Foreground-Background Aware Exemplar-based Image Colorization with CLIP-guided Reference Retrieval

Taisei Kishimoto
School of FSE
Waseda University
Tokyo, Japan
tai_390156@akane.waseda.jp

Kakeru Koizumi
Graduate School of FSE
Waseda Univesity
Tokyo, Japan
kkeverio@ruri.waseda.jp

Hiroshi Watanabe
Graduate School of FSE
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*— Automatic colorization of grayscale images is now widely available. However, exemplar-based colorization often suffers from color bleeding. This is mainly caused by ambiguous correspondence. It also suffers from suboptimal reference retrieval that does not reflect semantic content well. This paper proposes a foreground–background aware exemplar-based colorization pipeline. First, it explicitly separates the target image into foreground and background using U^2-Net. Second, it improves reference selection by replacing VGG-based similarity with CLIP-based semantic similarity. Experiments on COCO2017 grayscale targets with ImageNet references demonstrate improved visual fidelity. They also show consistent gains over a conventional exemplar-based baseline.**

Keywords*— *image colorization, exemplar-based colorization, reference retrieval, CLIP, saliency segmentation

I. INTRODUCTION

Recent deep learning approaches enable automatic colorization for grayscale images and videos. However, in exemplar-based colorization [1], where a reference color image guides the output, two issues remain. First, region correspondences between the target and reference can be ambiguous, causing background colors to leak into foreground objects. This phenomena is called color bleeding. Second, reference retrieval may prioritize low-level luminance or texture similarity over semantic relevance, reducing controllability for users.

Our goal is to achieve practical colorization that simultaneously preserves semantic consistency in an image and supports user-oriented control through reference images. We focus on clarifying foreground-background correspondences and strengthening content-aware reference selection. For background and foreground separation, we employ the deep learning model U^2-Net [2], which can generate high-quality masks.

II. RELATED WORK

A. Deep Exemplar-based Colorization

Deep Exemplar-based Colorization, which transfers colors from a reference to a target grayscale image via a deep network, has been proposed by He et al.[1]. While effective, colorizing an entire image at once can blur correspondences across regions, leading to color bleeding. Their Gray Image Retrieval (GIR) selects references by combining semantic and luminance similarity, but its semantic comparison relies on VGG features, which are not optimized for open-vocabulary semantic alignment. Such example is shown in Fig.1. The gray input image is shown in Fig.1 (b). Based on the VGG features of this gray image, the mushroom image shown in Fig.1 (c) is automatically selected. In this case, the cause is that the mushroom's pattern resembles the camouflage pattern on the clothing. As a result, the mushroom's colors have been applied to the boy's image as shown in Fig.1 (d), which is far different from the original Fig.1 (a). The obtained GIR result shows dot pattern color in the jacked and of mushroom in the quoted reference and grass color on the snow.

(a) Original (b) Gray input (c) Reference (d) Result

Fig.1. Semantically mismatched reference retrieved by VGG-based similarity. The obtained result shows dot pattern color in the jacked and of mushroom in the quoted reference and grass color on the snow.

B. CLIP

Vision-language models such as CLIP provide semantically meaningful embeddings aligned with natural language and can improve content-aware retrieval [3]. In addition, saliency segmentation networks (e.g., U^2-Net) can separate salient foreground from background, which is useful for reducing inter-region confusion.

III. PROPOSED METHOD

Our goal is to enforce clearer correspondence between target and reference while retrieving semantically appropriate references. The outline the proposed pipeline is shown in Fig.2. The method consists of five steps.

A. Target Preparation

We sample a color image from COCO2017 [4] and convert it to grayscale as the target input. The original color image is kept as ground truth for evaluation.

B. Foreground/Background Masking with U^2-Net

We apply U^2-Net salient object detection to obtain a binary mask that separates the main foreground object from the background. This constrains subsequent correspondence and color transfer to occur within each region. As a result, color bleeding can be suppressed.

C. Region-Specific Reference Retrieval with CLIP

For each masked region (foreground and background), we compute an image embedding using CLIP's image encoder. We precompute CLIP embeddings for candidate reference images in ImageNet [5]. We retrieve the top-1 reference per region by cosine similarity. This region-specific retrieval allows the background to borrow colors from scene-consistent references while the foreground borrows colors from object-consistent references.

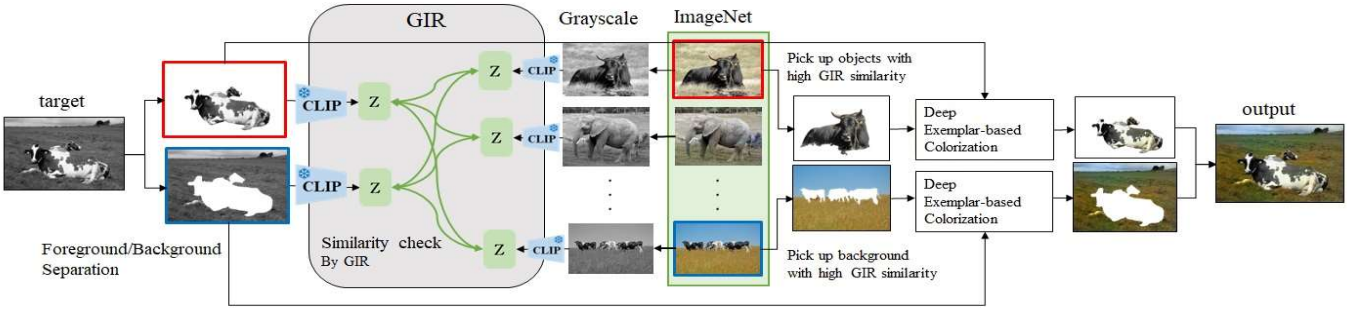

Fig.2. Overview of the proposed foreground–background aware exemplar-based colorization pipeline.

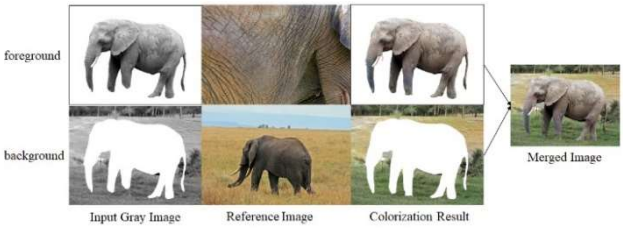

Fig. 3. Process of coloring and compositing image using colors from the mask region in the reference image, applied to the masked gray image. The system transfers color information from reference images to highly similar areas while applying pre-trained natural colorization to low-similarity areas.

D. Masking the Retrieved References

To further align the color transfer process, we run U²-Net on each retrieved reference to obtain corresponding foreground and background masks. This yields masked reference inputs for region-wise exemplar colorization.

E. Per-Region Exemplar Colorization and Composition

We perform exemplar-based colorization independently for the foreground and background using the masked target and masked reference. The two colorized outputs are then composited using the target masks to form the final color image.

IV. Experiments

We use COCO 2017 images converted to grayscale as targets and ImageNet images as the reference pool. We compare the baseline exemplar-based method with our approach, focusing on overall colorization quality and the impact of CLIP-based retrieval versus VGG-based retrieval.

Qualitative comparisons are shown in Fig. 4. The proposed method has been confirmed to suppress color bleeding, which was identified as an issue with conventional methods. The quantitative evaluation result by average PSNR, SSIM and LPIPS over 40 images is shown in Table I. In quantitative evaluation, the proposed method outperformed conventional colorization methods across all metrics.

The proposed method replaces the VGG feature extractor used in conventional GIR with CLIP, which has been trained on image-language correspondences. When using CLIP as the feature extractor, reference images selected are semantically more relevant to the target image as shown in Fig.5.

TABLE I Quantitative Evaluation Result

Method	PSNR↑	SSIM↑	LPIPS↓
Deep Exemplar-based [1]	19.14	0.8096	0.3102
Ours	**21.26**	**0.8426**	**0.2378**

(a) Ground Truth (b) Conventional [1] (c) Our approach

Fig. 4. Qualitative comparison of our approach and Deep Exemplar-based Colorization results.

(a) Original Input (b) VGG selection (c) CLIP selection

Fig. 5. Reference image selection by VGG feature and CIP in GIR.

V. Conclusion

We presented a foreground–background aware exemplar-based colorization method that reduces color bleeding and improves controllability. By combining U²-Net segmentation with CLIP-guided reference retrieval, the proposed approach selects semantically relevant references and produces more consistent colors. Future work includes extending the method to multi-object scenes with complex occlusions and to video colorization with temporal consistency constraints.

References

[1] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep Exemplar-based Colorization," ACM Transaction on Graphics, vol. 37, no. 4, pp.1-12, Aug. 2018.

[2] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jägersand, "U²-Net: Going Deeper with Nested U-Structure for Salient Object Detection," Pattern Recognition, vol. 106, no. 107404, May 2020.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, pp.8748-8763, Jul. 2021.

[4] T. Lin et al., "Microsoft COCO: Common Objects in Context," in Proc. ECCV, pp.740-755, Sep. 2014.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in Proc. CVPR, vol. 115, no. 3, pp.211-252, Apr. 2009.