# NATIONAL CONFERENCE ON INTEGRATING TECHNOLOGIES, IDEAS AND DISCIPLINES FOR ENGINEERING INNOVATION (NCITIDE 2025)

NCITIDE 2025

RAG-Enhanced Comic Panel Colorization Using GANs: A Retrieval-Augmented Approach for Context-Aware Automatic Colorization

**Paper ID - 17**

**Jaspreet Singh | MSIT | 18-19 Nov. 2025**

**Prakhar Chandra, Pulkit Kapur**

# Abstract


- This paper serves as a proof-of-concept for RAG-enhanced comic colorization using GANs for better contextual color guidance.
- Automatic comic colorization suffers from inconsistent colors due to missing contextual cues in grayscale panels.
- We propose a **Retrieval-Augmented Generation (RAG)** + **GAN** pipeline to provide contextual color guidance.
- A **U-Net GAN baseline** is first trained, followed by **RAG-enhanced fine-tuning** using CLIP-based visual similarity retrieval.
- A curated reference set (100 images) is created using **K-means clustering** on 1,900 paired, colored comic panels.
- Top-3 retrieved CLIP embeddings (1536-dim) are fused into the generator's bottleneck via learned embedding layers.
- RAG improves validation L1 loss from **0.42 → 0.37**, with visibly better color consistency and palette coherence.
- Establishes a foundation for context-aware creative AI for comics and visual storytelling.

# Introduction

NCITIDE 2025

- Comics rely on **consistent character colors**, emotional tones, and narrative-driven palettes.
- Manual colorization is labor-intensive and requires artistic expertise.
- Existing GAN-based colorizers often show:
- Arbitrary color choices
- Poor consistency across sequential panels
- Lack of semantic understanding
- **Key problem:** Grayscale panels contain *no color clues*.
- **Idea:** Add external context via *retrieval* to guide color generation.
- **Our contribution:**
- RAG + GAN hybrid for automatic comic colorization
- CLIP-based retrieval for semantic color reference
- Learned embedding fusion for context-aware colorization

# Literature Survey and Problem Statement

NCITIDE 2025

- **GAN-Based Image Colorization**

1. Pix2Pix: paired translation using U-Net + PatchGAN
2. CycleGAN: unpaired translation
3. Anime/comic colorizers exist but lack context-based consistency

- **CLIP for Visual Similarity**

1. CLIP embeddings capture **semantic + structural similarity**
2. Used in style transfer and creative generation
3. No major work applying CLIP retrieval to comic, or in general, image colorization.

- **Retrieval-Augmented Generation (RAG)**

1. Introduced for knowledge-intensive NLP tasks
2. Successful in diffusion models & image editing
3. Not yet explored for GAN-based colorization tasks

# Proposed Methodology

NCITIDE 2025

**1. Two-Stage Pipeline**

- **Stage 1:** Train baseline U-Net GAN
- **Stage 2:** Enable embedding fusion + fine-tune with retrieved CLIP context

**2. CLIP-Based Retrieval**

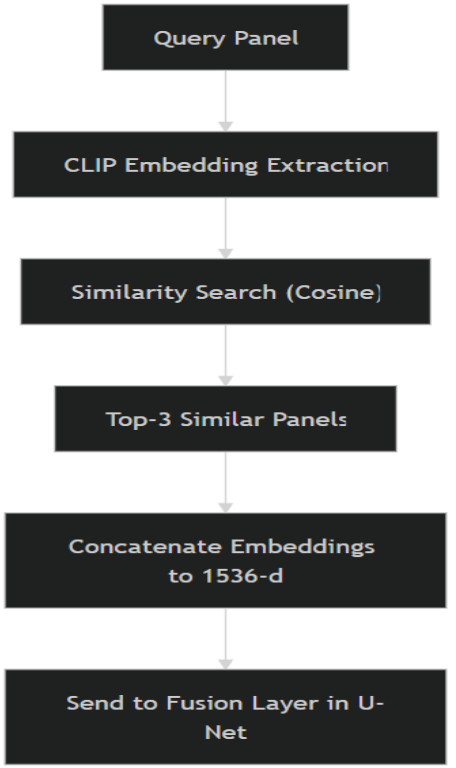
- Compute embeddings for all 1,900 colored panels
- K-means (k = 100) → representative reference set
- Retrieve top-3 matches using cosine similarity
- Concatenate embeddings → **1536-dim context vector**

**3. Embedding Fusion**

- Project context vector → 512-dim
- Inject into U-Net bottleneck via additive fusion
- Guides generator toward contextually plausible colors

**4. GAN Architecture**

- U-Net generator (64→1024 channels)
- PatchGAN discriminator
- Loss = L1 + perceptual + TV + adversarial




$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}$$

where

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\log D(x_{bw}, y_{color})\right] + \mathbb{E}\left[\log\left(1 - D(x_{bw}, G(x_{bw}))\right)\right]$$

$$\mathcal{L}_{L1} = \mathbb{E}\left[\|y_{color} - G(x_{bw})\|_1\right]$$

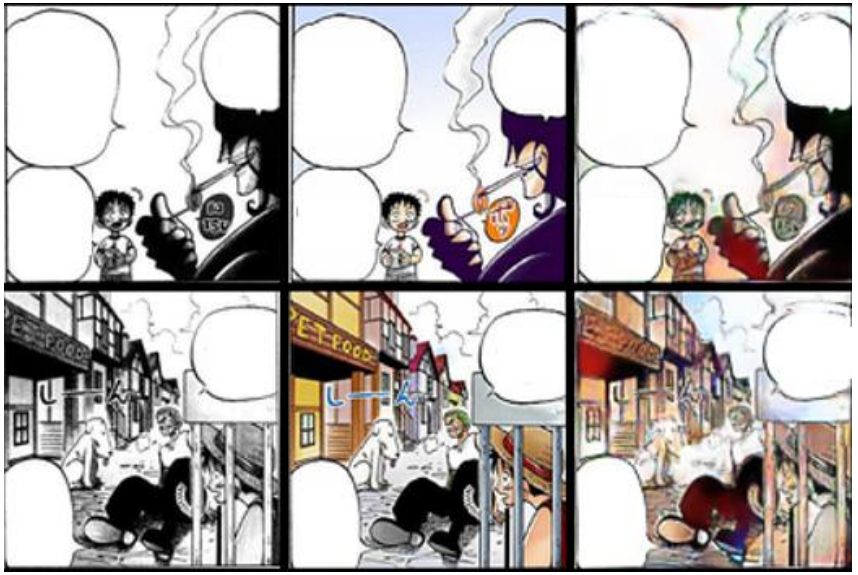
$$\mathcal{L}_{\text{perc}} = \mathbb{E}\left[\|\phi_{\text{VGG}}(y_{color}) - \phi_{\text{VGG}}(G(x_{bw}))\|_1\right]$$

$$\mathcal{L}_{\text{tv}} = \text{TV}\left(G(x_{bw})\right)$$

# Experimental Setup

NCITIDE 2025

- **Dataset**
  - 1,900 paired grayscale–color comic panels
  - 90/10 train/validation
  - 128×128 resolution
  - Augmentations: horizontal flip, light rotation
- **Training Setup**
  - PyTorch + AMP
  - Batch size: 16
  - Adam (LR = 2e-4)
  - 50 epochs total (baseline + fine-tuning)
  - Multi-GPU training (DataParallel)
- **Reference Database**
  - CLIP ViT-B/32 embeddings (512-d)
  - K-means = 100 clusters
  - Precomputed normalized embeddings for fast retrieval



# Results and Discussion

NCITIDE 2025

- **Quantitative Results**

1. **Baseline L1 Loss:** 0.42
2. **RAG-Enhanced L1 Loss:** 0.37
3. ≈12% improvement

- **Qualitative Observations**

1. Consistent character colors
2. Better background & environment realism
3. Improved hue/saturation stability
4. Semantic correctness from retrieved color cues

- **Training Behavior**

1. Smooth GAN convergence
2. No mode collapse
3. Embedding fusion integrates context without instability

- **Overhead**

1. +5% additional training time
2. Negligible extra inference overhead

# Conclusion and Future Scope

NCITIDE 2025

**Conclusion**

- RAG + GAN enables **context-aware comic colorization**.
- CLIP retrieval provides missing color cues, improving consistency.
- Achieves both quantitative & perceptual improvements.
- Demonstrates potential of retrieval-augmented creative AI.

**Future Scope**

- Higher-resolution colorization (512×512+)
- Attention-based embedding fusion
- Larger reference database
- Dynamic retrieval (variable top-k)
- Cross-panel temporal consistency
- Interactive artist-in-the-loop tools

# References

NCITIDE 2025


[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1125–1134.

[2] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Inf. Process. Syst., vol. 33, pp. 9459–9474, 2020.

[3] A. Blattmann et al., “Retrieval-augmented diffusion models,” in Advances in Neural Inf. Process. Syst., vol. 35, pp. 15309–15324, 2022.

[4] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2223–2232.

[6] Y. Ci et al., “User-guided deep anime line art colorization with conditional adversarial networks,” in Proc. ACM Int. Conf. Multimedia (MM), 2018, pp. 1536–1544.

[7] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “StyleCLIP: Text-driven manipulation of StyleGAN imagery,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 2085–2094.

[8] S. Sheynin et al., “KNN-diffusion: Image generation via large-scale retrieval,” arXiv preprint arXiv:2204.02849, 2022.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[10] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 694–711.

[11] I. Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241.