T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	参照画像とテキストを用いたスケッチの自動着色法
Title(English)	Automatic Sketch Colorization using Reference Image and Text
著者(和文)	エン ティコン
Author(English)	YAN Dingkun
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第477号, 授与年月日:2025年9月22日, 学位の種別:課程博士, 審査員:齋藤 豪,小池 英樹,篠田 浩一,下坂 正倫,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第477号, Conferred date:2025/9/22, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

Powered by T2R2 (Science Tokyo Research Repository)

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Computer Science Artificial Intelligence	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineer)
学生氏名： Student's Name	Dingkun Yan		審査員主査： Chief Examiner	Suguru Saito

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

This dissertation tackles the bottleneck of coloring line drawings in animation production, especially for anime-style illustrations. It proposes algorithms that transfer hue, shading, and even fine texture from a single reference image to an arbitrary sketch, while remaining robust when the sketch and reference diverge in viewpoint or content. By first diagnosing why existing GAN-based methods over-fit to training pairs, the work designs a frozen-encoder GAN that outperforms previous systems, then moves to a diffusion-based framework that decisively removes "spatial entanglement" artifacts. Along the way it curates large-scale datasets, invents novel attention blocks, and introduces practical evaluation protocols. Extensive experiments and user studies confirm that the final ColorizeDiffusion model delivers sharper edges, richer palettes, and easier local edits than competing approaches, advancing reference-based sketch colorization from fragile prototypes to a deployable, controllable toolchain ready for production use.

Chapter 1, as the introduction of the thesis, sets the stage by arguing that manual color filling is the last labor-intensive step in digital illustration pipelines. It contrasts traditional user-guided tools, as well as data-driven text/user-guided methods with reference-based ones and pinpoints the core challenge in reference-based methods: models trained on perfectly aligned sketch-reference pairs fail when that alignment is absent at inference. To guide the rest of the thesis, it articulates two goals: faithful color transfer and generalization across mismatched input pairs. Additionally, text-based editing algorithms are developed to further enhance the controllability of the proposed colorization systems.

Chapter 2 reviews the technical knowledge required to meet those goals. It walks from basic network architectures, multilayer perceptrons and convolutional networks to transformers, then categorizes generative paradigms—VAEs, GANs, diffusion models—highlighting their strengths and weaknesses for conditional generation. Then, it reviews the prior sketch-colorization systems in details. Specifically, related reference-based colorization methods are dissected to show how jointly trained reference encoders memorize training data, leading to significant overfitting deterioration. This chapter reveals that the overfitting issue in reference-based sketch colorization can be solved by a different training paradigm, where a pre-trained reference encoder is adopted and frozen during the colorization training.

Chapter 3 introduces the data structure and evaluation suite. Two datasets are produced: a 0.7-million-triple Danbooru subset for initial GAN experiments and a 7-million full set for diffusion training. Each color image is turned into a sketch image in various styles by jointly using different sketch extractors, yielding tightly aligned pairs. This chapter also defines edge-adherence and style-fidelity scores, adopts FID for global quality, and designs structured user surveys, establishing a consistent yardstick for all subsequent models.

Chapter 4 presents the first concrete solution. A two-step training strategy is proposed to train GAN-based colorization framework, where a pre-trained ResNet-34 is utilized as the reference encoder and frozen during the training. A generator equipped with spatial attention and a novel Reference-Based Channel-wise Attention block learns to inject color cues, while a self-adaptive MLP allows latent interpolation for tag-based edits. Ablation studies reveal that fixing the encoder cuts over-fitting and improves FID, and crowd-sourced comparisons show clear visual gains over other baselines. Yet the approach still struggles with complex scenes and occasional texture loss, hinting

at the limitations of GAN stability.

Chapter 5 introduces ColorizeDiffusion, a latent-diffusion pipeline tailored to reference transfer. It diagnoses spatial entanglement as a visible symptom of distribution shift toward reference semantics, where the reference overwhelms sketch semantics. This chapter proposes a three-pronged design: staged noisy pre-training that suppresses unintended semantic carry-over without compromising style transfer, split cross-attention that decouples foreground and background tokens to stop semantic bleeding across regions, and a hierarchical disentanglement of the reference representation so that high-level semantics and low-level style are routed and weighted independently. In combination, these mechanisms deliver artifact-free colorizations and unlock fine-grained, text-prompted editing of specific regions without retraining.

Chapter 6 focuses on training paradigms, contrasting conventional joint optimization with the frozen encoder philosophy. Locking the reference encoder compels the sketch pathway to align at a higher semantic level, avoiding low-level memorization and encouraging robust generalization. The discussion broadens these findings, showing that frozen multimodal encoders combined with controlled noise injection form a generally applicable recipe for many image-guided generation tasks beyond colorization.

The concluding chapter describes the contributions of this thesis: a training strategy that significantly improves the generalization and generation performance of reference-based sketch colorization algorithms and corresponding text-based manipulation algorithms. The proposed training paradigm effectively disentangles sketch and reference representations based on their representation levels by utilizing a frozen reference encoder that kept frozen during colorization training, locking reference representations at embedding level to prevent overfitting and associated perceptual deterioration. Together with the text-based manipulation methods, this thesis presents an effective reference-based colorization system.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東京科学大学リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Science Tokyo Research Repository Website (T2R2).