

Fine-Tuning the Zhang et al. Image Colorization Model on Indian Cultural Heritage Domains

Radheshyam Vishwakarma¹ Sudhanshu Shekhar Dadsena²;

¹Student, Department of CSE; ²Assistant Professor, Department of CSE

^{1,2}Shri Rawatpura Sarkar University, Raipur (C.G.)

Email-id- ssdadsena70@gmail.com, Email-id- itsradhes27@gmail.com

ABSTRACT

Automatic image colorization has demonstrated impressive results on natural scene datasets, yet pre-trained models struggle with domain-specific imagery such as Indian cultural heritage—vibrant festivals, classical dance forms, temple architecture, and traditional art. This paper presents a transfer-learning pipeline that fine-tunes the Zhang et al. (ECCV 2016) colorization convolutional neural network on a curated seven-domain Indian cultural dataset comprising 1,233 images across architecture, art, culture, dance, festivals, music, and other categories. Training was conducted entirely on Google Colab's free-tier T4 GPU using PyTorch with a two-stage fine-tuning strategy: early encoder layers were frozen to preserve low-level feature representations while deeper layers were adapted to Indian cultural colour palettes over 20 epochs. The fine-tuned model achieves an overall PSNR of **20.35dB** and SSIM of **0.8587** on a held-out test set of 107 images, with music (22.91dB) and architecture (21.94dB) achieving the highest per-domain scores. The complete pipeline is described in reproducible detail to enable adoption by resource-constrained academic researchers.

Keywords: Image Colorization, Transfer Learning, Indian Cultural Heritage, Convolutional Neural Networks, PyTorch, Google Colab, PSNR, SSIM, CIELAB, Zhang et al., Domain Adaptation.

Note: This manuscript is prepared for academic and project submission purposes only. Publication metadata such as DOI, issue number, page number, and journal indexing information will be assigned only after official journal acceptance and publication.

I. INTRODUCTION

Background

Colourization of grayscale images is a long-standing computer vision challenge with practical applications in historical photograph restoration, film archiving, and cultural heritage preservation [1]. Deep learning methods—particularly convolutional encoder-decoder networks operating in the CIELAB colour space—have demonstrated remarkable skill in predicting plausible colour distributions for natural images [2].

India possesses an extraordinarily rich visual heritage spanning temple architecture, classical dance costumes, folk art traditions, and vibrant festivals. Monochrome photographs of these subjects present unique challenges: highly saturated, culturally constrained palettes and fine-grained textures that generic models trained on Western datasets fail to reproduce faithfully [3].

Research Problem

State-of-the-art colorization networks trained on Image Net or Places365 exhibit colour desaturation and palette confusion when applied to Indian cultural imagery. No reproducible, resource-constrained pipeline exists for adapting these models to this specific domain [4].

II. OBJECTIVES

- Curate a seven-domain Indian cultural image dataset with 80/10/10 train/val/test splits.
- Fine-tune the Zhang et al. eccv 16 colorization CNN on Google Colab free-tier T4 GPU.
- Evaluate the model using PSNR and SSIM on a held-out test set of 107 images.
- Provide a fully reproducible six-cell Colab workflow for academic use.
- Analyse per-domain colorization quality across seven cultural categories.

III. LITERATURE REVIEW

Zhang et al. [1] introduced a fully automatic colorization framework Treating colour prediction as a multinomial classification problem over 313 quantised a^*b^* bins in CIELAB space. Their class-rebalancing loss and annealed-mean inference remain the foundation of modern colorization pipelines and serve as the base model in this work.

Iizuka et al. [5] proposed a joint global-local feature fusion network that leverages high-level semantic information alongside local texture cues to guide colour assignment, demonstrating that semantic context substantially improves colourization plausibility. Su et al. [6] extended this with instance-aware colorization, achieving state-of-the-art results on COCO benchmarks.

Isola et al. [7] demonstrated that conditional GANs can perform image-to-image translation including grayscale-to-colour synthesis, though GAN-based methods demand GPU memory that exceeds free-tier Colab limits. Vitoria et al. [9] applied chromatic adaptation to historical photographs but did not target Indian cultural domains—the gap this research directly addresses.

IV. RESEARCH METHODOLOGY

Dataset Curation

Images were collected from Creative Commons-licensed sources and organized into seven domain directories. PIL-based augmentation (horizontal flip and $\pm 15^\circ$ rotation) was applied to categories with fewer than 150 training images: art (108→150), dance (86→150), and music (70→150). Table 1 summarises the final split.

Table 1: Dataset split. *=augmented to 150.

Category	Train	Val	Test	Total
Architecture	185	23	24	232
Art	150*	13	14	177
Culture	140	18	18	176
Dance	150*	11	11	172
Festivals	120	15	15	150
Music	150*	9	9	168
Other	126	16	16	158
Total	1021	105	107	1233

Images were converted to CIELAB colour space using scikit-image. The L^* (lightness) channel serves as model input; a^*b^* chrominance channels are the prediction targets. All images were resized to 256×256 pixels.

Base Model

The Zhang et al. eccv16 model was obtained by cloning the official GitHub repository (richzhang/colorization) and inserting its path into sys.path—bypassing the unavailability of the colorizers package on PyPI. Pretrained ImageNet weights (≈ 130 MB) were loaded automatically on first import.

Fine-Tuning Strategy

Early encoder layers (model1, model2) were frozen to preserve low-level ImageNet feature representations (edges, gradients, textures). All remaining layers were trained with Adam optimiser ($\text{lr}=1\text{e-}4$, weight decay= $1\text{e-}5$) and cosine annealing scheduler over 20 epochs. Batch size=16. MSE loss was computed on unnormalised a^*b^* predictions against ground-truth chrominance channels. The best validation checkpoint was auto-saved to Google Drive after each improvement.

Colab Workflow

The complete pipeline was implemented in six sequential Colab cells:

- **Cell1:** git clone the colorization repo; insert path into sys.path.
- **Cell2:** Mount Google Drive; unzip dataset.
- **Cell3:** 80/10/10 split + PIL augmentation.
- **Cell4:** IndianColorDataset—RGB→LAB conversion; L^* input, a^*b^* targets.
- **Cell5:** Fine-tuning loop; best val check point saved.
- **Cell6:** PSNR/SSIM evaluation per domain.

V. RESULTS

Training Convergence

Table 2 shows selected epoch results. Training loss dropped sharply after epoch1 (95,999→378) as gradients stabilized, then converged steadily to 374.21 By epoch 20. The large initial value reflects unnormalised MSE on a^*b^* channels (range ± 110) before the model adapts. Validation loss reached its minimum of 381.03 at epoch14, at which point the best checkpoint was saved.

Table2: Selected training/validation losses over 20 epochs.

Epoch	Train Loss	Val Loss	Saved
01	95999.44	392.98	Yes
02	378.20	384.43	Yes
03	377.37	383.10	Yes
06	376.17	381.35	Yes
10	375.84	382.68	—
14	375.16	381.03	Best
18	374.66	382.45	—
20	374.33	382.65	—

Per-Domain PSNR and SSIM

Table 3 presents evaluation on the 107-image test set. PSNR and SSIM were computed by comparing the full RGB colorized output against the original colour ground truth at 256×256 resolution.

Table 3: Per-domain PSNR and SSIM on 107 test images.

Domain	PSNR(dB)	SSIM	n
Architecture	21.94	0.9216	24
Art	19.54	0.8478	14
Culture	21.82	0.8964	18
Dance	19.39	0.7629	11
Festivals	18.02	0.8139	15
Music	22.91	0.9128	9
Other	18.84	0.8553	16
Overall	20.35	0.8587	107

Music achieved the highest PSNR (22.91dB), reflecting consistent instrument tones; architecture achieved the highest SSIM (0.9216) due to uniform stone textures. Festivals scored lowest (18.02dB) owing to extreme saturation diversity across Holi, Diwali, and Durga Puja imagery. Dance recorded the lowest SSIM (0.7629) because costume colours vary widely across classical traditions (Bharatanatyam, Kathak, and Odissi).

VI. DISCUSSION

The overall PSNR of 20.35dB compares favourably with baseline colorization on domain-shifted data, where off-the-shelf models typically score 15–18dB on out-of-distribution imagery due to palette mismatch [1, 3]. Freezing early encoder layers proved effective: these layers encode universal low-level features that transfer well across domains, while unfrozen deeper layers adapted colour predictions to Indian cultural hues — saffron temple stones, indigo classical costumes, and earth-toned folk art.

A key limitation is dataset size: with 70–185 raw images per category, the model has limited exposure to intra-class colour variation. The festi- val domain illustrates this clearly—120 training examples cannot cover the full palette diversity of Indian festivals. PIL-based augmentation adds structural variation but not chromatic diversity, which future work should address through colour jitter augmentation and larger curated datasets [9].

VII. FUTURE SCOPE

- Expand the dataset to 500+ images per domain sourced from museum archives and ASI digital collections.
- Incorporate the siggraph17 user-guided colorization variant for culturally ambiguous regions.
- Adopt a perceptual loss function (VGG feature matching) to improve quality on highly saturated festival imagery.
- Extend to video colorization for archival Doordarshan and documentary footage of Indian cultural performances.
- Deploy a lightweight mobile model for on-device heritage photograph restoration.
- Collaborate with the Archaeological Survey of India (ASI) and the National Museum of India to build a curated benchmark dataset.

VIII. CONCLUSION

This paper presented a fully reproducible pipeline for fine-tuning the Zhang et al. image colorization model on Indian cultural heritage imagery, implemented entirely within Google Colab’s free T4GPU environment. A seven-domain dataset of 1,233 images was curated, split 80/10/10, and augmented for underrepresented categories. A two-stage fine-tuning strategy produced a model achieving an overall PSNR of **20.35dB** and SSIM of **0.8587** on a 107-image test set. Music and architecture domains achieved the strongest results (22.91dB and 21.94dB respectively), while festival imagery proved most challenging (18.02dB) due to extreme palette diversity. The complete six-cell Colab workflow is described in sufficient detail for full reproducibility without dedicated GPU hardware, making this approach accessible to resource-constrained researchers in cultural heritage computing.

REFERENCES

- [1] Zhang, R., Isola, P., & Efros, A.A. (2016). Colorful Image Colorization. *European Conference on Computer Vision (ECCV) 2016*. <https://arxiv.org/abs/1603.08511>
- [2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR2016*.
- [3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- [4] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://scikit-learn.org/>
- [5] Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2016). Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4), 1–11.
- [6] Su, J., Chu, H., & Jia, J. (2020). Instance-aware Image Colorization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020*.
- [7] Isola, P., Zhu, J.Y., Zhou, T., & Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR2017*.
- [8] Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. *ICLR2017*.
- [9] Vitoria, P., Raad, L., & Ballester, C. (2020). ChromaGAN: Adversarial picture colorization with semantic class distribution. *IEEE Winter Conference on Applications of Computer Vision (WACV) 2020*.
- [10] Wang, Z., Bovik, A.C., Sheikh, H.R., & Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.
- [11] McFee, B., et al. (2015). Librosa: Audio and Music Signal Analysis in Python. <https://librosa.org/>
- [12] Cho, K., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP 2014*.