Enabling Region-Specific Control via Lassos in Point-Based Colorization

Sanghyeon Lee, Jooyeol Yun, Jaegul Choo

Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea
{shlee6825, blizzard072, jchoo}@kaist.ac.kr

Abstract

Point-based interactive colorization techniques allow users to effortlessly colorize grayscale images using user-provided color hints. However, point-based methods often face challenges when different colors are given to semantically similar areas, leading to color intermingling and unsatisfactory results—an issue we refer to as color collapse. The fundamental cause of color collapse is the inadequacy of points for defining the boundaries for each color. To mitigate color collapse, we introduce a lasso tool that can control the scope of each color hint. Additionally, we design a framework that leverages the user-provided lassos to localize the attention masks. The experimental results show that using a single lasso is as effective as applying 4.18 individual color hints and can achieve the desired outcomes in 30% less time than using points alone.

Introduction

Point-based interactive colorization (Levin, Lischinski, and Weiss 2004; Yin, Gong, and Qiu 2019) on grayscale images aims to assist users in restoring colors by selecting and applying them to specific locations. The primary objective in training these models (Zhang et al. 2017; Huang, Zhao, and Liao 2022) is to generate colorized images with minimal user interaction by effectively propagating the user-selected colors to relevant areas. For instance, a model can significantly reduce the user's effort by automatically colorizing an entire apple given a single hint. These models are not only useful for restoring aged photographs but also for a wide range of tasks, such as recoloring images and creating artistic visuals.

However, existing methods often produce unsatisfactory images when multiple color hints are provided in closely related semantic regions. Specifically, Figure 1 illustrates cases when different colors are assigned to semantically identical but separate objects (*e.g.*, distinct apples or petals). As shown in the second column, even the state-of-the-art model (Yun et al. 2023) suffers from the irregular intermingling of colors, producing implausible results. This issue, which we refer to as the ***color collapse***, arises as the model attempts to spread different colors across areas that appear similar. Color collapse is observed in repetitive patterns consisting of different colors ($e.g.$, flower petals, tiles, and fruit baskets).

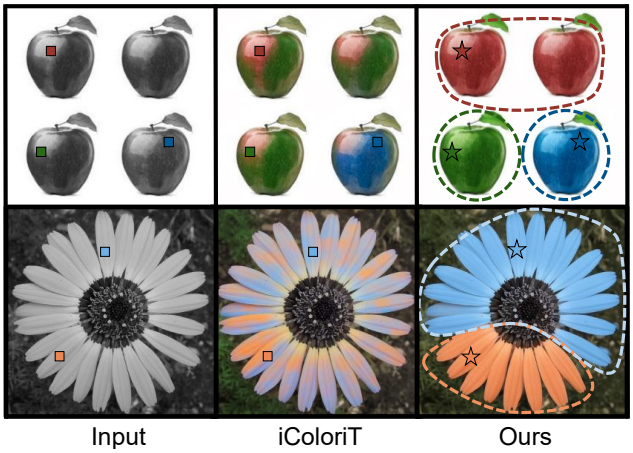

Figure 1: The examples of the color collapse. The start mark and the corresponding lasso in the same color describe the region designated for each color hint by the user. By specifying regions, users can better control how colors spread, thereby mitigating color collapse and leading to a more intentional colorization process.

The fundamental cause of color collapse during point-based interactive colorization is the absence of interactive tools that enable the user to determine the region of the color hint spread. Although using color points as a medium of interaction is simple, these alone provide insufficient guidance for the model on the limits of color spread, often requiring excessive color hints to obtain a satisfactory result. Iteratively providing additional color hints until the color collapse is resolved is not only time-consuming but also diminishes user-friendliness.

Addressing this inherent problem with point-based interactions, we introduce an additional interactive tool, the lasso, which allows users to roughly define the scope of the color they want to spread. As shown in the third column of Figure 1, our lasso tool is designed to operate with loosely defined boundaries, eliminating the need for users to provide strict contours, which are often challenging to define.

We first design a colorization model that uses cross-attention layers (Vaswani et al. 2017) to inject color hints into the image. By restricting the cross-attention map to only attend within the user-provided lasso, we effectively control the scope of each color hint influence. This approach allows our model to adapt effectively to different lassos provided by users, even when using the same color hint, ensuring consistent results that align with varying user preferences.

To demonstrate the effectiveness of our interactive tool, we evaluate our approach in commonly encountered but challenging colorization scenarios. Our extensive experiments demonstrate that our model can effectively assist users in resolving color collapse using lassos while also maintaining the ability to produce colorful images without using lassos. Furthermore, our user study reveals that a single lasso interaction is as effective as 4.18 color points, and users achieve the same quality results in approximately *30% less time*.

Our contributions are as follows:

- We introduce a lasso tool that enhances point-interactive colorization by enabling users to precisely control the region where colors propagate, effectively addressing the color collapse.
- We propose a framework incorporating a localization attention mask, effectively limiting the spread of color hints while adapting to various sizes of lassos.
- We demonstrate through experiments that our lasso tool reduces the number of interactions and time required for colorization tasks by effectively mitigating color collapse.

Related Work

Interactive Colorization

Interactive colorization models are designed to generate colorized images from grayscale input leveraging color conditions. Users can manipulate these conditions to tailor the colorized output to their preferences. Based on the precision with which the conditions are applied, these methods can be categorized into global and localized interactions. Global interactions alter the overall style of the image rather than focusing on specific locations.

Widely studied global interactions include the use of example images where users select an image with a desired style to influence the global style of the colorization (He et al. 2018; Li et al. 2019; Zhang et al. 2019; Xiao et al. 2020; Lu et al. 2020; Xu et al. 2020; Li et al. 2021; Yin et al. 2021; Bai et al. 2022). Another approach involves modifying the color palette by adjusting the color histogram to apply a consistent color theme throughout the image (Wang et al. 2022; Wu et al. 2023). Additionally, textual inputs allow users to specify color tones or themes globally using words that denote different colors (Chen et al. 2018; Bahng et al. 2018; Manjunatha et al. 2018; Weng et al. 2022b; Chang et al. 2022, 2023).

These global interactions are beneficial for their simplicity and minimal user effort, enabling changes in style with just a single global condition. However, they are designed primarily for global style modifications, thus limiting their capacity for detailed color editing on specific locations.

Conversely, localized interaction allows users to specify exact locations within an image to apply edits. A primary method for localized interaction is the use of points. Traditional point-based colorization approaches (Levin, Lischinski, and Weiss 2004; Yin, Gong, and Qiu 2019) employ hand-crafted image filters in an optimization-based approach. These methodologies require an optimization process tailored to each image, which prevents real-time modifications, significantly limiting their practicality. Learning-based methods have been developed to overcome the inefficiencies of inference time optimization. In early deep-learning-based research, Zhang *et al.* (Zhang et al. 2017) leverages a U-net structure to enable propagation based on semantic information. Recently, Yun *et al.* (Yun et al. 2023) achieved significant performance improvement by utilizing the long-range receptive fields of Vision Transformers (Dosovitskiy et al. 2020) to spread hint information to distant relevant areas.

There are also efforts to integrate global and localized interactions to enhance the effectiveness of the colorization process (Huang, Zhao, and Liao 2022; Liang et al. 2024). However, despite these advancements, localized interaction models lack the ability to control the area over which a color hint spreads, often necessitating exhaustive trial-and-error until the desired coloration is achieved.

Attention Manipulation

Recent studies (Hertz et al. 2022; Xiao et al. 2024; Park et al. 2022) have made advances in text-driven image editing by manipulating the attention maps of pre-trained large-scale text-to-image synthesis diffusion models (Rombach et al. 2022). Hertz et al. (Hertz et al. 2022) propose a method that directly utilizes the attention maps derived from a reference image during the diffusion process, leading the generated image to faithfully capture the style of the reference image, for image editing by injecting attention maps, copied from a target condition, into the diffusion generation process. FastComposer (Xiao et al. 2024) demonstrates similar findings in multi-subject personalization of diffusion model by fine-tuning a text-to-image generation model to create distinct attention map for individual subjects, resulting in successful generation across multiple subjects. Similarly, (Park et al. 2022) controls the shape of the generated image by directly masking within the attention map linked to the subject text token, influencing the spatial shape of the subject in the resultant image.

Method

Overall Workflow

Figure 2 illustrates the overall framework of our proposed methods, which utilize user-provided color hints and lasso input to colorize grayscale images. We leverage the L-channel from the CIELab image I_{Lab} as the grayscale image to initiate the process. Our model incorporates a decoder structure inspired by Transformer (Vaswani et al. 2017).

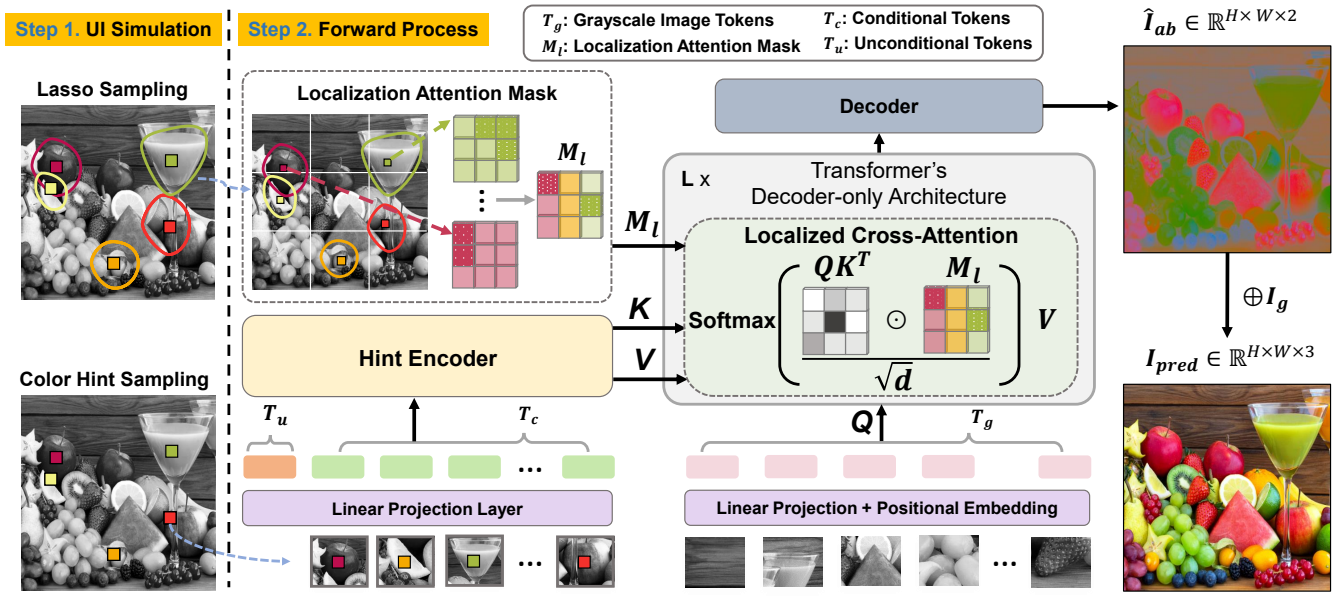

Figure 2: **The overview of our framework.** Our framework acquires color hints and corresponding lassos through a user interaction simulation process for training. The grayscale image is used as the query, and color hints as keys and values generate the cross-attention map QK^T. Subsequently, the attention map is modulated by an attention mask derived from the lassos to precisely control the influence of each color hint on the query image tokens.

Grayscale images and user-provided color hints are transformed into patches and serve as the network's inputs. We leverage the grayscale patches as the queries and the color hints as the keys and values in the cross-attention mechanism. This attention map shows which colors are propagated to which areas of the image. We use a lasso associated with each color hint to create an attention mask that controls the spread of colors based on the hints.

In the final stage, the model predicts the ab color image $\hat{I}_{ab}$, which is then concatenated with the input grayscale image I_g to produce the final output I_{pred}. Our framework employs a fixed-size pre-defined lasso to simplify the user's task and improve usability. We determined the pre-defined lasso size by testing point increments and measuring PSNR on the benchmark dataset, selecting the size with the highest performance. This strategy allows users to refine the results through lassos only for color hints that do not accurately reflect their intentions.

Simulating User Interactions During Training

Our framework requires user-provided hints for training, but manually collecting extensive human data is infeasible. Instead, we simulate color hints and lassos from the ground truth image to mimic user behavior.

Color hint simulation. For simulating point interactions, we follow the sampling process from previous studies (Zhang et al. 2017; Yun et al. 2023). Each color point consists of a hint location and ab color values, and the location is uniformly sampled from the image. During the training procedure, the number of hints h is sampled from a uniform distribution $\mathcal{U} \sim (0, 150)$.

Lasso simulation. Lassos determines the area affected by each point stroke. For each sampled color hint, we simulate a corresponding lasso. Specifically, the lasso is represented by an $H \times W$ binary mask, highlighting the regions needing attention. The intended scope of the hint can vary and be inaccurate due to individual differences. Thus, during training, we sample lassos from a randomly sized rectangle centered on the color hint's location. With this approach, the ground truth color for the hint is always enclosed within the lasso region.

Model Architecture

Our model architecture consists of a hint encoder, localized cross-attention, and a decoder.

Hint encoder. The hint encoder accepts a specified number of h color hints as input. To embed rich context information in each hint, we utilize 3-D cropped color hint patches of size $P \times P$, where P denotes the patch size. To obtain the color hint patches X_{hint}, we crop patches centered on the color hints from the I_{Lab}. Within these cropped patches, areas not including the ab values at the points are masked with zeros. These color hint patches are then embedded into conditional token T_c through a linear projection layer. In our design, positional embedding is not directly applied to X_{hint}. Instead, the gray-scale patch X_g provides positional information. The model can determine the precise locations for the color hints by leveraging the similarity of gray-scale between X_g and X_{hint}. Beyond the conditional token T_c produced by the linear projection layer, we incorporate an unconditional token T_u into the hint encoder's input. This unconditional token ensures the model's operation when color hints are not provided. Containing information about the en-

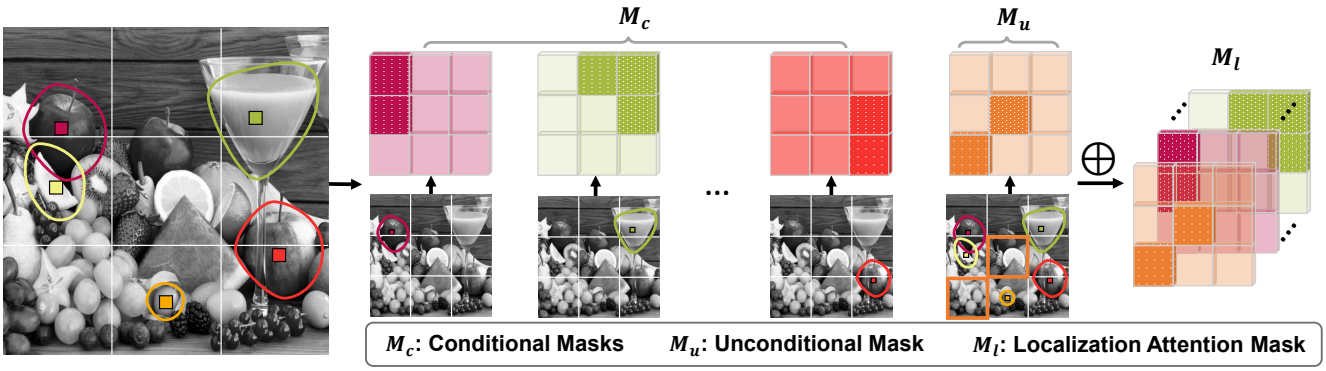

Figure 3: **Localization Attention Mask.** For each color hint, we apply a mask with a value of 1 to the tokens corresponding to patches interior of the lasso areas. Simultaneously, we construct an unconditional mask, M_u, based on regions not overlapped by lassos. The final localization attention mask, M_l, is produced by concatenating M_u and M_C.

tire image, this token assists in coloring areas where no color hints are provided.

Localized cross-attention. The localized cross-attention layer utilizes the structure of a transformer's decoder-only architecture. In this layer, attention computation involves constructing query, key, and value representations. These are derived from grayscale image tokens, T_g, and hint tokens, T_c and T_u. The grayscale image tokens T_g are obtained by applying a linear projection layer and positional embeddings to the input grayscale image patches.

Specifically, the query matrix Q is generated from these image tokens T_g, while the key and value matrices K and V are produced from the hint tokens. The dimensions of these matrices are defined as $Q \in \mathbb{R}^{N\times d}$ and $K, V \in \mathbb{R}^{(h+1)\times d}$, where N is the number of image tokens, $h+1$ is the number of hint tokens, and d is embedding dimension.

The localized cross-attention operation is formulated as

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d} \odot M_l)V, \quad (1)$$

where M_l is the localization attention mask from the sampled lassos.

Localization attention mask. To focus on the color-related region, as shown in Figure 2, localization attention mask M_l explicitly masking the attention map QK^T from the lassos. In the training process, these masks are driven from the simulated lassos.

First, given the number of h color hints, we resize each corresponding lasso $L \in \mathbb{R}^{H\times W\times h}$ into sizes with $H/P \times W/P$. Afterward, we define a conditional mask $M_c \in \mathbb{R}^{H/P\times W/P\times h}$ corresponding to the hint tokens T_c. As illustrated in Figure 3, these conditional masks originate from the lasso, where the interior of the lasso is set to 1, while all other areas are set to 0. Meanwhile, the unconditional mask M_u is a mask with the same spatial dimensions as M_c, designed to identify patches not specified by the lasso interaction. In this mask, areas not designated by the lasso are marked as 1, and all other areas are 0. Our final localization attention mask, $M_l \in \mathbb{N}^{(h+1)\times N}$, is constructed by concatenating M_c and M_h, and then reshaped to match the size of the cross-attention map, where N is the number of grayscale image tokens.

Decoder. The color image $\hat{I}_{ab}$ is then obtained using pixel shuffling (Yun et al. 2023), an efficient upsampling technique that rearranges the output feature map. Finally, $\hat{I}_{ab}$ is concatenated with the input grayscale image I_g, producing the predicted color image $I_{pred} \in \mathbb{R}^{H\times W\times 3}$.

Objective Function

The lasso provides an attention mask that guides color propagation, ensuring the model colorizes only within the defined region and preventing undesirable color spread. Therefore, with no additional regularization term, we rely solely on the Huber loss (Huber 1992) between I_{pred} and I_{gt} in the CIELab color space. The Huber loss L_h, a conventional loss function within colorization tasks (Zhang, Isola, and Efros 2016; Zhang et al. 2017; Yun et al. 2023; Weng et al. 2022a), is computed as follows:

$$\begin{aligned}\mathcal{L}_h =& \frac{1}{2}(I_{pred} - I_{gt})^2 \mathrm{I}_{|I_{pred}-I_{gt}|<1} \\ &+ (|I_{pred} - I_{gt}| - \frac{1}{2})\mathrm{I}_{|I_{pred}-I_{gt}|\geq 1}.\end{aligned} \quad (2)$$

Experiments

Datasets. For the training process, we utilize the ImageNet 2012 dataset (Russakovsky et al. 2015), which contains 1.3M images. We employ the ImageNet ctest (Larsson, Maire, and Shakhnarovich 2016) dataset, a commonly used benchmark in colorization research, to evaluate our approach. Also, we broaden our scope to diverse domains with the Oxford 102flowers (Nilsback and Zisserman 2008) and CUB-200 (Welinder et al. 2010) datasets. The ImageNet ctest (Larsson, Maire, and Shakhnarovich 2016) is a subset of the ImageNet validation set containing 10,000 images. The Oxford 102flowers dataset encompasses 1,020 images of flowers, and the CUB-200 dataset consists of 3,033 images representing 200 species of birds. Furthermore, to validate the effectiveness of our methods, we manually collect 98 samples from unsplash.com that exhibit repetitive patterns. We utilize the collected dataset for user studies.

Baselines. In our experiments, we compare our model with the point-interactive colorization approaches (Yin, Gong,

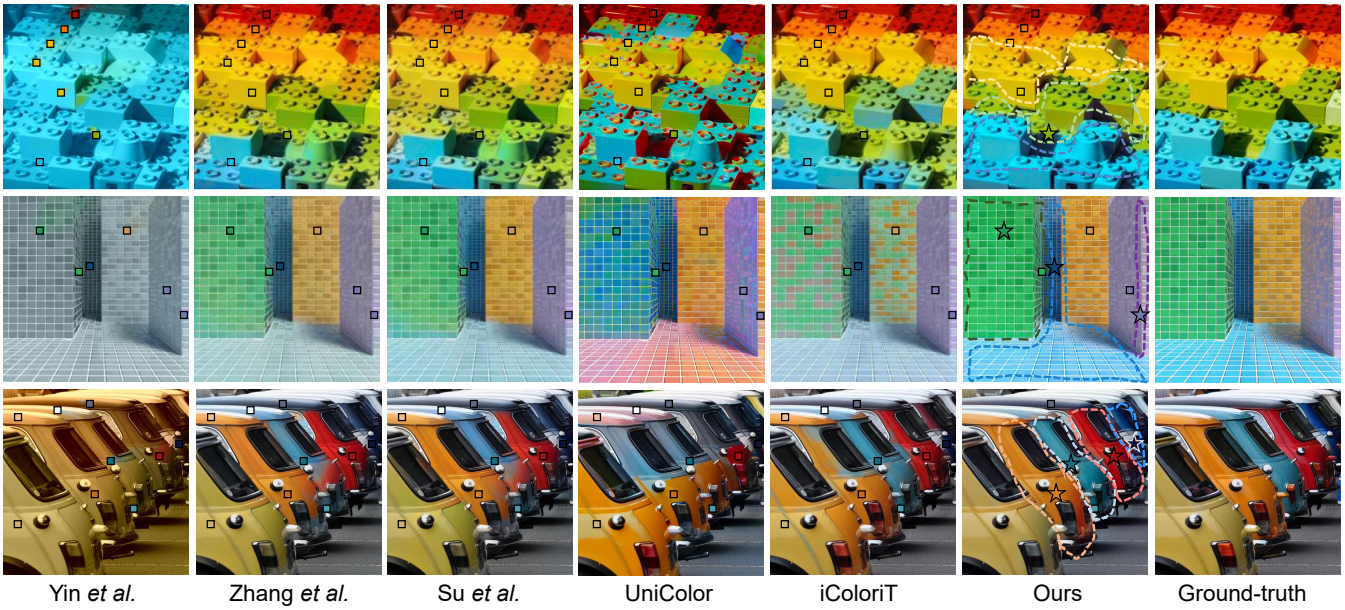

Figure 4: Qualitative results compare with baselines. Each star and its matching-colored lasso highlight the user-selected region for that color. The presented results from our method reflect the colorization achieved through user-directed applications of both lassos and points with pre-defined lasso.

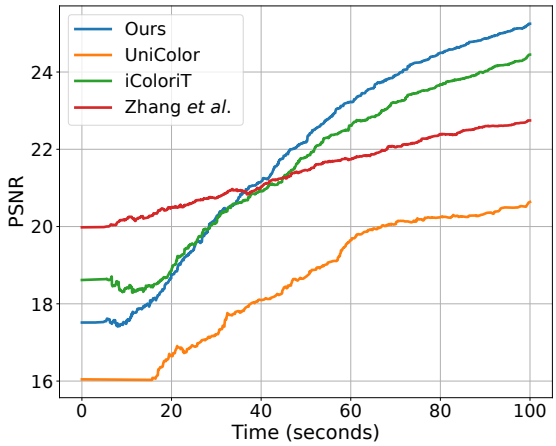

Figure 5: User study results on color collapse easy samples. We measure the average PSNR over the user interaction time, with the initial PSNR derived from each model's unconditional inference.

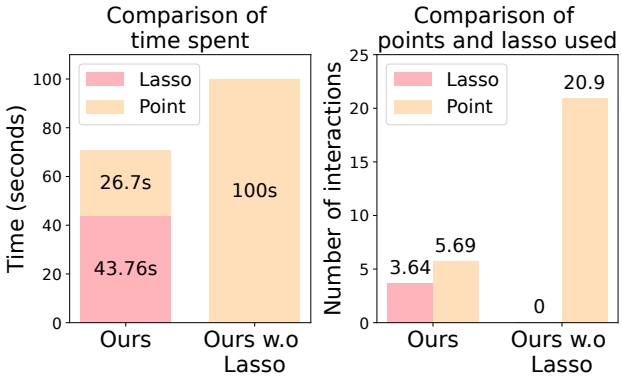

Figure 6: (Left) Time spent to achieve the same quality results with and without using lassos. (Right) The number of interactions required to achieve the same quality.

and Qiu 2019; Zhang et al. 2017; Yun et al. 2023). For iColoriT (Yun et al. 2023), we utilize the base model trained for 100 epochs on ImageNet (Russakovsky et al. 2015). Following the Kim *et al.* (Kim et al. 2021), we also modify an unconditional model by Su *et al.* (Su, Chu, and Huang 2020) to handle user-provided point strokes. Additionally, we compare against UniColor (Huang, Zhao, and Liao 2022), which relies on large (16×16 pixel) points to ensure a strong control signal. This inherently limits its fine-editing capabilities; therefore, we have included UniColor in user studies and selected qualitative comparisons.

Evaluation metrics. To evaluate the quality of the results, we employ the Peak Signal-to-Noise Ratio (PSNR), which measures the mean squared error between the ground truth and predicted images.

Effectiveness Evaluation of the Lasso Tool

User study on handling color collapses. We assess the efficacy of lasso interaction and investigate cases where it complements point-based interactions. For this study, we prepared a collection of 98 challenging samples characterized

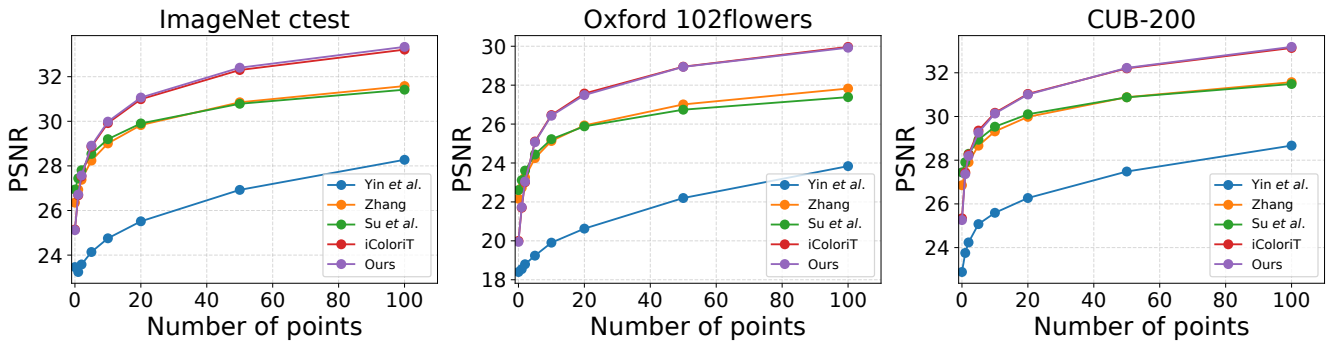

Figure 7: PSNR of the benchmark dataset according to the number of provided hints. Our method and the previous state-of-the-art iColoriT exhibit comparable performance using only point hints.

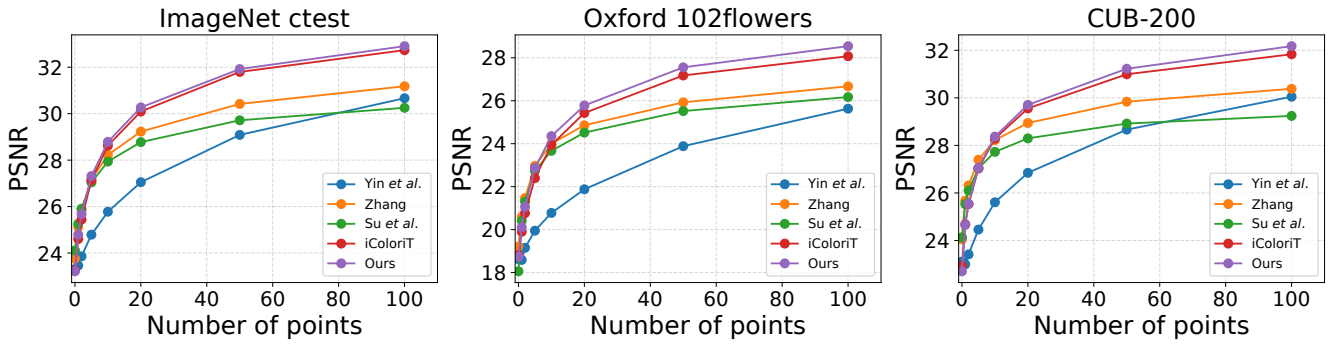

Figure 8: PSNR of the synthetic color collapse prone dataset according to the number of provided hints. A point pair refers to four points sampled from the same location across the 2×2 grid images.

by similar patterns but varied colors and conducted a user study using this dataset. Human participants are provided with a user interface for colorization and asked to restore the grayscale image to its original color. Each user colorizes the image within a time limit of 100 seconds. If users did not provide a lasso interaction for the corresponding color hint, our model used a fixed-size, pre-defined lasso. To evaluate our approach, we conducted a comparison with baseline models that provided a user interface for colorization.

Figure 5 presents the results of the user study, showing PSNR over time. The initial PSNR corresponds to the results of unconditional inference. In our models, the PSNR increases modestly as participants spend time drawing lassos. However, by the end of the 100-second period, our model with lasso interactions outperforms the baseline methods. In particular, UniColor (Huang, Zhao, and Liao 2022) achieves the lowest performance due to its longer inference times and difficulties with detailed editing. Although Zhang *et al.* (Zhang et al. 2017) demonstrates strong initial performance, its improvement plateaus as more color hints are introduced.

Furthermore, 85.7% of participants reported that the lasso interaction helped achieve better colorization results. Additionally, Figure 4 shows the qualitative results for the challenging samples. Notably, color collapse is observed in baseline models that do not leverage lasso interactions, particularly in the first row's green tile (bottom of the third column and inside the fourth column) and in the second row's differently colored cars.

Figure 6 shows how incorporating the lasso interaction improves user efficiency. We first conducted a user study in which participants were asked to colorize images using only points within 100 seconds, achieving a baseline PSNR of 23.94. We then measured how quickly and with how many interactions our method could reach the same PSNR when both points and lassos were available.

As the figure illustrates, using lassos enabled our method to reach the target PSNR 29.50 seconds faster than the point-only approach. Moreover, while the point-only setup required 20.9 points on average, the lasso-included setup used only 5.69 points and 3.64 lassos. Since the time taken per interaction is similar for both points and lassos, 3.64 lassos effectively replaced 15.12 points. This corresponds to one lasso providing the same effectiveness as approximately 4.18 $(4.18 = 15.12/3.64)$ points, clearly demonstrating the efficiency gains offered by integrating the lasso interaction.

Figure 9 presents an aesthetically appealing example generated using our model. To achieve a visually similar result, as judged by human perception, the process required three points and two lassos with corresponding points, and the

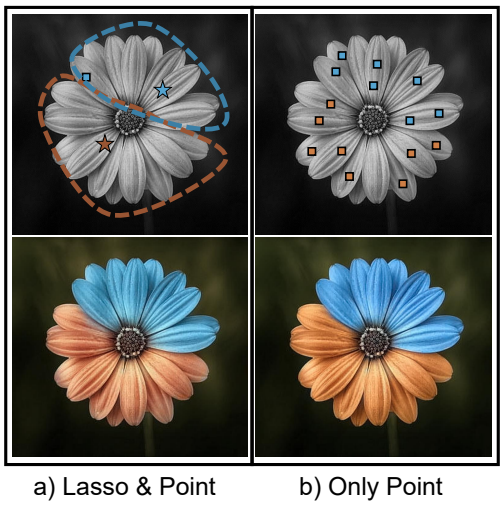

Figure 9: Comparison of colorization results using our model with lassos and points vs using only points.

user obtained the final output with only three inferences. In contrast, when using only color hints, as shown on the right, achieving the same result required 15 points, necessitating 15 iterations of model inference and correction.

Comparison on Benchmark Datasets

Evaluation on colorization benchmarks. We conduct a systemic quantitative comparison against existing point interactive colorization models by randomly sampling points on test images and assigning their corresponding ground truth colors as hints. For our approach, we use a pre-defined lasso for each point to ensure fair evaluation conditions. As shown in Figure 7, our model performs comparably to the previous state-of-the-art, even without the use of a lasso.

Evaluation on the synthetic color collapse-prone dataset. Color collapse is often encountered when distinct colors are provided to semantically similar objects within an image. To effectively simulate these scenarios, we synthesize a color collapse-prone dataset where the same pattern is repeated. As illustrated in the first row of Figure 1, we start by duplicating a single image to create four copies but with shifted colors. These are then organized into a 2×2 grid. While each image keeps its original grayscale channel, we randomly change the color components (ab channel values) of each.

To assess existing colorization approaches on this dataset, color hints sampled from the original image are uniformly applied to the corresponding locations in corresponding locations within the grid, assigning different color hints to semantically similar regions. We evaluated the model by sampling between 1 and 100 points from each grid. A point pair refers to four points sampled from the same location across the 2×2 grid.

As shown in Figure 8, our proposed approach outperforms in these challenging scenarios. Further qualitative results can be found in the supplementary material.

	Ours	w.o M_l	w.o L
PSNR@1	**26.722**	25.805	25.984
PSNR@10	**29.987**	26.728	28.465
PSNR@100	**33.349**	30.190	30.406

Table 1: Ablation study results. The size of the pre-defined lasso L used in the ablation studies is 16×16.

Ablation studies. We conduct an ablation study on the effect of the localization attention mask M_l during the training phase to demonstrate its beneficial impact on performance. We train the model without M_l and then apply the localization attention mask during inference. The results, shown in the second column of Table 1, indicate the model's performance trained without M_l but with lassos applied during inference. These results demonstrate that including a training phase yields better performance compared to training-free methods (Hertz et al. 2022; Park et al. 2022) that modify the cross-attention map of the pre-trained model. Meanwhile, the results displayed in the third column represent the inference of our model without the lasso, which is the same as using the image lasso. Severely inaccurate lasso sizes degrade the quality of the results.

Limitations

Our lasso interaction operates on the resolution of the attention maps, specifically $H/P \times W/P$, which limits the precision of lassos. This challenge can be effectively addressed by employing both lasso and point interactions in a complementary manner. In our framework, each point hint is processed by cropping a localized image patch centered on the point's coordinate, allowing fine-grained distinctions even within the same resolution patch. By using the lasso to specify the coarse area and then refining details with points, our approach balances efficiency and precision, effectively mitigating color collapse and enhancing user control in the colorization process.

Conclusion

In this study, we tackle the color collapse problem by integrating a lasso tool for point-based colorization. The lasso tool offers a practical solution to color collapse, an undesired behavior in point-based colorization models. Our framework currently employs a cross-attention mechanism to integrate user-provided color hints, and this architecture suggests a promising direction for future work, enabling the model to accommodate a broader range of user hints.

Acknowledgements

This work was supported by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korean government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City.

References

Bahng, H.; Yoo, S.; Cho, W.; Park, D. K.; Wu, Z.; Ma, X.; and Choo, J. 2018. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the european conference on computer vision (eccv)*, 431–447.

Bai, Y.; Dong, C.; Chai, Z.; Wang, A.; Xu, Z.; and Yuan, C. 2022. Semantic-sparse colorization network for deep exemplar-based colorization. In *European Conference on Computer Vision*, 505–521. Springer.

Chang, Z.; Weng, S.; Li, Y.; Li, S.; and Shi, B. 2022. L-CoDer: Language-based colorization with color-object decoupling transformer. In *European Conference on Computer Vision*, 360–375. Springer.

Chang, Z.; Weng, S.; Zhang, P.; Li, Y.; Li, S.; and Shi, B. 2023. L-CoIns: Language-Based Colorization With Instance Awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19221–19230.

Chen, J.; Shen, Y.; Gao, J.; Liu, J.; and Liu, X. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8721–8729.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

He, M.; Chen, D.; Liao, J.; Sander, P. V.; and Yuan, L. 2018. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4): 1–16.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Huang, Z.; Zhao, N.; and Liao, J. 2022. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16.

Huber, P. J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*, 492–518. Springer.

Kim, E.; Lee, S.; Park, J.; Choi, S.; Seo, C.; and Choo, J. 2021. Deep Edge-Aware Interactive Colorization against Color-Bleeding Effects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14667–14676.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning Representations for Automatic Colorization. In *European Conference on Computer Vision (ECCV)*.

Levin, A.; Lischinski, D.; and Weiss, Y. 2004. Colorization Using Optimization. *ACM Transactions on Graphics*, 23: 689–694.

Li, B.; Lai, Y.-K.; John, M.; and Rosin, P. L. 2019. Automatic example-based image colorization using location-aware cross-scale matching. *IEEE Transactions on Image Processing*, 28(9): 4606–4619.

Li, H.; Sheng, B.; Li, P.; Ali, R.; and Chen, C. P. 2021. Globally and Locally Semantic Colorization via Exemplar-Based Broad-GAN. *IEEE Transactions on Image Processing*, 30: 8526–8539.

Liang, Z.; Li, Z.; Zhou, S.; Li, C.; and Loy, C. C. 2024. Control Color: Multimodal Diffusion-based Interactive Image Colorization. *arXiv preprint arXiv:2402.10855*.

Lu, P.; Yu, J.; Peng, X.; Zhao, Z.; and Wang, X. 2020. Gray2colornet: Transfer more colors from reference image. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3210–3218.

Manjunatha, V.; Iyyer, M.; Boyd-Graber, J.; and Davis, L. 2018. Learning to color from language. *arXiv preprint arXiv:1804.06026*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Park, D. H.; Luo, G.; Toste, C.; Azadi, S.; Liu, X.; Karalashvili, M.; Rohrbach, A.; and Darrell, T. 2022. Shape-Guided Diffusion with Inside-Outside Attention. *arXiv preprint arXiv:2212.00210*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Su, J.-W.; Chu, H.-K.; and Huang, J.-B. 2020. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7968–7977.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Xia, M.; Qi, L.; Shao, J.; and Qiao, Y. 2022. PalGAN: Image colorization with palette generative adversarial networks. In *European Conference on Computer Vision*, 271–288. Springer.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Weng, S.; Sun, J.; Li, Y.; Li, S.; and Shi, B. 2022a. CT 2: Colorization transformer via color tokens. In *European Conference on Computer Vision*, 1–16. Springer.

Weng, S.; Wu, H.; Chang, Z.; Tang, J.; Li, S.; and Shi, B. 2022b. L-code: Language-based colorization using color-object decoupled conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2677–2684.

Wu, S.; Yang, Y.; Xu, S.; Liu, W.; Yan, X.; and Zhang, S. 2023. FlexIcon: Flexible Icon Colorization via Guided Images and Palettes. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8662–8673.

Xiao, C.; Han, C.; Zhang, Z.; Qin, J.; Wong, T.-T.; Han, G.; and He, S. 2020. Example-Based Colourization Via Dense Encoding Pyramids. In *Computer Graphics Forum*, volume 39, 20–33. Wiley Online Library.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 1–20.

Xu, Z.; Wang, T.; Fang, F.; Sheng, Y.; and Zhang, G. 2020. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9363–9372.

Yin, H.; Gong, Y.; and Qiu, G. 2019. Side window filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8758–8766.

Yin, W.; Lu, P.; Zhao, Z.; and Peng, X. 2021. Yes," Attention Is All You Need", for Exemplar based Colorization. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2243–2251.

Yun, J.; Lee, S.; Park, M.; and Choo, J. 2023. iColoriT: Towards Propagating Local Hints to the Right Region in Interactive Colorization by Leveraging Vision Transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1787–1796.

Zhang, B.; He, M.; Liao, J.; Sander, P. V.; Yuan, L.; Bermak, A.; and Chen, D. 2019. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8052–8061.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

Zhang, R.; Zhu, J.-Y.; Isola, P.; Geng, X.; Lin, A. S.; Yu, T.; and Efros, A. A. 2017. Real-time user-guided image colorization with learned deep priors. *ACM TOG*.