ChromaFusionNet (CFNet): Natural Fusion of Fine-Grained Color Editing

Yi Dong[1]*, Yuxi Wang[12], Ruoxi Fan[2], Wenqi Ouyang[3], Zhiqi Shen[2]*, Peiran Ren[3]*, Xuansong Xie[3]

[1]Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Institute for Intelligent Computing, Alibaba Group
{ydong004, ywang103, rfan002, zqshen}@ntu.edu.sg,
vinkeyoy@gmail.com, renpeiran@gmail.com, xingtong.xxs@taobao.com

Abstract

Digital image enhancement aims to deliver visually striking, pleasing images that align with human perception. While global techniques can elevate the image's overall aesthetics, fine-grained color enhancement can further boost visual appeal and expressiveness. However, colorists frequently face challenges in achieving accurate, localized color adjustments. Direct composition of these local edits can result in spatial color inconsistencies. Existing methods, including color style transfer and image harmonization, exhibit inconsistencies, especially at boundary regions. Addressing this, we present ChromaFusionNet (CFNet), a novel approach that views the color fusion problem through the lens of image color inpainting. Built on the Vision Transformer architecture, CFNet captures global context and delivers high-fidelity outputs, seamlessly blending colors while preserving boundary integrity. Empirical studies on ImageNet and COCO datasets demonstrate CFNet's superiority over existing methods in maintaining color harmony and color fidelity. Robustness evaluations and user studies have further validated the effectiveness of CFNet. In conclusion, CFNet introduces an innovative approach to seamless, fine-grained color fusion, paving the way for advancements in the domain of fine-grained color editing. Code and pretrained models are available at our project page: http://yidong.pro/projects/cfnet.

Introduction

Digital image enhancement strives for visually compelling and realistic outcomes, resonating with human perception. While global color enhancements (Yang et al. 2022a; Wang et al. 2019a) improve the aesthetic of the entire image, they falter when nuanced or region-specific adjustments are sought (Zhang, Gao, and Zhang 2023; Dong et al. 2020). Professional color grading often requires modifying specific regions or objects, ensuring the frame's overall integrity remains intact. However, these adjustments are based on spatial masks, such as object contours. When composed directly, they can inadvertently introduce artifacts, especially when inaccurate masks are used. Such imprecisions, as demonstrated in Figure 1, underscore the challenge of color coherence when implementing fine-grained color modifications.

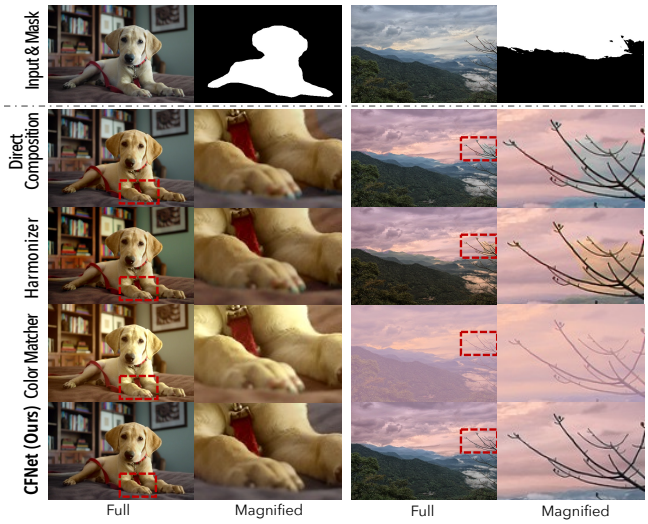

Figure 1: CFNet can effectively alleviate the spatial color inconsistency in direct composition of multiple color edits without changing the overall look and the color of the non-boundary regions. It outperforms both color style transfer methods (Hahne and Aggoun 2021a) and image harmonization approaches (Ke et al. 2022).

Tackling these challenges, colorists typically resort to meticulous edge refinement using interactive tools (Kang et al. 2023), such as power window in Davinci Resolve. This labor-intensive process not only consumes time but also limits the rapid application of granular color enhancements. Currently, an automated color fusion approach remains largely unexplored. Automatic color fusion demands spatial consistency for realism, smooth color transitions across intricate boundaries, and adaptability to varied lighting and textures.

As shown in Table 1, existing methods exhibit significant limitations. Color style transfer (Hahne and Aggoun 2021a) and Color Matcher (Hahne and Aggoun 2021b) techniques can achieve seamless color blending, but struggle to maintain precision at boundary regions, often altering the colors in non-boundary areas as well. Image harmonization methods (Ke et al. 2022; Liang et al. 2022a) can mitigate such artifacts, but they too typically fail to preserve precise boundaries and inadvertently change non-boundary colors. Image inpainting

*Corresponding author.

algorithms (Zhang et al. 2021; Zhao et al. 2020) can generate seamless blending, but they tend to alter the texture of the input image, which is undesirable in color editing tasks.

In contrast, we introduce a novel solution that reframes these discrepancies as an image color inpainting challenge. We identify inconsistencies, particularly prominent at the junctions of different color edits, as analogous to voids in the color space. Our approach, CFNet, innovatively recasts the color fusion task as color inpainting, which is predicting and filling these voids with the appropriate color values.

CFNet employs an encoder-decoder framework, with the encoder built upon the Vision Transformer (Dosovitskiy et al. 2021) architecture. This encoder leverages self-attention to excel in global context capture and inpainting accuracy, capitalizing on inherent scalability. In harmony with this, the pixel-shuffling decoder, preceded by a convolution layer that constrains its receptive field, efficiently produces high-fidelity outputs—prioritizing neighboring pixels for effective boundary filling. A refinement module, featuring Residual-in-Residual Dense Blocks, further complements the design. This refined architecture sustains robust performance, even with increasing model depth, by effectively harnessing rich feature representations for meticulous refinement. Collectively, our approach presents an effective solution to natural color fusion of fine-grained color editing.

In our comprehensive experimentation, CFNet was trained on the ImageNet (Deng et al. 2009) dataset and then assessed on both the ImageNet test set and the COCO (Lin et al. 2014) test set. Quantitative evaluations leveraged a plethora of metrics such as PSNR, SSIM, ΔE, and the B-PSNR considering just boundary areas. Inharmonious Metric, empowered by MadisNet (Jing et al. 2022), was adopted to discern inharmonious regions, demonstrating our method's superiority in preserving color harmony. For comparison, we benchmarked CFNet, our pioneering color fusion method, against closely related color manipulation techniques such as Color Matcher, image harmonization algorithms like Harmonizer and S2CRNet, and style transfer algorithms like MCC-Net (Deng et al. 2021), CAP-VSTNet (Wen, Gao, and Zou 2023), and StyA2K (Zhu et al. 2023). Remarkably, CFNet consistently outperformed all baselines in terms of the W_1 metric, indicating more congruent color with the direct composition image, and achieved comparable performance to Color Matcher in terms of the inharmonious regions metric. Ablation studies further underscored the potency of our refinement module, showing enhanced metrics performance when incorporated. Robustness tests revealed CFNet's capability to reliably reconstruct color in images, even when extended beyond its typical boundary-area use case. User studies and visual results further validated our method's efficacy. In summary, CFNet is a reliable, automated solution suitable for both expert and beginner users. It paves the way for the next advancements in fine-grained color editing.

Related Works

This section provides a concise overview of pertinent research, setting the context for our work within the broader field, though not exhaustively covering all related studies.

Method	Natural	Color	Texture
Color Matcher	✓	✗	✓
Image Harmoniztion	○	○	✓
Image Inpainting	✓	✓	✗
Color Inpainting (Ours)	✓	✓	✓

Table 1: Comparison of color inpainting with existing methods. Our formulation addresses spatial color inconsistencies at boundaries during multi-local color editing blending, preserving color and texture in non-boundary areas. Other methods fall short in one or more aspects.

Color Enhancement and Fine-grained Color Editing

Color enhancement aims to augment an image's visual appeal. Traditional approaches like Harmonizer (Ke et al. 2022), DLR (Park et al. 2018), and PSENet (Wang et al. 2019a) predominantly focus on global adjustments, often missing the finesse required for localized editing. In contrast, fine-grained color editing targets specific spatial regions or color spaces. Professional tools like DaVinci Resolve's qualifiers and 3D Look-Up Tables (LUTs) enable such precision but are limited by their inherent mapping constraints, as observed in SepLUT (Yang et al. 2022b) and AdaInt (Yang et al. 2022a).

Spatial mask-based techniques (RSFNet (Ouyang et al. 2023), DeepLPF (Moran et al. 2020), DCCF (Xue et al. 2022), LEDNet (Zhou, Li, and Loy 2022)) and panoptic segmentation methods (Mask2Former (Cheng et al. 2021), SegFormer (Bai et al. 2023), SAM (Kirillov et al. 2023)) offer more nuanced control but are constrained by their predefined filter shapes or manual refinement needs. Our proposed CFNet, however, autonomously addresses boundary discrepancies for seamless color consistency, outperforming these methods in both flexibility and precision.

Color Fusion

Color fusion involves isolating and blending colors in different regions while maintaining the image's original integrity. Existing works, despite their advancements in accurate segmentation, fall short in fine control and natural fusion, especially with complex textures. Common techniques like feathering subtly blur the boundary between the mask and the surrounding area, softening the transition. However, feathering may struggle to handle intricate image details or complex mask shapes, and overuse can lead to a loss of sharpness in the image. Lightness fusion algorithms like ReCoRo (Xu et al. 2022) offer fusion but are limited to the lightness channel. CFNet, in comparison, presents an unprecedented approach , adeptly handling fusion across color channels.

Image Harmonization and Inpainting

Image harmonization blends foreground and background elements using techniques like tone mapping (Harmonizer (Ke et al. 2022), S2CRNet (Liang et al. 2022b), etc.), but these methods can deviate from user-defined colors. Image inpainting (SPL (Zhang et al. 2021), ZITS (Dong, Cao, and Fu 2022), etc.) fills image gaps realistically but often alters texture undesirably. We innovatively formulate color fusion as

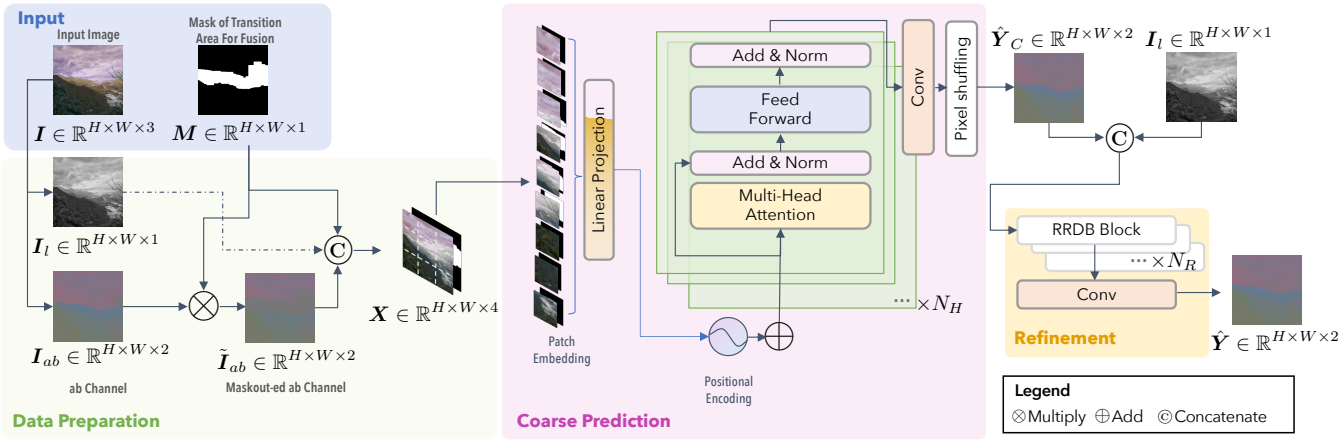

Figure 2: CFNet architecture overview. Tackling spatial inconsistencies from varied color edits, CFNet treats them as an image color inpainting problem. The process commences with an RGB to CIELab conversion, followed by mask application to produce an incomplete ab channel. The encoder, built on Vision Transformer (ViT), processes the input into global receptive fields, while the pixel-shuffling based decoder yields the coarse inpainting outcome. Lastly, the Refinement Module, rooted in the Residual-in-Residual Dense Block (RRDB), further refines the output, capitalizing on a deep and feature-rich design, ensuring boundary smoothing and enhanced color coherence.

a color inpainting problem, successfully maintaining spatial consistency and original texture.

Contribution Summary

In summary, as shown in Table 1, we innovatively formulate the color fusion task as a color inpainting problem. This innovative approach not only maintains spatial consistency but also preserves the original color manipulation across different regions without altering the texture. Our method surpasses existing techniques, facilitating natural multi-region color blending and overcoming previous constraints.

Approach

Motivated by the spatial inconsistency challenges arising from direct composition of varied color edits, we present a novel problem formulation. The spatial inconsistencies, predominantly noticeable at the boundary regions between different color edits, can be perceived as "holes" or gaps in the color spectrum. Naturally, this can be considered as an image color inpainting problem. **Color Inpainting Problem** Given an input image converted to the CIELab color space, its color channels can be represented as $\boldsymbol{I}ab \in \mathbb{R}^{H\times W\times 2}$. Here, H and W are the height and width of the original image, respectively, and we identify areas of potential color inaccuracies using a boundary mask $\boldsymbol{M}$. Multiplying $\boldsymbol{I}_{ab}$ and $\boldsymbol{M}$ results in an image $\tilde{\boldsymbol{I}}_{ab}$ with missing color information, primarily at region boundaries. The objective of the color fusion is to inpaint these gaps with suitable color values. This can be mathematically formulated as:

$$\hat{\boldsymbol{Y}} = \mathcal{C}\left(\tilde{\boldsymbol{I}}_{ab}; \boldsymbol{\theta}_C\right),$$

where $\hat{\boldsymbol{Y}}$ is the inpainted color channels, and $\boldsymbol{\theta}_C$ denotes the parameters of the color inpainting model $\mathcal{C}$. With effective training, the color inpainting network $\mathcal{C}$ can yield satisfactory outcomes. This claim is substantiated by the fact that the color inpainting formulation permits supervised learning on large-scale datasets like ImageNet, as the necessary data is inherently present within the image dataset itself.

Data Preprocessing

CFNet accepts an image $\boldsymbol{I}$ and its associated binary mask $\boldsymbol{M}$. The mask indicates regions requiring spatial color consistency. The goal is to produce an image where these regions exhibit cohesive color filling.

Image Resizing As a preprocessing step, images are resized to 224×224. This size was chosen based on the human visual system's reduced sensitivity to resolution changes in color channels. Maintaining the full resolution for the intrinsic lightness channel ensures high-quality results, as validated in the later experiments section. Higher resolutions in color channels show marginal quality improvements, as detailed in the supplementary materials.

Data Preparation Given the direct composition image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ (as depicted in Figure 2), we first convert it from the RGB to the CIELab color space. This conversion results in a lightness channel, $\boldsymbol{I}_l \in \mathbb{R}^{H\times W\times 1}$, and an ab channel image, $\boldsymbol{I}_{ab} \in \mathbb{R}^{H\times W\times 2}$. To create an incomplete ab channel, denoted as $\tilde{\boldsymbol{I}}_{ab} \in \mathbb{R}^{H\times W\times 2}$, we multiply $\boldsymbol{I}_{ab}$ by a transition area mask for fusion, $\boldsymbol{M} \in \mathbb{R}^{H\times W\times 1}$. This multiplication removes potentially incorrect color information from the ab channel. Finally, the encoder input, $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 4}$, is constructed by concatenating $\boldsymbol{I}_l$, $\boldsymbol{M}$, and $\tilde{\boldsymbol{I}}_{ab}$.

Network Overview

In CFNet, we utilize an encoder-decoder structure for coarse color inpainting, followed by an image refinement module.

Encoder We utilize Vision Transformer (ViT) (Dosovitskiy et al. 2021) as an encoder to achieve a global receptive field. We first reshape $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 4}$ into a sequence of tokens $\boldsymbol{X}_p \in \mathbb{R}^{N\times\left(P^2\times 4\right)}$, where H, W are the height and width of the original image, P is the patch size, and $N = HW/P^2$ is the number of input tokens. Thus, the $P \times P \times 4$ size image patches from the original input $\boldsymbol{X}$ are used as a sequence of input tokens. These tokens are passed through the ViT-based encoder, with sinusoidal positional encoding, multi-head self-attention, and layer normalization.

Decoder The decoder is mainly based on pixel shuffling, a lightweight upsampling technique that reshapes the output channel dimension into a spatial resolution. It rearranges the $(H/P, W/P, C \times P^2)$ feature map into a shape of (H, W, C) to obtain a full-resolution image. As the nature of color inpainting is to infer the holes from neighbouring pixels, to limit the receptive field to neighbouring areas, we add a convolutional layer before the pixel shuffling. This design, although simple, can effectively reconstruct the image to output the coarse result of color inpainting.

Refinement Module The CFNet's refinement module is architecturally grounded in the Residual-in-Residual Dense Block (RRDB) (Wang et al. 2019b). An RRDB is a stack of three Residual Dense Blocks (RDBs), each housing five convolutional layers. These layers are designed to learn features effectively, incorporating them from previous layers through a growth channel. Following convolution, a LeakyReLU activation function is employed, and the layer's output is merged with its input. The module accepts the concatenated coarse prediction $\hat{\boldsymbol{Y}}_C \in \mathbb{R}^{H\times W\times 2}$ and the lightness channel of the input image $\boldsymbol{I}_l \in \mathbb{R}^{H\times W\times 1}$, and refines it through a sequence of N_R RRDBs to produce the refined color inpainting results, $\hat{\boldsymbol{Y}} \in \mathbb{R}^{H\times W\times 2}$. This streamlined architecture guarantees robust performance even as model depth increases, effectively leveraging a rich feature set for precise refinement. The refinement module demonstrates improvement in color inpainting details, as evidenced in the experimental section.

Training Scheme

Mask Generation In the training phase, we need to construct the mask $\boldsymbol{M}$, indicating areas absent of colors. To mimic the data characteristics during inference time, we utilize Mask2Former (Cheng et al. 2021) to perform panoptic segmentation on the input image. Once we have acquired the segmentation masks, we make a random selection of object masks, typically between 1 to 3. Then, the edges of these masks are achieved through canny algorithm with edge dilation parameter. Specifically, we randomly select the number of dilation iterations from a range of 1 to 16. The result is a binary mask comprising all the dilated edges, which is used as the mask $\boldsymbol{M}$.

Loss and Learning strategy For the coarse prediction module, we employ the Huber loss (Huber 1964) to measure the discrepancy between the model prediction $\hat{\boldsymbol{Y}} \in \mathbb{R}^{H\times W\times 2}$ and the color-related channels of the input image $\boldsymbol{I}_{ab} \in \mathbb{R}^{H\times W\times 2}$. The Huber loss introduces a threshold parameter δ. The loss for the coarse prediction, $\mathcal{L}_{\text{coarse}}$, is given by:

$$\mathcal{L}_{\text{coarse}}(\hat{\boldsymbol{Y}}, \boldsymbol{I}_{ab}) = \begin{cases} \frac{1}{2}(\hat{\boldsymbol{Y}} - \boldsymbol{I}_{ab})^2 & \text{if } |\hat{\boldsymbol{Y}} - \boldsymbol{I}_{ab}| \le \delta, \\ \delta|\hat{\boldsymbol{Y}} - \boldsymbol{I}_{ab}| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases}$$

where δ is set to be 0.01, which is the threshold where the loss switches from quadratic to linear.

The inpainted image is then fed into the refinement module, which is trained with a combination of L1 loss, perceptual loss and GAN loss:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{percept}} + \mathcal{L}_{\text{GAN}}.$$

The final loss function is:

$$\mathcal{L}_{\text{CF}} = \frac{1}{2}(\mathcal{L}_{\text{coarse}} + \beta\mathcal{L}_{\text{refine}}),$$

where β is a weighting parameter. We simply fix $\beta = 1$ in all our experiments.

Inference and Post-Processing

During inference, we obtain a binary edge mask $\boldsymbol{M}$ from the segmentation mask using the Canny edge detection algorithm. For a given input image $\boldsymbol{I}$, CFNet takes both $\boldsymbol{I}$ and $\boldsymbol{M}$ as input, denoted as $\boldsymbol{X}$, and outputs the predicted color channels $\hat{\boldsymbol{Y}}$. In regions away from boundaries, the spatial color consistency of direct composition is preserved. For refining the boundary areas, we combine the CFNet output with the original image by calculating $\hat{\boldsymbol{Y}} \times \boldsymbol{M} + \boldsymbol{I}_{\text{ab}} \times (1 - \boldsymbol{M})$. The final full-resolution output image is formed by concatenating the unchanged lightness channel, $\boldsymbol{I}_l$, with this combined and upscaled color representation.

Application: Fine-grained Color Enhancement

We design a text-driven color enhancement tool utilizing our novel Chroma Fusion Network (CFNet). Unlike conventional global enhancement methods, our approach offers more varied and expressive outcomes. Emulating professional colorist procedures, our tool starts with exposure adjustment, applies color styles using a lookup table, and then fine-tunes specific object colors to match natural memory colors and enhance visual attractiveness. Specifically, the application tool comprises an exposure control module, a text-based lookup table generator, and an advanced color manipulation module consisting of a colorization model and a text-to-color mapping mechanism. Enhanced color regions are harmoniously integrated using our CFNet to rectify any spatial inconsistencies.

Figure 3 illustrates an example of fine-grained color enhancement using our application: the base image undergoes exposure adjustment, followed by global enhancement using a 3D LUT. Specific image segments are then polished via text-based fine-grained color editing. While this heightens finer details and enriches the color correction decision space, it can produce artifacts at segment boundaries. CFNet ensures a reasonable merge of these areas, reducing boundary anomalies and improving the overall visual allure.

For a comprehensive description, additional results, and a user study that features a colorist's feedback, please refer to the supplementary material.

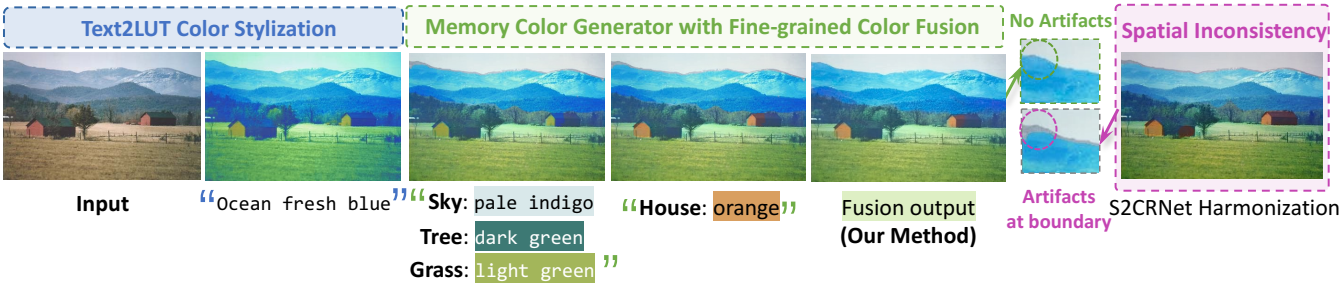

Figure 3: Fine-grained color enhancement. While region-specific color adjustments boost visual appeal and expressiveness, they can introduce boundary issues. CFNet effectively blends these areas, removing inconsistencies and improving visual quality.

	ImageNet				COCO			
	$W_1 \downarrow$	IP@0.85↓	IP@0.90↓	IP@0.9651↓	$W_1 \downarrow$	IP@0.85↓	IP@0.90↓	IP@0.9651↓
Direct composition	-	9.04	8.51	7.11	-	4.46	4.30	3.52
Harmonizer	5.44	8.38	7.86	6.50	12.52	3.24	3.03	2.46
S2CRNet	6.69	8.30	7.78	6.46	20.88	3.71	3.47	2.85
Color Matcher	50.05	5.74	5.05	**3.50**	41.04	**1.39**	**1.24**	**0.91**
CAP-VSTNet	3.15	8.93	8.72	8.20	1.67	3.60	3.35	2.72
MCCNet	6.67	**4.60**	**4.50**	4.18	5.35	2.31	2.16	1.85
StyA2K	3.39	6.29	5.97	5.38	1.98	3.32	3.10	2.62
CFNet	**1.66**	5.73	5.28	4.14	**0.21**	2.63	2.34	1.70

Table 2: Comparisons of color deviation W_1 and inharmonious region proportion IP (%) with different thresholds (IP@0.9651 reaches the best accuracy as stated in MadisNet) between direct composition, Harmonizer, S2CRNet, Color Matcher and our method. For both W_1 and IP metrics, lower value indicates better performance.

Experiments

Datasets

We train CFNet on the **ImageNet** (Deng et al. 2009) dataset, which is an expansive visual dataset containing millions of annotated images from a broad spectrum of categories. We use the ImageNet test set and the **COCO** (Lin et al. 2014) test set for evaluation and ablation study.

Evaluation Metrics

Similarity Metrics We employ the commonly used peak signal to noise ratio (**PSNR**), structural similarity (**SSIM**), and ΔE as the quantitative evaluation metric for all datasets. We also include the **B-PSNR** which only considers the boundary area. In general, the larger values of PSNR, B-PSNR and SSIM indicate better color inpainting results while smaller ΔE indicates less color deviation.

Color Deviation Similar to Color Matcher (Hahne and Aggoun 2021b), we compute normalized CDFs to obtain a Wasserstein metric, where lower values indicate higher similarity, to evaluate the consistency of overall color distribution between the result and direct composition image.

Inharmonious Metric To evaluate the efficacy of our color fusion method in maintaining color harmony within the image, we utilize MadisNet (Jing et al. 2022) to automatically detect the inharmonious region. We compute the proportion of inharmonious region (IP) under different thresholds to gain a comprehensive understanding of the extent of inharmonious regions in the image.

Implementation Details

Baselines To the best of our knowledge, we are the first to propose the color fusion method, and thus there are no direct methods for comparison in this subsection.Instead, we compare our method with closely related color manipulation: Color Matcher (Hahne and Aggoun 2021b); image harmonization algorithms: Harmonizer (Ke et al. 2022) and S2CRNet (Liang et al. 2022a); and style transfer algorithms: MCCNet (Deng et al. 2021), CAP-VSTNet (Wen, Gao, and Zou 2023), and StyA2K (Zhu et al. 2023). To fairly implement Color Matcher, we only consider the masked region as a reference to generate the result image. And for the image harmonization algorithms, we designate the masked region as the foreground and use the Harmonizer and S2CRNet algorithms to modify the background region accordingly.

Quantitative Results To assess the quality of our results compared to baseline methods, we employ two metrics: color deviation and the proportion of inharmonious regions. We begin by identifying the region of interest in the input image utilizing panoptic segmentation approach. Having a region mask, we then perform random color adjustments within this region to create a composite image with direct composition. This composite image serves as a basis for evaluating the performance of our proposed method and the baseline models. We analyze color deviation to evaluate the consistency of color distribution between result image and direct composi-

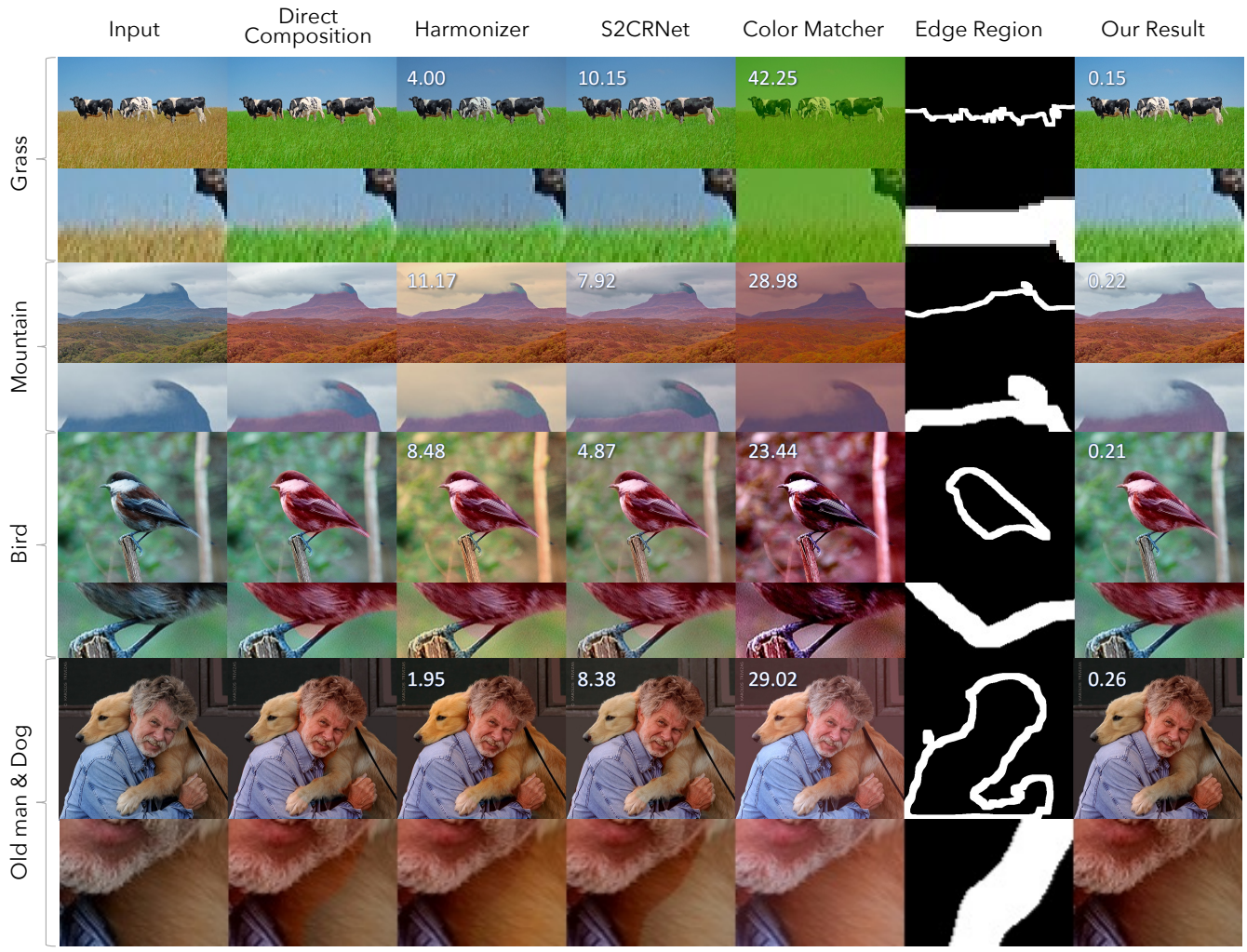

Figure 4: Comparison of results from Harmonizer, S2CRNet, Color Matcher, and our method. Each two-row set presents a full image and its zoomed region. Columns one and two show the original and directly composed images, respectively. The edge column highlights the enhanced region boundary. Each method's result includes the W_1 distance to the direct composition, with smaller values indicating closer color distribution.

tion image. Inharmonious region detection is used to evaluate the color fusion performance around the edge region. Table 2 compares the W_1 metric and inharmonious region proportion, denoted as IP, under different thresholds between our method and other baseline models. Our evaluations are conducted on both the ImageNet and COCO datasets. Our method outperforms all baselines by achieving the lowest W_1 value on both datasets, and achieves comparable performance with Color Matcher in terms of IP metric.

Visual Results

Figure 4 illustrates that our method is able to mitigate the spatial inconsistency problem around the edge region while preserving the original color distribution. CFNet can effectively blend color edits and preserve colors in the non-boundary regions. For additional visual results across various configurations, kindly consult the supplementary materials provided.

Ablation Studies

To validate the effectiveness of our proposed refinement modules, we carry out ablation experiments using the ImageNet test set and the COCO test set. As presented in Table 3, CFNet with the refinement module clearly outperforms the coarse model across all metrics. These results underscore the significance and impact of the refinement module.

Robustness Evaluation

Although CFNet is typically used to complete boundary areas, which usually account for no more than 15% of the whole image area, our algorithm is robust and can effectively reconstruct the color of images even when up to 50% of the area is masked. As demonstrated in Figure 5, despite a natural decline in PSNR and an increase in ΔE with higher mask ratios, CFNet achieves good results on both ImageNet and COCO datasets, indicating high fidelity and minor color

	ImageNet		COCO	
	Coarse	Coarse + Refine	Coarse	Coarse + Refine
PSNR ↑	34.40	**34.96**	35.67	**36.12**
B-PSNR ↑	25.61	**26.57**	30.75	**32.05**
SSIM ↑	**0.97**	**0.97**	**0.97**	**0.97**
ΔE ↓	0.68	**0.65**	0.74	**0.73**

Table 3: Ablation studies for reconstruction performance on ImageNet test set and COCO validation set. We conduct experiments on the coarse CFNet and CFNet with refinement module on PSNR, B-PSNR, SSIM and ΔE metrics.

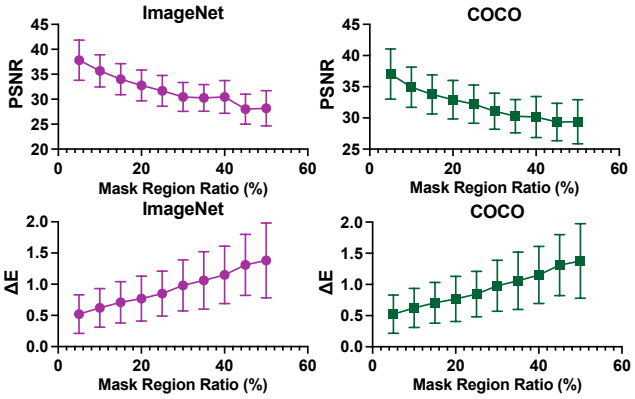

Figure 5: Robustness Study of CFNet on ImageNet and COCO datasets. This figure highlights CFNet's capability to accurately reconstruct image colors with mask ratios up to 50%. There is a natural decrease in PSNR and a rise in ΔE values with increasing mask ratios. However, CFNet demonstrates high fidelity and minimal color discrepancies between original and reconstructed images, showcasing its robustness in handling significant image areas.

discrepancies between the original and reconstructed images.

User Study

We conducted two user studies with 52 participants, including art school attendees and professionals, to evaluate CFNet's capabilities in color fusion and fine-grained enhancements. As shown in Figure 6, CFNet outperformed alternatives like Harmonizer, S2CRNet, and Color Matcher.

Color Fusion Participants ranked results from Harmonizer, S2CRNet, Color Matcher, and CFNet based on color fusion seamlessness and consistency. Feedback indicated that CFNet excelled, emphasizing its natural fusion and color accuracy.

Fine-grained Color Enhancement Participants compared CFNet against leading enhancement techniques like UniColor (Huang, Zhao, and Liao 2022), DeepLPF (Moran et al. 2020), Harmonizer, and AdaInt (Yang et al. 2022a). The consensus was clear: CFNet dominated, receiving praise for its seamless enhancement, confirming its leading edge.

Limitation and Failure Cases

As shown in Figure 7, CFNet faces challenges with transparent areas due to inherent complexities and potential training

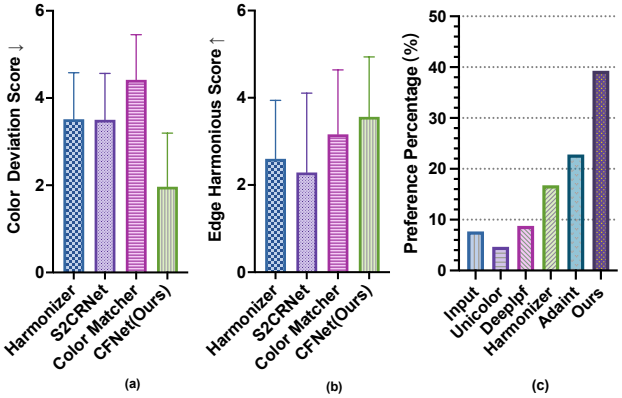

Figure 6: User study results showcasing the efficacy of CFNet in color fusion and fine-grained color enhancement. (a) and (b) display the color fusion scores for color deviation and edge harmony, respectively. (c) represents user preference percentages for color enhancement images. Across all figures, CFNet surpasses competitors.

Figure 7: Failure cases. CFNet's output illustrates color discrepancies within transparent regions.

data gaps. Despite producing generally acceptable results, there are color inconsistencies. Furthermore, as demonstrated in the robustness evaluation (Figure 5), CFNet's color inpainting ability is limited when the ratio of colors to be predicted is too high. In addition, CFNet assumes that the L channel remains unchanged. Although CFNet is specialized, it can effectively work alongside the lightness fusion algorithms, such as ReCoRo (Xu et al. 2022), which is limited to region-specific lightness modifications and natural fusion. Together, CFNet and ReCoRo are capable of color editing across all channels. Future endeavors will emphasize dataset diversity and algorithm refinements.

Conclusion

We propose CFNet, an innovative approach tailored for the precise fusion of fine-grained color edits, reconceptualizing the challenge as an image color inpainting task. Capitalizing on the Vision Transformer architecture, it adeptly identifies and addresses color discrepancies, guaranteeing uniformity and seamless transitions. Empirical results validate CFNet's superiority over current methodologies, showcasing its proficiency in upholding color integrity and harmony in various contexts. CFNet emerges as a robust, automated tool for both adept and novice users, filling the existing void in natural fusion for region-specific color enhancements. It will inspire further evolution of fine-grained color editing.

Acknowledgments

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. This research project was partly supported by Nanyang Technological University under the URECA Undergraduate Research Programme. We are grateful to the anonymous reviewers for their insightful and constructive feedback.

References

Bai, H.; Wang, P.; Zhang, R.; and Su, Z. 2023. SegFormer: A Topic Segmentation Model with Controllable Range of Attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 12545–12552.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021. Mask2Former for Video Instance Segmentation. *CoRR*, abs/2112.10764.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, Y.; Tang, F.; Dong, W.; Huang, h.; chongyang, M.; and Xu, C. 2021. Arbitrary Video Style Transfer via Multi-Channel Correlation. In *AAAI*.

Dong, Q.; Cao, C.; and Fu, Y. 2022. Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 11348–11358. IEEE.

Dong, Y.; Liu, C.; Shen, Z.; Gao, Z.; Wang, P.; Zhang, C.; Ren, P.; Xie, X.; Yu, H.; and Huang, Q. 2020. Domain Specific and Idiom Adaptive Video Summarization. In *Proceedings of the ACM Multimedia Asia*, MMAsia '19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368414.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hahne, C.; and Aggoun, A. 2021a. PlenoptiCam v1.0: A Light-Field Imaging Framework. *IEEE Transactions on Image Processing*, 30: 6757–6771.

Hahne, C.; and Aggoun, A. 2021b. PlenoptiCam v1.0: A Light-Field Imaging Framework. *IEEE Transactions on Image Processing*, 30: 6757–6771.

Huang, Z.; Zhao, N.; and Liao, J. 2022. UniColor: A Unified Framework for Multi-Modal Colorization with Transformer. *ACM Trans. Graph.*, 41(6): 205:1–205:16.

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101.

Jing, L.; Li, N.; Penghao, W.; Fengjun, G.; and Teng, L. 2022. Inharmonious Region Localization by Magnifying Domain Discrepancy. In *AAAI*.

Kang, X.; Yang, T.; Ouyang, W.; Ren, P.; Li, L.; and Xie, X. 2023. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 328–338.

Ke, Z.; Sun, C.; Zhu, L.; Xu, K.; and Lau, R. W. H. 2022. Harmonizer: Learning to Perform White-Box Image and Video Harmonization. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, 690–706. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *CoRR*, abs/2304.02643.

Liang, J.; Cun, X.; Pun, C.; and Wang, J. 2022a. Spatial-Separated Curve Rendering Network for Efficient and High-Resolution Image Harmonization. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, volume 13667 of *Lecture Notes in Computer Science*, 334–349. Springer.

Liang, J.; Cun, X.; Pun, C.; and Wang, J. 2022b. Spatial-Separated Curve Rendering Network for Efficient and High-Resolution Image Harmonization. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, volume 13667 of *Lecture Notes in Computer Science*, 334–349. Springer.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; and Slabaugh, G. G. 2020. DeepLPF: Deep Local Parametric Filters for Image Enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 12823–12832. Computer Vision Foundation / IEEE.

Ouyang, W.; Dong, Y.; Kang, X.; Ren, P.; Xu, X.; and Xie, X. 2023. RSFNet: A White-Box Image Retouching Approach using Region-Specific Color Filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12160–12169.

Park, J.; Lee, J.; Yoo, D.; and Kweon, I. S. 2018. Distort-and-Recover: Color Enhancement Using Deep Reinforcement Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 5928–5936. Computer Vision Foundation / IEEE Computer Society.

Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; and Shao, S. 2019a. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9336–9345.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2019b. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision – ECCV 2018 Workshops*, 63–79. Cham: Springer International Publishing. ISBN 978-3-030-11021-5.

Wen, L.; Gao, C.; and Zou, C. 2023. CAP-VSTNet: Content Affinity Preserved Versatile Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18300–18309.

Xu, D.; Poghosyan, H.; Navasardyan, S.; Jiang, Y.; Shi, H.; and Wang, Z. 2022. ReCoRo: Region-Controllable Robust Light Enhancement with User-Specified Imprecise Masks. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1376–1386. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Xue, B.; Ran, S.; Chen, Q.; Jia, R.; Zhao, B.; and Tang, X. 2022. DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, volume 13667 of *Lecture Notes in Computer Science*, 300–316. Springer.

Yang, C.; Jin, M.; Jia, X.; Xu, Y.; and Chen, Y. 2022a. AdaInt: Learning Adaptive Intervals for 3D Lookup Tables on Real-time Image Enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 17501–17510. IEEE.

Yang, C.; Jin, M.; Xu, Y.; Zhang, R.; Chen, Y.; and Liu, H. 2022b. SepLUT: Separable Image-Adaptive Lookup Tables for Real-Time Image Enhancement. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVIII*, volume 13678 of *Lecture Notes in Computer Science*, 201–217. Springer.

Zhang, L.; Gao, G.; and Zhang, H. 2023. Spatial-Temporal Federated Learning for Lifelong Person Re-identification on Distributed Edges. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, W.; Zhu, J.; Tai, Y.; Wang, Y.; Chu, W.; Ni, B.; Wang, C.; and Yang, X. 2021. Context-Aware Image Inpainting with Learned Semantic Priors. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 1323–1329. ijcai.org.

Zhao, L.; Mo, Q.; Lin, S.; Wang, Z.; Zuo, Z.; Chen, H.; Xing, W.; and Lu, D. 2020. UCTGAN: Diverse Image Inpainting Based on Unsupervised Cross-Space Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5740–5749.

Zhou, S.; Li, C.; and Loy, C. C. 2022. LEDNet: Joint Low-Light Enhancement and Deblurring in the Dark. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, 573–589. Springer.

Zhu, M.; He, X.; Wang, N.; Wang, X.; and Gao, X. 2023. All-to-Key Attention for Arbitrary Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 23109–23119.