic acquisition process. The sequence illustrates the dual analog–digital workflow implemented to obtain calibrated images of artworks by JTG and TTG. The first two rows show the setup of large-format analog cameras, lighting control, and film exposure, while the third row documents the calibration of distance, focus, and color targets, together with high-definition digital captures. Photographs by Ignacio Seimanas.

silver–based emulsions on film, where light exposure formed a latent image later developed chemically in a darkroom, yielding exceptional tonal depth and fidelity in grayscale. Complementary digital photographs were then taken with modern high–resolution cameras, ensuring chromatic accuracy and material detail.

The second session took place at the Torre de las Telecomunicaciones, coordinated by Alejandro Díaz from the MTG, in collaboration with ANTEL authorities. In this phase, Ignacio Seimanas and Jaime Vázquez conducted high–definition color captures of the murals created by the TTG.

All images from both sessions were carefully color–calibrated (for more details about calibration please see Section 2.1) by Luis Sosa, Ignacio Seimanas, and Jaime Vázquez, who also performed monitor calibration at the MTG and on the author's workstation, ensuring consistent chromatic accuracy across devices and viewing environments.

Special thanks are extended to Luis Sosa, Ignacio Seimanas, and Jaime Vázquez for their technical expertise and dedication throughout the acquisition process, and to Alejandro Díaz for coordinating access and logistics with ANTEL and the Torre de las Telecomunicaciones.

Destroyed artworks from MTG Archive, named as IMGD. The grayscale images shown in Figure 3.4 correspond to a dataset composed of 52 photographic records of the pieces destroyed in the 1978 fire. These materials were also provided by the MTG. Most originate from analog photographic records that were subsequently digitized by the MTG team, resulting in variable quality and contrast levels. The digitized files are in PNG format, averaging approximately 40 MB each. This source has more than one image per represented artwork. Some photographs focus exclusively on the artwork itself, while others include parts of the surrounding room or exhibition space where the piece was displayed.

Name	**Source / Archive**	**Number of Images**	**Format**	**Size**	**Color**	**3D**
IMGC	*Catalogue Raisonné*	2,427	JPEG	29 KB	Color / Grayscale	Yes
IMGCec	Cecilia de Torres Archive	21	JPEG, PNG	0.4–12 MB	Color	Yes
IMGM	MTG	2,590	TIFF, JPEG	up to 100 MB	Color / Grayscale	Yes
IMGP	MAM Paris	13	TIFF	14–67 MB	Color	Yes
IMGL	LAPA	21	JPEG	0.2–1 MB	Color / Grayscale	No
IMGG	Fundación José Gurvich	145	TIFF, PNG	≈100 MB	Color	Yes
IMGF	Gonzalo Fonseca Archive (MTG)	20	TIFF	≈30 MB	Color	Yes
IMGS	MTG & ANTEL	10	TIFF	up to 300 MB	Color / Grayscale	Yes
IMGD	MTG photographic archive	52	PNG	≈40 MB	Grayscale	Yes

Table 3.2: Summary of image sources: name of the set, their institutional sources, number of images, formats, image size, grayscale or color, and if they include images from 3D objects.

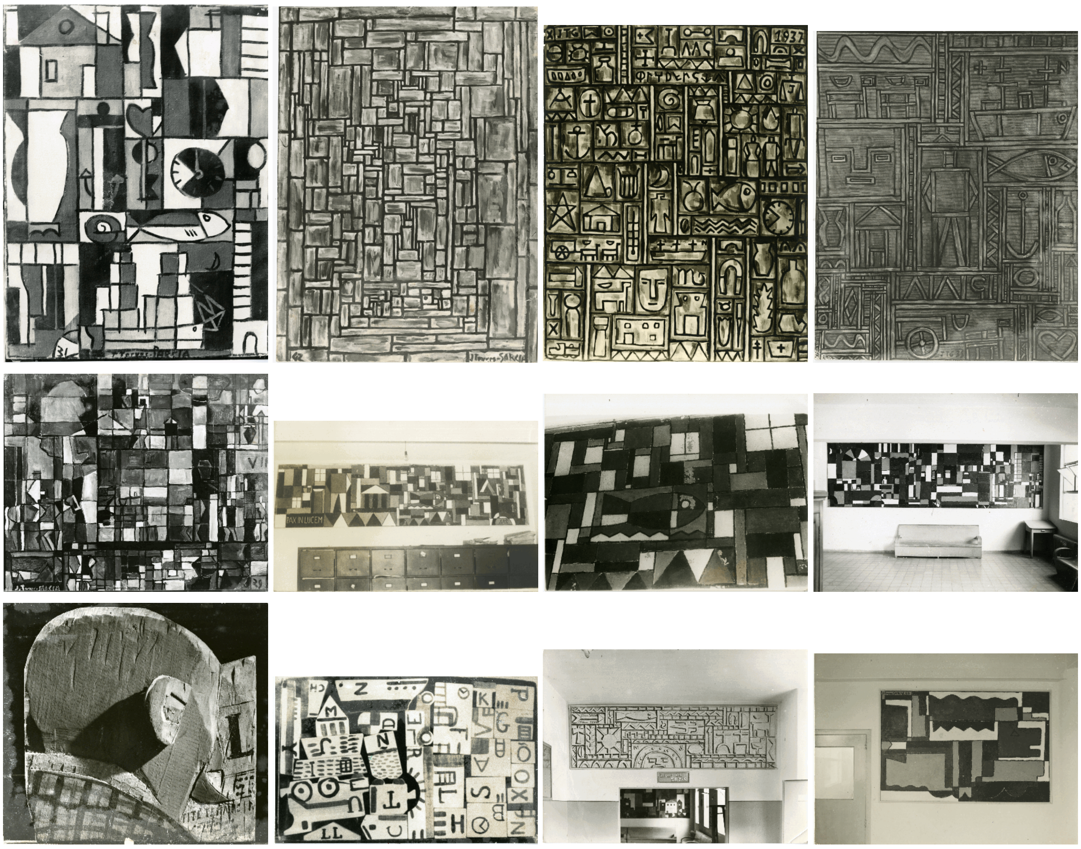


Figure 3.4: Grayscale photographic records from the MTG archive corresponding to some destroyed artworks. Titles and their respective catalog reference numbers (in parentheses) (Catalogue Raisonné). Row 1 — (1) TN5 (1931.09), (2) Infinito (1942.19), (3) Pintura (1937.40), (4) Composición (1938.34); Row 2 — (1) Pintura Constructiva (1929.18), (2) Pax in Lucem (1944.27), alternate view 1, (3) Pax in Lucem (1944.27), alternate view 2, (4) El Tranvía (1944.30); Row 3 — (1) Cabeza (1921.14), (2) ABC (T4.617), (3) Pacha Mama (1944.12), (4) Forma (1944.29). Courtesy of the Museo Torres García.

3.2. Construction of Datasets

The image sources described in the previous section are not entirely disjoint. While most of their files are distinct, several depict the same underlying artwork, producing intersections at the level of represented paintings rather than at the level of file identity. Given the heterogeneity of these files, identifying all versions of a given work was particularly challenging, especially considering the lack of standardized titles (as JTG did not assign names to most of his artworks) and the diversity of formats and colorizations across archives. Figure 3.2 shows an example of overlap at the artwork level. In this example, the same artwork appears in four different collections. The top row shows the IMGP and IMGM versions, from left to right. The bottom row shows the IMGM and IMGCec versions, from left to right.

Let P denote the set of all artworks by JTG, and O the set of artworks produced by members of the TTG. These two sets serve as the basis for defining the coverage, intersections, and overlaps of each set described in Table 3.3 and visualized in Figure 3.5. It should be noted that P corresponds to the artworks documented in the *Catalogue Raisonné* («Joaquín Torres García Catalogue Raisonné», 2003), while there is no complete inventory of O.

Using the already mentioned image source sets, four progressively inclusive datasets were constructed to train and evaluate the colorization model. Each dataset builds upon the previous one, ensuring a shared evaluation subset and allowing consistent comparison of model performance under equivalent test conditions. This hierarchical organization enables the analysis of how data quantity, curatorial precision, and chromatic diversity influence the model's behavior. Please note that grayscale and three-dimensional artworks were excluded from all training datasets. The resulting dataset configurations are as follows:

$\mathbf{D_1}$: This dataset contains high-resolution images of Joaquín Torres García's artworks. Here, high-resolution refers to digitized files with greater definition than those available in the public online catalogue. When multiple image versions of the same artwork existed, the most accurate and visually consistent reproduction was selected under the expert supervision of Carlos Serra at MTG. This dataset provides partial coverage of P. Let $P' \subset P$ denote the set

Name	Artworks in set	Coverage artworks	Multiplicity	Overlap (artworks / images)
IMGC	P	All P	1 per work	Artwork-level overlap with IMGCec, IMGM, IMGP, IMGL, and IMGS / Image overlap with IMGM
IMGCec	P	Subset $P^1 \subset P$	≥1 per work	Artwork-level overlap with IMGC and IMGM
IMGM	P	All P	≥1 per work	Artwork-level overlap with IMGC, IMGCec, IMGP, IMGL, and IMGS / Image overlap with IMGC
IMGP	P	Subset $P^2 \subset P$	1 per work	Artwork-level overlap with IMGC, IMGCec, IMGM, and IMGL
IMGL	P	Subset $P^3 \subset P$	≥1 per work	Artwork-level overlap with IMGC, IMGCec, IMGM, and IMGP
IMGG	O	Subset $O^1 \subset O$	≥1 per work	Artwork-level overlap with IMGF, and IMGS
IMGF	O	Subset $O^2 \subset O$	≥1 per work	Artwork-level overlap with IMGF, and IMGS
IMGS	P, O	Subsets $P^4 \subset P, O^3 \subset O$	≥1 per work	Artwork-level overlap with IMGC, IMGM, IMGCec, IMGD, IMGL and IMGF
IMGD	P	Subset $P^5 \subset P$	≥1 per work	Artwork-level overlap with all P sets

Table 3.3: Characteristics of the source image sources, including their art coverage, multiplicity (amount of images per work) in artworks representation, and possible overlap levels (either at the artwork or image level).

of artworks represented in D_1. Size: **546 images**.

$\mathbf{D_2}$: $D_1 \cup \{\text{IMGF}, \text{IMGG}, \text{IMGS}\}$. This dataset extends D_1 by incorporating high-resolution images of artworks produced by members of the Taller Torres García, while maintaining a single representative image per artwork, as in D_1. The inclusion of these additional sources aimed to increase dataset diversity while preserving image quality, expanding the range of visual compositions, textures, and color palettes available for training. Partial coverage of both P and O. Size: **708 images**.

$\mathbf{D_3}$: $D_2 \cup \text{IMGC}(P \setminus P')$. This dataset completes the coverage of all artworks in P that are not represented in D_1. For each artwork $p \in P \setminus P'$, the corre-

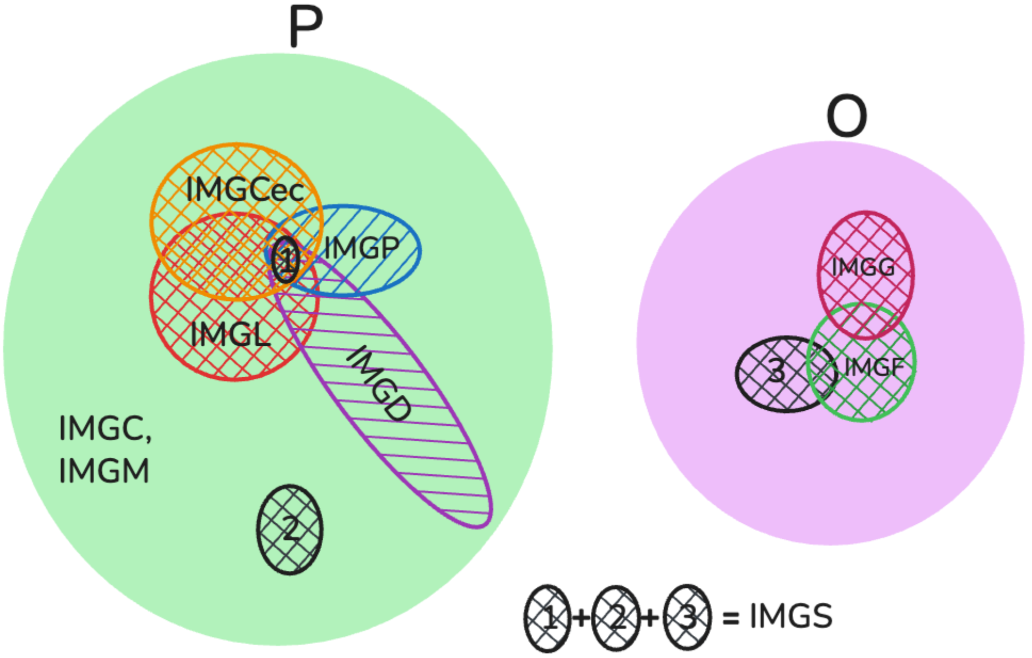


Figure 3.5: Visual representation of image sources intersections. Left: P is the set of Paintings of JTG. IMGC and IMGM have full representation of all P and for that reason, they overlap in representation. While IMGD, IMGP, IMGL and IMGCec include partial representations of P and they have partial overlap in representations as well. Right: O is the set of Paintings of TTG. There are partial overlapping between IMGF and IMGG. The Studio dataset IMGS bridges the domains (P and O) via the high-quality/studio images and has partial respresentation of both P and O, and overlaps IMGF at artwork-level.

sponding image was retrieved from IMGC and added to the dataset, ensuring full coverage of P. Each artwork remains represented by a single image. Full coverage of P; partial coverage of O. Size: **1,964 images**.

D$_4$: This dataset extends D_3 by incorporating all remaining image variants of each artwork from all available sources. Consequently, some artworks are represented by multiple images. Full coverage of P; partial coverage of O. Size: **2,410 images**.

3.2.1. Train, Validation and Test Partitions

For each dataset D_k, the data was divided into three subsets: training (Tr$_k$, 70%), validation (V$_k$, 15%), and test (T$_k$, 15%), preserving the hierarchical structure across datasets. This means that each subset is contained within the next one (e.g., $\mathrm{T}_1 \subseteq \mathrm{T}_2 \subseteq \mathrm{T}_3 \subseteq \mathrm{T}_4$), so that T_4 includes all previous test sets. This is illustrated in Figure 3.6.

For the dataset D_4, which includes multiple image representations of the same artwork, all images corresponding to a given work were assigned to the same split. In other words, if one image of an artwork was included in the training set, no other image of that artwork appeared in the validation or test sets. This constraint ensured per-artwork consistency across the hierarchy and prevented data leakage, guaranteeing that the model never encountered any version of a test artwork during training and vice versa.

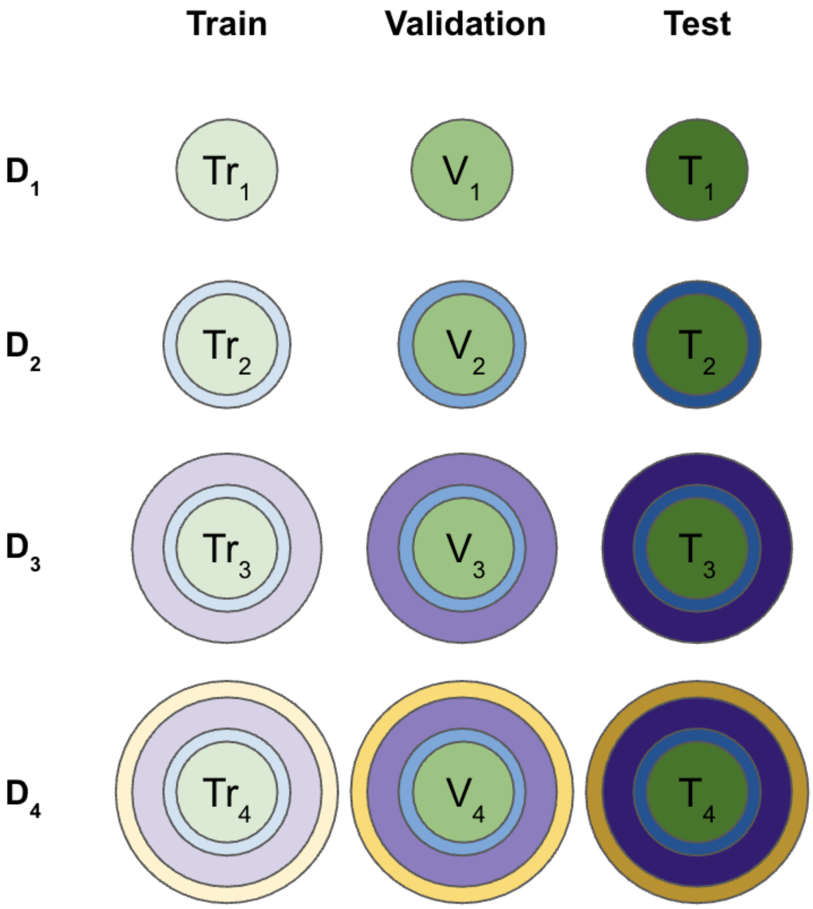


Figure 3.6: Schematic representation of the four datasets (D_1–D_4) and their hierarchical partitions into training (Tr_k), validation (V_k), and test (T_k) subsets (70%, 15%, 15%). Each dataset fully contains the splits of the previous one: the blue (T_1), green (V_1), and orange (Tr_1) subsets of D_1 are entirely included within the corresponding subsets of D_2; likewise, D_2 is nested in D_3, and D_3 in D_4.

3.2.2. Data pre-processing methodology

Due to differences among the image sources, it was not straightforward to determine which images depicted the same artwork. Therefore, a mixed approach combining automated matching techniques and expert human review was used.

3.2.2.1. Automatic Detection

First, the names and sizes of the files were compared to automatically detect images corresponding to the same artwork. A feature-similarity comparison was then performed on the remaining files to identify further duplicates.

Name and size. Files were listed by size (in bytes). Exact matches in both name and size were considered obvious duplicates.

Feature similarity. A pre-trained VGG16 (Simonyan & Zisserman, 2015), implemented in Keras (Chollet et al., 2015) and trained on ImageNet (Russakovsky et al., 2015), was used as a feature extractor. For a VGG overview, please see Section 2.3.1.1. Following the procedure outlined in the technical tutorial (Samarasinghe, 2023), each image was resized to 224×224 pixels, normalized, and fed into the network to obtain feature embeddings from the penultimate layer. Pairwise similarities between feature vectors were then computed using the cosine similarity between two vectors $\vec{u}$ and $\vec{v}$:

$$\mathrm{sim}(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \, \|\vec{v}\|}, \qquad \mathrm{dist}(\vec{u}, \vec{v}) = 1 - \mathrm{sim}(\vec{u}, \vec{v}).$$

For each image in the source dataset, the five most similar candidates in the target dataset were retrieved using this metric. This top-five retrieval strategy made it possible to identify potential duplicates or alternate versions of the same artwork across collections, without requiring a fixed similarity threshold. Experts subsequently validated visually close cases to confirm true correspondences.

3.2.2.2. Manual Review and Selection Criteria

All image pairs identified through the feature similarity process according to the following rules:

- If the same artwork appears more than once with similar quality, an MTG expert selects the best sample, considering the chromatic fidelity and detail.
- If quality differs, the best-quality image is retained.
- If images are identical, only one copy is kept.
- If both grayscale and color versions exist, the color version is retained.

3.2.2.3. Examples of data heterogeneity

To illustrate the data inconsistencies described above, representative examples of the main issues encountered during dataset construction are presented here. These include instances of the same artwork stored in multiple file resolutions within a single source (Figure 3.7), discrepancies in colorization across versions (Figure 3.8), and cases where several images of the same artwork appear under inconsistent naming conventions and color treatments (Figure 3.9). Additionally, Figure 3.10 highlights a common issue in comprehensive art catalogues, where distinct artworks corresponding to exploratory or study processes may appear visually very similar. In such cases, they may be incorrectly identified as duplicates, while expert curatorial review reveals them to be separate works, often differing in technique, color treatment, or material execution.

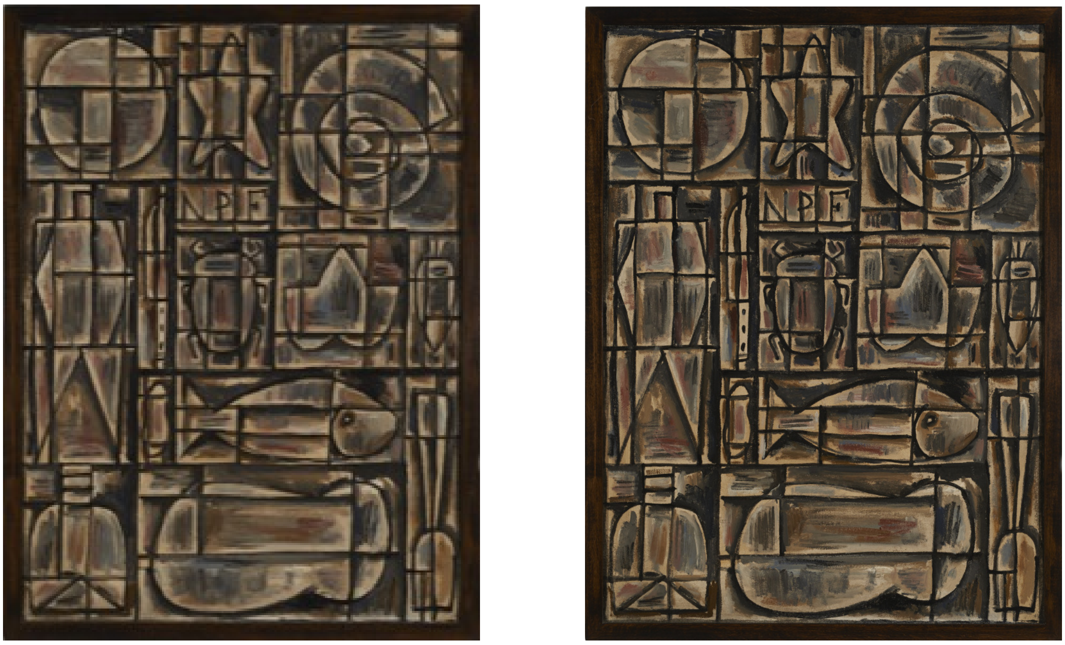


Figure 3.7: Two versions of the same artwork (1931.49) with different names (1931-492028new29 and 1931-492028new292) and sizes (37KB and 5.3MB), from IMGM.

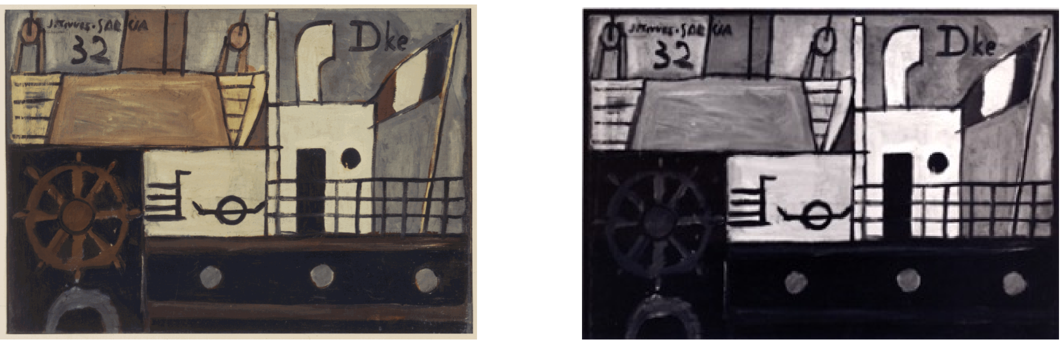


Figure 3.8: Same artwork (1932.09) with different colorization from different sources IMGC and IMGM.

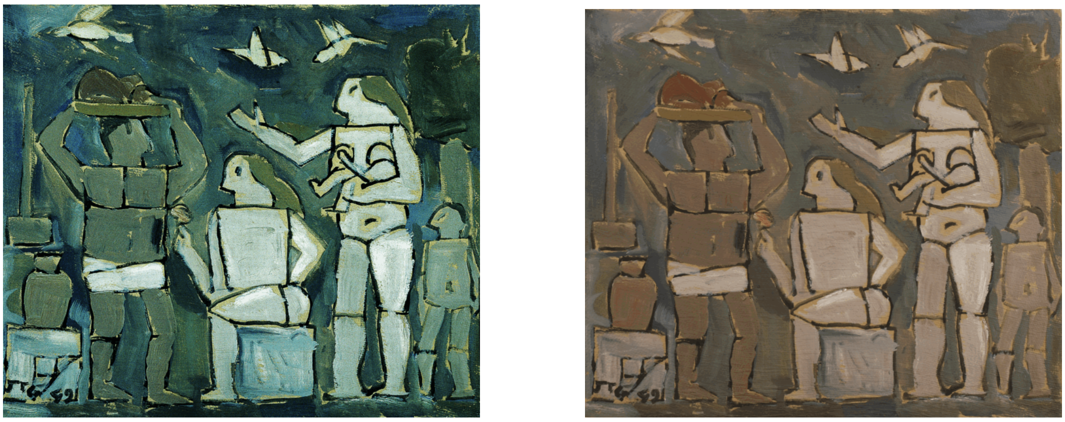
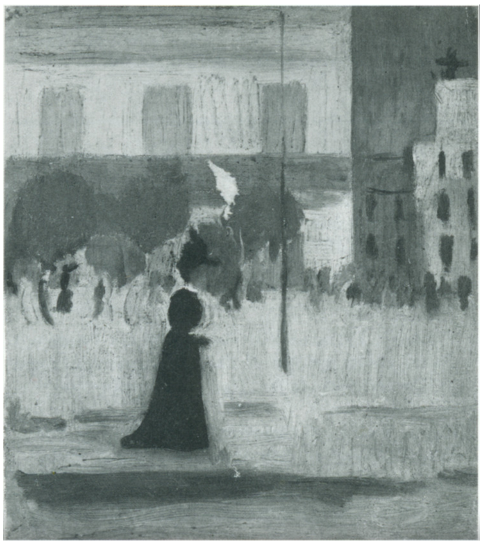


Figure 3.9: Same artwork (1949.11), different resolutions and colorization. From IMGM.



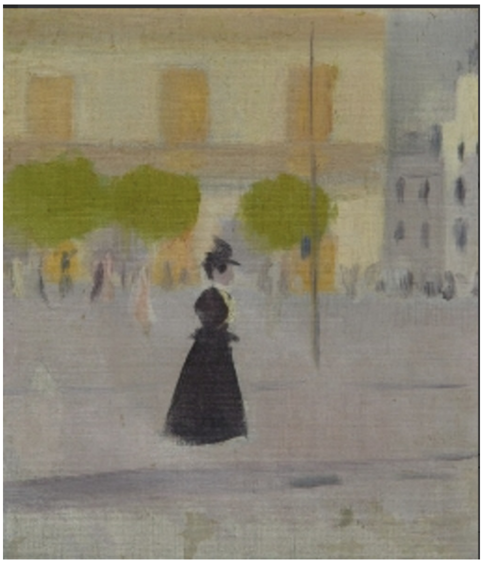


Figure 3.10: Different images from different artwork, 1898.01 left and 1898.02 right. From IMGM and IMGC.

Chapter 4

Methodology

This chapter presents the methodology followed throughout the work, from the initial model selection process to the fine-tuning experiments. The goal was to identify, adapt, and validate a colorization model capable of reconstructing plausible and consistent color representations for Joaquín Torres García's artworks.

The methodology builds upon and extends the research introduced in two ML-Brief papers: *A Short Analysis of BigColor for Image Colorization* (García et al., 2024b) and *A Brief Analysis of iColoriT for Interactive Image Colorization* (García et al., 2024a). These publications provided the preliminary benchmarking necessary to select the most appropriate architecture for the specific constraints of artistic restoration.

4.1. Analysis of Colorization Methods

Building on the terminology introduced in Section 2.2, the goal of this analysis was to compare three representative paradigms of color priors: (i) fully automatic deep learning methods that rely exclusively on dataset priors, (ii) exemplar-based approaches that incorporate into the dataset priors a reference image to guide chromatic transfer, and (iii) hybrid interactive methods that integrate dataset priors and user-provided scribbles to resolve ambiguous regions.

Although more recent methods have since emerged, the selected models constitute the most relevant and technically mature approaches available during the early phases of this work.

To determine the most suitable approach for adaptation to artistic image restoration, two state-of-the-art models were initially evaluated: BigColor (Kim et al., 2022) and iColoriT (Yun et al., 2023). In the taxonomy introduced above, BigColor corresponds to paradigm (i), as it is a fully automatic method that relies exclusively on dataset priors learned from large-scale image collections. And iColoriT aligns with paradigm (iii), since it augments learned priors with user-provided color hints, placing it within the hybrid interactive family of methods.

A third model, Color2Embed (Zhao et al., 2021), representing paradigm (ii), was initially evaluated by running the inference pipelines available in the authors' public GitHub repository[1] and using the released pretrained weights. While this enabled a qualitative inspection of the model's exemplar-based colorization behavior, the lack of detailed instructions prevented further model adaptation. As a result, Color2Embed was excluded from the subsequent experiments.

This chapter describes iColoriT (Yun et al., 2023), the method on which our approach to color restoration is based. For a detailed description of BigColor (Kim et al., 2022), please refer to the MLBrief publication (García et al., 2024b), which analyzes the original work.

4.1.1. iColoriT overview

iColoriT performs interactive image colorization using a Vision Transformer (ViT; Dosovitskiy et al., 2021) to propagate sparse color constraints across the image. The architecture comprises three main components, illustrated in Figure 4.1: a transformer encoder responsible for propagating chrominance information, a pixel

[1] Official implementation of Color2Embed available at https://github.com/zhaohengyuan1/Color2Embed

shuffling module that reshapes the patch-level output into the image grid, and a local stabilization layer that smooths color transitions between neighboring patches after the shuffling stage.

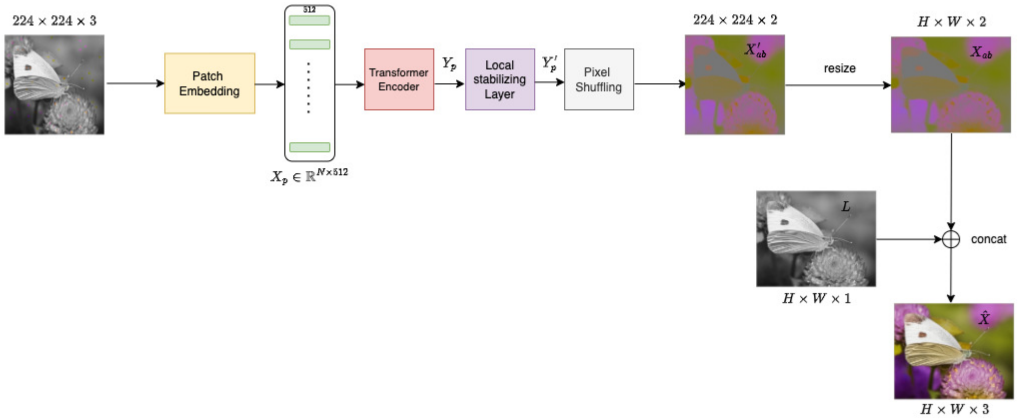


Figure 4.1: Overview of iColoriT. A patch embedding layer transforms the input X into a sequence of patch tokens X_p. These tokens are processed by the transformer encoder, producing Y_p, which is refined by a local stabilization layer to yield Y'_p. The stabilized tokens are then reshaped by a pixel shuffling module, producing a chrominance prediction X'_{ab} that is resized to the original image resolution and concatenated with the luminance channel L.

The model operates on images represented in the CIELAB color space. In practice, we adopt the RGB-to-CIELAB conversion code provided by the authors, which follows the RGB–CIEXYZ–CIELAB pipeline and the sRGB specification, as described in Section 2.1.2. In addition to this standard color space transformation, the authors introduce an *ad hoc* normalization tailored to the learning setting. After conversion, the lightness component L and the chromatic components a and b are normalized.

Starting from an input image of size $H \times W \times C$, possibly grayscale ($C = 1$) or color (RGB, $C = 3$), the image is converted to the normalized CIELAB color space and resized to 224×224. The luminance channel L is taken as the grayscale image to be colorized and is provided to the network together with a constraint representation, whose construction depends on the context in which iColoriT is used, leading to non-interactive and interactive input configurations.

Non-interactive mode. In the non-interactive configuration, primarily used for training, evaluation, and reproducible experimentation, color constraints are specified offline, for instance, as a list of pixel coordinates stored in a text file. These coordinates are remapped to a 224×224 grid. In this setting, the input image is required to be available in color in order to extract ground-truth chrominance values at the specified locations. These chrominance values, together with their spatial positions, are used to construct a constraint representation. The input X is formed by concatenating the luminance channel L with the corresponding constraint representation and is fed to the colorization network. An illustration of this process is shown in Figure 4.2.

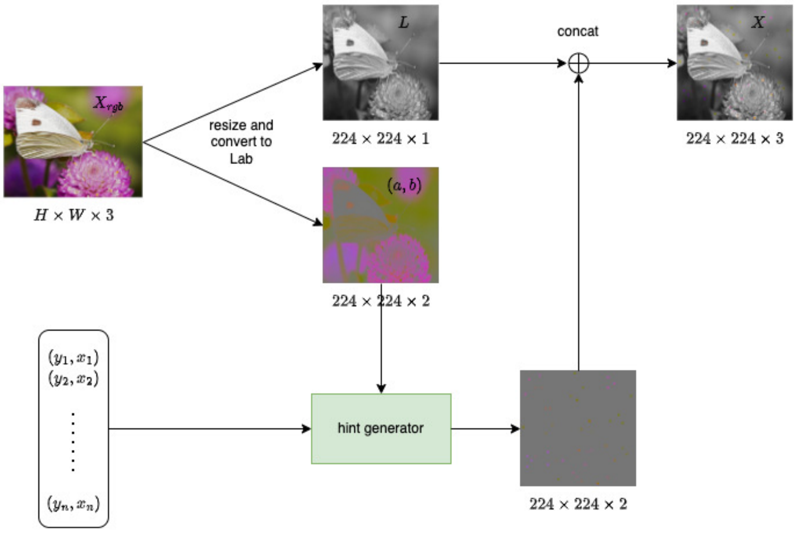

X_{rgb}
$H \times W \times 3$
resize and convert to Lab
L
$224 \times 224 \times 1$
(a, b)
$224 \times 224 \times 2$
concat
X
$224 \times 224 \times 3$
(y_1, x_1)
(y_2, x_2)
(y_n, x_n)
hint generator
$224 \times 224 \times 2$

Figure 4.2: Process of converting the color input image and the list of position hints into the three-channel input image (non-interactive mode). The ground-truth image X_{rgb} is resized to $224 \times 224 \times 3$ and converted to the CIELab color space. A list of coordinates $\{(y_1, x_1), ..., (y_n, x_z)\}$, with Z the number of hints provided, and the resized chrominance channels (a, b) are passed to the hint generator module, resulting in a 2-channel hint mask. The hint generator rescales the coordinates relative to an image of size 224×224. The resized luminance L is concatenated with the hint mask, resulting in the 3-channel input X.

Interactive mode. In the interactive setting, color constraints are specified through a graphical user interface, where the user places sparse color hints directly on the image. These hints are remapped to the 224×224 grid and converted to the normalized CIELAB representation using the same conversion and normalization pipeline as the input image. The chromatic components of the hints define a two-channel constraint representation (a_h, b_h), which represents sparse chrominance values at the specified spatial locations. The resized luminance channel L is concatenated with the chrominance hint channels (a_h, b_h), yielding the three-channel input X that is fed to the colorization network (Figure 4.3).

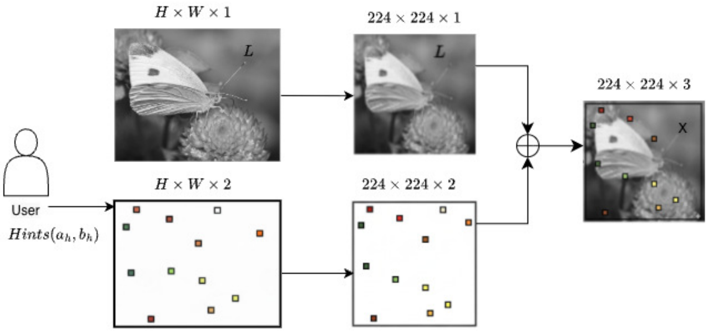

$H \times W \times 1$
$224 \times 224 \times 1$
$224 \times 224 \times 3$
L
L
X
User
$Hints(a_h, b_h)$
$H \times W \times 2$
$224 \times 224 \times 2$

Figure 4.3: Construction of the input X in the interactive mode. The luminance channel L is obtained from the resized CIELAB representation of the input image. User-provided color hints, specified as a set of pixel coordinates together with their associated chrominance values, are converted into a two-channel chrominance map (a_h, b_h). The final input X is formed by concatenating L and (a_h, b_h).

Transformer encoder The input X is first embedded into X_p, a sequence of N tokens of dimension $d = P \times P \times C$, where $N = H'W'/P^2$ is the number of tokens, P is the size of the patch side, and $C = 2$. This sequence is obtained by passing X in a patch embedding layer composed of one convolutional layer of d kernels of size $P \times P$ and a stride of size P. The resulting sequence of tokens X_p is the input to the encoder transformer, and its output is Y_p, of the same shape as X_p. The different stages of the transformer encoder are as follows. First, as explained in Section 2.3.2, a sinusoidal positional encoding E_p (Dosovitskiy et al., 2021) is added to the input X_p yielding

$$z_0 = X_p + E_p.$$

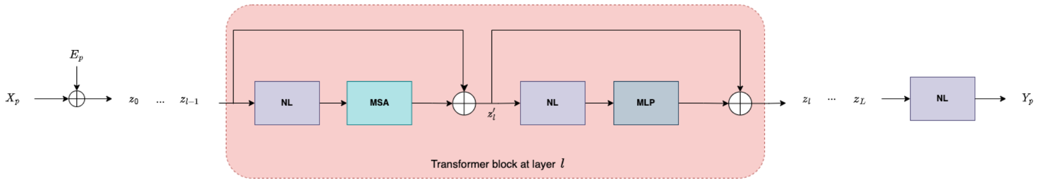


Figure 4.4: iColoriT. Encoder transformer block at layer l. The intermediate result $z_l^{'}$ is the concatenation of the output of layer l-1 and the same signal processed by a normalization layer NL and a Multi-headed self-attention layer MSA. A second step normalizes $z_l^{'}$ and passes it through an MLP that comprises one hidden layer to obtain z_l.

As shown in Figure 4.4, a cascade of L multi-headed transformer blocks follows. These blocks are described in Section 2.4 and are computed here as

$$z'_l = \mathrm{MSA}(\mathrm{NL}(z_{l-1})) + z_{l-1} \tag{4.1}$$

and

$$z_l = \mathrm{MLP}(\mathrm{NL}(z'_l)) + z'_l. \tag{4.2}$$

The last stage yields the output of the transformer

$$Y_p = \mathrm{NL}(z_L). \tag{4.3}$$

Please recall that $\mathrm{MSA}(\cdot)$ denotes the multi-headed self-attention layer, while $\mathrm{NL}(\cdot)$ denotes the normalization layer. The MSA layer computes the attention layer output as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + B\right) V, \tag{4.4}$$

where $Q, K, V \in \mathbb{R}^{N \times d}$ are the query, key, and value matrices. The term $B \in \mathbb{R}^{N \times N}$ represents the relative positional bias, which is added to supplement the input positional encoding E_{pos}, compensating for the fact that self-attention does not inherently capture position-related information.

Please note that iColoriT adopts a Pre-Norm configuration as described in 2.4. The Layer Normalization (NL), implemented via the `torch.nn` module, is applied *before* the Multi-Head Self-Attention (MSA) and MLP blocks, rather than after. Specifically, the NL operation standardizes the input features across the channel dimension to stabilize gradients during training, while retaining learnable affine parameters to preserve representational capacity (Ba et al., 2016). Furthermore, the MLP block (comprising one hidden layer of $4 \times d$ dimensions) utilizes a GELU activation function instead of the standard ReLU.

Decoder: Local Stabilization and Pixel Shuffling This stage functions as a decoder, reshaping the Transformer encoder's output to restore the original image's spatial resolution and yielding the two chrominance channels. This is efficiently achieved using two sequential operations: a Local Stabilization and Pixel Shuffling.

Pixel Shuffling The pixel shuffling layer reshapes an image of size $H' \times W' \times (P^2 \times C)$ into an image of size $(H' \times P) \times (W' \times P) \times C$. The output Y_p of the transformer can be viewed as an image of size $(H'/P) \times (W'/P) \times (P^2 \times 2)$. Then, by passing Y_p through the pixel shuffling stage, every channel in Y_p is resized to a patch of size $P \times P \times 2$, yielding an image of size $H' \times W' \times 2$.

Local Stabilizing Layer Yun et al. (2023) point out that, for a patch side size P greater than eight, the resulting images, after applying pixel shuffling, have visible artifacts along the image patch boundaries. To avoid this, a stabilization layer is applied before the pixel-shuffling stage.

This stabilization layer consists of a convolutional layer with d filters of size 3×3 and stride 1. The input to the stabilizing layer is Y_p of size $H'/P \times W'/P \times d$, and the output is Y_p' of the same size. Y_p' is then passed to the pixel shuffling stage, which produces the chrominance image X_{ab}' of size $H' \times W' \times 2$, which is finally resized to an image of size $H \times W \times 2$ and labeled X_{ab}. It is important to emphasize that this output corresponds solely to the chrominance channels (a, b).

Predicted color image The predicted color image $\hat{X}$ is the result of concatenating the input luminance channel $L \in \mathbb{R}^{H \times W \times 1}$ and the two predicted chrominance channels, $X_{ab} \in \mathbb{R}^{H \times W \times 2}$,which are obtained from the decoder.

4.1.1.1. Original Training

To train the iColoriT network, the Huber loss is employed as the objective function. While the comparison is formally defined between the predicted image $\hat{X}$ and the ground truth image X in the CIELab color space, note that the luminance channel L is identical in both. Consequently, the loss effectively quantifies the reconstruction error solely between the predicted chrominance channels $\hat{X}_{ab}$ and the ground truth chrominance X_{ab}.

The Huber loss, between $\hat{X}_{ab}$ and X_{ab}, is defined as follows (Yun et al., 2023):

$$\mathcal{L}_{\text{recon}} = \begin{cases} \frac{1}{2}\left(\hat{X}_{ab} - X_{ab}\right)^2, & \text{if } \left|\hat{X}_{ab} - X_{ab}\right| < 1, \\ \left|\hat{X}_{ab} - X_{ab}\right| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{4.5}$$

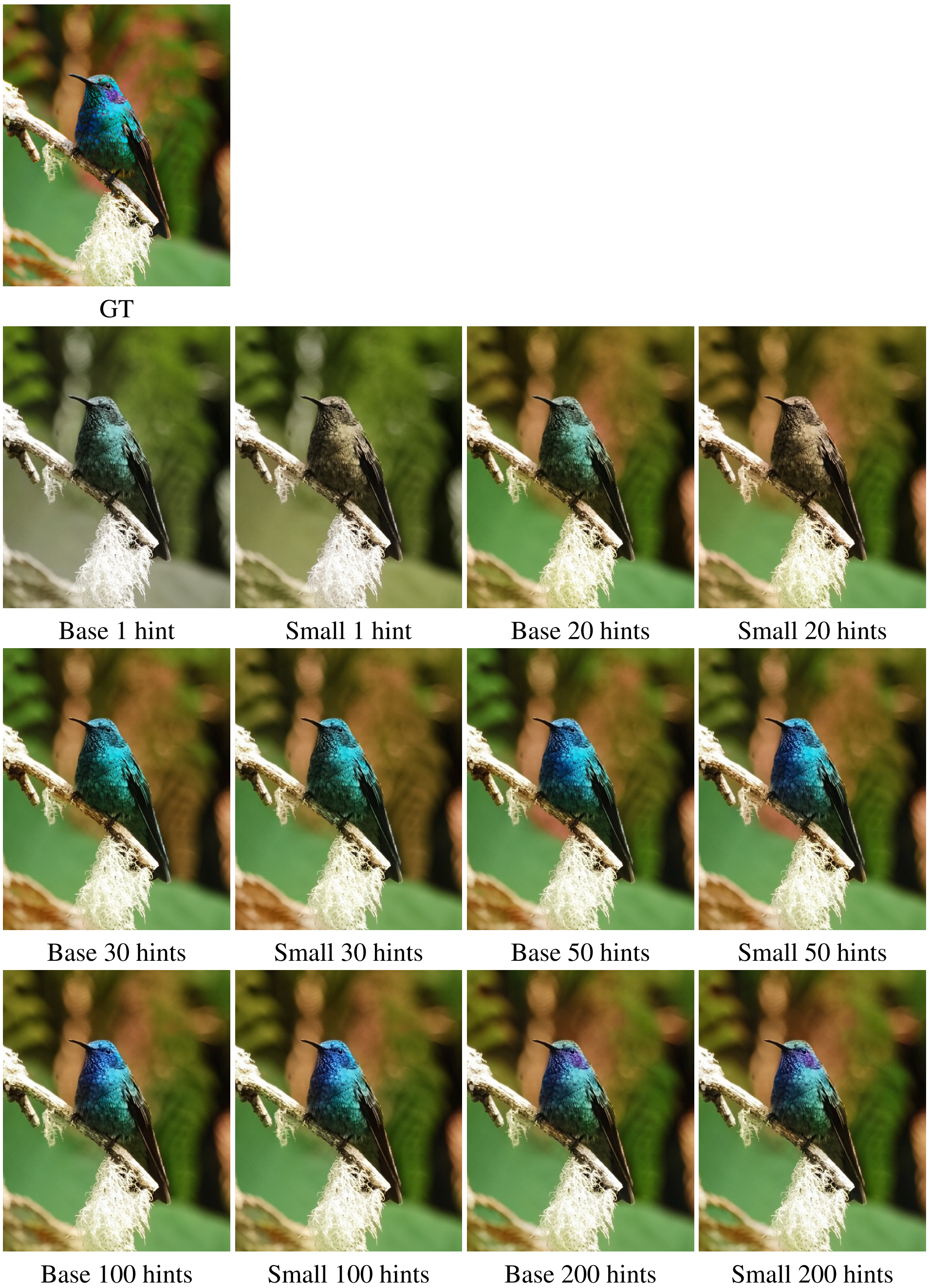
The network was trained on the ImageNet 2012 train split images (Russakovsky et al., 2015). The image set consists of 1281167 images. During training, the images were resized to 224×224. The patch size was set to $P = 16$; hence $N = H'/P \times W'/P = 196$ and $d = 512$. The number of transformer blocks was set to $L = 12$, and the number of heads per transformer block was set to $H_T = 12$. In training, the hints were generated by randomly selecting a position and number of hints, and assigning the average color from the ground-truth color image to each hint. Specifically, the number of hints was uniformly sampled from 0 to 128 for each batch and epoch. The AdamW (Loshchilov & Hutter, 2019) optimizer was used with a cosine annealing learning rate schedule (Loshchilov & Hutter, 2017), as described in 2.5.3.

4.1.2. Experiments

The purpose of these experiments was to verify some of the performance results presented by the authors (Yun et al., 2023) by comparing the Base and Small models (two of the three models provided by the authors), the number and location of hints, and how much the size of the hint affects the results. After finding that the Base model and a hint size of 2 were the best choices, several experiments with different datasets were conducted with those two parameters fixed. We used images from ImageNet1k (Russakovsky et al., 2015), a subset with 1000 labels from the original ImageNet dataset (Russakovsky et al., 2015), from CUB (Welinder et al., 2011), and from the Oxford 102flowers («Automated Flower Classification over a Large Number of Classes», 2008) datasets (mentioned by the authors in the original paper). Some experiments were also done with artistic images. These results are discussed in Section 4.1.2.4.

4.1.2.1. Performance experiments

Small model vs Base model The authors provided three model types: Tiny (the smallest), Small (the medium), and Base (the largest). Base vs. Small models were tested. The paper claims that with 10 hints, the performance of the three models was similar. In our tests, the qualitative performance of Base seems to be better than Small, as shown in Figure 4.5 for up to 30 hints. After this experiment, the Base model was used for all subsequent experiments.



GT

Base 1 hint

Small 1 hint

Base 20 hints

Small 20 hints

Base 30 hints

Small 30 hints

Base 50 hints

Small 50 hints

Base 100 hints

Small 100 hints

Base 200 hints

Small 200 hints

Figure 4.5: Comparison between Base and Small iColoriT models for varying number of hints. The first row shows the ground truth (GT) image. The number of hints varies from 1 to 200. Note how the results become (subjectively) similar after 30 hints. Images from Imagenet (Russakovsky et al., 2015)

Number and location of hints One challenge with this method is determining where to place the color hints and how many to use. Is it possible to achieve the same results with the same amount of hints but in different locations? To answer this question, the following experiment was conducted. For a fixed number of hints, two different results were generated: one where the positions of the hints were manually selected and another where the positions of the hints were randomly generated by sampling from a uniform distribution. This procedure was repeated for different numbers of hints. As the number of manually selected hints increases, the previously chosen hints are kept, and new ones are added. In the case of random hint positions, each result is obtained by randomly generating all hints anew, which does not guarantee that the hint positions of one experiment are included in the set of positions of the next one. As shown in Figure 4.6 and 4.7, location has a notorious effect on the results, but it diminishes as the number of hints grows. For example, in Figure 4.6, for an example taken from the ImageNet dataset (Russakovsky et al., 2015), it is possible to see that the colorization is much more accurate when the hint positions are manually defined than in the randomly generated hint positions case, for the five-hint experiment. Once the number of used hints exceeds 20, the results are very similar. The same observation applies to the experiment depicted in Figure 4.7, where random generation cannot achieve accurate colors with fewer than 20 hints.

Hints size As in the original paper (Yun et al., 2023), the effect of the hint size on the colorization results was analyzed. Since hints are generally larger than a single pixel and the colors of all pixels contained in the hint are used to determine the unique color of the entire hint, it seems relevant to analyze the influence of this parameter on the colorization performance. Different square hint sizes were tested: 1×1, 2×2, 4×4, 7×7, and 8×8. Note that the hint is a square patch in the resized 224×224 image, and that the shape of the respective hint in the original image has the same aspect as the original image and is not necessarily square. As shown in Figure 4.8, sizes 2×2 and 4×4 give the best results. Consequently, the square hint size was set to 2×2. Using a larger hint size will result in inaccurate coloring. The process of determining a hint color is as follows: first, the input color image, resized to 224×224, is down-sampled by the hint size. The chrominance channels of the resulting image are then multiplied by a binary mask with ones at the hint locations and zeros otherwise. Finally, these masked chrominance are upsampled

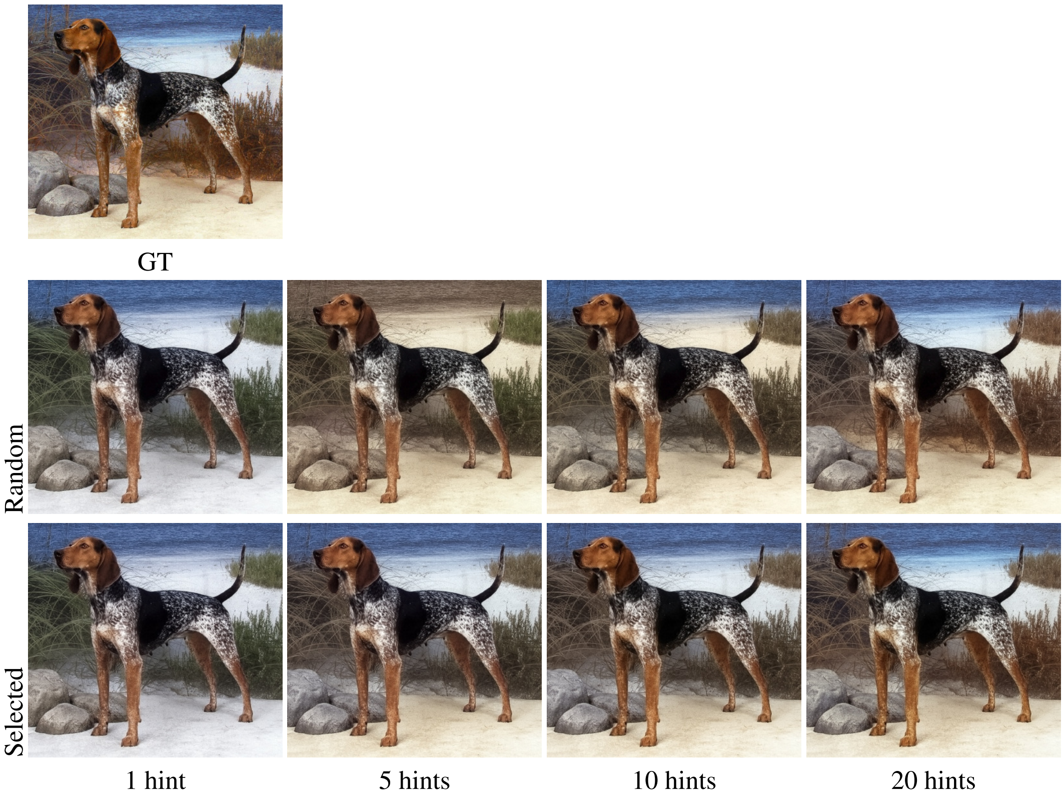


Figure 4.6: iColoriT results for different numbers of hints and hint locations on a dog image from the ImageNet dataset (Russakovsky et al., 2015). First row: Ground truth (GT) input image. First to fourth columns: 1, 5, 10, and 20 hints. Second row: Random hint location. Third row: manually selected hint location.Images from Imagenet (Russakovsky et al., 2015)

back to 224×224 using nearest-neighbor interpolation. This step effectively assigns the sampled color value to the entire spatial extent of the hint, resulting in a uniform color for the whole $k \times k$ patch passed to the network.

4.1.2.2. Natural image datasets

As it was mentioned in 4.1.1.1 the ImageNet 2012 train split (Russakovsky et al., 2015) was the training dataset used for iColoriT. ImageNet ctest10k validation split was used as a standard benchmark for evaluating colorization models. The ImageNet ctest10K is a subset of the ImageNet 2012 validation split. Additionally, the authors tested two other image datasets, CUB (Welinder et al., 2011) and Oxford 102flowers («Automated Flower Classification over a Large Number of Classes», 2008). The iColoriT performance was tested on Imagenet1K (Rus-

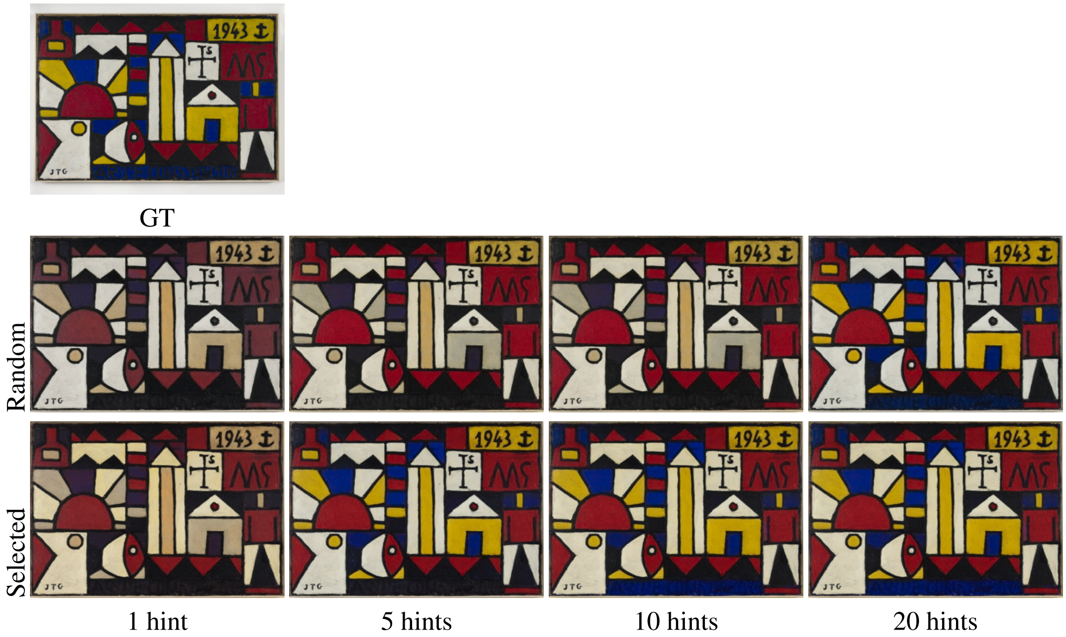


Figure 4.7: iColoriT results for an image from Torres García's painting with different numbers of hints and random hint generation vs. selected by user location. First row: Ground truth (GT) input image. Second row: Random location. Third row: selected location. In the last two rows, first to fourth columns: 1, 5, 10, and 20 hints. Image from IMGM.

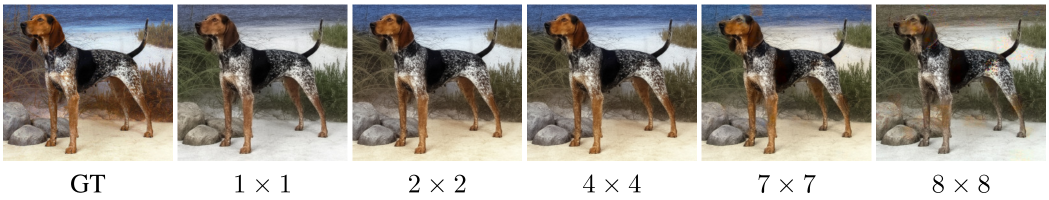


Figure 4.8: Influence of the hint size on iColoriT. From left to right: the ground truth input (GT) and different colorization results with varying hint sizes: 1×1, 2×2, 4×4, 7×7, and 8×8. The same number and position of hints are used in each image. Image from Imagenet (Russakovsky et al., 2015)

sakovsky et al., 2015) -a 1000-label subset of the original ImageNet-, CUB, and the Oxford 102flowers datasets. The following experiments used the Base model and a square-shaped hint of size 2×2.

Imagenet First, the performance was tested on images from Imagenet1K, assuming this is the best scenario. The results are shown in Figures 4.9 and 4.10. It can be observed that for images with straightforward semantic content (as in the first two rows of Figure 4.9), iColoriT doesn't need many hints to achieve a good result; between 5 and 10 hints will be enough, and even 0 hints is a very good result. As the complexity of the semantic content increases, the method requires more hints to display color details, as is evident in the flowers in the third and fourth rows of Figure 4.9 or in both examples of Figure 4.10. In the case of the flowers example, more than 50 hints are needed for a satisfactory colorization result, and even more hints improve the finer details. For the concert example in Figure 4.10, at least 50 hints are needed to color the girl's shirt, and up to 200 hints are required to avoid the greenish background guitarist. For the piano example in Figure 4.10, up to 100 hints are needed to start colorizing the little girl's pants. It is important to note that the method can colorize with zero hints in a fully automatic mode. For these experiments, hints were generated randomly; however, as the number of hints increased, previously generated hints were retained, and new ones were added. For example, to generate 50 hints, the first ten hints were retained, and 40 new ones were randomly generated.

CUB Figure 4.11 shows the iColoriT results for the Caltech-UCSD Birds-200-2011 (CUB)(Welinder et al., 2011) dataset, a public bird image dataset mainly used for classification tasks. The method produces satisfactory colorization above 50 hints in both examples. The observations are similar to those in the ImageNet experiment.

Oxford 102flowers dataset The Oxford 102flowers dataset («Automated Flower Classification over a Large Number of Classes», 2008) consists of 102 different categories of flowers common in the UK. Images in this dataset exhibit large-scale, pose, and lighting variations. This dataset was used for testing, as described in the original paper. Flower images typically exhibit high levels of detail at multiple scales, making them of great interest for evaluating the algorithm's performance on this type of image. The results are shown in Figure 4.12. The number of hints needed to achieve good results in both examples lies between 50 and 100.

Trade-off between priors The goal of these experiments was to evaluate the trade-off between the color hints provided by the user and the color priors learned by the model from the training dataset, and their impact on colorization results. To accomplish this, the demo provided by the authors was tried, which allows users

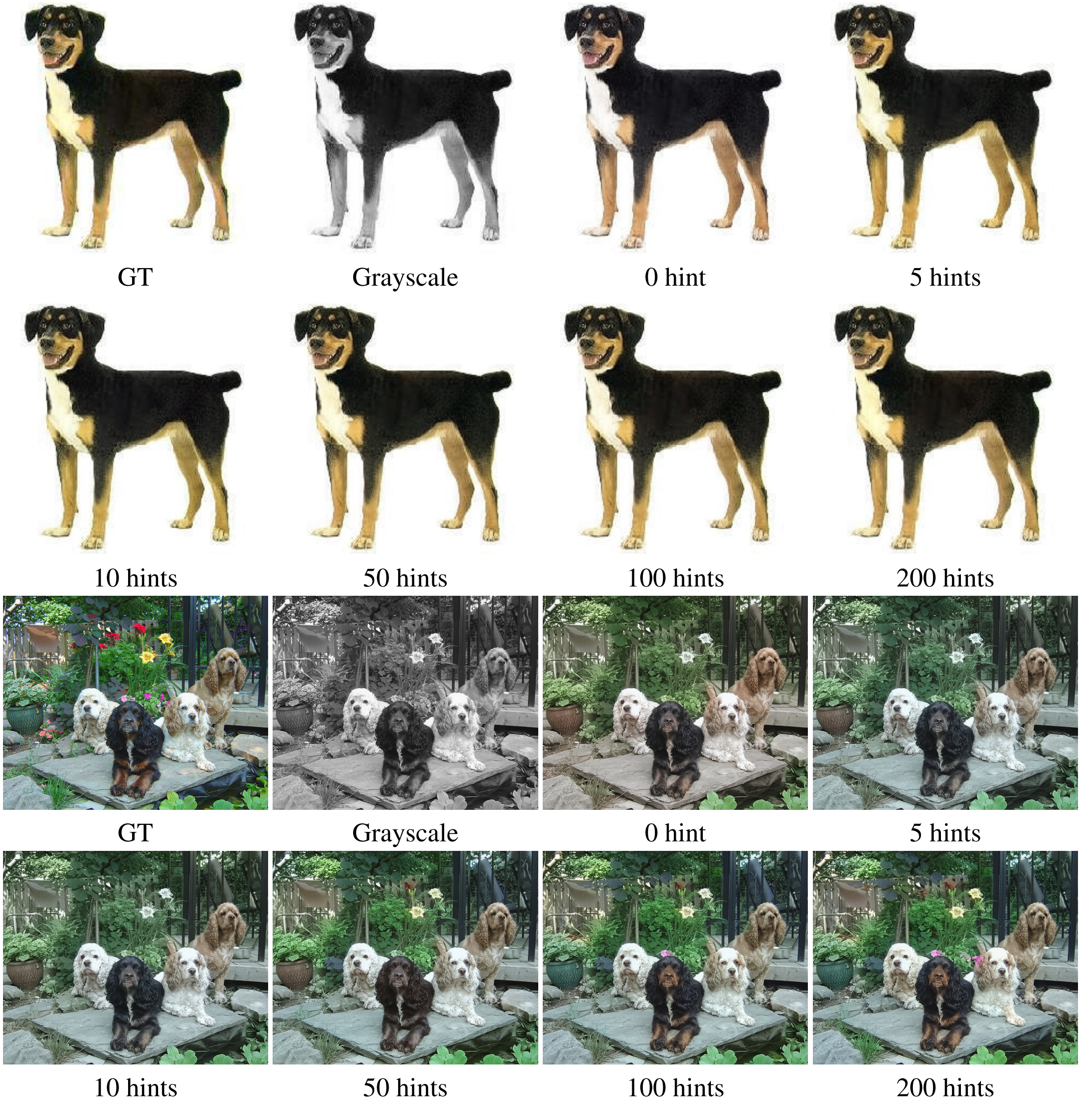
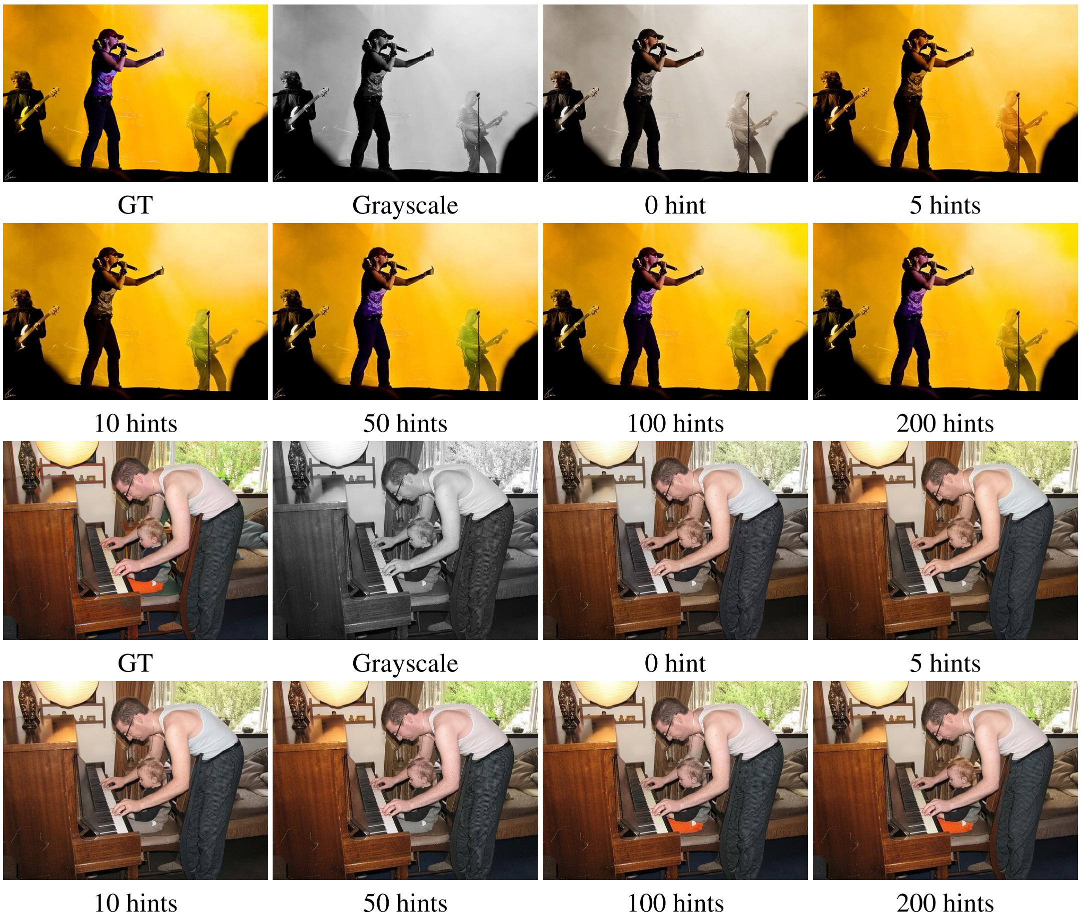


Figure 4.9: Experiment with iColoriT on the Imagenet1K dataset (Russakovsky et al., 2015). The first (rows 1 and 2) shows a very simple semantic image. The second (rows 3 and 4) shows a more complex scene. For each example, several colorization results are shown for an increasing number of hints (0, 5, 10, 50, 100, and 200).

to select different hint colors for a given grayscale input. In Figure 4.13, different hint colors were tried to colorize the same input image. In the left example, three hints were placed to be coherent with the color priors of the image dataset used for training. In the middle example, the same hint locations were kept, and the hint color in the dog's ear was changed from brown to purple. There are no purple-dog color priors in the training dataset, so the propagation of this color hint is limited; the other ear is not colored purple. In the right example, several hints were added to the grass using colors such as purple and pink. However, these colors are inconsis-



GT Grayscale 0 hint 5 hints

10 hints 50 hints 100 hints 200 hints

GT Grayscale 0 hint 5 hints

10 hints 50 hints 100 hints 200 hints

Figure 4.10: iColoriT experiments with the Imagenet1K dataset (Russakovsky et al., 2015). The first (rows 1 and 2) shows a concert scene with fog and blurred people in the background. The second (rows 3 and 4) shows a scene with multiple object instances. For each example, several colorization results are shown for an increasing number of hints (0, 5, 10, 50, 100, and 200).

tent with the color priors learned from the dataset and are therefore not propagated to the rest of the grass. The grass color changes to a less saturated brownish color without clearly accepting these peculiar colors.

4.1.2.3. Paintings

This section focuses on experiments designed to explore how iColoriT colorizes abstract paintings. This is an interesting test, as the colors and semantics of the abstract paintings differ from those represented in the training set.

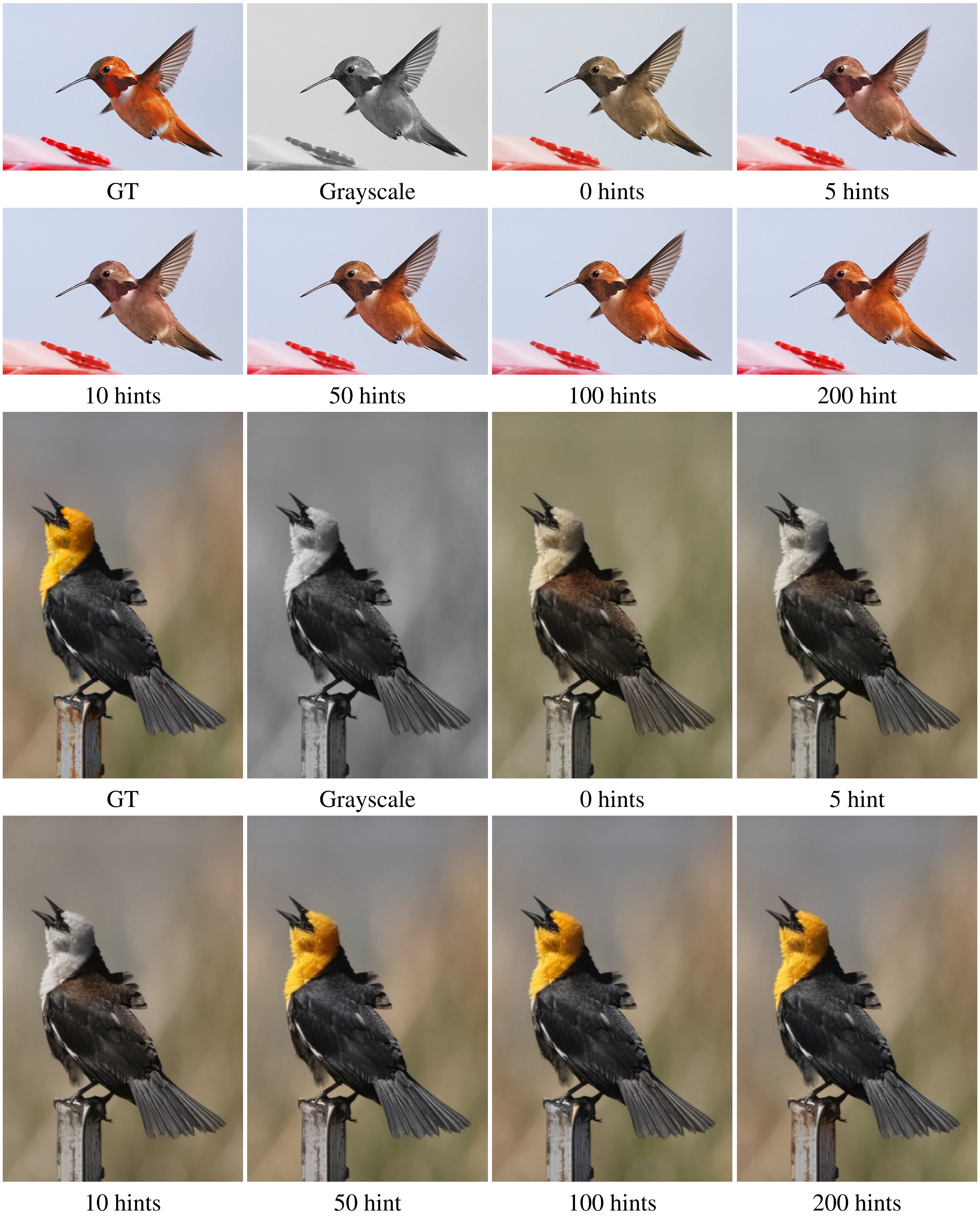
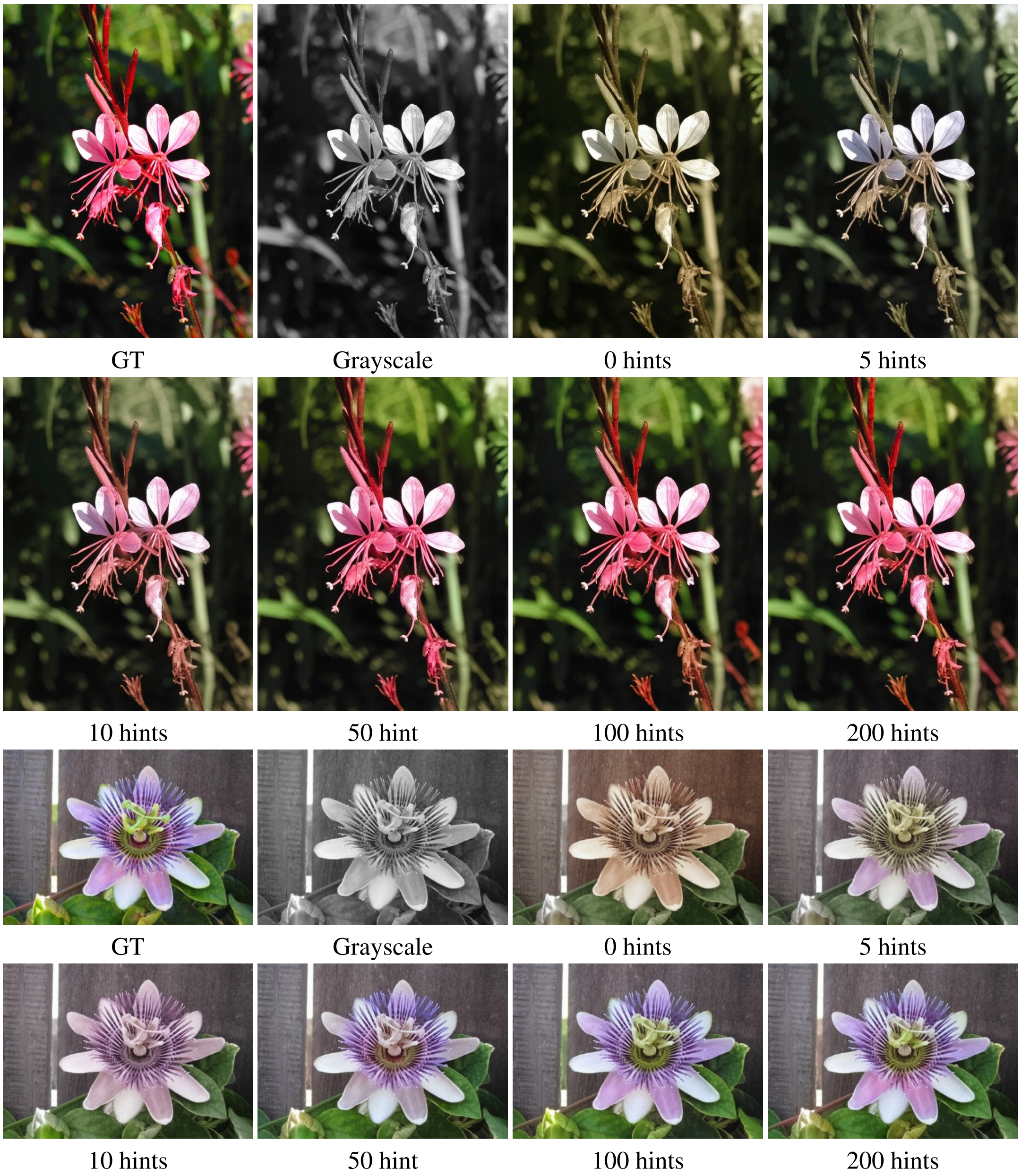


Figure 4.11: iColoriT experiments with the CUB dataset (Welinder et al., 2011) for two bird examples. For each example, several colorization results are shown for an increasing number of hints (0, 5, 10, 50, 100, and 200).

Joaquín Torres García's paintings. Figure 4.7 and the first two rows of Figure 4.14 show experiments using color photographs taken from existing paintings that belong to the Museo Torres García. In the last two rows of Figure 4.14, a Mondrian



GT Grayscale 0 hints 5 hints

10 hints 50 hint 100 hints 200 hints

GT Grayscale 0 hints 5 hints

10 hints 50 hint 100 hints 200 hints

Figure 4.12: iColoriT experiments with Oxford 102flower dataset («Automated Flower Classification over a Large Number of Classes», 2008) for two flower examples. For each example, several colorization results are shown for an increasing number of hints (0, 5, 10, 50, 100, and 200).

painting [1] provides another example of coloring abstract artworks. In general, it can be observed how well this method colorizes the paints. In this case, both Torres García's and Mondrian's paintings yield fairly good results with 50 or more hints.

[1] picryl.com.

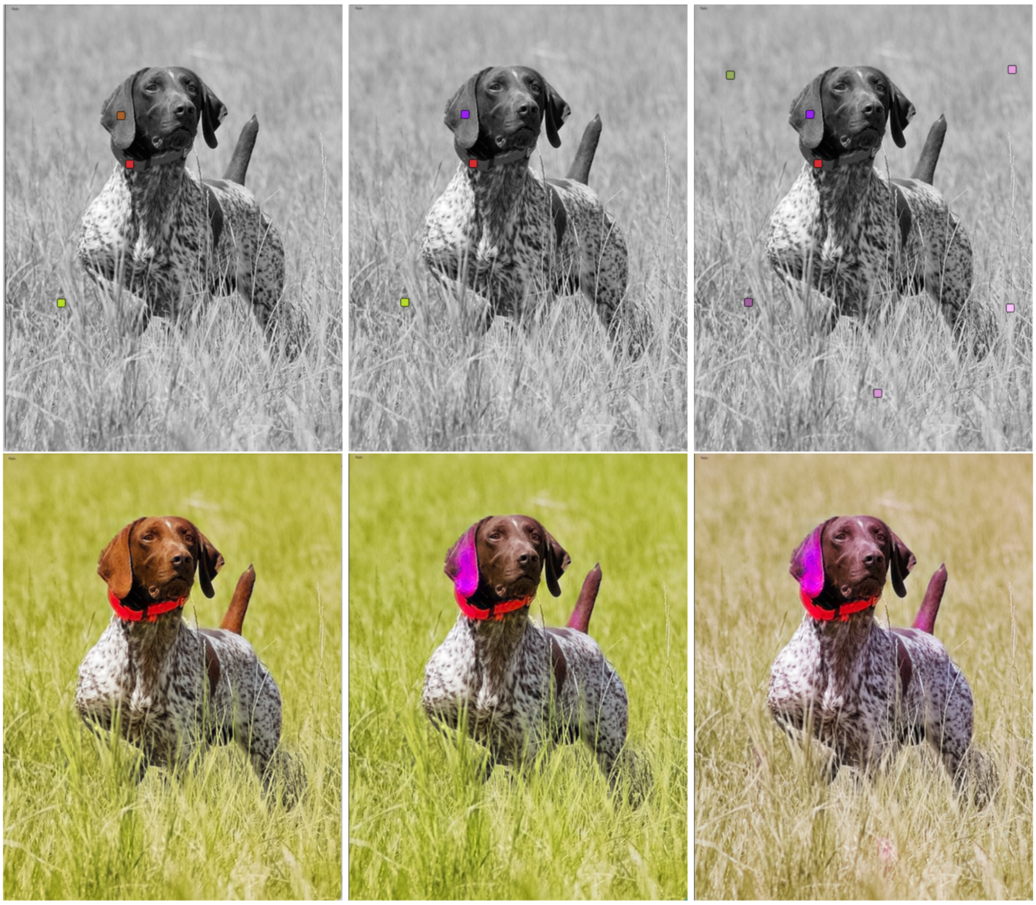
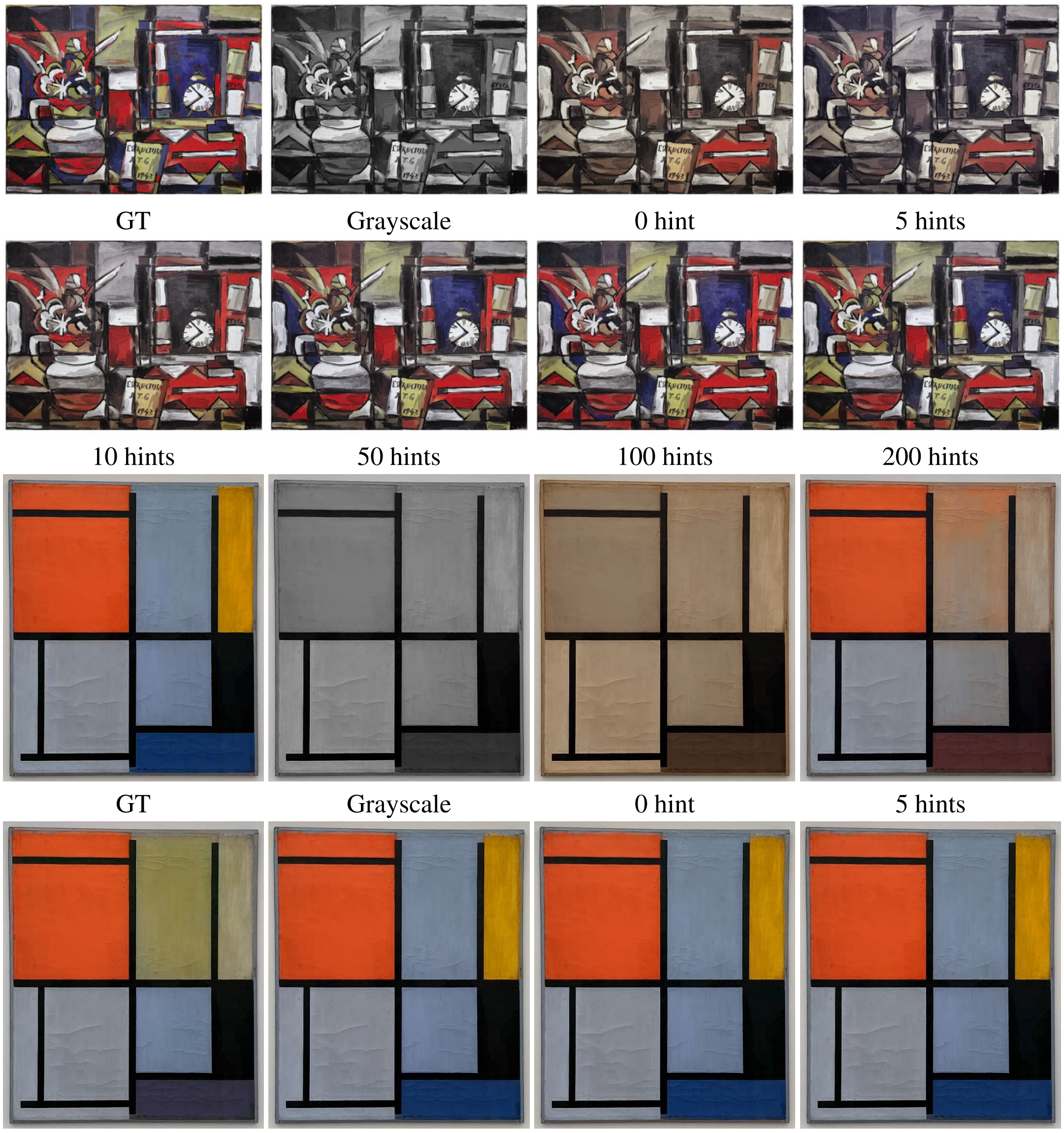


Figure 4.13: iColoriT experiment. Trade-off between color hints and color priors learned from the dataset. Each column displays the grayscale image to be colorized, its color hints at the top, and the resulting colorization at the bottom. Left column: three "normal" color hints. Middle column: unusual purple hint on the dog's ear. Right column: unusual purple and pink hints on the grass. Image from Imagenet (Russakovsky et al., 2015)

4.1.2.4. Discussion

The iColoriT method is a hybrid colorization method that combines color priors learned from a large dataset with color priors provided as color hints. This gives better control over the colorized output. This method generally achieves very good results with fewer than 20 hints for a not-too-complex image when the hints' locations and colors are accurate. Results for images with simple semantics require very few hints, as shown in Figures 4.9, and 4.11, in contrast to images with more complex semantics, as the ones shown in Figures 4.10 and 4.12, which require twice as many hints.

Because of the method's interactive nature, some typical drawbacks of image



GT Grayscale 0 hint 5 hints

10 hints 50 hints 100 hints 200 hints

GT Grayscale 0 hint 5 hints

10 hints 50 hints 100 hints 200 hints

Figure 4.14: iColoriT experiments with Torres García and Mondrian Paintings. Torres García paint: rows 1 and 2. Mondrian paint: rows 3 and 4. From top to down, left to right: Ground Truth (GT), Grayscale, 0, 5, 10, 50, 100 and 200 hints.) Image from IMGM and Public domain images

colorization methods, such as color bleeding and the misrepresentation of the color palette in an image, can be quickly addressed and corrected. Figure 4.15 shows a color bleeding artifact (note the reddish zone in the upper right part of the yellow triangle, produced by the red triangle on top of it). As the second row of Figure 4.15 shows, this is corrected by adding a blue hint in this zone.

Figure 4.16 illustrates a case of color misrepresentation. The background of the

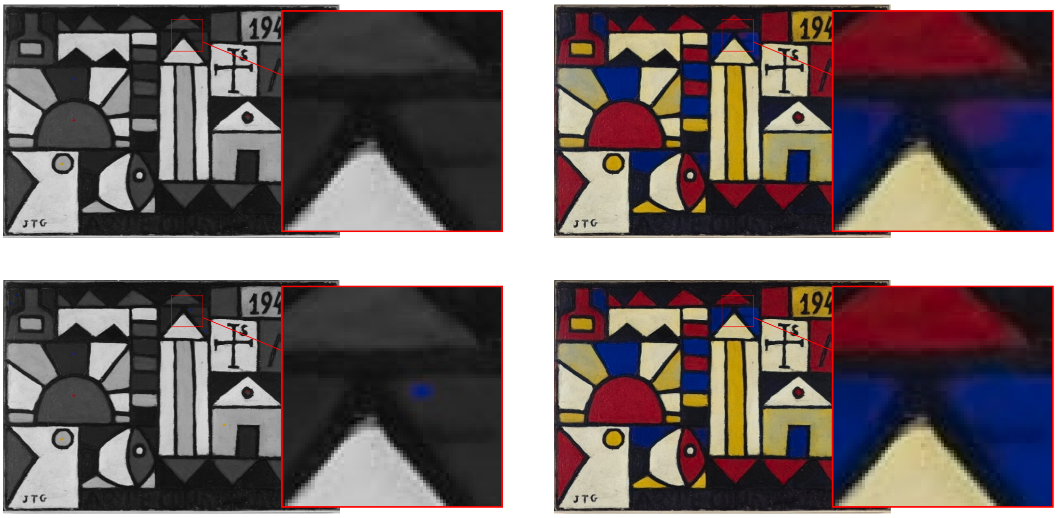


Figure 4.15: iColoriT experiment showing a color bleeding example. Top left: grayscale image to be colorized with a close-up showing no hint of bleeding in the area. Top right: colorized image and close-up of the bleeding area. Bottom left: grayscale image to be colorized with a close-up showing that a blue hint was added in the bleeding area. Bottom right: colorized image and close-up showing the removal of the bleeding artifact. Image from IMGM.

red structure must be blue on both sides, but the left part is gray. This is corrected again by adding an additional blue hint in that region, as seen in the second row of Figure 4.16. This is particularly helpful for tasks that require precise colorization, such as image restoration.

Note the method's behavior when no hints are provided. This case is equivalent to an automatic colorization method. The results obtained, even if the exact colors are not recovered, serve as a good starting point and demonstrate how the model utilizes color priors from the training dataset and color hints when provided.

It's important to note that the results for all experiments in this analysis are obtained using color hints that are almost exactly the ground truth colors (since the average of all pixels in the hint region is used as the hint color). In a real-world colorization problem, it will be necessary to specify the positions of the hints and the color of each hint (without information about the actual colors), which can make the task more difficult and degrade the colorization results.

Another crucial point to consider is the significance of the priors learned by the model, which play an essential role in determining the colorization outcomes. As

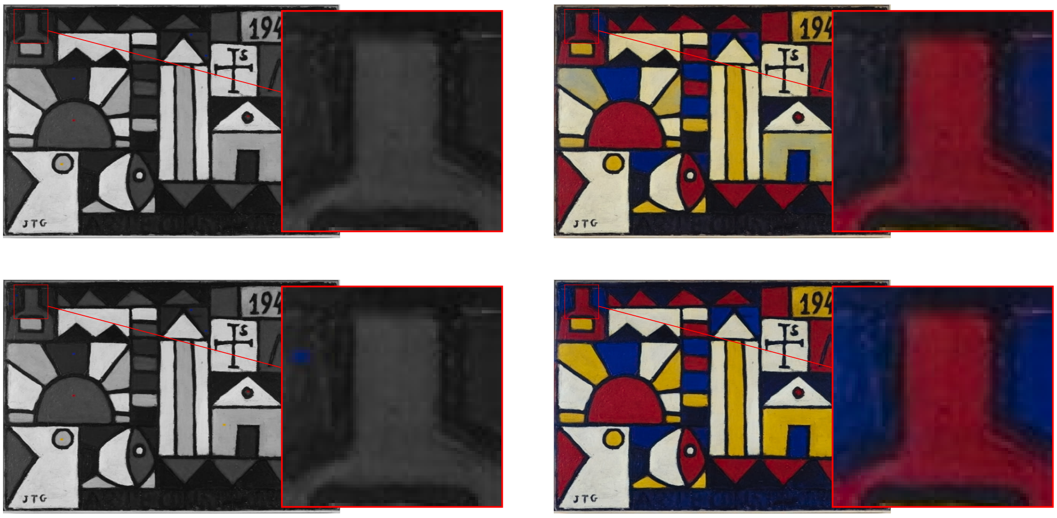


Figure 4.16: iColoriT experiment showing a color inconsistency example. Top left: A grayscale image to be colorized, along with a close-up showing the missing color region. Top right: colorized image and close-up showing that both sides of the bottle are inconsistent; both should be blue. Bottom left: Grayscale image to be colorized and a close-up showing that a blue hint was added in the inconsistent area. Bottom right: Colorized image and close-up showing that both sides of the bottle are now consistent. Image from IMGM.

depicted in Figure 4.13, the number of hints required to achieve satisfactory results varies depending on the color coherence with the color prior learnt from the training dataset. For instance, in the experiment illustrated in Figure 4.13, the purple color isn't typically associated with elements within the dog image. Therefore, iColoriT restricts the propagation of the violet hint even when a large number of violet hints are provided. This aspect is especially relevant for art image restoration problems, where the semantic–color relationships often differ from those observed in natural image datasets.

While the results on artistic images are visually promising (Figures 4.7 and 4.14), and the model demonstrates capability in automatic colorization (Figures 4.9 and 4.10 with 0 hints), it is important to note the high degree of user intervention required. As observed in the experiments, achieving satisfactory results on abstract paintings typically necessitated 50 or more hints. This dependency on dense guidance is a limitation for restoration tasks. Consequently, adapting this colorization model to artistic datasets is desirable not only to incorporate painting semantics, which differ significantly from those of natural images, but also to improve the model's internal priors. This adaptation aims to achieve high-fidelity restorations with significantly fewer hints, reducing the burden on the expert user.

4.2. Model Adaptation and Transfer Learning

Encouraged by the results presented in the previous section, the next step was to adapt iColoriT to the artistic domain through targeted fine-tuning. To this end, several transfer learning strategies were explored to adapt the pretrained colorization model to the artworks datasets defined in Chapter 3.

4.2.1. Transfer Learning Strategies

To explore transfer learning strategies, an additional dataset was assembled from available materials. It was constructed by combining *partial IMGM*, *IMGCec*, *IMGL*, and *IMGP*, resulting in a total of 1,400 images. The same data–split ratios defined in Section 3 were applied: 70% for training, 15% for validation, and 15% for testing. This dataset is not described in the main dataset chapter 3, as it was constructed at an earlier stage of the project, prior to the completion of the final data collection process. Although it includes images from sources that were later expanded, it does not reflect the full extent or quality of the final datasets. In particular, it contains no images from the Taller Torres García and includes fewer high-quality reproductions than those ultimately obtained. As such, it is used exclusively for exploratory experiments and not for the core evaluation of restoration performance.

Three Transfer Learning strategies were evaluated:

- **Full Fine-Tuning (FFT):** all parameters were updated during training, enabling maximum adaptation but with high computational cost and high risk of overfitting.
- **Partial Fine-Tuning or Frozen Blocks (FBFT):** early transformer layers were frozen, while later layers were retrained to adapt higher-level representations.
- **LoRA:** Low-Rank Adaptation (Hu et al., 2022) was applied to the attention blocks, significantly reducing the number of trainable parameters while maintaining adaptability.

To evaluate these three strategies, several models were trained based on the Base iColoriT architecture, which comprises 12 transformer blocks. For strategies 1 and 2, the original training script provided by the authors was used with one modification: the ability to freeze specific blocks during the training. For the third strategy, which does not rely on freezing model blocks, the code was adapted to introduce additional low-rank trainable matrices within the attention layers. The pretrained Base colorization model provided by the authors was the starting point for all strategies. Each configuration corresponds to a model in which a defined number of transformer blocks remain frozen, while the rest are updated during training.

Strategies 1 and 2 are described and analyzed in Section 4.2.2 and strategy 3 is analyzed separately in Section 4.2.4 because it differs substantially from the previous two approaches.

4.2.2. Strategies 1 and 2: Full and Partial Fine-Tuning.

In the first strategy (FFT), no transformer blocks were frozen (i.e., zero blocks frozen), so all model parameters were updated during training. Preliminary experiments revealed that updating all parameters did not improve performance relative to the Base model; in fact, performance degraded.

A visual inspection of the colorized outputs revealed critical limitations of the first strategy. Both the FFT model (from scratch and from the Base model weight) exhibited degradation in image quality. As shown in Figure 4.17, the generated images with the FFT from scratch lacked chromatic coherence, appearing sepia when no hints were applied (automatic) or presenting chaotic color distributions when 20 hints were applied.

The FFT from the Base model generally introduced a distinct type of artifact when subjected to user guidance (20 hints). As illustrated in Figure 4.17, the output displayed visible red "color patches" in the pan handle. This phenomenon suggests a failure in the *local stabilizing layer*'s ability to smooth the transitions produced by the *pixel shuffling* upsampling operation. While the stabilization layer is designed to mitigate blocking artifacts, the domain shift between natural images and Torres García's geometric abstractions—combined with the constraints of 20 hints—appears to result in disjointed patch predictions.

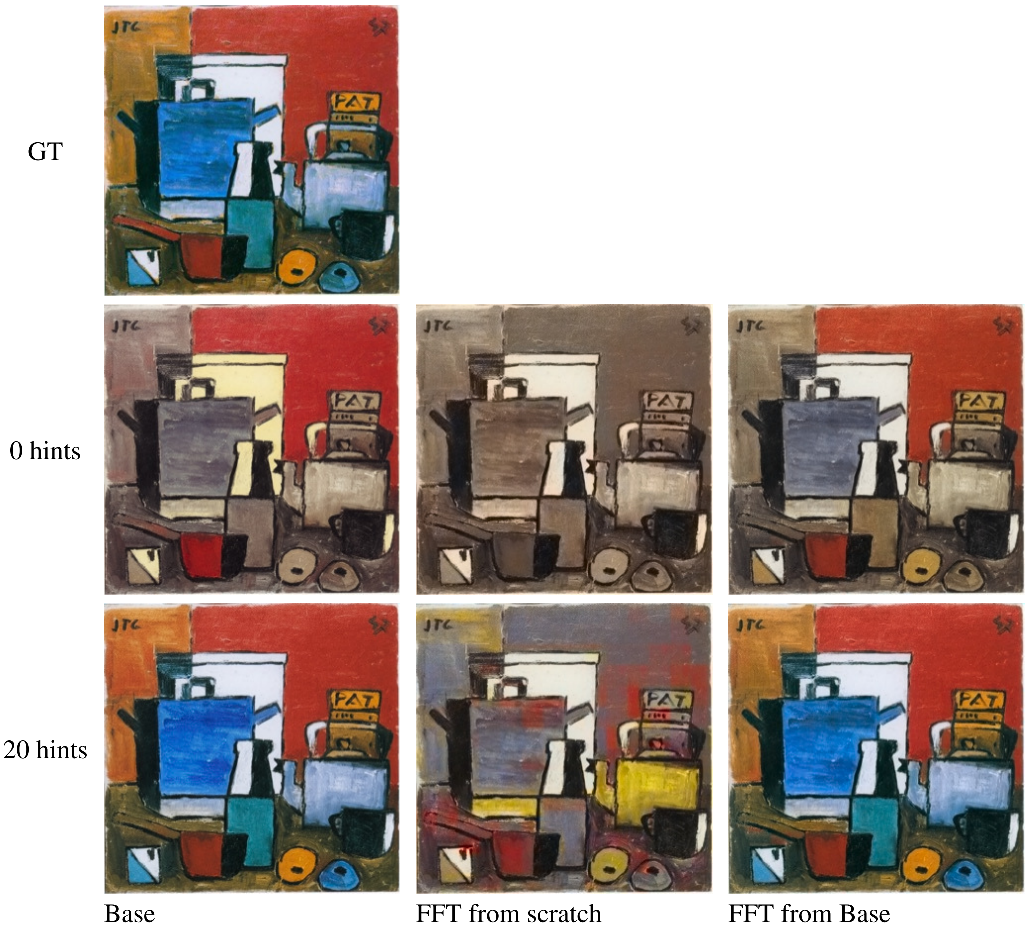


Figure 4.17: Visual comparative colorization results. Painting 1947.14, at different hint levels, 0 (automatic) and 20 hints (where hints were manually placed by a user). Rows from top to bottom: Ground Truth (GT), 0, and 20 hints. Columns show inferences: Base (left), Full fine-tuning from scratch (middle), and full fine-tuning from the base model (right). Image from IMGM

Nevertheless, this configuration—representing the most challenging optimization landscape due to the high number of trainable parameters was utilized to benchmark the learning rate schedulers. As detailed in Section 4.2.3, analyzing the loss evolution under this setting allows us to isolate and assess the impact of the chosen learning rate scheduler.

For the second strategy (FBFT), several configurations were explored by freezing different numbers of early transformer blocks, while retraining the remaining ones. These configurations are summarized in Table 4.1. To evaluate this strategy, we used the PSNR between the ground-truth and predicted chrominance values, computed as the average PSNR across a subset of 12 test images. The selected images were: 1947.14, 1928-129, 1929-54, 1929-64, 1930-14, 1930-30, 1932-38, 1932-49, 1932-51, 1933.01, 1933-29, and 1936-0220; image identifiers follow the original catalog nomenclature. Two experimental conditions were considered: fully automatic colorization (0 hints) and interactive colorization (20 hints), in which hints were manually placed by a user, allowing their spatial locations and chromatic values to be deliberately controlled. This reduced subset was intentionally selected for an exploratory analysis, as the manual nature of the interactive evaluation made large-scale testing impractical at this stage.

FT-Frozen Blocks	PSNR 0 hints	PSNR 20 hints
Base	24.31	30.39
6	24.51	26.24
8	25.04	26.89
10	25.38	27.93
11	25.98	28.20

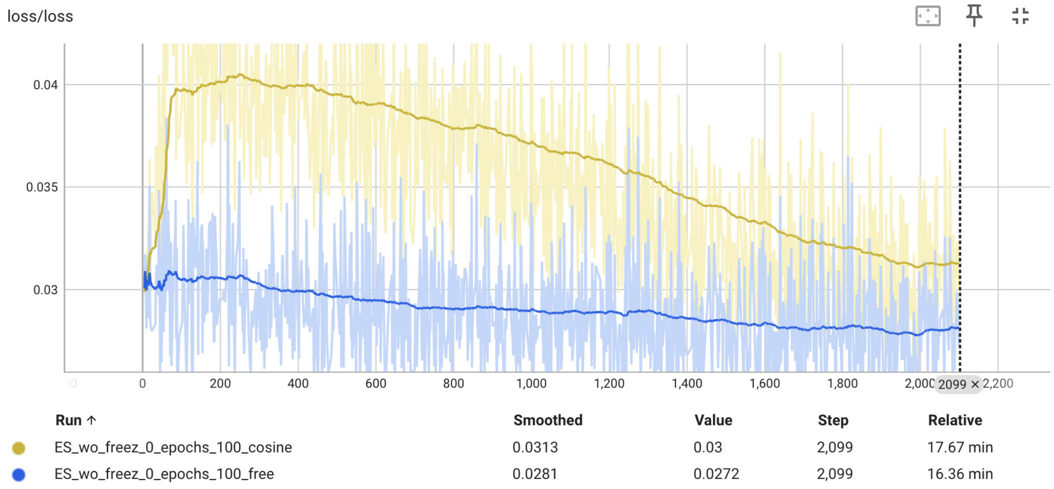
Table 4.1: Comparison of different Fine-Tuning configurations. From top to bottom, the Base model serves as a baseline, followed by models with an increasing number of frozen blocks, from 6 to 11. The metric corresponds to the PSNR values in the cases with 0 and 20 hints.

The results in Table 4.1 suggest that, in the fully automatic setting (0 hints), fine-tuned models generally outperform the Base pretrained model. However, in the interactive setting (20 hints), the Base model without fine-tuning achieves higher PSNR values, indicating a better propagation of user-provided color hints. This behavior suggests that fine-tuning improves automatic colorization, though it may slightly reduce the model's ability to generalize from sparse external guidance given the amount of available training data.

Given that transformer-based architectures typically require large amounts of data to be effectively trained, a more parameter-efficient approach was considered. Specifically, a LoRA-based fine-tuning strategy was applied to this transformer, as it enables model adaptation with significantly fewer training samples while preserving representational capacity. This is described in Section 4.2.4.

4.2.3. Effect of Learning Rate Scheduling.

As we mentioned in Section 2.5.3, the learning rate and its scheduling play a critical role in shaping the optimization trajectory and convergence behavior of deep models. In this subsection, we empirically analyze the effects of different learning rate schedules on the model's fine-tuning stability. The original iColoriT training setup uses a cosine learning rate schedule, as we mentioned in Section 4.1.1. In this experiment, we compare this default configuration with the schedule-free alternative, initializing both from the same learning rate value reported by the authors (lr= 1×10^{-5}.)



	Run ↑	Smoothed	Value	Step	Relative
●	ES_wo_freez_0_epochs_100_cosine	0.0313	0.03	2,099	17.67 min
●	ES_wo_freez_0_epochs_100_free	0.0281	0.0272	2,099	16.36 min

Figure 4.18: Loss evolution during 0 fixed blocks fine-tuning runs starting from the Base pretrained weights (wo). Both experiments utilized Early Stopping (ES) and 100 epochs, with an initial learning rate of 1×10^{-5}. The yellow curve corresponds to the *cosine* learning rate scheduler, while the blue curve shows the *free* schedule.

As illustrated in Figure 4.18, the *free* schedule achieved a lower and more stable loss throughout training, indicating faster and more consistent convergence. Additional primary tests, initialized from scratch and/or a pretrained base model with partially frozen transformer blocks, confirmed the same trend. Based on these results, the *free* schedule was adopted for all subsequent fine-tuning experiments.

4.2.4. Strategy 3: Low-Rank Adaptation (LoRA)

4.2.4.1. LoRA adaptation

As mentioned in Section 2.6.1, LoRA introduces trainable low-rank matrices that adapt the pre-trained weights of a model without directly modifying the original parameters. Formally, each weight matrix $W \in \mathbb{R}^{d \times k}$ is redefined as:

$$W' = W + \Delta W, \quad \text{where } \Delta W = BA,$$

with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$ denoting the rank of the adaptation (Lora rank). Only the low-rank matrices A and B are optimized during fine-tuning, while the original weights W remain frozen. This approach substantially reduces the number of trainable parameters while preserving the model's representational capacity ($\frac{\text{params}_{\text{LoRA}}}{\text{params}_{\text{full}}} = \frac{r(d+k)}{dk} \leq \frac{2r}{\min{(d,k)}} << 1$).

The implementation used in this work relies on the open-source `loralib` library released by Microsoft Research,[1] which provides a practical implementation of the method described by Hu et al. (Hu et al., 2022). LoRA was integrated into the attention mechanism of iColoriT by replacing the linear layers responsible for the query, key, and value projections (corresponding to weight matrices W_Q, W_K, W_V) and the output projection (W_O, denoted as `proj` in the implementation) with their low-rank `lora.Linear` counterparts. This modification was encapsulated in the `LoRAAttention` class, enabling all Transformer blocks to leverage LoRA-based adaptation while preserving the original architecture and pretrained weights.

4.2.4.2. LoRA experiments

In the LoRA experiments, multiple configurations were tested to analyze the effects of the amount and quality of the Train Data, the rank (r), and the learning rate (lr) on model performance. For each dataset (D_1–D_4) defined in Section 3.1, several LoRA fine-tuning runs were conducted. Lower-rank configurations ($r \in \{1, 4\}$) significantly reduce the number of trainable parameters and computa-

[1] https://github.com/microsoft/LoRA

tional cost, but they may also limit the model's ability to capture more complex feature interactions. Conversely, higher ranks ($r \in \{8, 32\}$) increase the adaptive capacity at the expense of greater memory and training time requirements. This trade-off was systematically evaluated to determine the most effective configuration for colorization tasks. The number of trainable parameters for the considered ranks r is detailed in Table 4.2.

Technique	Trainable Parameters
Full fine-tuning	89.478 million
LoRA rank 32	1.769 million
LoRA rank 8	0.442 million
LoRA rank 4	0.221 million
LoRA rank 1	0.055 million

Table 4.2: Comparison of trainable parameters between full fine-tuning of IColorit and LoRA configurations for different ranks r.

4.2.5. Hyperparameter Tuning and Early Stopping.

Hyperparameters were tuned using the validation split of D1 (see Section 3.2.1). In particular, three initial learning rates were identified via a random search (sampling from predefined ranges) to locate promising regions—defined as those exhibiting a rapid and stable decrease in training loss during the initial epochs—and subsequently refined using a grid search (systematic exploration over a fixed set) based on validation performance. Since the original training procedure did not implement early stopping, the pipeline was extended to include it. Early stopping is driven by the average PSNR computed over three validation regimes corresponding to 1, 10, and 100 number of hints. This aggregated metric is evaluated every five epochs, and training is terminated if no improvement is observed for ten consecutive validation checks. The best model checkpoint was automatically saved whenever a new maximum of the monitored metric was reached.

Fine-Tuning. In total, 48 LoRA models were trained to systematically evaluate the adaptation performance. The training configurations were defined as follows:

- **Datasets:** Four distinct datasets (D1, D2, D3, and D4) were utilized to assess the impact of data quantity and domain specificity on the model's ability to learn the artist's palette. As described in Chapter 3, these datasets represent a hierarchy of progressively inclusive corpuses, ranging from a curated core of Torres García's artworks to a broader collection incorporating diverse archival sources.

- **LoRA ranks:** The rank $r \in \{1, 4, 8, 32\}$ controls the number of trainable parameters and, consequently, the adaptive capacity of the fine-tuning process. Lower ranks ($r = 1, 4$) were tested to evaluate parameter efficiency and regularization, while higher ranks ($r = 8, 32$) were included to assess whether a higher intrinsic dimension enables the capture of more complex stylistic features in the artistic domain.
- **Learning rate range:** The learning rates were selected from the set $\{1 \times 10^{-2},\ 1 \times 10^{-3},\ 1 \times 10^{-5}\}$. These values were identified through the previously described preliminary random and grid search process.

4.2.6. Training Settings

Computational Resources. All fine-tuning experiments were performed on the national high-performance computing platform of Uruguay, hosted at the *Centro Nacional de Supercomputación* («ClusterUY: Centro Nacional de Supercomputación», n.d.) and located in the Datacenter Ing. José Luis Massera (ANTEL, Canelones). ClusterUY aggregates multiple computing nodes under a unified management system, providing over 10,000 traditional CPU-equivalent cores for scientific computation across the nation. The infrastructure was funded by the *Agencia Nacional de Investigación e Innovación* (ANII) and the *Comisión Sectorial de Investigación Científica* (CSIC), and is operated collaboratively through the Fundación Ricaldoni.

Specifically, the experiments in this work were conducted on DELL EMC PowerEdge 7525 nodes (`node40`–`node46`), each equipped with 128 AMD EPYC 7763 CPU cores and ranging in RAM from 256 GB to 512 GB. Nodes `40` and `46` included dual NVIDIA A40 GPUs (48 GB VRAM each, 10,752 CUDA cores, 336 Tensor cores, and 84 RT cores), providing high-performance acceleration for the fine-tuning experiments.

I gratefully acknowledge Gregory Randall and the Signal Processing Department (DPS) of the Electrical Engineering Department (IIE) in the Facultad de Ingeniería (FIng) of the UdelaR for enabling access to these resources.

Training Configuration. All LoRA experiments utilized the Base-pretrained model. Only the LoRA additional parameters were updated, while all pretrained weights remained frozen, corresponding to a parameter-efficient fine-tuning (PEFT) setup.

The original input resolution (224×224) was used, incorporating the dynamic hint generation strategy defined by the authors (Yun et al., 2023) during training to simulate diverse user interactions. This process was controlled by three key parameters:

- `num_hint_range` $= [0, 128]$: Defines the bounds of the uniform distribution $\mathcal{U}(0, 128)$ from which the number of hints is sampled for each training instance. This forces the model to learn color propagation across varying levels of information density, from fully automatic colorization (0 hints) to dense guidance (128 hints).
- `hint_size` $= 2{\times}2$: Specifies the spatial extent of each hint on the grid. As analyzed in Section 4.1.2.1, this size was selected because larger patches (e.g., $7{\times}7$) can introduce color inaccuracies, while $2{\times}2$ offers an optimal trade-off between visibility and precision.
- `avg_hint` $=$ True: Indicates that the color value assigned to a hint is computed as the average of the ground truth chrominance values within the $2{\times}2$ patch, rather than sampling a single pixel. This ensures the hint represents the local region robustly.

While training involved stochastic hint generation, validation was performed with fixed hint counts of $\{1, 10, 100\}$. These discrete levels were intentionally selected to benchmark the model's behavior across distinct regimes: minimal, moderate, and dense guidance.

Optimization was performed using AdamW as in the original work. Instead of a cosine scheduler, the schedule-free strategy was employed (see Sections 2.5.3 and 4.2.3), enabling the optimizer to dynamically adjust the learning rate without an explicit decay schedule. Training was run for 1000 epochs, with Early Stopping (ES) added. Other parameters included a batch size of 32 and automatic checkpoint saving every 10 epochs.

To complement the PSNR distortion metric used in the original code, the perceptual LPIPS metric (R. Zhang et al., 2018) was incorporated into the validation pipeline. LPIPS (Learned Perceptual Image Patch Similarity) evaluates the perceptual distance between corresponding image patches by comparing deep feature activations extracted from pretrained neural networks. Higher LPIPS values indicate greater perceptual differences, whereas lower values indicate greater similarity. The metric can be computed using features from different backbone architectures (SqueezeNet, AlexNet, or VGG), each calibrated with a learned linear layer; in this work, LPIPS was computed using the VGG-based variant (version 0.1).[1]

All metrics, PSNR, LPIPS, and the generalization gap (defined as the difference between training and validation PSNR) were logged via TensorBoard for both the training and validation subsets.

Train Dataset Construction. The training dataset was built using the `build_pretraining_dataset` function provided by the authors, which applies a data augmentation pipeline `DataAugmentationForiColoriT`, as defined by the authors, to each image. This pipeline performs two key operations: dynamic image augmentation and random hint generation.

Each image is first augmented through random spatial transformations to increase data diversity and prevent overfitting. Specifically, the function `RandomResizedCrop` randomly crops and resizes the image to the target input size, introducing scale and position variability, while `RandomHorizontalFlip` (added in this work) randomly mirrors the image along the vertical axis to promote invariance to object orientation. These operations are followed by tensor conversion and normalization to prepare the data for the model's processing. They ensure that the same image appears differently in each iteration, thereby improving generaliza-

[1] https://github.com/richzhang/PerceptualSimilarity

tion and robustness to composition changes.

In parallel, the `RandomHintGenerator` function creates a new hint mask every time, with a variable number and spatial distribution of color hints according to the defined range (`num_hint_range` = [0,128]). This process simulates user input during training, forcing the model to learn color propagation across diverse positions of hints.

Because the `DataLoader` retrieves data dynamically from the dataset rather than from a precomputed cache, both the augmented images and the generated hints change on each epoch. Consequently, training is effectively performed on a continuously changing sample distribution.

Chapter 5

Evaluation and Results

This chapter presents the evaluation framework and the quantitative and qualitative analyses of the experimental results. The analysis aims to assess how different LoRAs' fine-tuning strategies impact colorization performance across varying levels of hint guidance and identify which configurations achieve the most stable and perceptually convincing results.

The first part explains the evaluation framework designed for the training models. The second part focuses on the quantitative evaluation of models using the PSNR and LPIPS metrics, examining their behavior across different configurations of hints, training datasets, LoRA ranks, and learning rates. The third part presents a visual inspection and comparison of the colorized outputs, highlighting color plausibility, structural coherence, and the effect of hints on image fidelity.

5.1. Evaluation Pipeline

To ensure a controlled and reproducible comparison between models and hint configurations, a fixed-hint evaluation pipeline was implemented. This process combines deterministic hint generation, systematic inference, and metric aggregation across multiple realizations. Both the validation (V_1 and test splits (T_1 of D_1

were used, with 63 images in each.

Hint Generation. For each ground truth (GT) image, 100 independent realizations of random hint coordinates were pre-generated, considering six cases of hint numbers: $\{0, 1, 10, 20, 50, 100\}$. Each realization, indexed by an integer $k \in \mathbb{Z}$ and $0 <= k < 100$, corresponds to a distinct random sampling of hint positions stored in separate subdirectories of the form `k/h2-nZ`, where `h2` indicates that each hint covers a $2{\times}2$ pixel region, and `Z` denotes the number of hints. Each file contains pixel coordinates in the format `[y,x]`, one per line, while files for $Z = 0$ are empty.

This step guarantees that all models are evaluated using identical hint sets, eliminating stochastic variation during inference and ensuring fair model-to-model comparisons. Moreover, performing k independent realizations with different hint locations improves the robustness of the results by mitigating biases associated with specific spatial configurations of the hints.

Inference. For each of the 49 models (48 trained models and the Base model) and for each case of the number of hints, the inference script processed the 100 random realizations of hint positions, producing the corresponding set of colored outputs. A total of 1.852.200 test (validation) results were generated for the 63 test (validation) images, six cases of number of hints, and 100 random realizations of hint positions. Each inference used the corresponding precomputed coordinate file so that hint positions remained consistent across models. As mentioned earlier, the model operated in the Lab color space, where the colorized image is the concatenation of the input luminance channel L and the predicted chrominance channels (a, b). Outputs were subsequently converted back to RGB. All color space transformations followed the implementation provided by the original authors in `utils.py`, consistent with the color representation principles described in Section 2.1.

Metric Evaluation. Each predicted image was compared against its ground truth using both quantitative and perceptual metrics. PSNR measured pixel-wise fidelity, while LPIPS (R. Zhang et al., 2018) quantified perceptual similarity through deep feature distances. Metrics were computed independently for each model, random realization of hint positions, and number of hints, and later aggregated to analyze the statistical behavior of each fine-tuning setting under varying levels of user guidance.

Metrics Aggregation. After inference, all per-image metric results were aggregated into summary files, enabling direct comparison. For each model, metrics were averaged across realizations and grouped by number of hints, yielding two levels of aggregation: per-image averages and global averages per hint. This produced compact CSV summaries for each model containing the mean *PSNR* and *LPIPS* values.

A final global aggregation step combined all model results from both validation and test splits into unified summary tables. Each entry was annotated with the corresponding model identifier and split type, enabling cross-model and cross-split comparisons. These consolidated files serve as the quantitative foundation for the analyses presented in the following section.

In addition to numerical evaluation, a qualitative visual comparison was performed to assess:

- Color plausibility and coherence.
- Preservation of structural and semantic elements.
- Effect of hints.

The visual analysis was conducted on the best-performing models identified from the quantitative results: Best PSNR, Best LPIPS, and compared with the Base model inference and Ground Truth images. For this purpose, both the validation (V_1) and test (T_1) subsets of D_1 (see Section 3.2.1) were used. Unlike the automatic evaluations described previously, the visual inspection employed manually placed hints for each image (independent of the model), generated using the interactive demo developed as part of the MLBriefs paper (García et al., 2024a). This setup allowed a direct qualitative assessment of the models' ability to propagate user-guided color information while preserving the original structural and artistic construction of each image. Representative examples are presented in Section 5.3.

5.2. Quantitative Analysis

The quantitative analysis is based on the metrics described in Section 5.1. Each model was evaluated using 100 random realizations of hint positions for six cases of hint numbers $\{0, 1, 10, 20, 50, 100\}$. This was done using both the validation (V_1) and test (T_1) splits of the D_1 dataset. The average PSNR and LPIPS values were computed for each random position, image, and hint number. These values, mean PSNR and mean LPIPS, were then compared across models to identify overall trends. Models are defined using the notation $r = i$, with $i \in \{1, 4, 8, 32\}$ indicates the LoRA rank, $lr = k$, with $k \in \{1e^{-5}, 1e^{-3}, 1e^{-2}\}$ the learning rate, and D_j, with $j \in \{1, 2, 3, 4\}$ the training dataset.

5.2.1. Performance Trends Across Configurations

5.2.1.1. Hints variation

Table 5.1 summarizes the best-performing LoRA configurations for each case of hint number on the T_1. For each case, the corresponding LoRA rank (r), learning rate (lr), and training dataset (D_i) variant are reported along with the resulting metric values. The results confirm that the performance improves consistently with the number of hints, with mean PSNR increasing and mean LPIPS decreasing as additional chromatic information is provided. However, the table also reveals some heterogeneity among the top-performing models: no single configuration dominates across all hint levels or metrics. The most frequent winner corresponds to the configuration r = 32, lr = 1e^{-3} and training dataset D_4, which achieves the best results in most cases. *PSNR-best-Z* denotes the model with the highest mean PSNR per hint number, while *LPIPS-best-Z* denotes the model with the lowest mean LPIPS per hint number. Reported PSNR and LPIPS values correspond to averages over the 100 random spatial realizations of hint positions per image.

In addition to the analysis by number of hints, an aggregated evaluation was performed by averaging the PSNR and LPIPS metrics of each model evaluated on T_1 across all configurations of quantity and position of hints. In other words, the cor-

Hints	PSNR-best-Z				LPIPS-best-Z			
	r	lr	Training set	PSNR [dB]	r	lr	Training set	LPIPS
0	32	1e^{-3}	D_1	26.2261	8	1e^{-3}	D_4	0.1175
1	32	1e^{-3}	D_4	27.4241	8	1e^{-3}	D_4	0.0962
5	32	1e^{-3}	D_4	29.5134	32	1e^{-3}	D_4	0.0781
10	32	1e^{-3}	D_4	30.5664	32	1e^{-3}	D_4	0.0696
20	32	1e^{-3}	D_4	31.7324	32	1e^{-3}	D_4	0.0612
50	32	1e^{-5}	D_4	33.2380	32	1e^{-3}	D_2	0.0518
100	32	1e^{-5}	D_4	**34.2662**	32	1e^{-5}	D_4	**0.0455**

Table 5.1: The best-performing LoRA configurations for each number of hints, evaluated on T_1. For each hint number (Z), the table reports the LoRA rank (r), learning rate (lr), the training dataset D_i, and the corresponding metric value. PSNR-best-Z and LPIPS-best-Z are denoted in bold.

responding metric was averaged for each model over all images, 100 random positions, and all cases of the number of hints. These scores are denoted as *PSNR-mean* and *LPIPS-mean*, respectively. Table 5.2 lists the top five model configurations for PSNR-mean and LPIPS-mean. The left values present the results in descending order of PSNR-mean, while the right values presents the results in ascending order of LPIPS-mean. Although PSNR and LPIPS capture different aspects of quality and are not expected to rank models identically, the results reveal a strong alignment between the two.

Moreover, the model with r = 32 and lr = 1e^{-3}, trained on D_4, achieves the best overall quantitative performance, with a PSNR-mean of 30.298 dB and the second-lowest LPIPS-mean value of 0.0749. Hereafter, the model configuration r = 32, lr = 1e^{-3}, and D_4 is referred to as *Best-Global*. Very close results are observed for lower-rank models (r = 8)), particularly when trained on the same dataset (D_4) and with the same learning rate, suggesting that the difference in performance between ranks 8 and 32 remains marginal in this setup.

Note that, unlike the PSNR-mean ranking, the top-five LPIPS-mean configurations include one additional configuration (r = 4, D_3, lr = 1e^{-3}) that does not appear among the top-5 PSNR-mean performers. The configuration that achieved the best LPIPS-mean score is (r = 8, D_4, lr = $1e^{-3}$), followed by the *Best-Global* as was mentioned before.

Top-5 PSNR-mean					Top-5 LPIPS-mean				
r	Training set	lr	PSNR [dB]	LPIPS	r	Training set	lr	LPIPS	PSNR [dB]
32	D_4	1e^{-3}	**30.298**	0.0749	8	D_4	1e^{-3}	**0.0746**	30.207
32	D_3	1e^{-3}	30.237	0.0767	32	D_4	1e^{-3}	0.0749	**30.298**
32	D_1	1e^{-3}	30.219	0.0774	8	D_3	1e^{-3}	0.0757	30.206
8	D_4	1e^{-3}	30.207	**0.0746**	32	D_3	1e^{-3}	0.0767	30.237
8	D_3	1e^{-3}	30.206	0.0757	4	D_3	1e^{-3}	0.0767	30.174

Table 5.2: Left: LoRA models evaluated on T_1, ranked by top-five PSNR-mean. LPIPS-mean is included for cross-metric comparison. Higher PSNR indicates better color fidelity. Right: LoRA models evaluated on T_1, ranked by top-five LPIPS-mean. PSNR-mean is included for cross-metric comparison. Lower LPIPS indicates better perceptual fidelity. Both sides: Values correspond to the average across all hint configurations (position and quantity) and all images. The best values are denoted in bold.

Unlike the PSNR ranking, the LPIPS top-five includes one additional configuration r = 32, D_3, lr = 1e^{-3} that does not appear among the top-5 PSNR performers. The configuration achieving the best LPIPS (Best-LPIPS) score is r = 8, D_4, lr = $1e^{-3}$, folowed by the *Best-Global* as was mentioned before.

5.2.1.2. Impact of the Rank

To assess the influence of the LoRA rank (r), all LoRA-based models were grouped according to their LoRA rank parameter $r \in \{1, 4, 8, 32\}$. Table 5.3 reports the average PSNR and LPIPS values for each LoRA rank over T_1. That is, for a fixed rank, the average is performed over twelve configurations, varying the dataset D_k and learning rate lr, as well as all images, hint positions, and hint quantities. We denote the respective scores as *PSNR-best-mean-R* and *LPIPS-best-mean-R*.

The results reveal no clear monotonic trend with rank. Although the configuration with r=32 achieves the highest average PSNR (29.969 dB), the differences across LoRA ranks are negligible —less than 0.05 dB in PSNR and 0.001 in LPIPS. This suggests that, within the tested range, increasing the rank does not systematically improve performance.

In practice, this indicates that lower-rank settings such as r=8 or even r=4 are already sufficient to capture the necessary adaptations for color propagation, achieving nearly identical perceptual quality while remaining more computationally efficient (see Table 4.2). To further examine this observation, qualitative comparisons between models with r=8 and r=32 will be included in Section 5.3.

r	PSNR-best-mean-R [dB]	LPIPS-best-mean-R
1	29.956	0.080
4	29.960	0.079
8	29.951	0.079
32	**29.969**	**0.079**

Table 5.3: Average PSNR and LPIPS for each LoRA rank evaluated on T_1. Best values are denoted in bold.

5.2.1.3. Impact of the learning rate

To evaluate how sensitive the fine-tuning process is to the learning rate (lr), all LoRA-based models were grouped by the values of lr considered: $lr \in \{1e^{-5}, 1e^{-3}, 1e^{-2}\}$. Table 5.4 presents the average PSNR and LPIPS values for each learning rate value over T_1. That is, for a fixed lr, the average is performed over sixteen configurations, varying the dataset D_k and LoRA rank r, as well as all images, hint positions, and hint quantities. We denote the respective scores as *PSNR-best-mean-LR* and *LPIPS-best-mean-LR*.

The results indicate that the configuration with a learning rate of 10^{-3} yields the highest average PSNR (30.133 dB) and the lowest LPIPS (0.078), outperforming both higher and lower rates. These results show that a moderate learning rate provides an effective balance between adaptation and stability, enabling the LoRA layers to refine the pretrained priors. In contrast, the most aggressive setting (10^{-2}) produces the lowest average performance (29.729 dB, LPIPS 0.080), and shows that excessive updates degrade consistency. The smallest rate (10^{-5}) remains competitive but shows marginally reduced gains, suggesting slower convergence. Overall, these results reinforce that the 10^{-3} configuration offers the most robust trade-off for fine-tuning.

Learning Rate	PSNR-best-mean-LR [dB]	LPIPS-best-mean-LR
1e^{-5}	30.016	0.080
1e^{-3}	**30.133**	**0.078**
1e^{-2}	29.729	0.080

Table 5.4: Average PSNR and LPIPS for each learning rate configuration evaluated on T_1. Best values are denoted in bold.

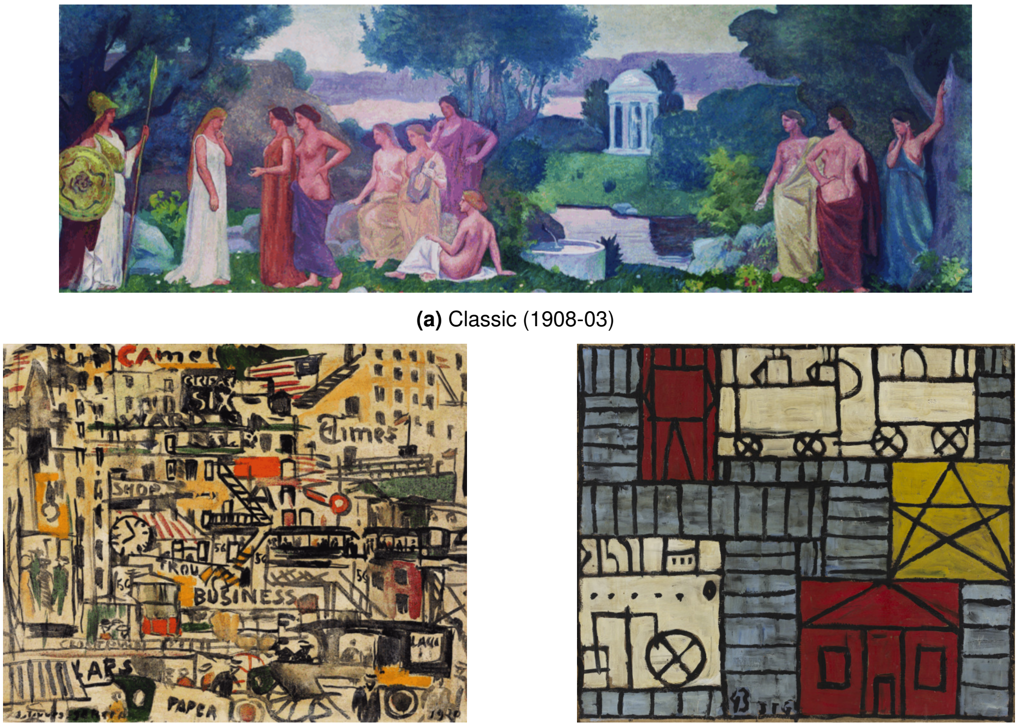
5.2.1.4. Datasets comparison

To analyze the influence of the training sets, the metrics were averaged across all LoRA model configurations within each dataset, as was done for the rank and learning rate influence (Table 5.5). This yielded the corresponding scores, *PSNR-mean-D* and *LPIPS-mean-D*, for each dataset D_k. Although the numerical differences are subtle—within 0.1 dB in PSNR and 0.002 in LPIPS—the ranking is consistent across both metrics. D_4, which combines a broader, more heterogeneous collection of artworks, achieves the highest mean PSNR (30.015 dB) and the lowest LPIPS (0.078), confirming its advantage in generalization and perceptual stability. D_3 shows the lowest average performance (29.914 dB/0.080).

Dataset	PSNR-mean-D [dB]	LPIPS-mean-D
D_4	**30.015**	**0.078**
D_3	29.914	0.080
D_2	29.954	0.079
D_1	29.952	0.079

Table 5.5: Average PSNR and LPIPS across all model configurations for each training dataset evaluated on T_1. Best values are denoted in bold.

Beyond these aggregated results, a deeper analysis of the training data could consider not only the number of samples and image quality, but also the underlying color and style distributions in each dataset. In the specific case of Torres García, the artist's production spans distinct stylistic phases—often grouped into "Classic", "Modern", and "Universal" (Museo Torres García, 2024)—which differ markedly in palette, geometry, and compositional structure as shown in Figure 5.1. Exploring dataset partitions that follow stylistic criteria, rather than purely chronological or source-based groupings, may offer a better understanding of how training distributions influence colorization performance. Alternatively, partitions aligned with the artist's geographical periods (e.g., Montevideo, Barcelona, New York, Paris) could also be relevant, as these contexts are associated with noticeable shifts in chromatic tendencies and thematic focus. Preliminary clustering experiments were conducted to investigate this direction; however, the results were inconclusive and are therefore not included in this thesis. This line of research remains open as a promising direction for future work.



(a) Classic (1908-03)

(b) Modern (1920-03)

(c) Universal (1943-68)

Figure 5.1: Examples of three distinct stylistic phases of Joaquín Torres García.

5.2.1.5. Comparison with the Base Model

Table 5.6 summarizes the metrics of the Base iColorit model evaluated on the test split T_1, namely the *PSNR-mean-Z-Base* and *LPIPS-mean-Z-Base*, which are the averages of the PSNR and LPIPS, respectively, over all images, hint positions, and hint quantities.

Table 5.7 contrasts Table 5.6 and Table 5.1 by showing the quantitative differences, Δ*PSNR-Z*, between PSNR-best-Z and PSNR-mean-Z-Base, defined as ΔPSRN-Z $=$ PSNR-best-Z $-$ PSNR-mean-Z-Base. Equivalently, the quantitative LPIPS difference, Δ*LPIPS-Z*, is reported.

Hints	PSNR-mean-Z-Base[dB]	LPIPS-mean-Z-Base
0	24.63	0.1355
1	26.82	0.1075
5	29.21	0.0822
10	30.33	0.0730
20	31.61	0.0633
50	33.21	0.0525
100	**34.25**	**0.0458**

Table 5.6: Mean PSNR and LPIPS values of the Base iColoriT model for each hint number evaluated on the test split T_1. Values correspond to the average of 100 randomly generated hints per image and across all images. Best values are highlighted in bold.

The results reveal a clear pattern: the fine-tuned models outperform the Base model in both metrics, particularly at low hint levels. With no hints (automatic), the LoRA configuration trained on D_1 achieves a gain of +1.60 dB in PSNR and -0.018 in LPIPS, indicating an enhancement in automatic colorization. As the number of hints increases, the performance gap progressively narrows, stabilizing around +0.02 dB and -0.0003 in LPIPS at 100 hints. This convergence suggests that when user guidance dominates the colorization process, the pretrained iColoriT already propagates chroma effectively, while LoRA fine-tuning primarily benefits the unguided or weakly guided regimes. Overall, the improvements confirm that the proposed adaptations strengthen the model's internal color priors without compromising its interactive behavior.

Hints	**Δ PSNR-Z**				**Δ LPIPS -Z**			
	r	lr	Training Dataset	ΔPSNR [dB]	r	lr	Training Dataset	ΔLPIPS
0	32	1e−3	D_1	+1.60	8	1e−3	D_4	−0.0180
1	32	1e−3	D_4	+0.61	8	1e−3	D_4	−0.0113
5	32	1e−3	D_4	+0.30	32	1e−3	D_4	−0.0041
10	32	1e−3	D_4	+0.24	32	1e−3	D_4	−0.0034
20	32	1e−3	D_4	+0.12	32	1e−3	D_4	−0.0021
50	32	1e−5	D_4	+0.03	32	1e−3	D_2	−0.0007
100	32	1e−5	D_4	+0.02	32	1e−5	D_4	−0.0003

Table 5.7: Performance improvements over the Base iColoriT model for each hint number. The left block reports the PSNR-best-Z configuration per hint number and its corresponding Δ PSNR-Z with respect to the Base model. The right block reports the LPIPS-best-Z configurationand its corresponding Δ LPIPS-Z.

To quantify the overall benefit of the best LoRA trade-off model (Best-Global, with r = 32, D_4, and lr = 1e^{-3}), its performance was compared to that of Base iColoriT across all hint numbers: PSNR-mean-Z-Base and LPIPS-mean-Z-Base. These results are shown in Table 5.8.For the Best-Global configuration, the corresponding PSNR and LPIPS scores, averaged over all images and hint positions, are denoted *PSNR-Best-Global-Z* and *LPIPS-Best-Global-Z*, respectively. The choice of using the Best-Global configuration reflects the fact that, in the context of color restoration, PSNR provides a more direct measure of fidelity to the ground truth, while LPIPS captures perceptual similarity. Since the difference between the best and second-best LPIPS-mean configurations was minimal, prioritizing the best PSNR-mean model yielded a more accurate assessment of reconstruction accuracy without sacrificing perceptual quality.

The results Table 5.8 confirm that fine-tuning with LoRA improves the performance in both automatic and weakly guided regimes. At 0 hints (fully automatic colorization), PSNR increases by nearly +0.9 dB, and LPIPS decreases by 0.015, indicating better color reconstructions and perceptual gains even without user input. This improvement is the result of fine-tuning with data from JTG and his disciples. The advantage remains consistent for low to moderate hint counts (1–20), where color propagation and structural coherence benefit from the adapted priors. However, the trend reverses slightly at higher hint levels (above 50), where the hint guidance itself dominates the inference, and the pretrained model already propagates color effectively. This behavior aligns with the earlier observations from the full fine-tuning experiment where the improvment was only fot the automatic inference (0 hints) and from partial fine-tuning experiments, where the fine-tuned variants outperformed the baseline in fully automatic colorization but reduced benefits with 20 hints. Overall, these findings suggest that LoRA fine-tuning primarily enhances the model's internal chromatic priors—improving autonomy and stability under sparse user guidance—while maintaining comparable performance when color hints are abundant. Furthermore, it indicates that the task of propagating many hints may require more data to outperform the Base model. Note that with the LoRA strategy, only the projection matrices of the attention layers are updated. However, the decoder's stabilization layer was not updated.

In this context, exploring how to further enhance the performance by the addition of color hints becomes particularly relevant. In future work, it would be useful to analyze how these adapted models respond to user-provided color cues and whether additional conditioning strategies could further improve color propagation in interactive scenarios.

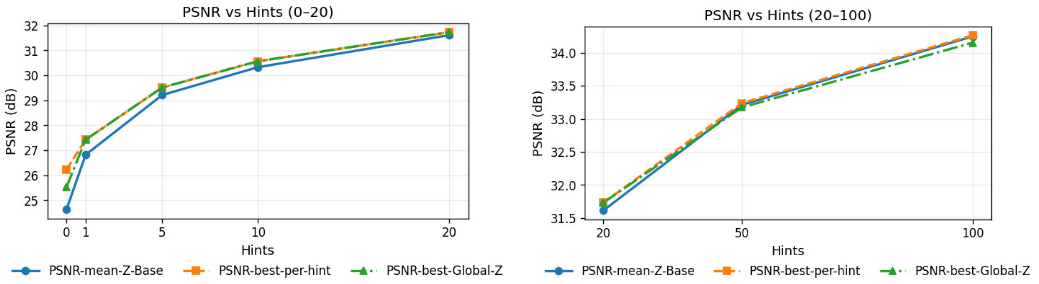
Hints	PSNR-Best-Global-Z	LPIPS-Best-Global-Z	ΔPSNR-Global	ΔLPIPS-Global
0	25.52	0.1201	+0.8940	−0.0154
1	27.42	0.0972	+0.6059	−0.0103
5	29.51	0.0781	+0.2991	−0.0041
10	30.57	0.0696	+0.2385	−0.0034
20	31.73	0.0612	+0.1192	−0.0021
50	33.17	0.0520	−0.0385	−0.0005
100	34.15	0.0460	−0.0957	+0.0002

Table 5.8: Comparison between the Base and LoRA Best-Global model across hint configurations on T_1. ΔPSNR-Global and ΔLPIPS-Global indicate the absolute differences in PSNR and LPIPS between the Best-Global and the Base models for each hint number case. Positive ΔPSNR and negative ΔLPIPS values correspond to improvements in reconstruction quality and perceptual similarity, respectively.

Figure 5.2 provides a compact, visual summary of the quantitative results reported in Tables 5.6 and 5.8, showing how the Base, PSNR-best-Z, and Best-Global models behave as the number of hints increases. The two plots separate the analysis into two regimes: sparse guidance (0–20 hints) and dense guidance (20–100 hints), matching the distinct behaviors observed in the tabulated metrics.

In the 0–20 hint range (left plot), both LoRA-based curves remain consistently above the Base model, summarizing the improvements already reflected numerically in the tables. This regime highlights the advantage of the adapted priors: LoRA fine-tuning enhances automatic colorization (0 hints) and continues to provide benefits under sparse user guidance (1–20 hints). The PSNR-best-Z models form the upper envelope, but the proximity to the Best-Global configuration indicates that a single, well-trained LoRA model can perform robustly across all low-to-moderate hint levels without requiring specialized tuning per number of hints.

The 20–100 hint range (right plot) as hint density increases, the user-provided chromatic cues dominate the inference, and the impact of the learned data priors diminishes. In this regime, LoRA models and the Base model behave similarly, with only marginal differences and occasional inversions at a very high number of hints.



(a) Number of hints between 0 and 20.

(b) Number of hints between 20 and 100.

Figure 5.2: Mean PSNR comparison across hint ranges. Base model (blue), PSNR-best-Z model (orange, Table 5.1), and Best-Global model (green, Table 5.8).

5.3. Qualitative Analysis

This qualitative evaluation complements the numerical metrics by assessing visual plausibility and fidelity. Selected examples from inference with the Base model, the best PSRN-mean (r = 32, D_4, lr = 1e^{-3}) and the best LPIPS-mean (r = 8, D_4, lr = 1e^{-3}), were examined under two criteria:

- Color plausibility and coherence with respect to the ground truth.
- Visual effect of increasing hint density on color propagation.

As mentioned in section 5.2, manually placed hints were used instead of random ones, allowing a controlled examination of user-guided colorization behavior. To this end, a set of representative images was selected to span different artistic styles and levels of scene complexity, enabling a qualitative comparison of the model behavior across varied visual conditions

5.3.1. Examples

Case 1. Semantic low complexity Images. In Fig. 5.3, we can see a global comparison of color perception across the three evaluated models: the Base iColoriT, the best PSNR-mean , and the best LPIPS-mean. Given the relatively low structural complexity of this artwork, the main differences between models emerge through their chromatic inference.

With 0 hints, both LoRA variants successfully infer the presence of yellow regions, and the best LPIPS-mean model additionally predicts a reddish tone consistent with the ground truth. However, the yellow inferred by the best PSNR-mean more closely matches the ground-truth hue, whereas the best LPIPS-mean tends toward a warmer, more orange yellow. The Base model, in contrast, fails to infer any color.

With 1 hint placed on a burgundy region, both LoRA models refine their predictions: the best PSNR-mean further improves its match to the true palette, while the best LPIPS-mean corrects the burgundy but maintains a more orange-biased yellow. The Base model only propagates the provided burgundy hint, still without inferring the yellow areas.

By the time 5 hints are provided, all three models converge toward highly similar solutions, successfully reconstructing the global palette. At this stage, remaining differences become subtle and require a more fine-grained inspection.

Case 2. Medium complexity semantic image. Figure 5.4 compares the color inferences produced by the Base model, the best PSNR-mean , and the best LPIPS-mean. This artwork is dominated by subtle variations of browns and greys, which makes small chromatic deviations especially noticeable.

With 0 hints, the three models diverge substantially: The Base model produces a noticeably more reddish–brown palette, the best LPIPS-mean shifts toward a cooler, blue–grey appearance, and the best PSNR-mean falls approximately in between these two extremes. As hints are introduced (5), the three models progressively converge toward a more consistent chromatic structure. Intermediate tones of grey and desaturated browns begin to appear across all models, and the global palette stabilizes.

By the time 50 hints are provided, the reconstructions become highly similar. However, a closer inspection reveals that the best LPIPS-mean mofrl still tends to preserve fewer reddish tones than the other two models. In contrast, both the Base and the best PSNR-mean retain a warmer appearance, more faithfully reflecting the ground truth's reddish accents.

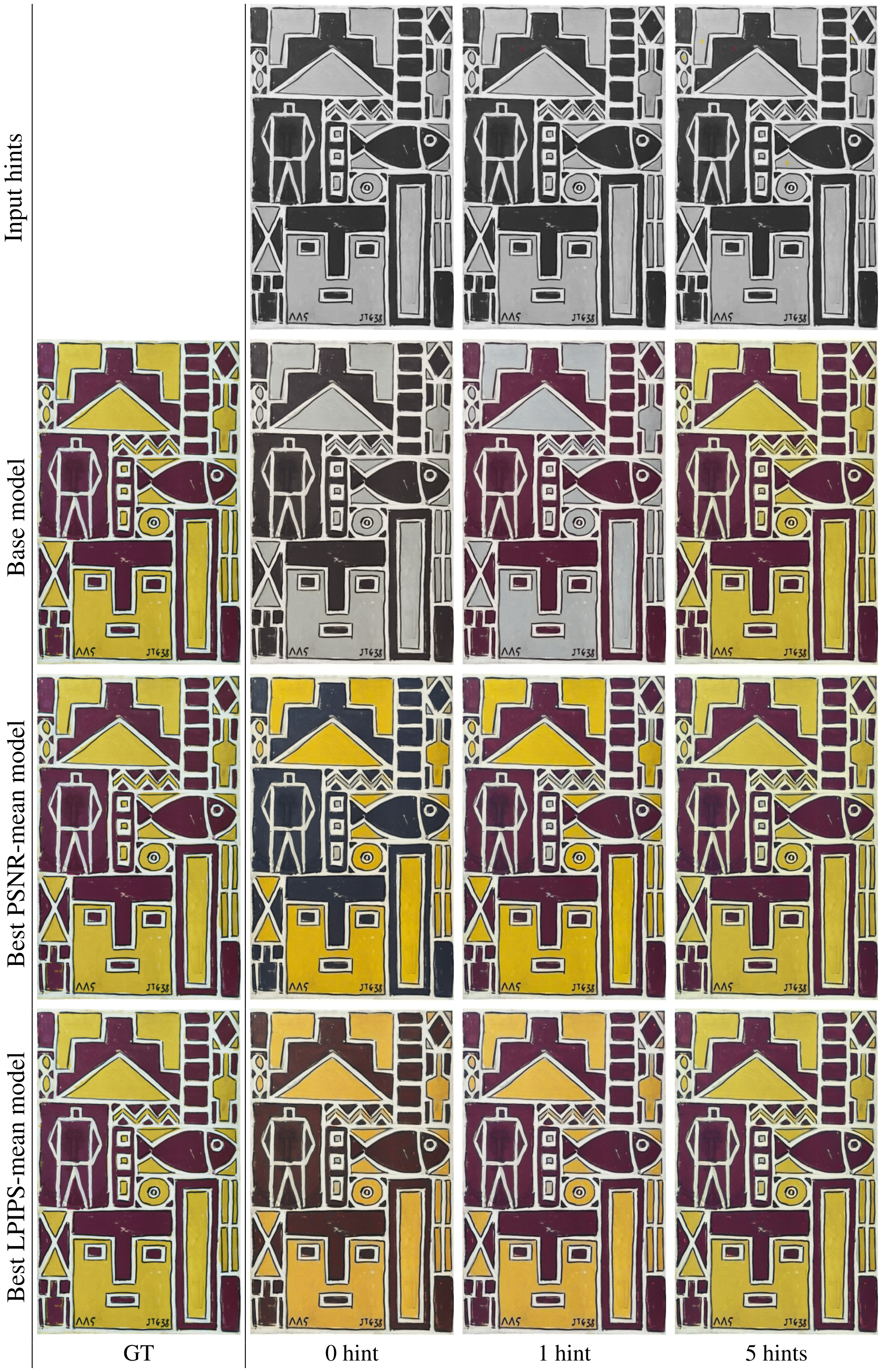


Figure 5.3: Comparative colorization results for a simple semantic composition (Painting 1938.12) at increasing hint densities. The first row shows the corresponding inputs with 0, 1, and 5 hints. The second, third, and fourth rows show, respectively, Ground Truth (GT) and the outputs of the Base, best PSNR-mean, and best LPIPS-mean models for 0, 1, and 5 hints. Please zoom in to better see the locations and colors of the hints.

Overall, this example illustrates two key points: best LPIPS-mean and best PSNR-mean induce different tonal biases when no or few hints are provided; and under high levels of external guidance, the best PSNR-mean tends to resemble the behavior of the Base model more closely than the best LPIPS-mean variant, suggesting that PSNR aligns better with the Base inference dynamics when sufficient hints are available.

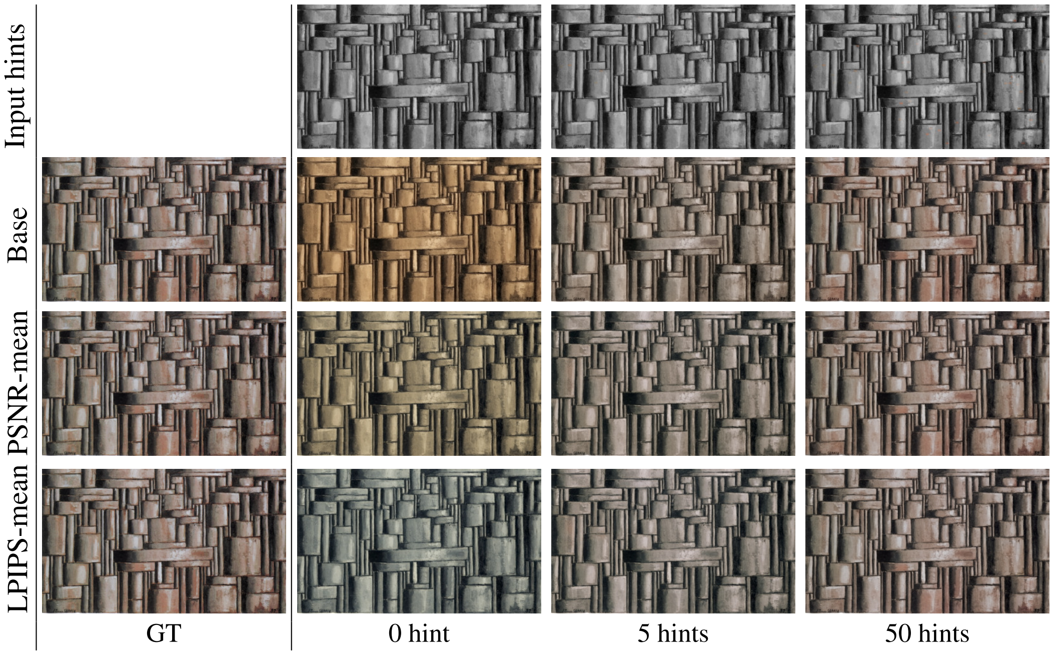


Figure 5.4: Comparative colorization results for a medium semantic composition (Painting 1937.26) at increasing hint densities. The first row shows the corresponding inputs with 0, 5, and 50 hints. The second, third, and fourth rows show, respectively, Ground Truth (GT) and the outputs of the Base, best PSNR-mean, and best LPIPS-mean models for 0, 5, and 50 hints. Please zoom in to better see the locations and colors of the hints.

Case 3. Medium-High complexity semantic image. As shown in Fig. 5.5, a comparison was conducted between the inferences produced by the Base model, the best PSNR-mean, and the best LPIPS-mean on a third painting. This artwork presents greater chromatic complexity, with multiple shades of green, blue, and terracotta. Despite this increased variability, the overall behavior of the three models remains consistent with the tendencies observed in the previous examples.

With 0 hints, the models diverge considerably. As in Figure 5.4, the best LPIPS-mean model exhibits the coldest tendency (bluish tones), The Base model leans toward a uniform sepia-like palette, and the best PSNR-mean model occupies an intermediate position but introduces noticeable reddish accents.

Between 1, 5, and 10 hints, a stable pattern emerges: The best-PSNR model tends to infer more saturated reddish tones in the central and lower areas, even though the global palette remains largely driven by greenish hues. This tendency is stronger than in both the Base and best LPIPS-mean models.

As the number of hints increases (50 hints), all three models gradually converge toward the blues, terracotta, and green hues present in the ground truth. At higher hint levels, the reconstructions become nearly indistinguishable across models.

Case 4. Medium-High complexity semantic image. As shown in Fig. 5.6, this painting presents a palette composed almost entirely of primary colors (white, red, blue, yellow) with black contour lines. In this case, a behavior already observed in Fig. 5.3 reappears: The best PSNR-mean model tends to infer yellow regions from the very beginning.

Although the three models differ little at 0 hints, their chromatic tendencies remain noticeable. The best PSNR-mean model shows early traces of yellow, The best LPIPS-mean leans toward cooler, blue–gray hues, and the Base model exhibits a characteristic tendency toward reds. These tendencies become even clearer with 1 hint.

With 10 hints, an interesting phenomenon emerges: The best LPIPS-mean model appears to follow the external guidance (the hints) more faithfully than the other models. A particularly striking example occurs at 10 hints, where the Base and the Best PSNR models produce a large red square interrupted by a central blue area that does not correspond to the ground truth. Neither the ground truth nor the best LPIPS-mean model contains this blue region, suggesting that LPIPS is better aligned with the hint specification in this configuration.

Consequently, for this painting, the best-LPIPS model provides the most accurate reconstruction with 50 hints, outperforming both the Base and best PSNR-mean models. Meanwhile, the best PSNR-mean model continues to behave similarly to the Base model, which reinforces the idea that best PSNR-mean variants tend to

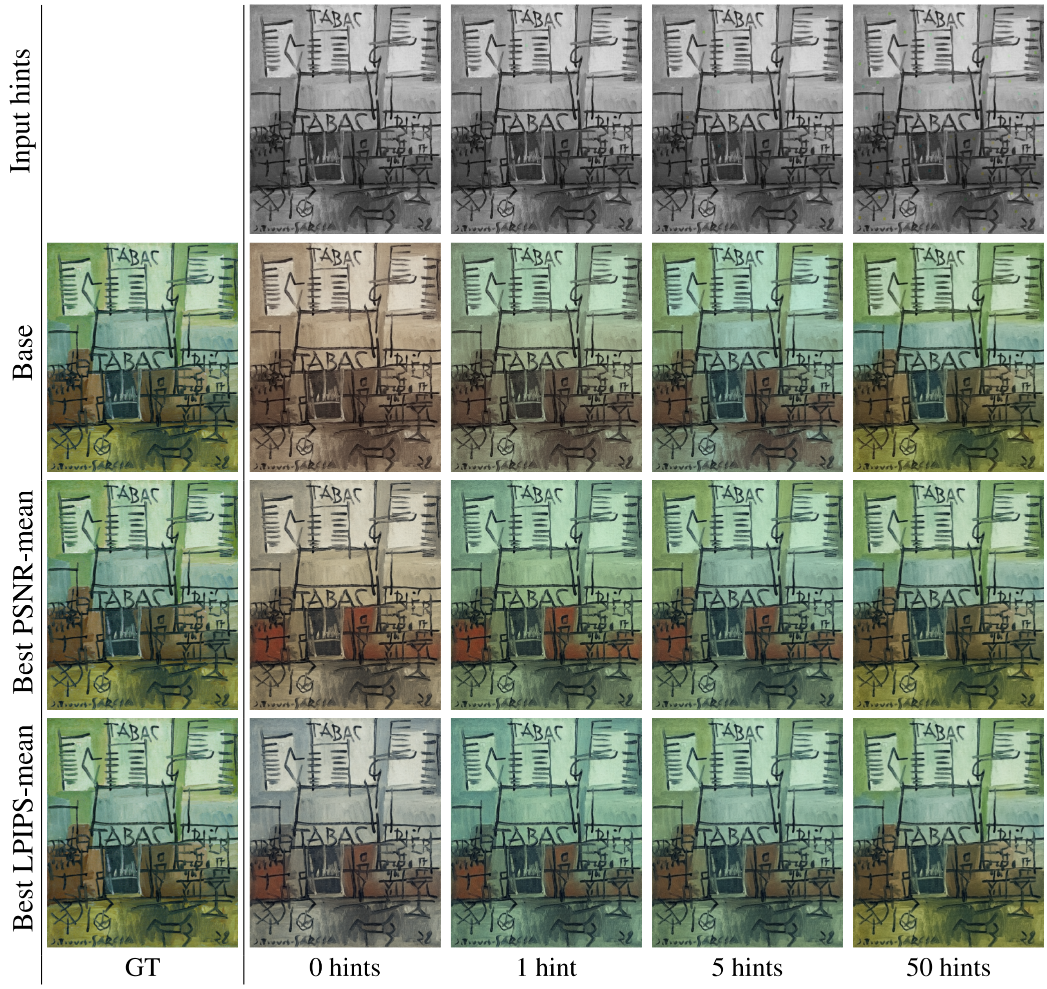


Figure 5.5: Comparative colorization results for a medium-high semantic composition, Painting 1928.177, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 5, and 50 hints. The second, third, and fourth rows show, respectively, Ground Truth (GT) and the outputs of the Base, best PSNR-mean, and best LPIPS-mean models for 0, 1, 5, and 50 hints. Please zoom in to better see the locations and colors of the hints.

inherit the base model's inference dynamics when given enough hints

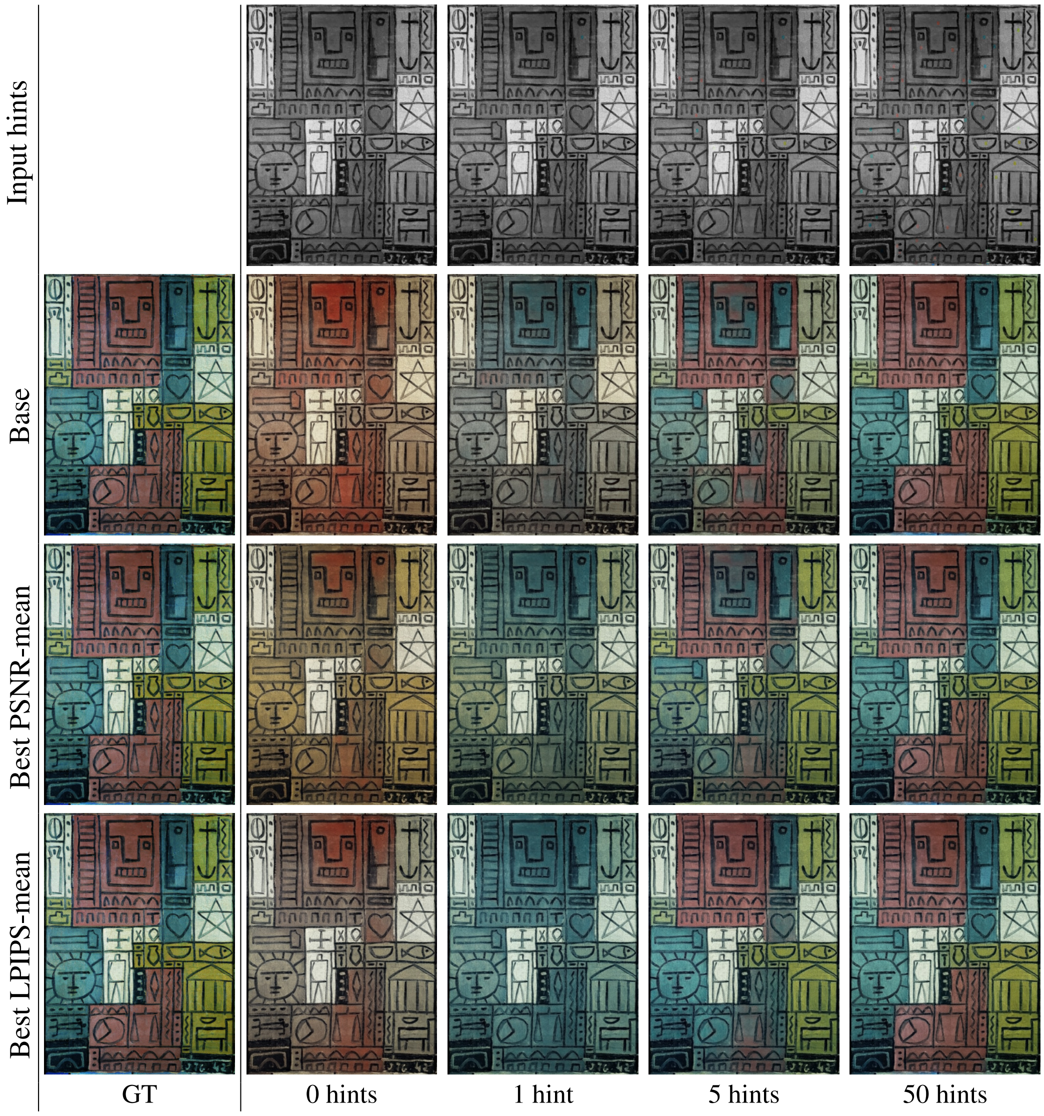


Figure 5.6: Comparative colorization results for a medium-high semantic composition, Painting 1942.37, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 10, and 50 hints. The second, third, and fourth rows show, respectively, Ground Truth (GT) and the outputs of the Base, best PSNR-mean, and best LPIPS-mean models for 0, 1, 10, and 50 hints. Please zoom in to better see the locations and colors of the hints.

Case 5. High complexity semantic image. From the results in Section 5.2, a single configuration provides the best trade-off between PSNR and LPIPS, ranking first in mean PSNR and second in mean LPIPS. The best PSNR-mean, $r = 32$, D_4, $lr = 1\text{e}{-3}$, was therefore selected for further visual analysis.

Figure 5.7 presents an artwork with high semantic and chromatic complexity. Even with 50 hints, neither model accurately reproduces the ground-truth color distribution. This example highlights the inherent difficulty of reconstructing multiple interacting color regions in scenes with dense, heterogeneous semantic structure and motivates systematic examination of these limitations under challenging conditions.

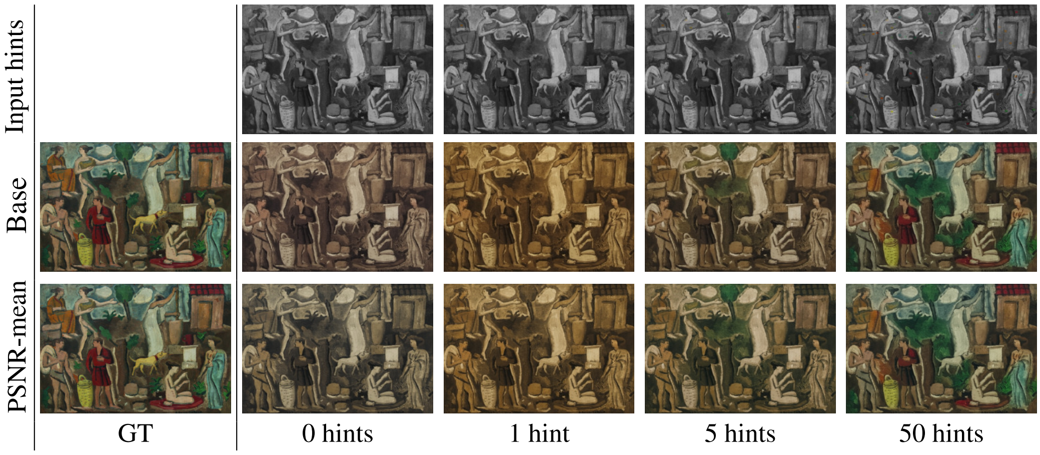


Figure 5.7: Comparative colorization results for a medium-high semantic composition, Painting 1927.65, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 5, and 50 hints. The second and third rows show, respectively, Ground Truth (GT) and the outputs of the Base and best PSNR-mean. Please zoom in to better see the locations and colors of the hints.

Case 6. Another medium complexity semantic image. Figure 5.8 presents another example, where the behavior of the best-PSNR LoRA model and the Base iColoriT diverges in an unexpected way at 5 hints. With 1 hint, the LoRA model performs noticeably better than the Base model in the sky region, correctly inferring a light blue tone closer to the ground truth.

However, at 5 hints, a singular phenomenon emerges. Due to the tendency of the best PSNR-mean model to propagate reddish hues more aggressively, the right lapel of the green jacket becomes fully tinted red in the LoRA reconstruction. In contrast, the Base model produces a mixture of green and red patches: its propagation of the hint is less coherent, but this unintentionally results in a more faithful approximation of the ground truth, whose true color for this region is green. Beyond this point, as the number of hints increases (50 hints), both models gradually converge to similar solutions, and the discrepancy disappears.

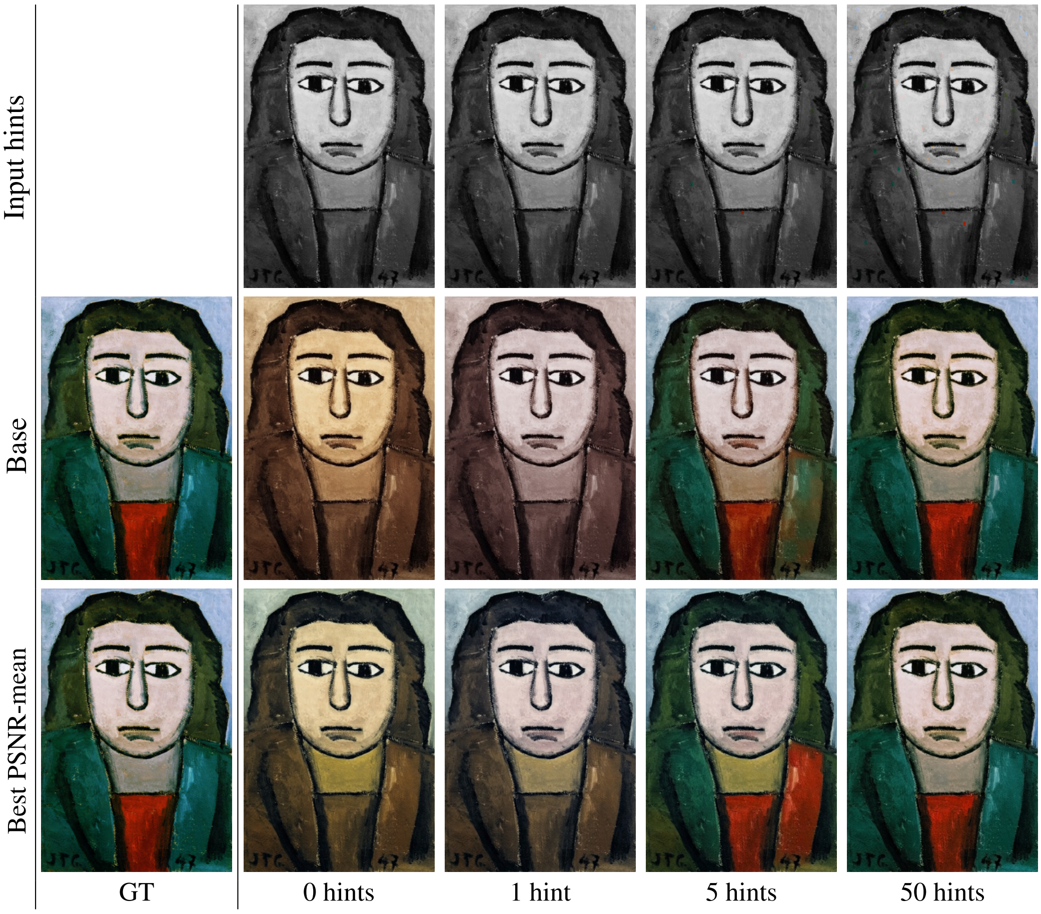


Figure 5.8: Comparative colorization results for a medium-high semantic composition, Painting 1947.34, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 5, and 50 hints. The second and third rows show, respectively, Ground Truth (GT) and the outputs of the Base and best PSNR-mean. Please zoom in to better see the locations and colors of the hints.

5.3.2. Failures cases.

Case 1 Figure 5.9 provides yet another example of a semantically complex scene, where both the Base iColoriT model and the best PSNR-mean model struggle to reconstruct the correct chromatic distribution, even when supplied with 50 hints. This still life contains multiple interacting objects with subtle, overlapping color transitions—such as the wicker basket, wine bottles, fruit, metallic pan, and fish—making it particularly challenging for hint-based propagation. Although the LoRA model shows a more coherent first shot, the final renderings at high hint levels remain far from the ground truth. Both models tend to converge toward overly uniform warm palettes, failing to recover the nuanced reds, blues, and ochres present in the original painting. This example highlights that, in highly heterogeneous scenes, even dense hint configurations may be insufficient for achieving accurate color inference.

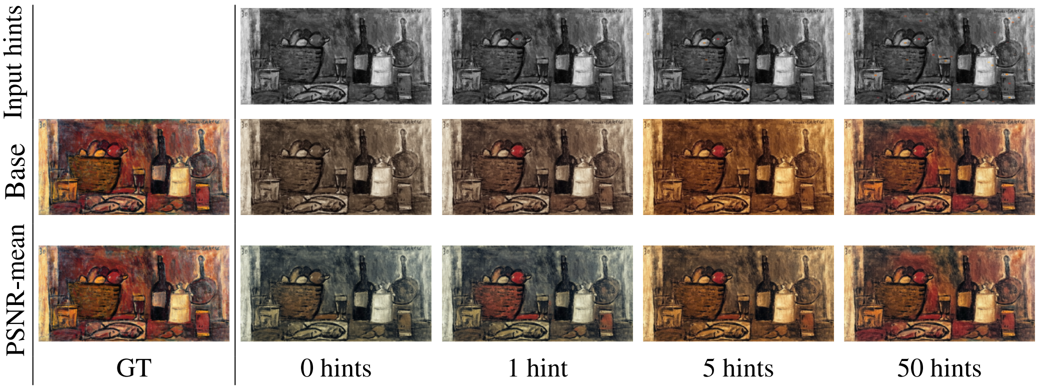


Figure 5.9: Comparative colorization results for a medium-high semantic composition, Painting 1930.37, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 5, and 50 hints. The second and third rows show, respectively, Ground Truth (GT) and the outputs of the Base and best PSNR-mean. Please zoom in to better see the locations and colors of the hints.

Case 2 Figure 5.10 shows a late work by Torres García, created during the final year of his life. This painting presents a noticeably different stylistic direction, with strong perspective structure and the inclusion of colors such as pink, which are uncommon in his earlier production. In this example, the best PSNR-mean model consistently outperforms the Base iColoriT model across all hint levels. Although neither model achieves a fully accurate reconstruction even at 50 hints, the LoRA predictions more faithfully approximate the ground-truth palette, particularly in the pink façades and blue architectural elements. The Base model, in contrast, tends to drift toward desaturated browns and yellows in the first inferences, failing to capture several of the distinctive chromatic traits of this late-period work. Overall, this case shows that the LoRA adaptation is better than the Base model at handling stylistic deviations and atypical color choices within the Torres García corpus.

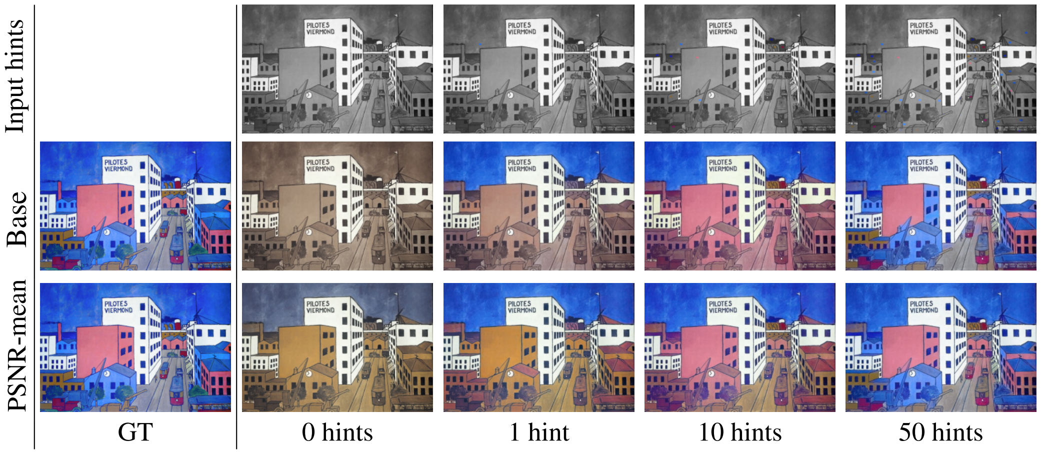
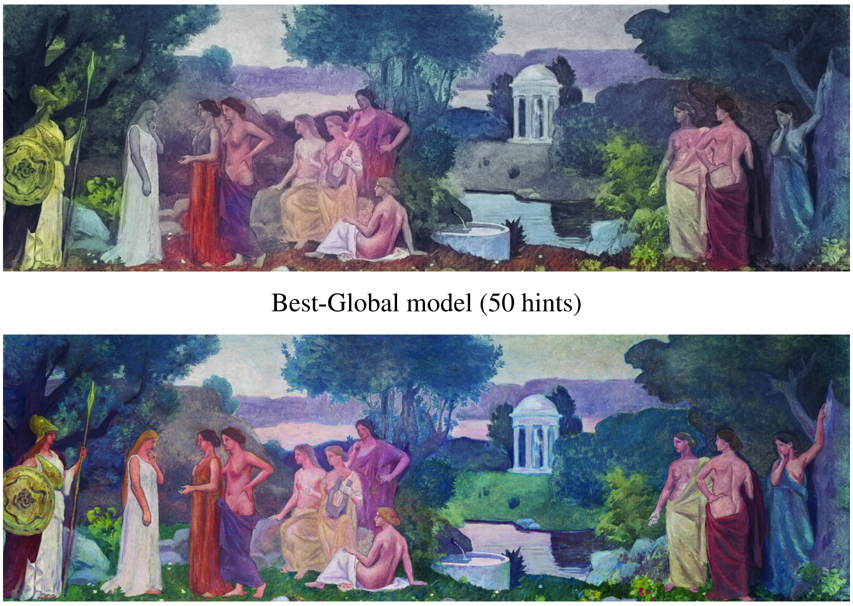


Figure 5.10: Comparative colorization results for a medium-high semantic composition, Painting 1947.09, at increasing hint densities. The first row shows the corresponding inputs with 0, 1, 5, and 50 hints. The second and third rows show, respectively, Ground Truth (GT) and the outputs of the Base and best PSNR-mean. Please zoom in to better see the locations and colors of the hints.

5.3.3. Further analysis

To further investigate the performance of the selected model, the best PSNR-mean, additional qualitative examples are presented that compare the LoRA model's inference results with 50 user-provided hints against the corresponding ground truth images. These examples aim to shed light on cases where, despite extensive user guidance, the inferred colorization still diverges from the reference. Three of the selected examples belong to the split T_1, while one example is drawn from the split V_1 (image `1934_02203.png`).

Despite the use of 50 user-provided hints in Figure 5.11, the resulting colorization remains visually inconsistent. The model assigns uniform green tones to the leftmost human figure and predominantly blue tones to the rightmost figure, while failing to colorize the second human figure and the grass beneath the white monument. A spurious red patch is also visible on the violet dress of the central figure.



Best-Global model (50 hints)

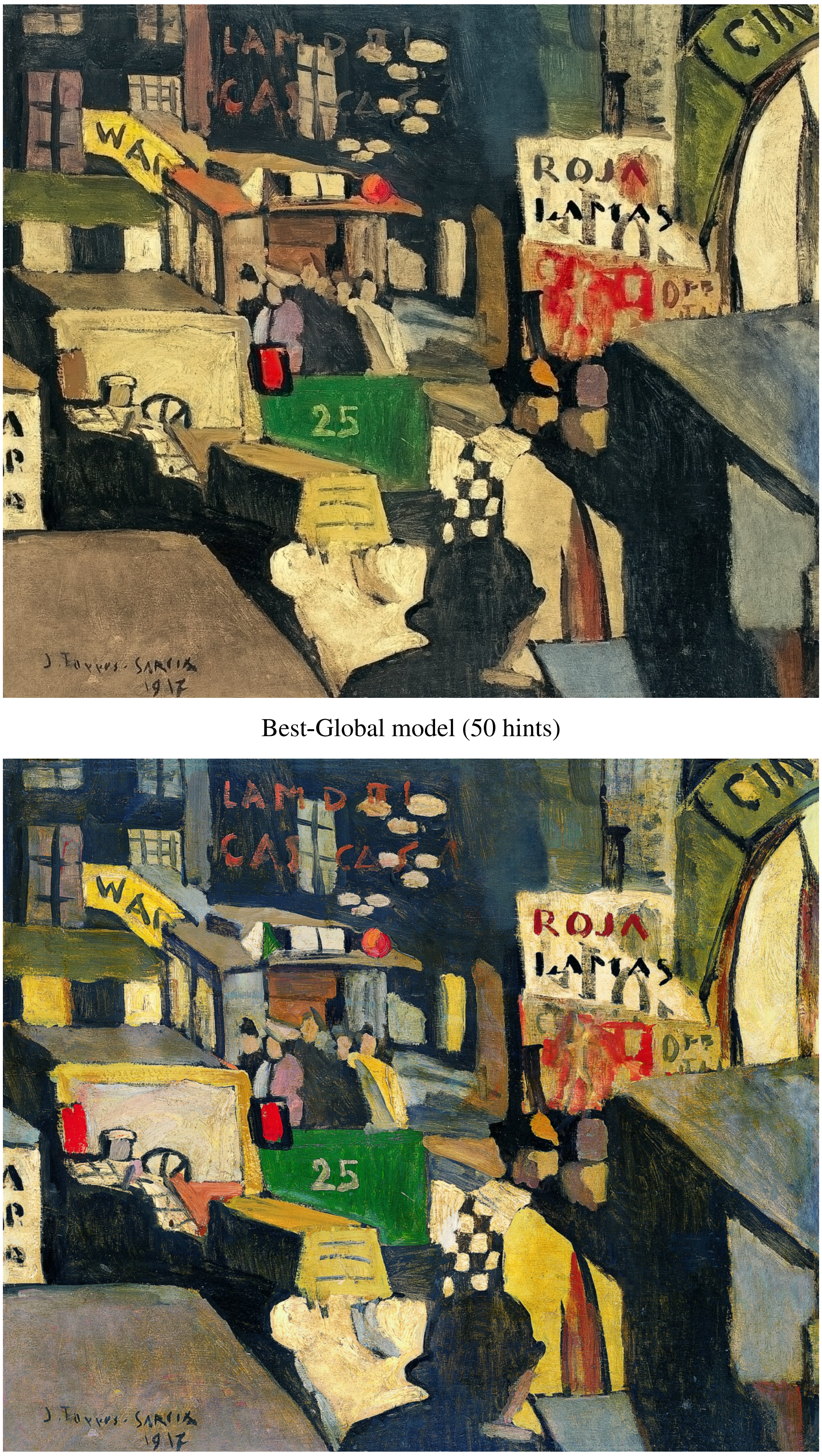
Ground Truth

Figure 5.11: Comparison between the Best-Global model inference at 50 hints (top) and the ground truth (bottom) for Painting 1908.03.

In the Figure 5.12, several color patch artifacts are observed. Red color patches appear both in the word *"ROJA"* on the right side and in the red circle at the center, where the color, although chromatically consistent, induces a noticeable bleeding effect into neighboring regions. In addition, the model fails to generalize red tones coherently across the scene, resulting in inconsistent colorization of the rectangular shape on the left, which is rendered brown despite being red in the ground truth.

In the Figure 5.13, which presents a higher level of scene complexity than the previous two examples, several failure cases can be identified upon closer inspection. The striped awning at the center does not fully converge to a consistent red coloration and introduces noticeable bleeding artifacts, while the checkered pattern above the word *"BUSINESS"* is erroneously colorized. These issues indicate that, although the model approximates the desired chromatic distribution reasonably well, it still requires a high density of user guidance to achieve visually coherent results in complex scenes.

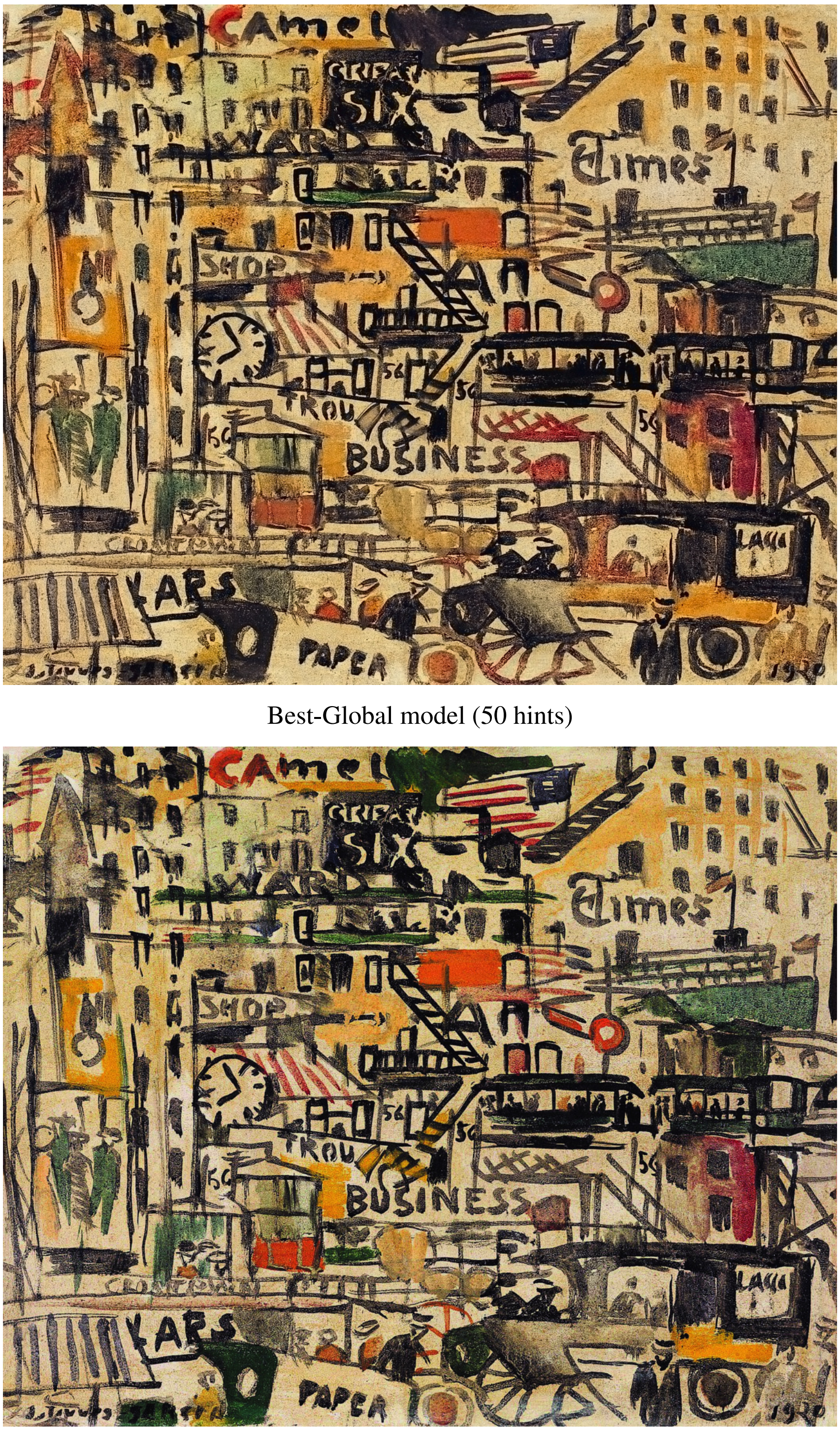
Although the example shown in Figure 5.14 appears less complex than the previous ones, the model still exhibits colorization errors and localized artifacts. In particular, miscoloration can be observed around the mouth of the fish, as shown in the figure in the lower-right corner. Similar issues appear in the blue-toned fig-



Best-Global model (50 hints)

Ground Truth

Figure 5.12: Comparison between the Best-Global model inference at 50 hints (top) and the ground truth (bottom) for Painting 1917.12.

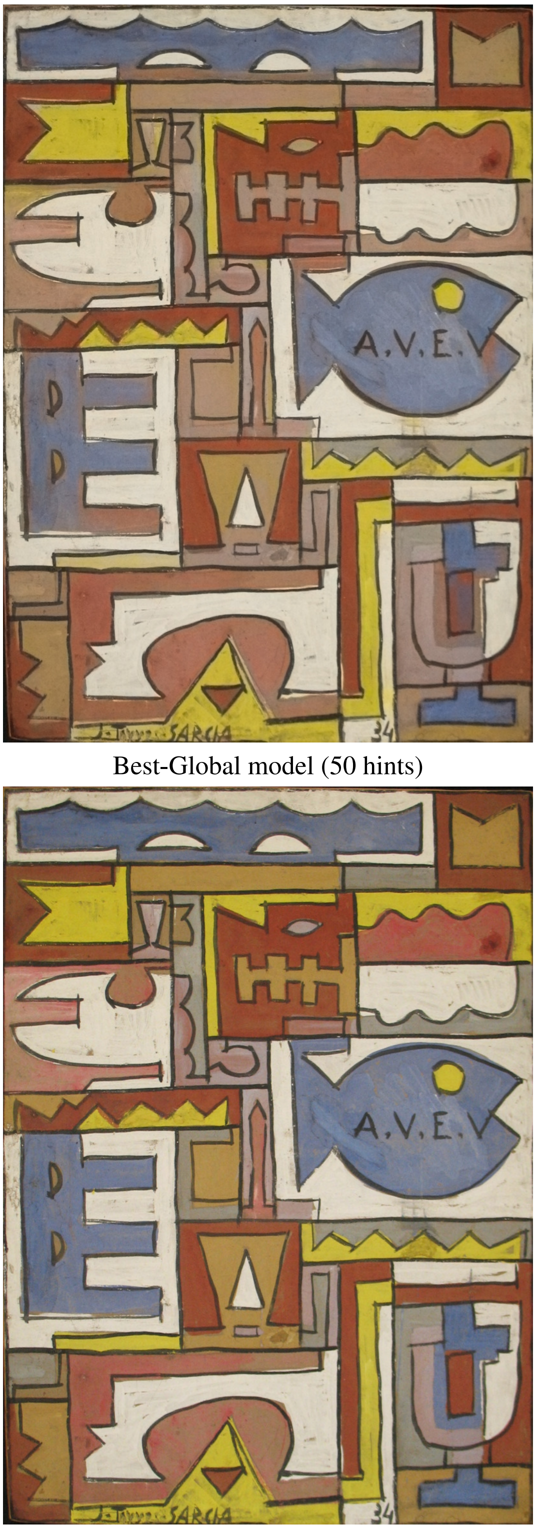


Best-Global model (50 hints)

Ground Truth

Figure 5.13: Comparison between the Best-Global model inference at 50 hints (top) and the ground truth (bottom) for Painting 1920.03).

ure on the left side of the image (left-central area), indicating persistent difficulties in achieving coherent color propagation even in comparatively simpler scenes and



Best-Global model (50 hints)

Ground Truth

Figure 5.14: Comparison between the Best-Global model inference at 50 hints (top) and the ground truth (bottom) for Painting 1934.02203.

with a high number of hints.

To better understand the residual gap between the selected LoRA model and the ground truth, selected regions from Section 5.3.1 were magnified as shown in Figure 5.15. At a global level, the LoRA-adapted predictions are visually coherent and closely aligned with the target palette, as already seen in Figure 5.3, supporting the quantitative improvements reported in Tables 5.7 and 5.8. However, the zoomed views reveal several subtle yet relevant discrepancies.

In the first zoom, the original painting has a small unpainted gap, revealing an underlying warm, off-white layer. The model recognizes that the area is not pure white and paints it with a slightly grey or burgundy-tinted tone, depending on the nearby colors. As a result, the warm undertone present in the ground truth is not reproduced.

In the second zoom, a small burgundy triangle appears uncolored in the inferred image, and the natural yellow paint bleeding at the bottom, visible in the ground truth, is not reproduced. The yellow area at the top is also slightly uneven, showing some artificial yellow bleeding that does not correspond to the original brushwork on the left and some grey area on the right.

These examples show that, while global colorization is highly consistent, some fine details of texture and painting technique remain simplified, even with 50 manually selected hints.

Based on the qualitative behaviour observed across all artworks, the r=32, lr=1e^{-3} and training dataset D_4 LoRA model still provides a stable and coherent performance. This model was selected to adapt the authors' original demo pipeline. The implementation of this adapted system is presented in the next chapter.

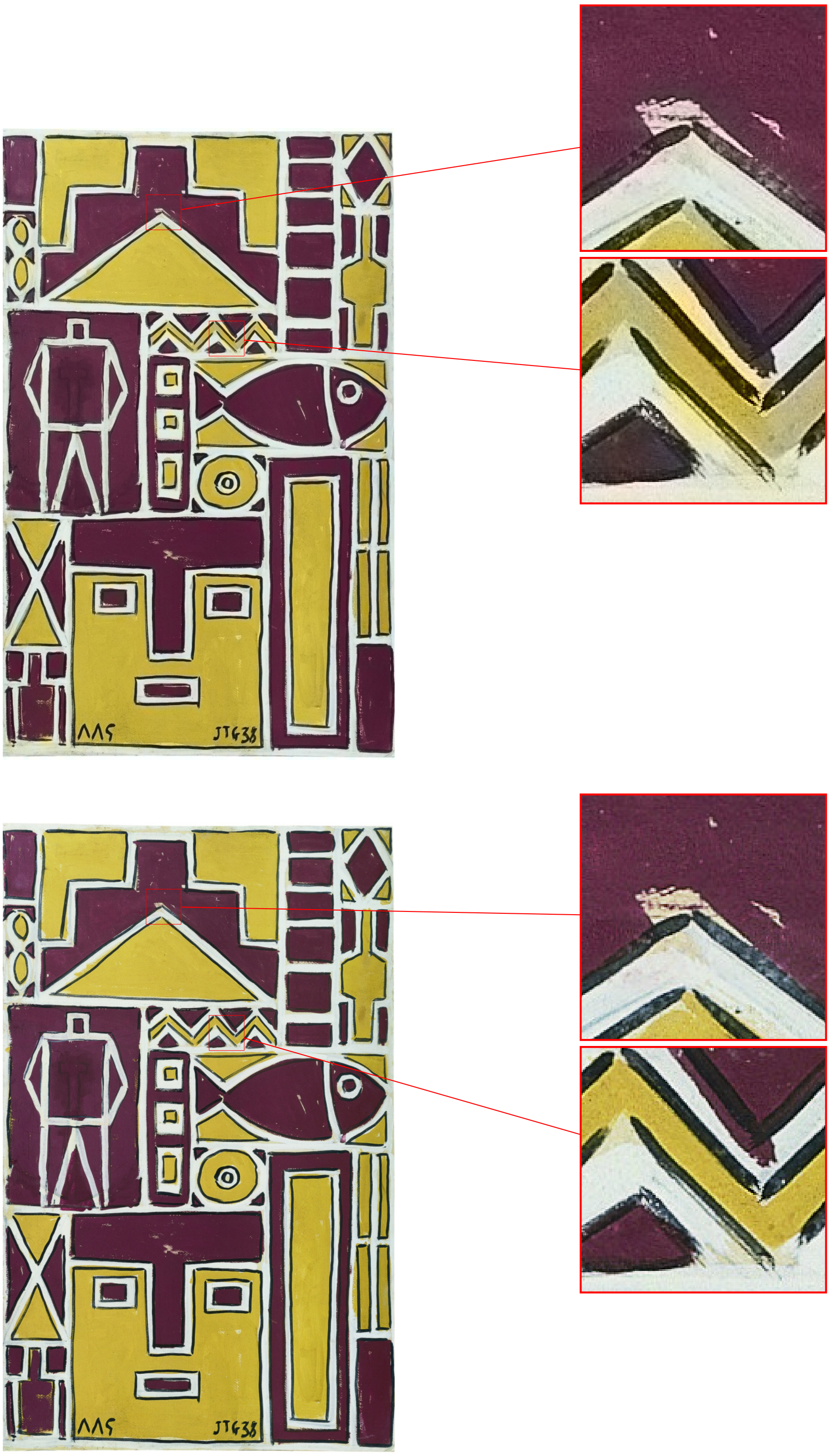


Figure 5.15: Comparison between the Best-Global model inference at 50 hints (top) and the ground truth (bottom). Two zoomed regions, placed at identical relative coordinates for each image. Painting 1938.12

Chapter 6

Interactive demo

This chapter describes the modifications made to the official iColoriT demo software provided by the authors (Yun et al., 2023) to integrate the LoRA model selected in Chapter 5. It also describes the additional functionalities and adjustments introduced to improve its usability within the context of the Torres García restoration project.

6.1. Icolorit Demo adaptation

The original demo consists of a PyQt5-based (Riverbank Computing, 2023) interactive GUI designed for CPU and GPU compatible devices, allowing users to load an image, paint color hints on a drawing pad, select chromatic values from an ab gamut, and generate a colorized output using the pretrained iColoriT checkpoints. The interface builds upon earlier user-guided colorization tools, in particular the real-time interactive framework of Zhang et al. (R. Zhang et al., 2017).

Since the LoRA logic had already been incorporated at the modeling layer (as detailed in Section 4.2.4), the base integration consisted of updating the internal workflow accordingly. However, beyond this integration, several additional functionalities were introduced to enhance the interactive experience and better support the project's use cases. Thus, although the GUI layout remains largely unchanged, the resulting application extends the original demo's capabilities rather than merely replicating it.

Before describing these new functionalities, it is useful to illustrate how the original application behaved, as seen Figure 6.1. The default interface displays the grayscale input image on the center panel (the Drawing pad) and the model's first inference — computed without any user hints — on the right panel. When the user clicks on the grayscale image, the system retrieves the corresponding Luminance value L of the selected pixel and presents, on the left panel, the set of possible (a, b) chromatic values compatible with that luminance in the CIELAB gamut as can be seen in Figure 6.2. Then, the user selects the color of the hint, which the model uses to color the image. For each new selected hint (color and position), the colorization image is updated with the new hint information.

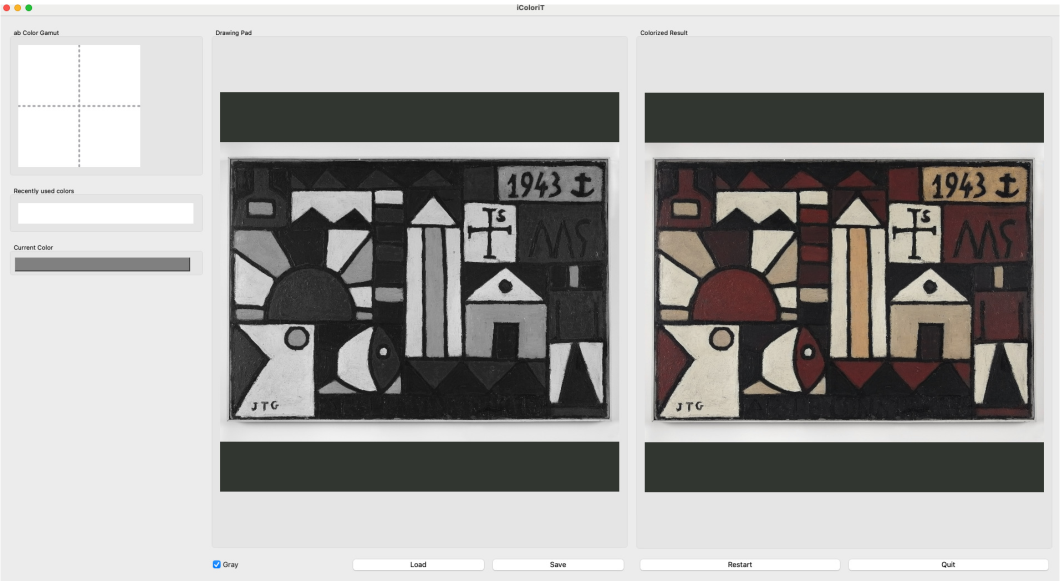


Figure 6.1: Original iColoriT demo interface - initial display. Left: interactive panel displaying the CIELAB gamut for the selected luminance (initially empty).Center: drawing pad with the grayscale image. Right: colorization output with 0 hint (Base model). Painting 1943.26.

The original demo provided the following interactive functionalities:

- **Left click on the Drawing Pad:** select the spatial location where the hint will be placed. A small square appears in that location.
- **Left click on the *ab* Color Gamut:** select the chromatic pair of values (a, b) associated with the selected luminance.

- **Restart button:** clear all existing hints.
- **Save button:** export the currently colorized image.
- **Load button:** load a new image into the interface.

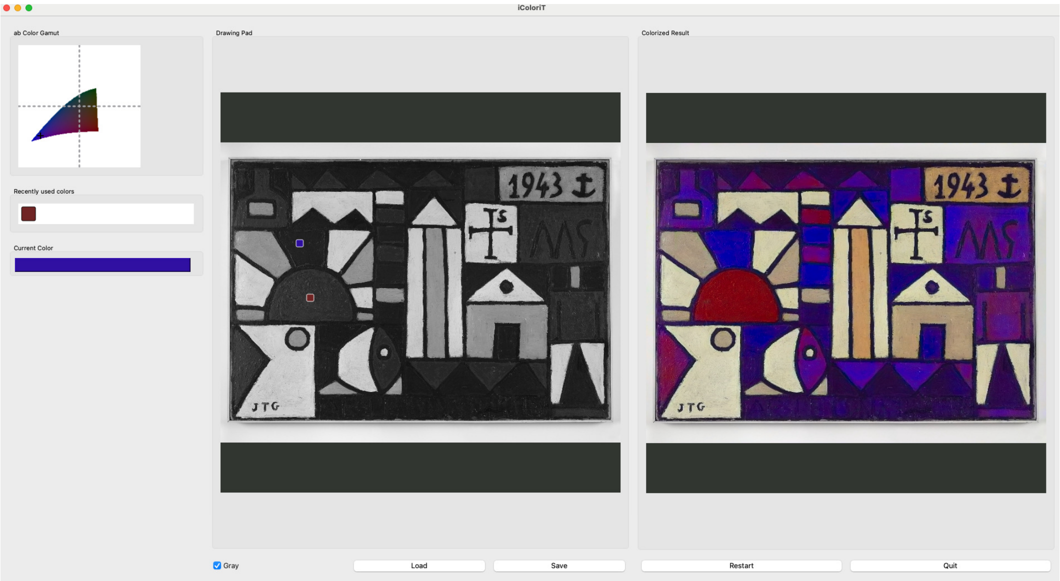


Figure 6.2: Original iColoriT demo interface - hint selection. Left: interactive panel displaying the CIELAB gamut for the selected hint's luminance. Center: grayscale input image with selected hints. Right: colorization output updated using the new selected hint (Base model). Painting 1943.26

6.1.1. New Functionalities Added to the Demo

In addition to integrating the selected LoRA model, several new functionalities were added to extend the capabilities of the original demo. First, the application now allows saving not only the final colorized image but also a JSON file containing the full set of user-provided hints, including spatial coordinates and their corresponding RGB and CIELAB values. This enables exact reproduction of any interactive session and supports downstream analysis. The system also saves a grayscale version of the input image with the hints overlaid, as well as a binary hint mask marking the exact pixel positions of all hints. Users can also specify a custom output directory where all generated images and JSONs files are stored.

Beyond these output-related additions, several interactive improvements were introduced. The grayscale image panel now supports zoom-in and zoom-out operations, enabling precise hint placement in regions with fine detail. A single-step undo mechanism (`Ctrl+Z`) was implemented to revert only the most recently placed hint, improving usability during dense annotations. Furthermore, the *ab* color gamut also includes a zoom functionality, allowing users to inspect the gamut at higher resolution and select chromatic values more accurately, especially in areas where many similar hues cluster tightly. Together, these extensions transform the original interface into a more flexible, precise, and reproducible colorization environment.

Figure 6.3 illustrates the extended user interface developed for this project. Using the selected Best Global LoRA model to infer, instead of the Base one. The model's first inference—computed without any user hints—is in the right panel. As we can see, it is considerably better than in Figure 6.1.

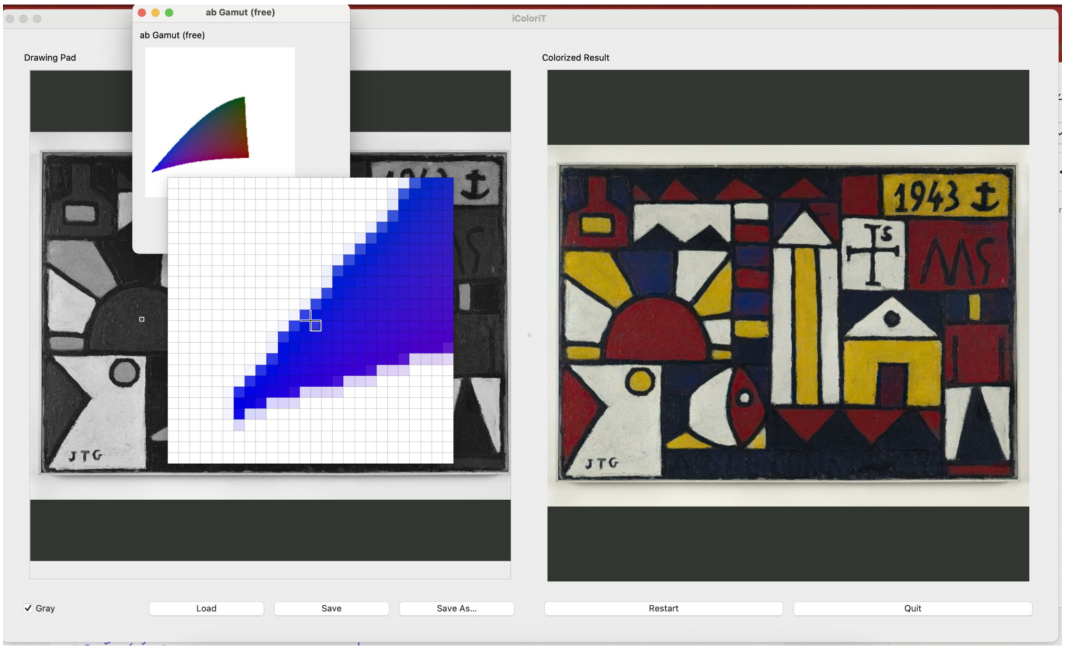


Figure 6.3: Updated iColoriT demo interface. The interactive panel now provides a zoomed-in view of the CIELAB gamut, allowing for more precise selection of hints. It is also detached from the drawing and result panels. Painting 1943.26

6.1.2. Reference-based color selection

While the original iColoriT demo is designed to explore user-guided colorization in a generic setting, the target application of this work is substantially more constrained: restoring the chromatic appearance of artworks that no longer exist. In this context, curators and domain experts often have access to related pieces, such as works created during the same period or by other painters of the same school, that provide priors on plausible palettes, typical pigment choices, and characteristic chromatic relationships. To exploit this expert knowledge more directly, the demo was extended to support reference-based color selection.

Concretely, the interface now allows the user to load an additional reference color image, which is displayed alongside the grayscale target. Instead of selecting chromatic values solely from the CIELAB gamut, the expert can click on colors from the reference image to sample colors from regions that they deem to be stylistically or materially compatible with the lost work.

Note that extracting the color directly from the reference image without considering the luminance of the selected point in the grayscale image will result in a different color than desired. In the iColoriT architecture, the selected luminance L_t from the target grayscale image and the user's hint chromaticity (a, b) are used as the model input, as was explained in Section 4.1.1. At inference time, the model predicts the (a, b) values and concatenates those values with the target's original L_t channel. As a result, transferring a color from a specific point in the reference image (L_r, a_r, b_r), to the target image by using the pair (a_r, b_r) does not guarantee that the preceived color will be the same. In other words, the color perceived from the triplet (L_r, a_r, b_r) differs from the color perceived from the triplet (L_t, a_r, b_r). With these phenomena two issues arise: (1) the reference (a_r, b_r) may not exist in the valid CIELAB gamut for that luminance L_t; (2) even if it exists, the perceived color can differ significantly when paired with a different luminance; and the colors sampled from the reference may therefore appear distorted, darkened, or overly saturated when placed on the target.

To mitigate this, the system filters the reference image based on the luminance structure of the target image. . Only reference pixels whose luminance is equal to or close enough to the target pixel's luminance $L_t 1.0$ are considered. The corresponding (a_r, b_r) values are extracted and intersected with the original iColoriT-valid gamut at that luminance. The result is a *reference-conditioned gamut*: a set of chromatic options that are simultaneously plausible according to the reference and perceptually coherent when combined with the target luminance.

Figure 6.4 illustrates this phenomenon by showing two visualizations of two different CIELAB gamuts: one for $L = 70$ on the left and one for $L = 50$ on the right. A reference color sampled at $(a = 15, b = 70)$ appears as an in-gamut sample in the left panel and an out-of-gamut sample in the right panel. In the latter, the pair $(a = 15, b = 70)$ is invalid for luminance $L = 50$. In this context, an (a, b) pair is considered invalid or out of gamut when it does not correspond to any physically realizable color within the sRGB color space for a given luminance L. Consequently, chromatic coordinates that are valid at one luminance level may fall outside the sRGB gamut at another.

Figure 6.5 on the top row shows how a single and fixed chromaticity (a, b) changes perceptually when combined with different luminance values. Even though the (a, b) pair is kept fixed $(a = 50, b = 50)$, the resulting sRGB colors show strong variations, ranging from dark browns at low L to desaturated oranges and pastel tones at high L.

Figure 6.5, bottom row, further highlights the issue of invalid values by showing which luminance values produce an in-gamut color for a fixed $(a = 15, b = 70)$. In sRGB, only a narrow band of luminance values yields a physically valid color, while most L-levels fall outside the displayable gamut.

Together, these visualizations illustrate why direct color transfer from the reference is not feasible: A color that is chromatically valid at its original luminance may become perceptually inconsistent or even out-of-gamut when forced to adopt the luminance structure of the target painting. This motivates the use of a luminance-conditioned filtering strategy when constructing the reference-based gamut.

Figure 6.6 illustrates the effect of applying the luminance-conditioned filtering procedure to a real reference painting. The left panel shows a representative work by Torres García used as a chromatic reference. The right panel displays the subset of pixels in the reference image whose luminance equals the target luminance $L \approx 48$. Only the (a, b) chromaticities of these pixels are shown.

The filtered chromaticities form a coherent region of CIELAB space, corresponding to chromaticities that are actually plausible at $L \approx 48$ in the reference image. Note, in Figure 6.6, how only certain values of the gamut are allowed.

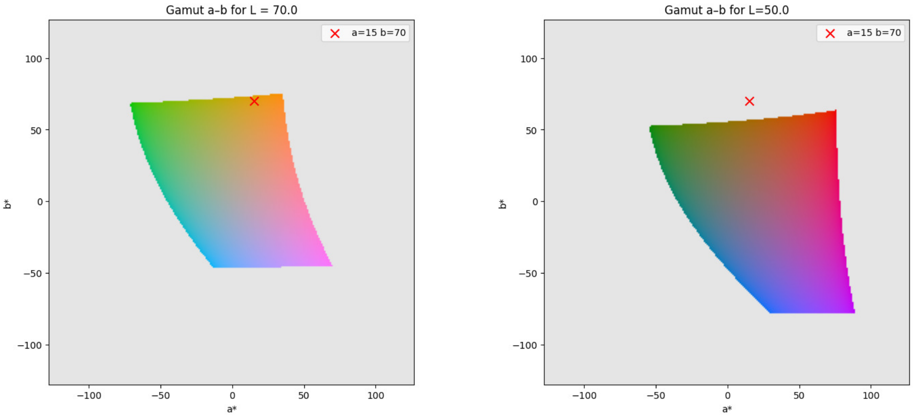


Figure 6.4: Effect of luminance L on chromatic validity and perceptual appearance in the CIELAB space. Left: Valid (a, b) gamut for $L = 70$ with $(a = 15, b = 70)$ as an in-gamut sample. Right: Valid (a, b) gamut for $L = 50$ with $(a = 15, b = 70)$ an out-of-gamut sample.

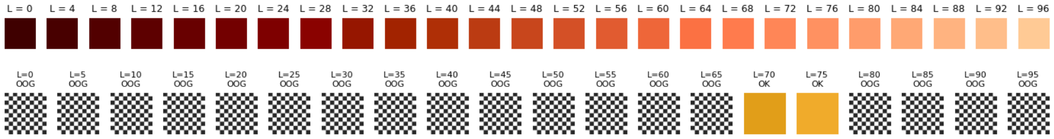


Figure 6.5: The effect of luminance variation on the CIELAB space. These visualizations illustrate why it is not possible to transfer reference colors directly unless their luminance structure matches that of the target image. The first row shows the same fixed $(a = 50, b = 50)$ values combined with different L values. Second row: For a fixed $(a = 15, b = 70)$, only some luminance levels are in-gamut for the considered (a, b) values.

This luminance-conditioned sampling forms the basis of the reference-conditioned gamut: a set of chromatic options that are valid for the target (hint) luminance and stylistically grounded in the reference artwork.

The application now includes a dedicated reference module that allows experts to load an external artwork and extract chromatic information from it. Upon selecting a reference image, the interface displays both the painting itself and the luminance-filtered (a, b) gamut derived from its pixels, enabling users to visualize only those chromaticities that are compatible with the luminance of the target region, as shown in Figure 6.7. To support precise sampling, a pixel-level zoom tool allows experts to inspect small regions of the target grayscale image and select specific areas. At the same time, a zoomable version of the gamut is provided, allowing fine-grained hint selection (for both reference and original gamut). Together, these interface components transform the demo into a flexible color–selection environment where the user can alternate between free exploration of the color gamut and

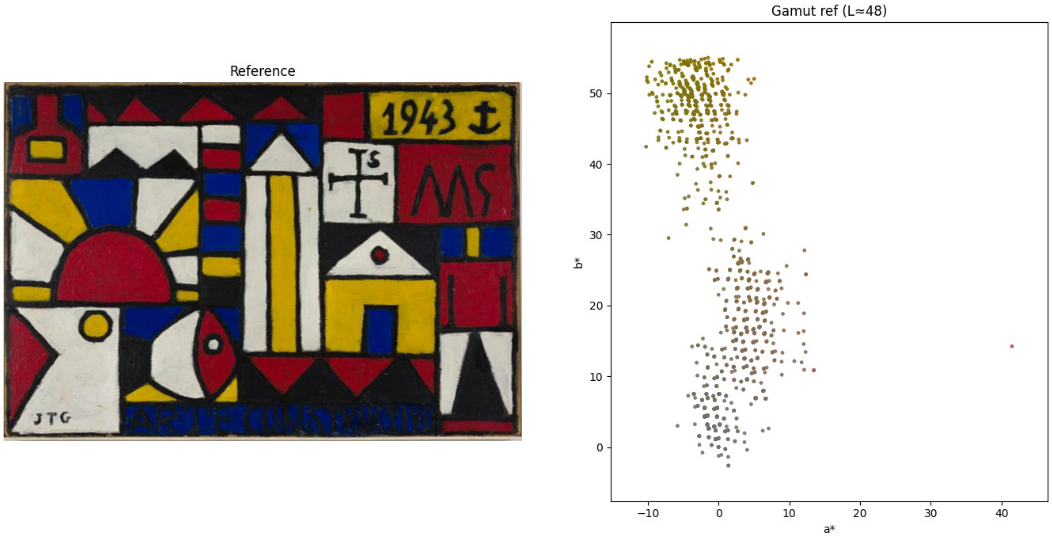


Figure 6.6: Luminance-conditioned chromatic sampling from the reference painting. The left side shows the selected reference artwork, Painting 1943.26, while the right side displays the (a, b) chromaticities of all reference pixels whose luminance matches the target luminance level ($L \approx 48$).

reference-driven sampling, resulting in more informed and historically grounded chromatic reconstructions.

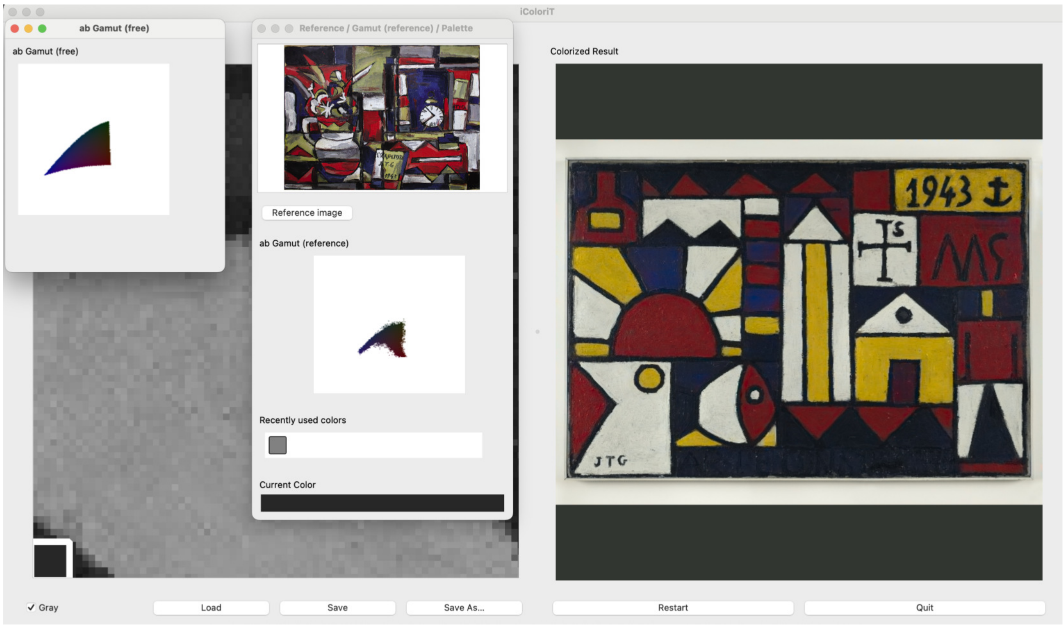


Figure 6.7: Interaction between the partial gamut (reference-based gamut) panel and the full CIELAB gamut panel (denoted by Free-gamut in the demo). Note that there are only a few pairs of values, (a, b), in the partial gamut that are compatible with the given hint's luminance value. Painting 1943.26.

6.1.3. Example with an image of a destroyed painting.

The image shown in Figure 6.8 corresponds to a scanned photographic reproduction of a painting lost in the 1978 fire at the MAM Rio. The original acquisition conditions of this photograph are unknown: there is no available information regarding the camera, film type, year of capture, or whether the image was originally recorded in grayscale or later converted. As a result, the chromatic content of the original artwork is entirely absent, and the scan constitutes the only visual record we have. The first result obtained with the selected LoRA-adapted model is a fully automatic colorization without user intervention, as illustrated in Figure 6.8.

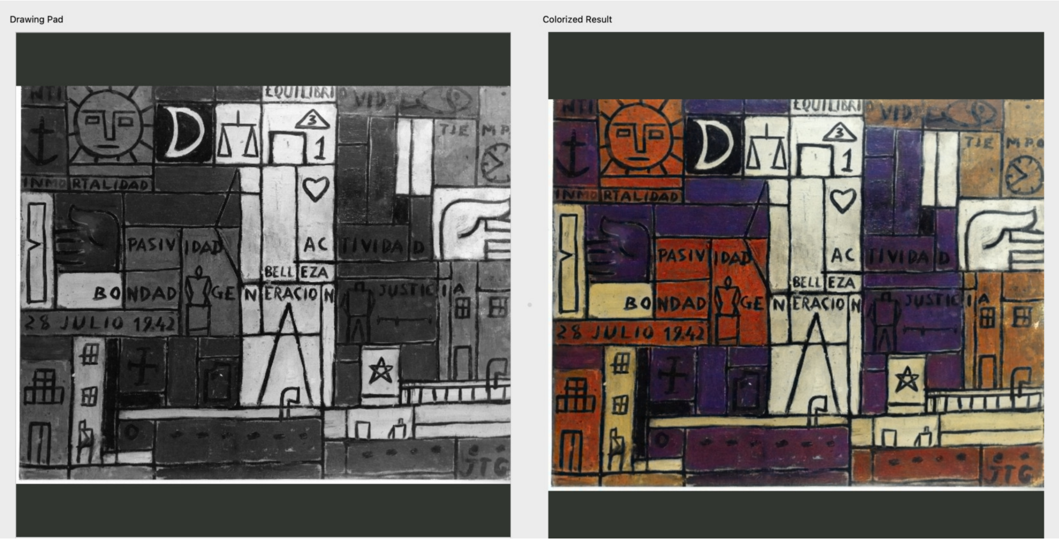


Figure 6.8: The scanned grayscale photographic reproduction of the lost Painting 1942.25 was automatically colorized using the selected Best-Global LoRA-adapted model. No user-provided hints were applied to this initial inference.

Given the stylistic characteristics of the scanned image and its estimated period of production, the artwork 1943.26 constitutes a plausible chromatic reference. Although no definitive correspondence between the two works can be established, their temporal proximity and visual similarities suggest using the 1943.26 painting as a reference image for color inference. Both images are shown in Figure 6.9.

After manually placing 50 color hints, the resulting colorization corresponds to the image in Figure 6.10. Each hint was sampled from the chromatic gamut of the reference image and constrained to be compatible with the luminance of the corresponding region in the target image. This procedure restricts the admissible

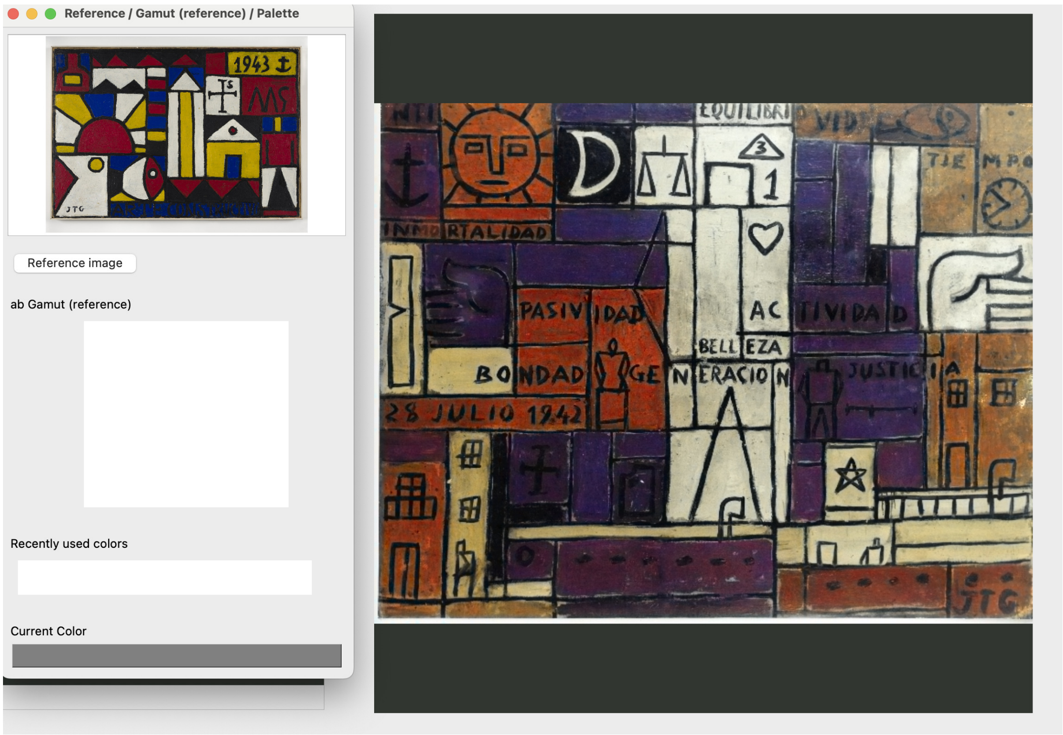


Figure 6.9: First automatic inference of the selected LoRA-adapted model for a scanned grayscale image of a lost artwork, incorporating a reference image from a 1943.26 artwork to guide color inference.

color space at each location, enabling the model to propagate chromatic information in a perceptually plausible and consistent manner with the underlying luminance structure.

As shown in Figure 6.10, achieving a satisfactory restoration may require several iterations of user interaction, gradually refining the colorization through the addition of further hints. It should be noted that the example shown here was generated by the author of the thesis rather than by expert practitioners. Improved results are therefore expected when the system is used by domain experts with in-depth knowledge of Joaquín Torres García's painting, particularly through the selection of more appropriate reference images.

To support this workflow, the interactive tool is made available to the museum as part of an ongoing chromatic restoration process for images of Joaquín Torres García's lost works. Through iterative feedback sessions, curators have identified potential extensions to the workflow—such as preprocessing steps for luminance calibration between target and reference images—that could further improve chromatic

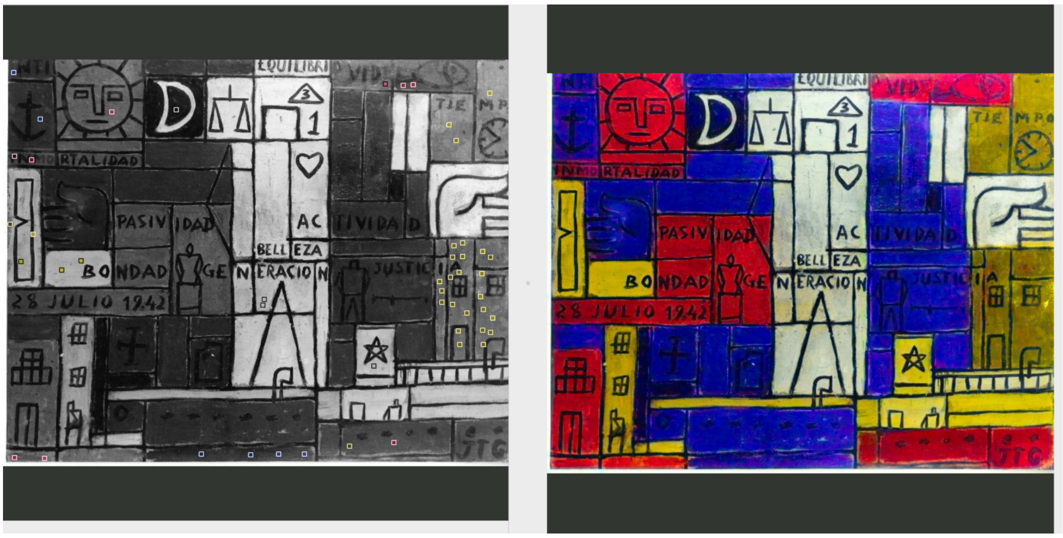


Figure 6.10: Colorization result obtained after providing 50 user-defined color hints sampled from a reference image and constrained by the luminance-dependent chromatic gamut of the target image.

consistency and alignment. These extensions were not addressed in the present work and are therefore left as directions for future research.

Chapter 7

Conclusion and Future work

This work explores adapting deep learning models for chromatic restoration of artworks, especially when the originals are unavailable.

Building on this premise, different possibilities and limitations of image colorization methods were examined, including fully automatic approaches, automatic methods with reference images, and automatic methods with color hints/scribbles. Based on this analysis, a system was developed and evaluated that combines the efficient adaptation of an automatic–interactive model, a systematic study of its behavior under different levels of guidance, and the adaptation of an interactive tool designed to support curatorial work.

7.1. Methodological Adaptation and Data Curation

Need for Interactivity. It became evident that purely automatic colorization models, such as BigColor (Kim et al., 2022), are not suitable for artistic restoration tasks. These methods rely exclusively on color priors learned from large-scale image datasets and produce plausible colorizations based on those priors. Depending on the training strategy, this colorization can sometimes lack vividness, and it offers no control whatsoever over the colorization result. Furthermore, since the color priors are derived solely from the training dataset, these automatic models always require additional training.

Color restoration of lost artwork cannot be automated for two fundamental reasons. First, automatic methods do not provide sufficient control over the output to account for stylistic, symbolic, or historically based color choices. Second, due to the nature of the problem, there is no way to ensure that any inferred colorization corresponds to the original artwork because the true chromatic information is irretrievably lost. Incorporating a human in the loop directly addresses the first limitation by enabling guided control through user hints. This motivated the selection of iColoriT (Yun et al., 2023), a hybrid, interactive method combining learned priors and user hints, as a more suitable basis for artistic restoration workflows.

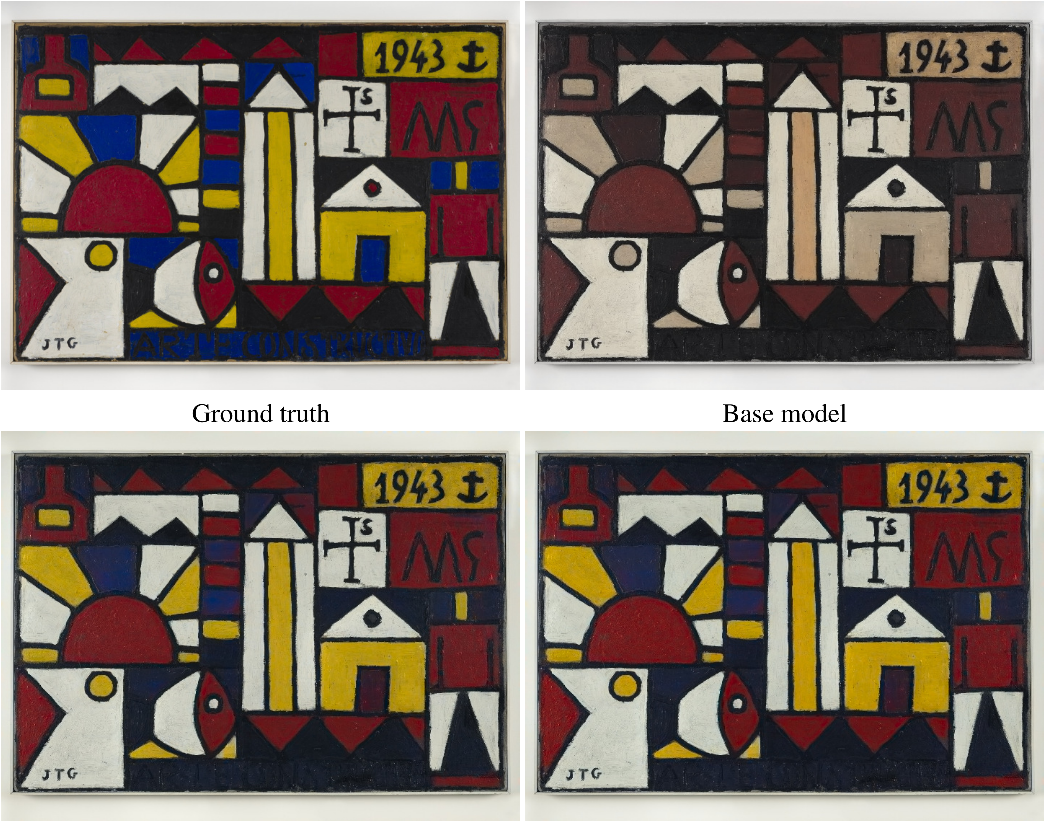
Efficiency of LoRA. The Low-Rank Adaptation (LoRA) strategy (Hu et al., 2022) proved to be the most efficient approach for fine-tuning the Base model, as it adapts the network by training only a small subset of parameters while keeping the backbone frozen. By substantially reducing the number of trainable parameters, LoRA lowers the computational cost and mitigates overfitting of fine-tuning, making it well-suited for scenarios with limited training data. Unlike full fine-tuning, which updates all network weights and requires large datasets for reliable generalization, LoRA enables effective adaptation under data-constrained conditions.

Impact of the Data. The effort invested in curating diverse data sources (IMGM, IMGCec, IMGP, IMGL, etc.) revealed that a broader range of data types improves performance. Quantitatively, dataset D_4, which incorporated a broader and more diverse collection of the artist's works and those of his disciples, produced the best overall PSNR and LPIPS results.

7.2. Adapted Model: Key Results

Quantitative and qualitative analyses confirmed that the model adapted through LoRA — specifically the Best-Global configuration achieving the highest mean

PSNR ($r = 32$, D_4, $lr = 1\mathrm{e}{-3}$) — outperformed the Base model in the specific context of Joaquín Torres García's artworks. Importantly, the performance gap between this configuration and the best LPIPS-mean setup ($r = 8$, D_4, $lr = 1\mathrm{e}{-3}$) was relatively small. Given the substantial reduction in trainable parameters when moving from rank 32 to rank 8, this result suggests that, under computational constraints, a LoRA rank of 8 can achieve performance close to the best-performing configuration while requiring significantly fewer resources. Figure 7.1 illustrates these effects by comparing the Base model inference with the best PSNR-mean and LPIPS-mean LoRA configurations under the zero-hint setting.



Ground truth

Base model

Best PSNR-mean

Best LPIPS-mean

Figure 7.1: 0 hint colorization of "Arte constructivo" (1943). Top row, from left to right: ground truth, result using the Base model. Bottom row, from left to right: result using the best PSNR-mean LoRA model and result using the best LPIPS-mean LoRA model.

Improvement in Colorization. The most significant improvement was observed in the regime of low or no user hints (0-20). With zero hints the best overall results, the PSNR-best-0 model outperformed the Base by approximately 1.60 dB in PSNR and $-$ 0.0162 in LPIPS, meanwhile the Best Global outperformed the Base by +0.8940 dB in PSNR and -0.0154 in LPIPS, indicating that fine-tuning effectively reinforced the internal chromatic priors to better align with the artist's palette.

Convergence Under Hint Guidance. As the number of hints increases (beyond 50), the performance of the LoRA models and the Base model converges. This shows that user-provided hints dominate the inference process in densely guided scenarios and that the Base model already propagates color effectively with a large number of hints.

A comparison of distortion and perceptual metrics. The comparison between models optimized for PSNR and those optimized for LPIPS revealed not many distinct chromatic tendencies. However, the PSNR-optimized model was selected as the best trade-off between the two metrics.

Qualitative Limitations. Despite strong overall chromatic coherence, the reconstructions still oversimplify fine paint textures, natural pigment bleeding, and fail to recover subtle transitions or semantically dense regions, even under high hint density as was analyzed in 5.3.3.

7.3. Practical Contribution: The Interactive Tool

The work culminated in the adaptation of an interactive software demo that integrates the best-performing LoRA model (Best-Global configuration) and adds functionality specifically designed for expert-guided restoration.

Usability and Traceability. The new features enable experts to save complete hint sessions, including coordinates and chromatic values. This ensures full traceability and reproducibility of colorization decisions.

Luminance-Conditioned Color Selection. A reference-based color sampling method conditioned on luminance was implemented. This method addresses an inherent issue in the CIELAB color space by preventing the transfer of colors that would appear incoherent or invalid. This functionality enables color guidance to remain style- and history-informed.

Real-World Impact. The adapted tool is currently being used by the Museo Torres García in the chromatic restoration workflow for images of lost artwork. Beyond its current use as an exploratory, expert-guided restoration instrument, the long-term objective is to use this tool to create colorized versions of surviving photographic records of destroyed works. These results will be disseminated through a dedicated book publication. This would make the reconstructions accessible to a broader audience while preserving a clear distinction between the original works and the digitally inferred restorations.

7.4. Future Work

Based on the results obtained and the limitations observed, several avenues for future research emerge that could enrich and expand upon the approach presented in this work.

7.4.1. Improving Hint Integration and Understanding Hint Propagation

Experiments show that although hints effectively guide chromatic propagation, the model does not always incorporate them with the expected strength, especially in specific or small regions of the image where colors are either not propagated or propagated incorrectly, as seen in Figure 4.16. This phenomenon deserves deeper investigation. Promising directions include:

- Analyzing hint propagation across network layers by identifying where hint information weakens or transforms throughout the network. Explainable AI techniques (Kashefi et al., 2023) could help visualize the contribution of each hint to the chromatic decision process and reveal regions where the model ignores or dilutes the provided guidance. A preliminary exploration in this direction was conducted in this thesis by inspecting the intermediate outputs of the 12 transformer blocks; however, a more quantitative analysis is left for future work.
- Exploring explicit reinjection of hints in deeper layers: for example, through recurrent injections. In this direction, inspired by Modular Co-Attention (MCAN) (Yu et al., 2019), iColoriT's standard attention mechanism could be replaced with Guided Attention (GA) units. In this setup, user color hints act as the guiding modality for the grayscale image features, modeling dense hint-to-region interactions to ensure more precise color propagation.
- Investigating interactive segmentation methods: such as those inspired by Lebon et al., 2023. These methods could help automatically identify areas where color propagation is difficult, suggesting new hint placements.

7.4.2. Exploring Dataset Chromatic Representativity and New Data Augmentation Strategies

Another important direction is to examine the chromatic representativity of the dataset, assess which tones are overrepresented or missing, and explore augmentation strategies oriented toward chromatic diversity.

A promising line of work could be to explore the use of synthetic images as regularity priors, following the ideas of Achddou et al., 2021 for image denoising and super-resolution. Synthetic images generated under controlled distributions could help improve generalization and reduce overfitting to idiosyncratic features of the real dataset.

Additionally, style-specific models could be explored by training on subsets organized by stylistic criteria (e.g., Classical, Modern, Universal Constructive phases), allowing the network to capture chromatic distributions more faithfully.

Finally, now that four curated datasets have been constructed and analyzed in detail (see Chapter 3), an interestinge experiment would be to fine-tune alternative colorization models such as BigColor Kim et al., 2022. Fine-tuning such models on stylistically and chromatically constrained subsets could provide a complementary perspective to the interactive approach explored in this thesis, and help disentangle the role of dataset composition from that of user guidance in artistic color restoration.

References

Achddou, R., Gousseau, Y., and Ladjal, S. (2021, April). Synthetic images as a regularity prior for image restoration neural networks.

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. (2021). Intrinsic dimensionality explains the effectiveness of language model fine-tuning [Originally posted as arXiv:2012.13255 (2020)]. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7319–7328.

Automated flower classification over a large number of classes. (2008). *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, 722–729.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. https://arxiv.org/abs/1607.06450

Bai, Y., Dong, C., Chai, Z., Wang, A., Xu, Z., and Yuan, C. (2021). Semantic-sparse colorization network for deep exemplar-based colorization. *arXiv preprint arXiv:2112.01335*.

Bergasa, J. (2014). Tvs. https://www.noticiasdenavarra.com/comunicacion/2014/05/05/cierre-9-canales-tdt-afecta-3005176.html

Bertero, M., and Boccacci, P. (1998). *Introduction to inverse problems in imaging*. Institute of Physics Publishing.

Bishop, C. M., and Bishop, H. (2023). *Deep learning: Foundations and concepts*. Springer.

Chollet, F., et al. (2015). *Keras*. https://keras.io

CIE. (1932). *Commission internationale de l'éclairage proceedings, 1931* (tech. rep.). CIE. Cambridge, UK.

CIE. (2022a). *Cie 248:2022 — an open colour appearance model (cam16-ucs) and cam16-lch* (tech. rep. No. CIE 248:2022). CIE. Vienna, Austria.

CIE. (2022b). *Cie standard illuminant d65* (International Standard No. ISO/CIE 11664-2:2022). CIE. Vienna, Austria.

Clusteruy: Centro nacional de supercomputación. (n.d.). Retrieved November 2, 2025, from https://cluster.uy/

Defazio, A., Mishchenko, K., and Cutkosky, A. (2024). The road less scheduled. *arXiv preprint arXiv:2405.15529*. https://arxiv.org/abs/2405.15529

Deshpande, A., Lu, J., Yeh, M.-C., Chong, M. J., and Forsyth, D. (2017). Learning diverse image colorization. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding [Originally posted as arXiv:1810.04805 (2018)]. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Fabregat, A. L., and Acosta, W. (1979). Joaquín torres garcía, su vida y obra. 1874–1949. https://archivosdocumentales.udelar.edu.uy/index.php/joaquin-torres-garcia-su-vida-y-obra

Ford, A., and Roberts, A. (1998). Colour space conversions. https://poynton.ca/PDFs/coloureq.pdf

Gabrieli, F., Delaney, J. K., Erdmann, R. G., Gonzalez, V., van Loon, A., Smulders, P., Berkeveld, R., van Langh, R., and Keune, K. (2021). Reflectance imaging spectroscopy (ris) for operation night watch: Challenges and achievements of imaging rembrandt's masterpiece in the glass chamber at the rijksmuseum. *Sensors*, *21*(20).

García, R., Randall, G., and Raad, L. (2024a). A Brief Analysis of iColoriT for Interactive Image Colorization. *Image Processing On Line*, *14*, 127–143.

García, R., Randall, G., and Raad, L. (2024b). A Short Analysis of BigColor for Image Colorization. *Image Processing On Line*, *14*, 144–158.

Giner, F. (2003). *Los murales de torres garcía*. Entrelíneas Editores.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS).*

He, M., Chen, D., Liao, J., Sander, P. V., and Yuan, L. (2018). Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, *37*(4).

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS).*

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning (ICML).*

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR).*

Hunt, R. W. G., and Pointer, M. R. (2011). *Measuring colour* (4th). John Wiley; Sons, Ltd.

HunterLab. (2015). *The basics of color perception and measurement* [Application Note]. Retrieved November 1, 2025, from https://www.hunterlab.com/basics-of-color-theory/

Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let there be color! joint end-to-end learning of global and local image priors. *ACM Transactions on Graphics (TOG)*, *35*(4).

International Electrotechnical Commission. (1999). *Multimedia systems and equipment - colour measurement and management - part 2-1: Colour management - default rgb colour space - srgb* (Standard No. IEC 61966-2-1:1999). International Electrotechnical Commission. Geneva, Switzerland. https://webstore.iec.ch/publication/6169

ISO/CIE. (2019). *Iso/cie 11664-4:2019. colorimetry – part 4: Cie 1976 l*a*b* colour space* (tech. rep.). ISO/CIE. Vienna, Austria.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Jawade, B. (2023, December 22). *Understanding lora – low rank adaptation for finetuning large models* [Accessed: 2025-01-10]. https://towardsdatascience.com/understanding-lora-low-rank-adaptation-for-finetuning-large-models-936bce1a07c6

Joaquín Torres García Catalogue Raisonné. (2003). Retrieved January 2, 2025, from https://www.torresgarcia.com/

Kamal, K., and Ez-Zahraouy, H. (2023, April). A comparison between the vgg16, vgg19 and resnet50 architecture frameworks for classification of normal and clahe processed medical images.

Kang, S. H., and March, R. (2007). Variational models for image colorization via chromaticity and brightness decomposition. *IEEE Transactions on Image Processing*, *16*(9), 2251–2261.

Kashefi, R., Barekatain, L., Sabokrou, M., and Aghaeipoor, F. (2023). Explainability of vision transformers: A comprehensive review and new perspectives. https://arxiv.org/abs/2311.06786

Kim, G., Kang, K., Kim, S., Lee, H., Kim, S., Kim, J., Baek, S.-H., and Cho, S. (2022). Bigcolor: Colorization using a generative color prior for natural images. *European Conference on Computer Vision (ECCV)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, *25*.

Kumar, M., Weissenborn, D., and Kalchbrenner, N. (2021). Colorization transformer. *arXiv preprint arXiv:2102.04432*.

Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. *European Conference on Computer Vision (ECCV)*.

Lebon, Q., Lefevre, J., Cousty, J., and Perret, B. (2023, November). Interactive segmentation with incremental watershed cuts.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Yere, Y., and Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–44.

Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. *ACM Trans. Graph.*, *23*(3), 689–694.

Liang, Z., Li, Z., Zhou, S., Li, C., and Loy, C. C. (2025). Control color: Multimodal diffusion-based interactive image colorization: Z. liang et al. *International Journal of Computer Vision*, *133*(11), 7897–7923.

Loshchilov, I., and Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations (ICLR)*.

Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Marrocchesi, A., and Erdmann, R. G. (2024). Empowering cultural heritage photography: A low-cost, automated, open-source approach. In *Ahm conference 2024: Heritage, memory and material culture*. Amsterdam University Press.

Mitric, J., Radulovic, I., Popovic, T., Scekic, Z., and Tinaj, S. (2024). Ai and computer vision in cultural heritage preservation. *28th International Conference on Information Technology (IT)*, 1–4.

Museo Torres García. (2018). *Tiempo de mirar (1978–2018)*. Retrieved February 23, 2025, from https://www.torresgarcia.org.uy/novedad/tiempo-de-mirar-1978-2018

Museo Torres García. (2024). *Joaquín torres garcía: 150 aniversario*. Retrieved November 16, 2025, from https://www.torresgarcia.org.uy/exposiciones/150/

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 46–54.

Pucci, R., Micheloni, C., and Martinel, N. (2021). Collaborative image and object level features for image colourisation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2160–2169.

Riverbank Computing. (2023). *Pyqt5* (Version 5.15.9). https://pypi.org/project/PyQt5/

Royer, A., Kolesnikov, A., and Lampert, C. H. (2017). Probabilistic image colorization. *arXiv preprint arXiv:1705.04258*.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, *60*(1), 259–268.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3).

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. (2022). Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Samarasinghe, M. (2023). *Feature extraction and reverse image search using vgg16* [Medium tutorial]. https://medium.com/@madawasamarasinghe/

Sarkar, B., and Singh, P. (2025, October). Chapter-1 ai and humanities: Reimagining cultural heritage preservation.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: An astounding baseline for recognition. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR).*

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2022). How to train your vit? data, augmentation, and regularization in vision transformers [Originally posted as arXiv:2106.10270 (2021)]. *Transactions on Machine Learning Research (TMLR).* https://openreview.net/forum?id=4nCV0R4r8g

Su, J.-W., Chu, H.-K., and Huang, J.-B. (2020). Instance-aware image colorization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Thickstun, J. (2022). Transformers [Available at https://johnthickstun.com/docs/transformers.pdf]. Retrieved November 2, 2025, from https://johnthickstun.com/docs/transformers.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS).*

Vitoria, P., Raad, L., and Ballester, C. (2020). Chromagan: Adversarial picture colorization with semantic class distribution. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.*

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2011). *The caltech-ucsd birds-200-2011 dataset* (tech. rep. No. CNS-TR-2011-001). California Institute of Technology.

Welsh, T., Ashikhmin, M., and Mueller, K. (2002). Transferring color to greyscale images. *ACM Trans. Graph.*, *21*(3), 277–280.

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., and Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. https://arxiv.org/abs/2312.12148

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems (NeurIPS).*

Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering. https://arxiv.org/abs/1906.10770

Yun, J., Lee, S., Park, M., and Choo, J. (2023). Icolorit: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1787–1796.

Zeiler, M. D., and Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision (ECCV).*

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. *European Conference on Computer Vision (ECCV).*

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., and Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors. https://arxiv.org/abs/1705.02999

Zhang, Y., and Bunyasakseri, T. (2025). Artificial intelligence-based color reconstruction of mogao grottoes murals. *International Journal of Computational and Experimental Science and Engineering*.

Zhao, H., Liu, Y., and He, D. (2021, June). Color2style: Real-time exemplar-based image colorization with self-reference learning and deep feature modulation.