




Towards Flat Color Prediction for Comics

Marnix Verduyn^{1,2}, Thomas Winters¹, and Tinne Tuytelaars²

¹ KU Leuven, Dept. of Computer Science (DTAI), Belgium

² KU Leuven, Dept. of Electrical Engineering (ESAT), Belgium

Abstract. This extended abstract explores the partial automation of flat colorization in comic books, utilizing a ResNet-based approach and a Transformer-based approach with positional embeddings. Existing research in this field primarily focuses on models that fully colorize black line drawings, enabling amateurs to color comics. However, professional colorists require support that offers controllability and interactivity in the coloring process. A critical step towards this support is the prediction of colors based on previously published comic books from the same series. This study investigates the efficacy of these advanced computational models in facilitating a semi-automated coloring process that aligns with the needs of professional colorists, aiming to enhance productivity while maintaining artistic integrity.

1 Introduction

Applying flat colors in comic books is a labor-intensive, repetitive, and minimally creative task. Partial automation of this task can accelerate the comic book production process and allow artists to spend more time on creative work. In this extended abstract, we explore two primary models for predicting flat colors in comic book panels: a ResNet-based [7] approach and a Transformer-based [4, 14] approach.

The comic book creation process typically involves several consecutive steps: scripting, storyboarding, penciling, inking, and coloring. The coloring stage is often divided into two steps: applying flat colors and adding shadows, light effects and ambiance. In practice, flat coloring is done using drawing programs such as Clip Studio Paint or Adobe Photoshop. The task involves repeatedly selecting a color and then clicking on a region of white pixels within closed inked contours, typically using the bucket tool. A flood fill algorithm then assigns the selected color to all pixels in that region. If contours in the ink drawing are not closed, the artist will first close them with a brush tool in the correct color and then fill them with the bucket tool. The main objective here is to ensure that colors are consistent throughout the entire comic book.

Partial automation of this process yields significant time savings and makes the artist’s job more interesting with more time for creative sub-tasks.

The main objective of this extended abstract is to automate the prediction of flat colors for regions within comic book panels, which involves correctly assigning colors assuming regions have been identified.

2 Background and Related Work

Coloring with flat colors distinguishes two subtasks: detecting regions and predicting colors for these regions. Most existing research [5,8,9,12,13,15,17] focuses primarily on the first subtask, often leading to oversegmentation, which significantly increases the number of regions to be colored, making the repetitive task of clicking even more burdensome. Studies that do predict color [1,2,6,8,9,13,17] primarily aim to enable amateurs to color comics with minimal effort, turning line-art in colored drawings in a few steps, skipping the flat color stage and enabling limited steerability. The real challenge, however, lies in a partially automated approach that supports the work of a professional artist in an interactive way with maximum steerability, without adding extra workload. Accurate color prediction is crucial, which is the subject of this study. The majority of studies [1,2,6,8,9,13,15,17] rely on publicly available datasets [3,16] consisting of separate panels, usually in manga style, with mainly characters in the foreground and almost no background. Our study will use Flemish comic series datasets where multiple albums have been released with consistent colors across different albums, featuring panels that depict diverse scenes with characters, backgrounds, and speech balloons. In this pilot study, we restrict our analysis to a single comic book from the complete dataset.

3 Model Implementation

3.1 Data Set

The dataset comprises 255 comic panels from a single album (*cf.* Fig. 1a). Each panel is extracted and segmented (*cf.* Fig. 1b) using a contour detection algorithm [10,11], with each region labeled according to the original artwork’s colors, resulting in 37.578 regions of 248³ colors. Regions are isolated into rectangles with an additional 20-pixel padding. Input images retain black outlines, non-regional colors are converted to white, and the target color within each region is marked green (0,255,0) (*cf.* Fig. 1c). The panels are then shuffled and split into an 80%-20% train-validation ratio.

3.2 ResNet-based Approach

- **Objective:** Utilize a ResNet [7] architecture to predict the color of individual segments within comic book panels.
- **Architecture:** Employ ResNet-18, modified at the output layer to support 248-way color classification. The model is trained using cross-entropy loss to match the predicted colors with the actual segment colors.
- **Results:** Achieved a prediction accuracy of 0.41.

³ Vector quantization reduces the original count of 3,282 colors to 248. The high initial count results from anti-aliasing effects on the edges of color regions and gradient-filled areas.

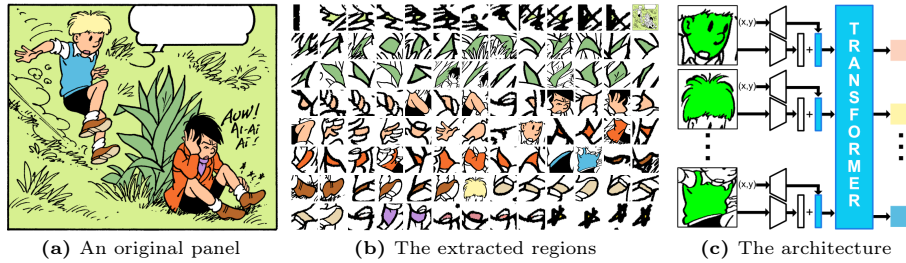


Fig. 1: The dataset consists of comic panels from the Flemisch comics series Jommeke © Jef Nijs, Standaard Uitgeverij. A transformer-based architecture is trained to predict the color of the panels, with the original color changed to $(0,255,0)$.

3.3 Transformer-based Approach with Positional Embeddings

- **Objective:** Enhance accuracy by incorporating positional embeddings (PE) in a Transformer model.
- **Data Preparation:** Additionally, positional information is encoded to help the model understand the spatial context of each region. The center of the rectangle containing a region is normalized with the height and width of the panel and scaled within the values $(-1, -1)$ and $(1, 1)$.
- **Architecture:** A Vision Transformer (ViT) [4] architecture processes entire panels, focusing on inter-region relationships using regions as tokens, unlike the typical ViT that uses image patches. Each region is represented by a single token, using feature embeddings from the initially trained ResNet. Panel centers’ x, y coordinates are converted into the transformer’s internal dimension via an MLP with one hidden layer, then added to the token embeddings (*cf.* Fig. 1c). To standardize sequence lengths across panels, sequences are padded to 200 tokens. For panels with more than 200 regions, we apply random sampling. The model, structured with four self-attention layers and four heads each, employs cross-entropy loss for classification tasks.
- **Results:** This approach achieved a prediction accuracy of 0.55.

4 Findings

Although this study is still in a pilot phase, promising results have been observed (*cf.* Tab. 1). The transformer-based approach outperforms the ResNet with 14%. Despite the modest accuracy rate of 0.55, the application of this method to colorize black and white comic panels from the validation set yields relatively satisfactory results (*cf.* Fig. 4). It appears that misclassifications predominantly occur in regions with a limited number of pixels. In Fig. 4, the left panel contains part of a jacket of the sitting character erroneously colored in skin color. We hypothesize that the transformer has learned that identical colors come in spatially grouped regions, and the jacket is considered to be part of leg. Information about the clustering of regions added to the input tokens could benefit the performance of the model. An ablation study demonstrates that the use

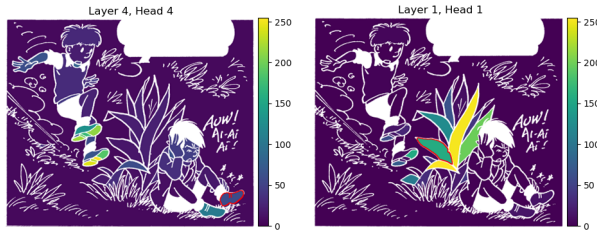


Fig. 2: Attention weights from the shoe region toward other regions.

Fig. 3: Attention weights from the plant leaf region toward other regions.

Table 1: Ablation study

ResNet	ViT PE	Acc.
yes	no	0.41
frozen	yes	0.51
frozen	yes	0.54
finetuning	yes	0.55

of positional embeddings yields a 3% increase in accuracy. This aligns with the findings reported in [4]. Fine-tuning ResNet weights in a Transformer improved accuracy by 1%. A visualization of some of the attention maps suggests that the transformer picks up visual and semantic relationships between regions. This can be seen in Fig. 2 and Fig. 3, where each region is filled with a color corresponding to the magnitude of the attention value for the query region shown with a red outline. In both cases, regions that are visually and semantically similar exhibit higher attention than other regions.

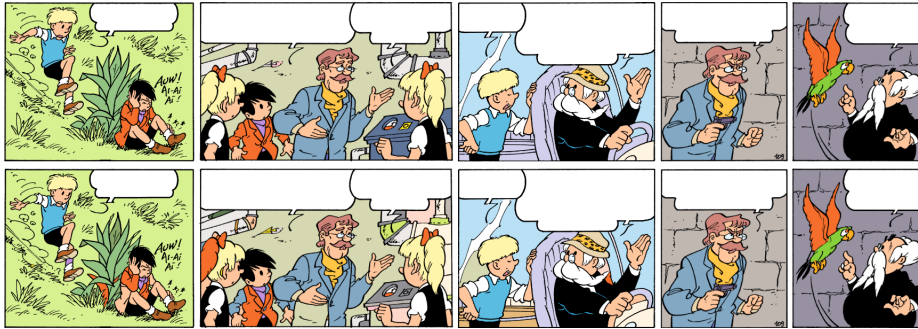


Fig. 4: Random validation samples: the top row shows original panels and the bottom, model-predicted recolorings. Accuracies (L to R): 0.60, 0.47, 0.48, 0.60, 0.63.

5 Conclusion and Future Work

The Transformer-based approach, augmented with positional embeddings, demonstrates a promising solution for the automation of flat colorization of comic books. Future research will concentrate on region segmentation and enforcing consistency across similar regions throughout various panels and comic books. Additionally, we aim to explore further aspects of interactivity.

Acknowledgements

MV is supported through VLAIO Baekeland, Standaard Uitgeverij and Studio 100. TW received a grant from Internal Funds KU Leuven (PDMT2/23/050).

References

1. Cao, Y., Meng, X., Mok, P.Y., Lee, T.Y., Liu, X., Li, P.: Animediffusion: Anime diffusion colorization. *IEEE Transactions on Visualization and Computer Graphics* pp. 1–14 (2024). <https://doi.org/10.1109/TVCG.2024.3357568>
2. Carrillo, H.: Guiding neural networks for image colorization through user interactions. *Theses, Université de Bordeaux* (Feb 2024), <https://theses.hal.science/tel-04446168>
3. Devs, N.: Danbooru2023: A large-scale crowdsourced and tagged anime illustration dataset (2023), <https://huggingface.co/datasets/nyanko7/danbooru2023>
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), <https://arxiv.org/abs/2010.11929>
5. Fourey, S., Tschumperlé, D., Revoy, D.: A fast and efficient semi-guided algorithm for flat coloring line-arts. *Vision, Modeling & Visualization* pp. 1–9 (2018). <https://doi.org/10.2312/vmv.20181247>
6. Gao, R., Jie, L., U, K.T.: Complex manga coloring method based on improved pix2pix model. In: 2023 International Conference on Machine Learning and Cybernetics (ICMLC). pp. 582–587 (2023). <https://doi.org/10.1109/ICMLC58545.2023.10328003>
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), <https://arxiv.org/abs/1512.03385>
8. Kim, H., Lee, C., Lee, J., Kim, D., Lee, K., Oh, M., Kim, D.: Flatgan: A holistic approach for robust flat-coloring in high-definition with understanding line discontinuity. In: *Proceedings of the 31st ACM International Conference on Multimedia*. p. 8242–8250 (2023). <https://doi.org/10.1145/3581783.3613788>
9. Lee, S., Park, E.: Autocaconet: Automatic cartoon colorization network using self-attention gan, segmentation, and color correction. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 403–411 (2024). <https://doi.org/10.1109/WACVW60836.2024.00050>
10. njean42: Kumiko, the comics cutter, <https://github.com/njean42/kumiko>
11. Suzuki, S., be, K.: Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing* **30**(1), 32–46 (1985). [https://doi.org/https://doi.org/10.1016/0734-189X\(85\)90016-7](https://doi.org/https://doi.org/10.1016/0734-189X(85)90016-7), <https://www.sciencedirect.com/science/article/pii/0734189X85900167>
12. Sýkora, D., John, D., Steven, C.: Lazybrush: Flexible painting tool for hand-drawn cartoons. *Computer Graphics Forum* **28**(2), 1–9 (2009). <https://doi.org/10.1111/j.1467-8659.2009.01400.x>
13. TaiZan: Paintschainer tanpopo (2016), preferredNetwork
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), <https://arxiv.org/abs/1706.03762>

15. Yan, C., Chung, J.J.Y., Kiheon, Y., Gingold, Y., Adar, E., Hong, S.R.: Flatmagic: Improving flat colorization through ai-driven design for digital comic professionals. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3491102.3502075>
16. Zhang, L., Ji, Y., Liu, C.: Danbooregion: An illustration region dataset. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 137–154. Springer International Publishing, Cham (2020)
17. Zhang, L., Li, C., Simo-Serra, E., Ji, Y., Wong, T.T., Liu, C.: User-guided line art flat filling with split filling mechanism. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)