ChromaFusionNet (CFNet): Natural Fusion of Fine-grained Color Editing Supplementary Materials

Abstract

In the supplementary material, we delve deeper into the workings and advantages of CFNet, which we introduced in the main document as a novel solution for the natural fusion of fine-grained color enhancement. Building upon the main paper's focus, we provide a detailed overview of how CFNet stands out in the realm of digital art post-processing. We accentuate the edge CFNet has over existing methods discussed in the main document, such as image harmonization, color transfer, and image inpainting techniques. The ablation study section, complementary to our main findings, offers an in-depth exploration of our network design decisions, the impact of the refinement module's size on pivotal metrics, and a thorough evaluation of different mask generation methodologies. Our user study, aligning with the main paper's assertion of CFNet as a novel tool, affirms its potential in color fusion. We supplement the main document's experiments with visual examples, showcasing CFNet's unparalleled performance in single and multi-object color fusion tasks, thus emphasizing its proficiency in achieving impeccable and aesthetically pleasing color blends.

Contents

1. Color Fusion and the Practical Applications of CFNet

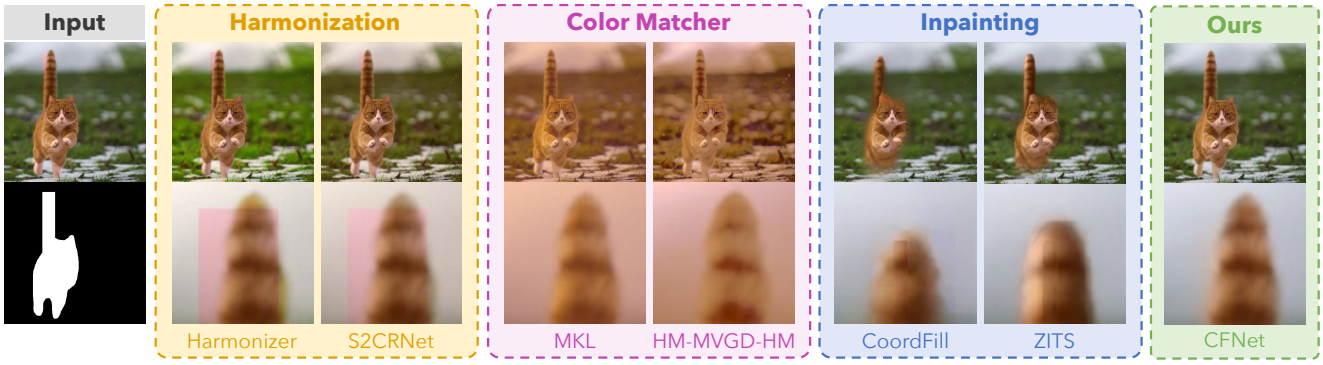
In this section, we compellingly showcase CFNet's practical utility and superiority in real-world applications.

1.1. Color Fusion: A Vital yet Underexplored Task in Fine-grained Color Enhancement

The domain of digital art post-processing has witnessed growing interest in fine-grained color enhancement [9]. Yet, within computer vision and deep learning, this niche remains under-represented. Afifi et al. [1], for instance, delved into localized color adjustments, applying color transfer between an image and its semantic segmentation. Notably, even such methodologies face a persistent challenge: direct composition of region-specific color alterations introduces spatial inconsistencies, particularly pronounced at boundary regions.

Addressing seamless color fusion is not just pivotal but an uncharted territory within fine-grained color enhancement. With our novel approach, CFNet, we aim to redress these spatial irregularities, heralding a paradigm shift in fine-grained color editing for computer vision research.

1.2. CFNet: Pioneering Color Fusion beyond Existing Method Limitations



	Harmonization		Color Matcher		Inpainting		Ours
Color Deviation	○		×		√		√
Natural Fusion	○		√		√		√
Texture Unchanged	√		√		×		√

Figure 1. A comparative visual analysis of color fusion techniques across three related domains: image harmonization, color transfer, and image inpainting. The figure juxtaposes results from leading methods—Harmonizer, S2CRNet, two variants of Color Matcher (MKL and HM-MVGD-HM), CoordFill, and ZITS—against our CFNet. Note the discrepancies in boundary blending, color consistency, and texture preservation, with CFNet exemplifying superior performance in all aspects.

Due to the lack of research specifically focused on the problem of color fusion, we made a comparative analysis with methods from three related areas: image harmonization, color transfer, and image inpainting. For each category, we select two state-of-the-arts methods and compare their visual performance with CFNet. In the image harmonization category, we choose Harmonizer [6] and S2CRNet [7]. We use two kinds of Color Matcher [5] methods: Monge-Kantorovich-Linearization (MKL) and Multi-Variate Gaussian Distributions in combination with Histogram Matching (HM-MVGD-HM). For image inpainting techniques, we compared our method with CoordFill [8] and ZITS [2].

To effectively address the color fusion challenge, an ideal algorithm should seamlessly blend boundaries, maintain the original color in non-boundary areas, and preserve texture. Figure 1 highlights the shortcomings of existing methods: image harmonization techniques produce inconsistent color blends at edges, while the Color Matcher leads to marked color discrepancies. Image inpainting alters texture noticeably—for instance, shortening the cat's tail. In stark contrast, our CFNet proficiently satisfies all these prerequisites, marking it as a superior solution.

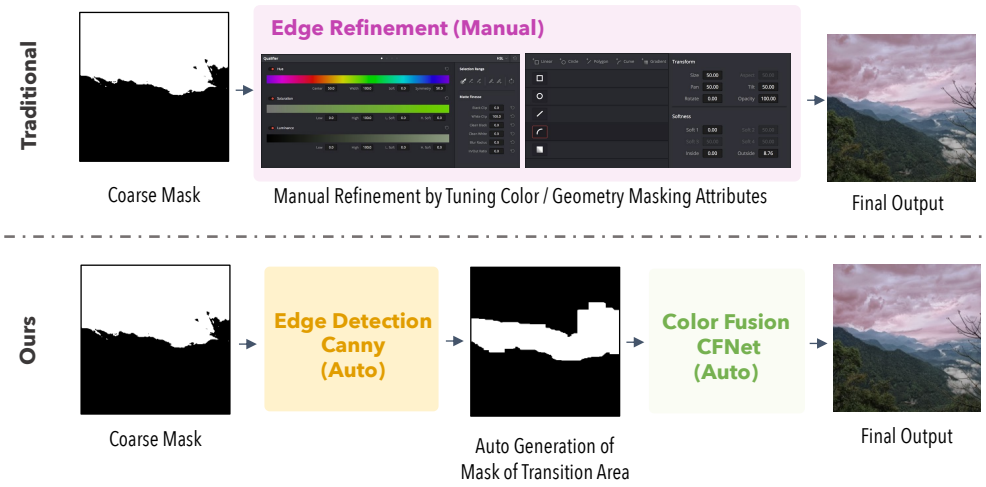


Figure 2. Comparison of CFNet with Traditional Methods: While traditional approaches demand extensive manual adjustments using over 20 controls for mask refinement, CFNet automates this process. Using edge detection (e.g., Canny), it identifies and rectifies potential inconsistencies, delivering a seamless region-specific color edit.

1.3. Seamless Integration of CFNet into Post-production: Eradicating Manual Edge Refinement and Saving Significant Time

CFNet can be easily integrated into current pipelines. In conventional approaches for region-specific color enhancement, refining a coarse mask requires intricate manual adjustments, employing over 20 controls for color and geometric attributes, followed by meticulous visual inspections. As shown in Figure 2, the proposed CFNet radically transforms this workflow by automating edge refinement for these tasks. With CFNet, upon receiving a coarse object mask, potential regions prone to spatial color inconsistencies are swiftly detected using renowned edge detection techniques, such as Canny. Subsequently, CFNet processes the mask and its direct composition, delivering a region-specific color edit that's both consistent and seamless.

2. Demo Application of CFNet for Fine-grained Color Enhancement

As described in the main paper, we have developed a demo application of CFNet showcasing its capability of empowering fine-grained color editing system.

2.1. Method

We present a text-driven image enhancement method that seamlessly combines style and memory colors, facilitating natural fusion of regional adjustments. This approach facilitates both global and fine-grained color enhancement, enabling effective establishment of desired aesthetics while preserving object-specific memory colors. Additionally, it capably translates user preferences expressed in text into color enhancement operations, creating an efficient text-to-enhancement mapping.

The demo application comprises three core components: the Text2LUT module, the Semantic-Aware Memory Color module, and ChromaFusionNet (CFNet). The Text2LUT module interprets user input and translates it into a corresponding LUT capable of executing the requested operations. This presents a challenge, given the inherent ambiguity of natural language. However, we construct this module based on the premise of a shared understanding of LUTs and related natural language descriptions. We carefully curate a dataset and design a ViT-based [4] model capable of effectively handling 3DLUTs, enabling the Text2LUT module to function effectively. The Memory Color module uses the textual input to generate natural colors that persist in a high-quality dataset. We have gathered a large-scale dataset comprised of 1,245,469 high-aesthetic-quality images for this purpose. The ChromaFusionNet can adeptly combine all adjustments, avoiding the artifacts commonly associated with regional adjustments. We innovatively formulate the region fusion as a color inpainting problem and designed a neural network that trained on ImageNet [3] to achieve natural blending and good generalization. The combined operation of these three modules produces a superior end result.

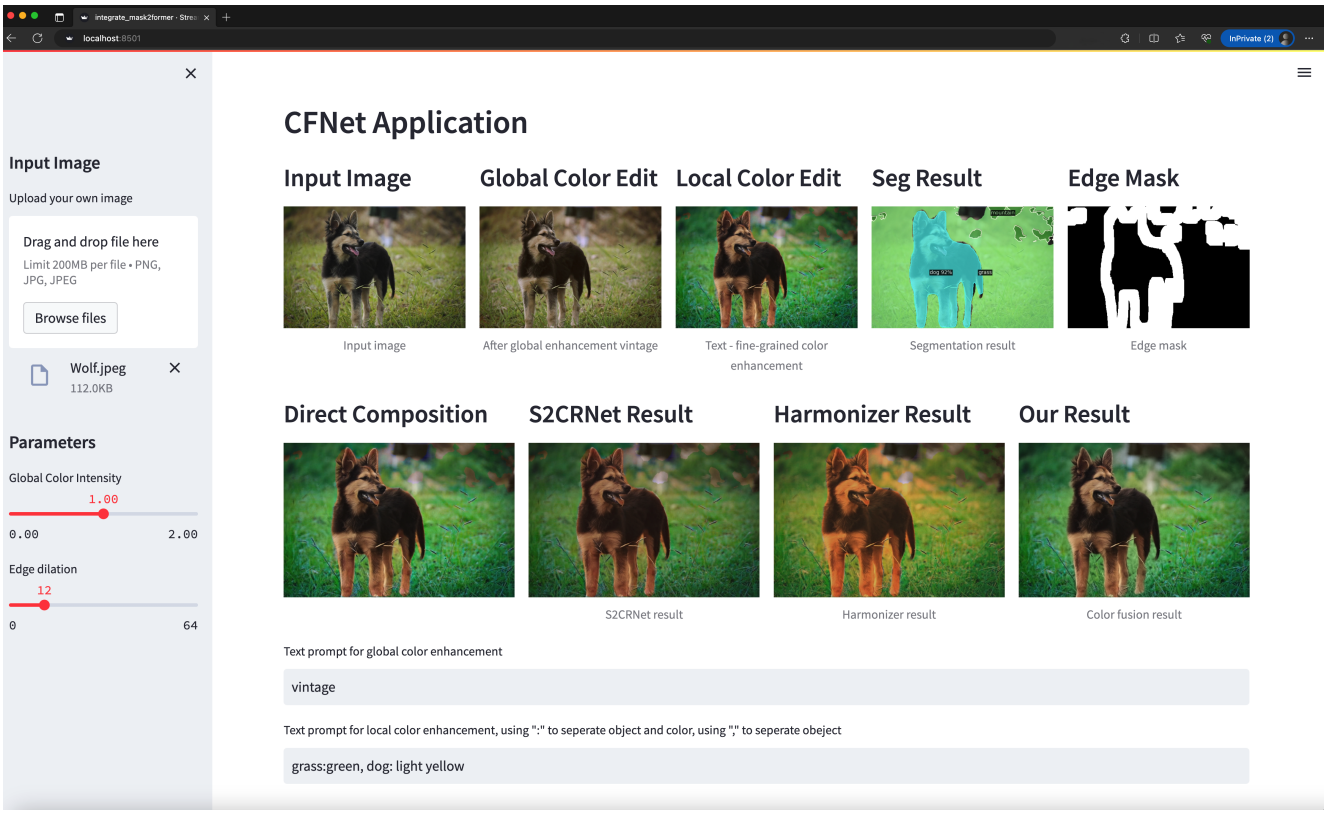


Figure 3. User Interface of the Text-based Fine-grained Color Enhancement: Left sidebar facilitates image uploading. Below the image review, main area provides input boxes for global color grading and object-specific color adjustments. The demonstrated image combines vintage grading with object-specific tones (grass:green, dog:light yellow). CFNet seamlessly integrates regional edits, avoiding spatial inconsistencies or color deviations.

2.2. User Interface

In Figure 3, the UI comprises a left sidebar for image uploads and a main area on the right for image preview and text-based color adjustments. This area provides two text boxes for global color grading and object-specific adjustments. An illustrated example applies a vintage global grade, complemented by object-specific tones (‘grass:green, dog:light yellow‘). CFNet adeptly fuses regional color edits, maintaining spatial consistency and avoiding color deviations even in boundary regions.

2.3. Example Outputs

We present visual results demonstrating the versatility and robustness of the demo application. As shown in Fig. 4, our text-driven image enhancement system exhibits exceptional capacity in producing diverse and personalized color enhancements on the same image using varying text descriptions, showcasing its flexibility. Further, we illustrate the consistency of the system in modifying similar content across different images, utilizing the same text input (Fig. 5). This consistency is indicative of the system’s robustness in handling a variety of input scenarios, demonstrating its potential for practical applications. Multiple images, depicting the results of our method in comparison to state-of-the-art solutions, illustrate the superiority of our approach in maintaining an effective balance of style color and memory color, while simultaneously avoiding artifacts.

3. Experiment Configuration and Computational Efficiency

3.1. Experiment Configuration Details

3.1.1 Training Hardware and Duration

Our model was trained on 4 Tesla V100 GPUs. The training process spanned approximately two days.

Figure 4. Demonstration of our system's ability to produce diverse and personalized color enhancement results using different text descriptions on the same image.

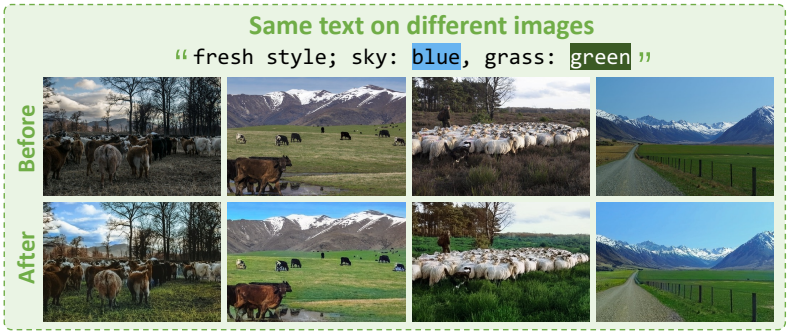


Figure 5. Demonstration of system consistency in enhancing similar content across different images using the same text description.

3.1.2 Data Preparation

- **Image Dataset**: We utilized the ImageNet dataset, resizing images to 224x224 pixels. This resizing step is crucial for standardizing input dimensions and ensuring computational efficiency.
- **Segmentation and Masking**: Using Detectron2, we performed panoptic segmentation on the dataset. To generate input masks, we randomly selected 1-3 objects per image. The edges of these objects were dilated with kernel sizes ranging from 3 to 5 and dilation rates between 2 and 32.

3.1.3 Training Settings

- **Epochs and Batch Size**: The training was conducted over 1,000,000 steps with a batch size of 16. This batch size was chosen to balance between computational efficiency and the need for stochasticity in gradient descent.
- **Optimizers**: We employed the AdamW optimizer for both generator and discriminator networks. The learning rate for the generator was set at 5e-4 and 1e-5 for the discriminator, with beta values of [0.9, 0.95] for both. AdamW was chosen for its effectiveness in handling sparse gradients and adaptive learning rates.
- **Scheduler**: A Cosine Annealing Restart Learning Rate scheduler was used, with an θ_{min} of 1e-6. This scheduler helps in avoiding local minima and ensures better convergence by periodically resetting the learning rate.

3.2. Computational Efficiency

Our benchmark results confirm that CFNet, tested on varying hardware, exhibits commendable computational efficiency. On an NVIDIA V100, CFNet operates at 13.07 FPS, while its coarse module reaches 50.25 FPS; on an NVIDIA RTX 2080Ti, these figures are 11 FPS and 45.66 FPS, respectively. The output of the coarse module, which exceeds the real-time preview standard of 25 FPS, offering more than twice the needed speed for previews. For output rendering, CFNet functions at half the real-time rate, which is acceptable. We have yet to pursue engineering optimizations, indicating substantial potential for

further efficiency enhancements. In essence, CFNet strikes an effective trade-off between speed and performance, aligning it well with practical deployment scenarios.

4. More Ablations

To support the decisions in network design, we have conducted a set of ablations. In these ablations, experiment settings and metrics are the same as reported in the main submission.

4.1. Size of Network Input and Output

Table 1. Ablation studies for reconstruction performance on ImageNet test set and COCO test set. We conduct experiments on the CFNet with refinement module of varying sizes on PSNR, B-PSNR, SSIM, and ΔE metrics.

Size	ImageNet				COCO				Time
	PSNR ↑	B-PSNR ↑	SSIM ↑	ΔE ↓	PSNR ↑	B-PSNR ↑	SSIM ↑	ΔE ↓	FPS ↑
224 × 224	34.96	**26.57**	**0.97**	0.65	36.12	**32.05**	**0.97**	0.73	**12.91**
448 × 448	**35.22**	25.95	**0.97**	**0.62**	**37.03**	31.49	**0.97**	**0.63**	4.11

A comprehensive ablation study was conducted to gauge the reconstruction performance of the CFNet on two prominent test sets: ImageNet and COCO. The study focused on discerning the influence of the refinement module's size on key metrics: PSNR, B-PSNR, SSIM, and ΔE. Instead of evaluating the whole image, the B-PSNR specifically measures the PSNR of the boundary areas to be filled, providing a more targeted assessment for color fusion. The results of these experiments, as showcased in Table 1, are discussed below.

- **Impact of Refinement Module Size**:
  - For both ImageNet and COCO datasets, increasing the refinement module's size from 224x224 to 448x448 led to a discernible improvement in PSNR (from 34.96 to 35.22 for ImageNet and from 36.12 to 37.03 for COCO). This suggests that a larger refinement module size aids in obtaining a more precise image reconstruction.
  - However, while the B-PSNR value for the COCO dataset decreased marginally with the larger size, the ImageNet dataset exhibited a minor reduction. Clearly, increasing the input size doesn't necessarily enhance the color fusion capability.
- **Consistency in SSIM**:
  - Remarkably, the SSIM value, a metric which indicates the structural similarity between the tested image and the reference, remained consistent at 0.97 for both sizes and datasets. This highlights that irrespective of the refinement module's size, the structural integrity of the reconstructed images is preserved.
- **Color Accuracy**:
  - The ΔE values, indicative of the color difference between the reconstructed and reference images, improved slightly with the increase in refinement module size. A decrease from 0.65 to 0.62 for ImageNet and from 0.73 to 0.63 for COCO emphasizes that the 448x448 size achieves better color accuracy in the reconstructions.
- **Processing Speed**:
  - A critical trade-off was observed in the FPS (frames per second) metric, which measures the processing speed. While the 224x224 size achieved a superior speed at 12.91 FPS, increasing the size to 448x448 resulted in a significant reduction to 4.11 FPS. This highlights that while the larger refinement module offers qualitative advantages, it does so at the cost of computational speed.

In conclusion, our ablation study on CFNet underscores the importance of choosing the appropriate module size based on the specific requirements of the task at hand. For applications demanding higher image quality, especially in terms of PSNR and color accuracy, the 448×448 size appears favorable. However, when processing speed is paramount, the 224×224 size would be the more prudent choice.

Table 2. Efficacy of Various Mask Generation Schemes. The table juxtaposes the performance metrics of different mask generation schemes, ranging from selective point densities to utilizing the full available image points. The results compellingly illustrate that denser point configurations lead to superior performance, thereby underscoring the robustness of our approach that leverages the entirety of image points for mask generation.

Point Number	ImageNet				COCO			
	PSNR↑	B-PSNR↑	SSIM↑	ΔE↓	PSNR↑	B-PSNR↑	SSIM↑	ΔE↓
14×14	29.93	24.23	0.94	1.29	31.25	29.57	0.95	1.23
28×28	31.81	25.10	0.95	1.02	33.29	30.95	0.95	1.01
56×56	32.18	25.14	0.95	0.99	33.61	30.75	0.95	0.99
Full	**34.96**	**26.57**	**0.97**	**0.65**	**36.12**	**32.05**	**0.97**	**0.73**

4.2. Training Mask Generation Schemes

We conducted an analysis to evaluate the performance of CFNet using different mask generation schemes, especially on two leading datasets: ImageNet and COCO. The core aim was to comprehend how the number of image points used for mask generation impacts essential metrics like PSNR, B-PSNR, SSIM, and ΔE. Notably, the B-PSNR provides a unique perspective by specifically evaluating the PSNR in the boundary regions requiring fusion. The findings from this analysis, presented in Table 2, are elaborated below.

- **Role of Image Points in Performance**:
  - A trend is evident across both ImageNet and COCO datasets: as the point density for mask generation increases, so does the PSNR value. Specifically, transitioning from a 14x14 to a Full scheme led to substantial improvements in PSNR (from 29.93 to 34.96 for ImageNet and from 31.25 to 36.12 for COCO). This highlights that employing a denser set of image points results in more accurate reconstructions.
  - B-PSNR values, which focus on boundary areas, also depict a consistent improvement with increasing point density, further advocating for the use of a denser point scheme.
- **Texture Integrity with SSIM**:
  - The SSIM values, which determine the structural consistency between the test and reference images, stay resilient across different point schemes, consistently hovering around 0.95-0.97. This suggests that the underlying texture remains consistent, regardless of the point density used in mask generation.
- **Color Fidelity and ΔE**:
  - Observing the ΔE values, which quantify the color difference between test and reference images, it's clear that a Full mask generation scheme provides the most color-accurate results, with the lowest ΔE values across both datasets.

To sum up, this analysis strongly endorses our approach of utilizing all available image points for mask generation in CFNet. It not only enhances the overall image quality in terms of PSNR but also ensures optimum color accuracy and texture consistency.

4.3. Network Structure

Our ablation study, as summarized in table 3, scrutinizes the performances of CNN and ViT models. While CNN excels in detail capturing, ViT shines in addressing long-range dependencies. Notably, both present comparable results on ImageNet and COCO, with a consistent SSIM of 0.97.

The combination of architectures reveals intriguing insights. Specifically, the Coarse (ViT) followed by the Refine (CNN) sequence outperforms other configurations, registering the highest PSNR and B-PSNR values. This superior performance can be attributed to ViT establishing a broad structure, subsequently refined by the CNN for detailed representations, ensuring minimal ΔE metrics.

Conclusively, the juxtaposition of these architectures amplifies their inherent strengths, emphasizing the efficacy of combining neural structures for enhanced results.

Table 3. Network Structure Ablation. By harnessing the detailed prowess of CNNs and the long-range strengths of ViT, our ablation highlights the superiority of combining both. Notably, positioning the CNN-based refinement module after the ViT-based coarse module emerges as the optimal configuration.

Network Structure	ImageNet				COCO			
	PSNR ↑	B-PSNR ↑	SSIM ↑	ΔE ↓	PSNR ↑	B-PSNR ↑	SSIM ↑	ΔE ↓
Refine (CNN)	34.57	25.76	**0.97**	0.66	35.88	31.29	**0.97**	0.73
Coarse (ViT)	34.40	25.61	**0.97**	0.68	35.67	30.75	**0.97**	0.74
Refine (CNN) + Coarse (ViT)	34.64	25.81	**0.97**	**0.65**	36.06	31.58	**0.97**	**0.72**
Coarse (ViT) + Refine (CNN)	**34.96**	**26.57**	**0.97**	**0.65**	**36.12**	**32.05**	**0.97**	0.73

Table 4. Performance Comparison of Image Harmonization Methods on ImageNet and COCO: This table contrasts CFNet with established methods like Harmonizer and S2CRNet, including their lightness-constant versions (denoted as -$\overline{L}$). Adjustments were made using the iHarmony4 dataset for consistent brightness. CFNet consistently outperforms others.

Dataset: **ImageNet**	W_1 ↓	IP@0.85↓	IP@0.90↓	IP@0.9651↓
Harmonizer	5.44	8.38	7.86	6.50
Harmonizer-$\overline{L}$	4.82	7.88	7.55	6.68
S2CRNet	6.69	8.30	7.78	6.46
S2CRNet-$\overline{L}$	6.71	8.09	7.73	6.83
CFNet (Ours)	**1.66**	**5.73**	**5.28**	**4.14**
Dataset: **COCO**	W_1 ↓	IP@0.85↓	IP@0.90↓	IP@0.9651↓
Harmonizer	12.52	3.24	3.03	2.46
Harmonizer-$\overline{L}$	10.91	3.07	2.93	2.53
S2CRNet	20.88	3.71	3.47	2.85
S2CRNet-$\overline{L}$	21.26	3.52	3.37	2.95
CFNet (Ours)	**0.21**	**2.63**	**2.34**	**1.70**

5. Comparison with Existing Algorithms in the Fixed L Channel Setting

Modification of existing methods to preserve brightness, particularly in inpainting, style transfer, and color matching, is fraught with substantial challenges or remains impractical. Even enhanced harmonization approaches that preserve the brightness channel still fall short of CFNet's performance. CFNet contrasts with inpainting that alters the lightness channel, compromising texture integrity despite successful fusions. Although style transfer involves intricate content and style disentanglement, color and lightness are intertwined; disentangling these elements presents a promising yet untapped research challenge. Traditional color matching, based on histogram matching in RGB space, will produce images with seriously deviated color when extended to spaces like Lab or HSV that preserve lightness due to properties of those color spaces. In summary, the revision of these methods to preserve brightness is either infeasible or a complex challenge.

As for enhanced harmonization methods, we've preprocessed the training data in iHarmony4 dataset for unchanged brightness to retrain those models and made the modification that incorporating the input's lightness directly into the output to ensure a valid comparison. As shown in Table 4, the results are clear: harmonization methods, when adjusted for constant brightness (Harmonizer-$\overline{L}$ and S2CRNet-$\overline{L}$), display marginal improvements or maintain performance, as evidenced by the table below. CFNet outperforms all benchmarks, affirming its superiority even over those adjusted for constant lightness.

6. Additional User Interview for CFNet with a Professional Colorist

6.1. Overview

A comprehensive user study was conducted with an experienced colorist to evaluate our demo application based on CFNet for fine-grained color enhancement. The study aimed to gather insights regarding the user experience, performance, and overall effectiveness of the application and the underlying CFNet.

6.2. Text-based Fine-grained Color Enhancement Application Evaluation Results

The participant successfully completed all tasks designed to test the functionality of our system. He was able to:

- Create a LUT from descriptive text. The participant noted that the model was intuitive, and the LUTs generated were

remarkably accurate and visually pleasing.

- Leverage the Memory Color Enhancement Module to generate a variety of color outcomes. He praised the module's ability to accurately translate memory colors, making color enhancement a straightforward process.
- Utilize the Text-driven Color Mapper to determine color attributes from the descriptive text. He found the module innovative and accurate in interpreting color attributes.
- Use the CFNet to preserve original color attributes while maintaining spatial consistency. He was impressed by the module's ability to retain original color attributes, observing that it significantly reduced time spent on manual color grading.

6.3. Color Fusion via CFNet Evaluation Results

The participant conducted a thorough review of the performance of CFNet in terms of its visual performance, speed, and ease of use.

- **Visual Performance:** The participant commended the CFNet's ability to consistently generate visually pleasing results, even in challenging color scenarios.
- **Speed:** The participant noted the impressive speed of CFNet. The model was able to produce color fusion outcomes rapidly, making real-time color grading feasible. CFNet's efficiency was observed in its ability to preserve original color attributes while maintaining spatial consistency, significantly reducing time spent on manual color grading. The participant emphasized that CFNet's speed was a significant advantage, especially for large-scale color grading projects.
- **Ease of Use:** The participant appreciated the user-friendliness and intuitiveness of CFNet. The model was found to be easy to work with, even for users with limited technical knowledge. CFNet's ease of use was emphasized as a crucial factor in its potential adoption in the color grading industry.

6.4. Conclusion

The results of the user study with an experienced colorist confirm the success and potential of our system. It delivers on its promise of a revolutionary color fusion approach that is accurate, efficient, and user-friendly. We appreciate the constructive feedback from our participant and will use his insights for further improvement and optimization of our system. The future of color fusion certainly looks bright with our solution.

7. More Visual Examples

In this section, we present a series of visual examples that underscore the efficacy of CFNet in color fusion tasks. Through these diverse examples, we highlight CFNet's ability to adeptly address the shortcomings inherent in existing methods. Notably, while our main paper focused predominantly on comparisons with image harmonization and color transfer approaches, here we further extend the analysis to encompass comparisons with image inpainting techniques. This extended evaluation, previously not presented due to page constraints in the main paper, provides a more thorough assessment of CFNet's performance and reveals its wide-ranging applicability in the context of color fusion tasks.

7.1. Image Inpainting

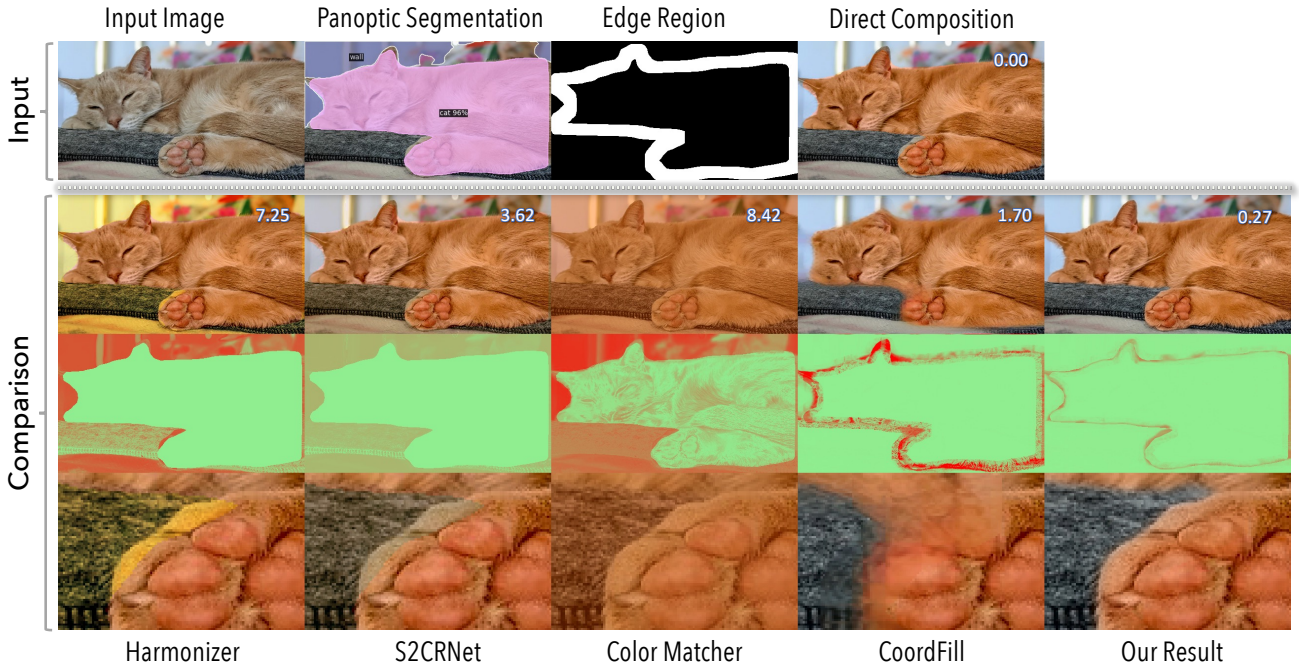
Image inpainting refers to the process of reconstructing missing or corrupted areas within an image by borrowing information from surrounding regions. While this technique can be effective for image restoration, object removal, or image completion, it exhibits several limitations when applied to color fusion.

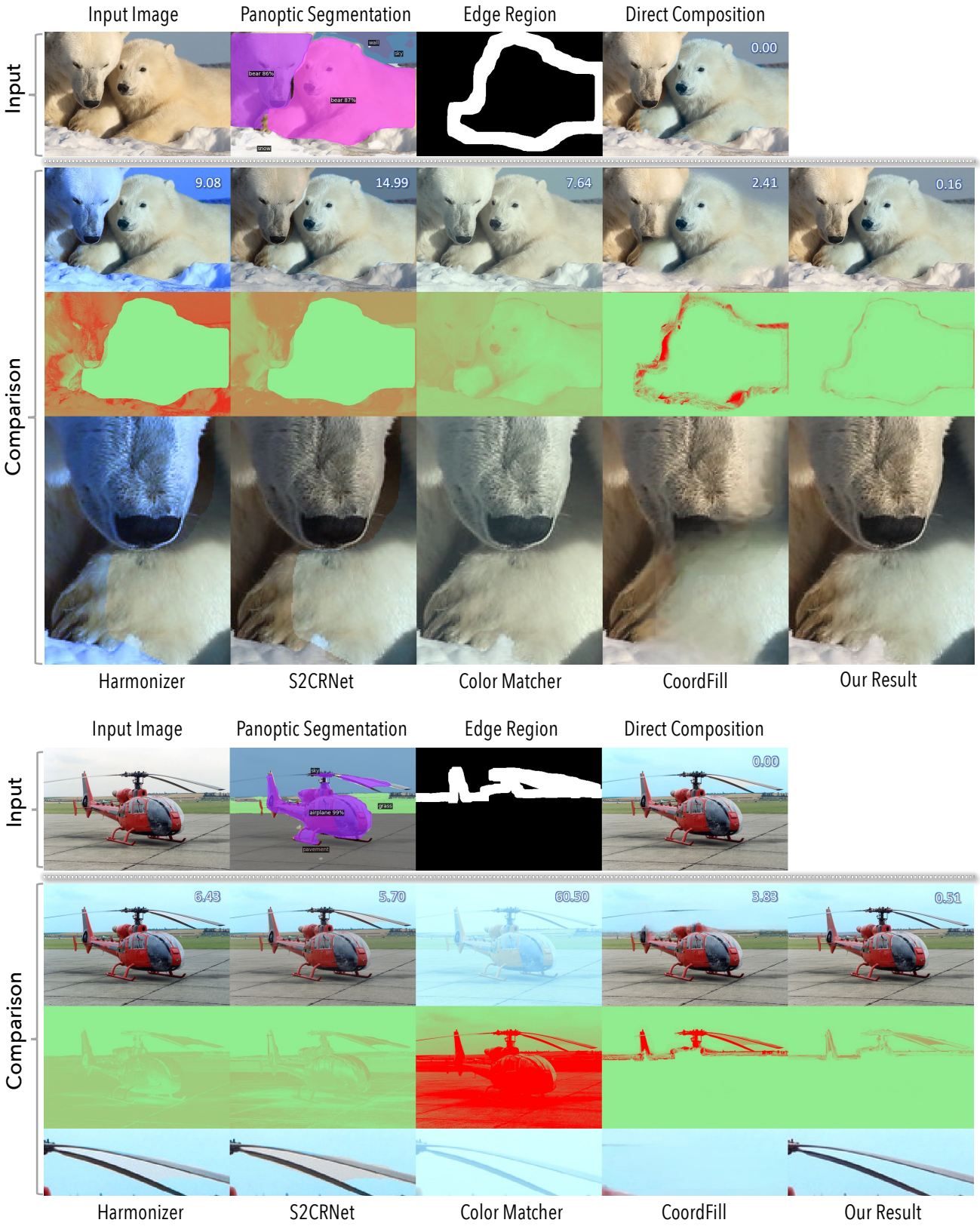
Inpainting methods often result in altered texture in the filled areas, given their reliance on adjacent pixels, producing an inconsistent and unnatural appearance. Furthermore, they may cause the loss of original context within the masked regions, detracting from the overall image quality and fidelity. Even if we can post-processing the result image with original L channel to preserve the image texture, it may introduce inconsistency between color and texture.

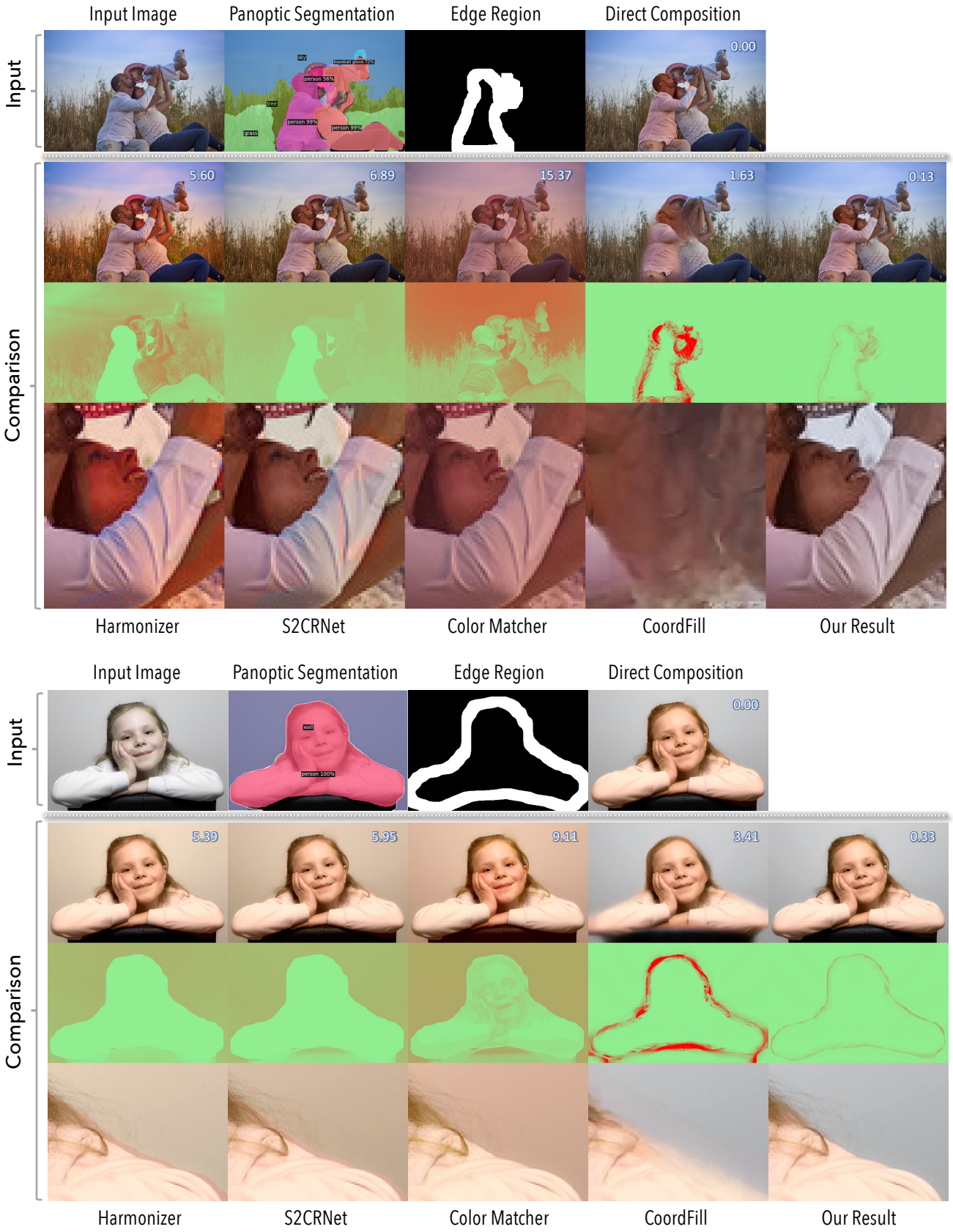
In contrast, CFNet is meticulously designed to circumvent these challenges. As the following visual examples demonstrate, CFNet ensures the preservation of texture, accurate color blending, and the elimination of color deviations, resulting in superior color fusion results.

7.2. Single Object Color Fusion

Figure 6 shows the visual results in single object color fusion cases. We compared CFNet with harmonization methods: Harmonizer [6] and S2CRNet [7], color transfer method: Color Matcher, and image inpainting method: CoordFill [8].




Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
0.00
Comparison
9.08
14.99
7.64
2.41
0.16
Harmonizer
S2CRNet
Color Matcher
CoordFill
Our Result
Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
0.00
Comparison
6.43
5.70
60.50
3.83
0.51
Harmonizer
S2CRNet
Color Matcher
CoordFill
Our Result


Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
0.00
Comparison
5.60
6.89
15.37
1.63
0.13
Harmonizer
S2CRNet
Color Matcher
CoordFill
Our Result
Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
0.00
Comparison
5.39
5.95
9.11
3.41
0.33
Harmonizer
S2CRNet
Color Matcher
CoordFill
Our Result

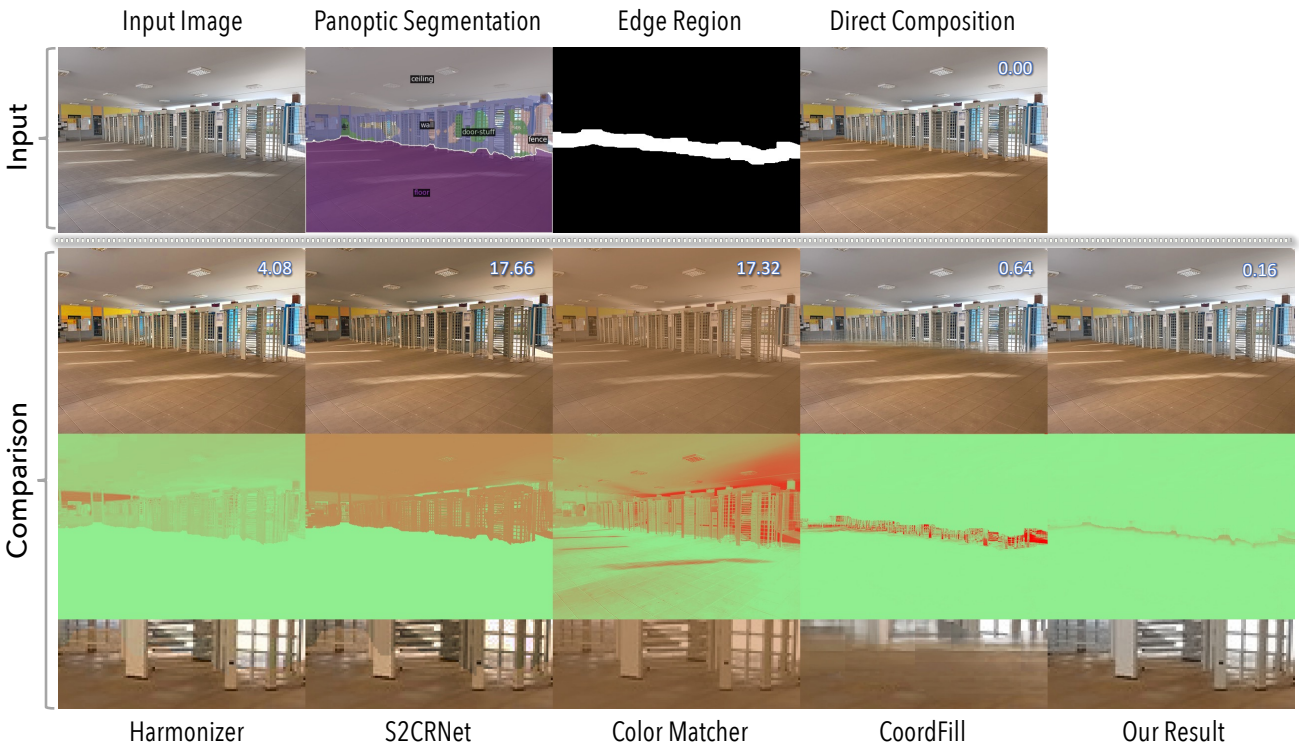
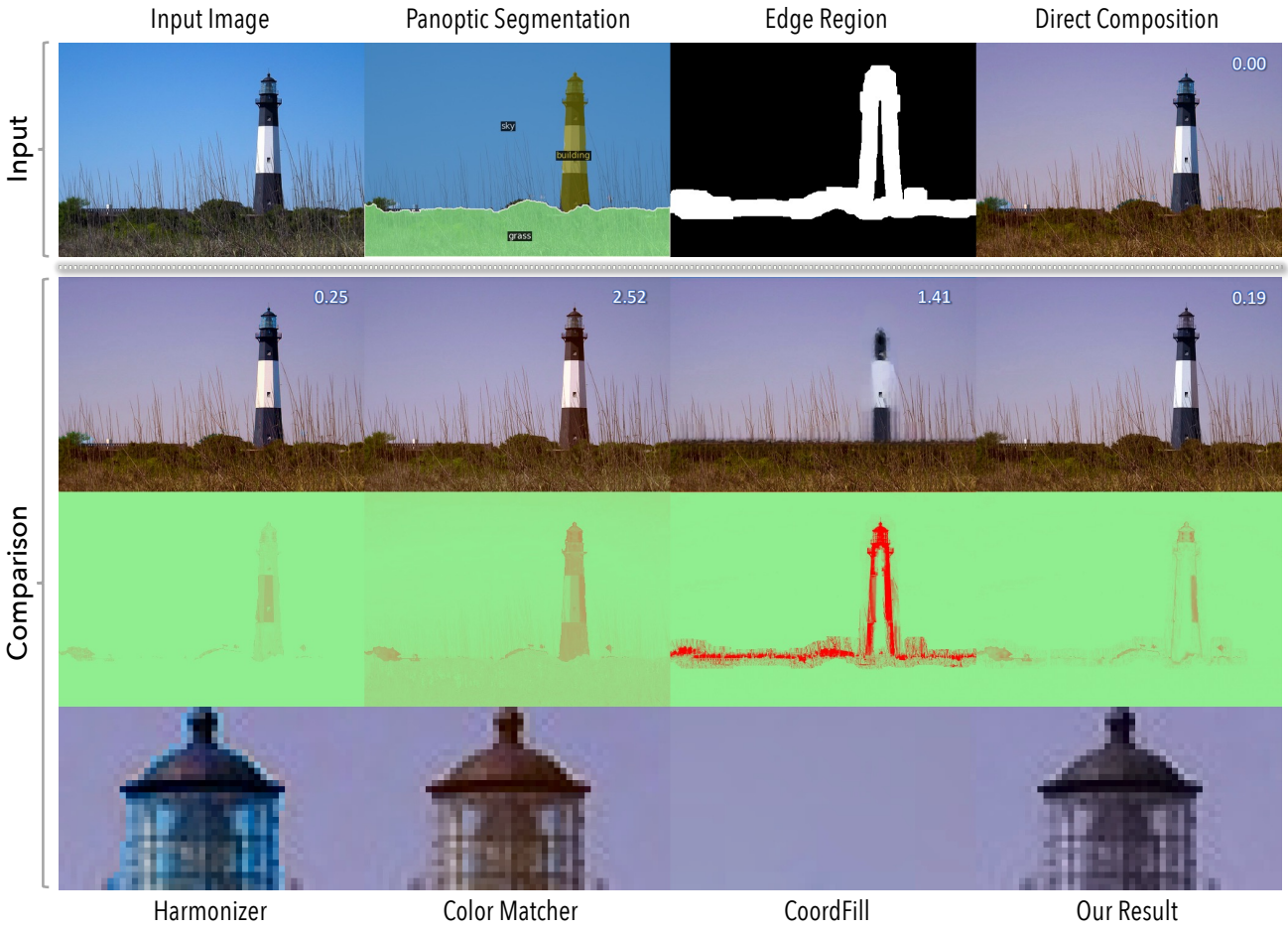
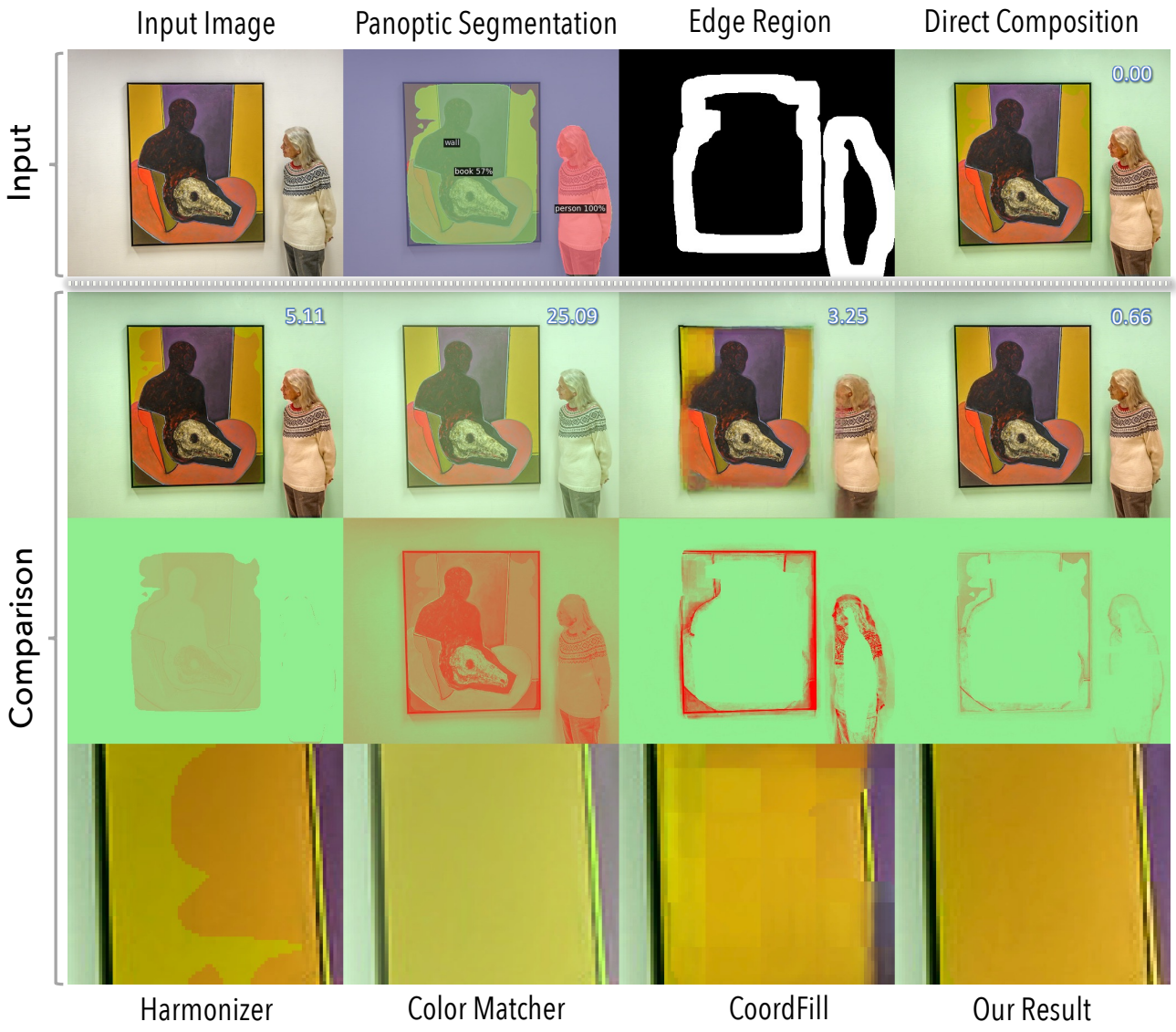


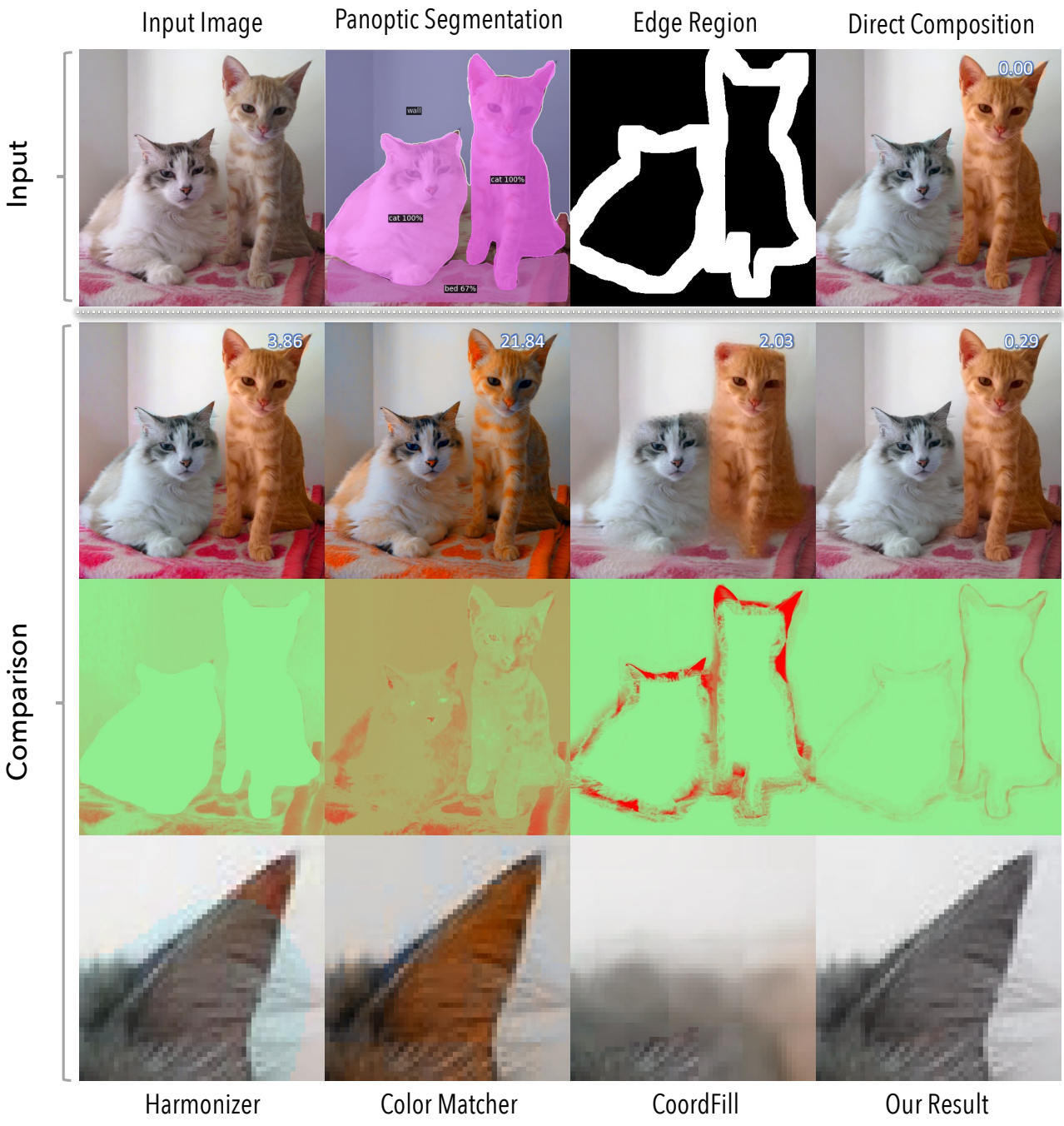
Figure 6. Visual results of single object color fusion cases, comparing CFNet with Harmonizer [6], S2CRNet [7], Color Matcher [5] and CoordFill [8]. For each set of four rows, the first row presents the input image, panoptic segmentation result, its corresponding edge region and direct composition result produced by single object color enhancements. The second row shows the result images generated by the respective methods. The third row presents the difference maps between the results and direct composition images, highlighting the color discrepancies. The fourth row offers zoomed-in views of the corresponding regions, providing a closer look at the details. The W_1 distance between each method's result and the direct composition is also included, where smaller values indicate a closer color distribution to the direct composition.

7.3. Multi-object Color Fusion

CFNet can also perform color fusion for multiple objects, while existing methods such as image harmonization and color transfer encounter difficulties in such setting. Figure 7 illustrates the remarkable capacity of CFNet to handle complex multi-object color fusion tasks, paving the way for more versatile and visually pleasing color blending.


Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
sky
building
grass
0.00
Comparison
0.25
2.52
1.41
0.19
Harmonizer
Color Matcher
CoordFill
Our Result


Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
Comparison
wall
book 57%
person 100%
0.00
5.11
25.09
3.25
0.66
Harmonizer
Color Matcher
CoordFill
Our Result


Input Image
Panoptic Segmentation
Edge Region
Direct Composition
Input
wall
cat 100%
cat 100%
bed 67%
0.00
Comparison
3.86
21.84
2.03
0.29
Harmonizer
Color Matcher
CoordFill
Our Result

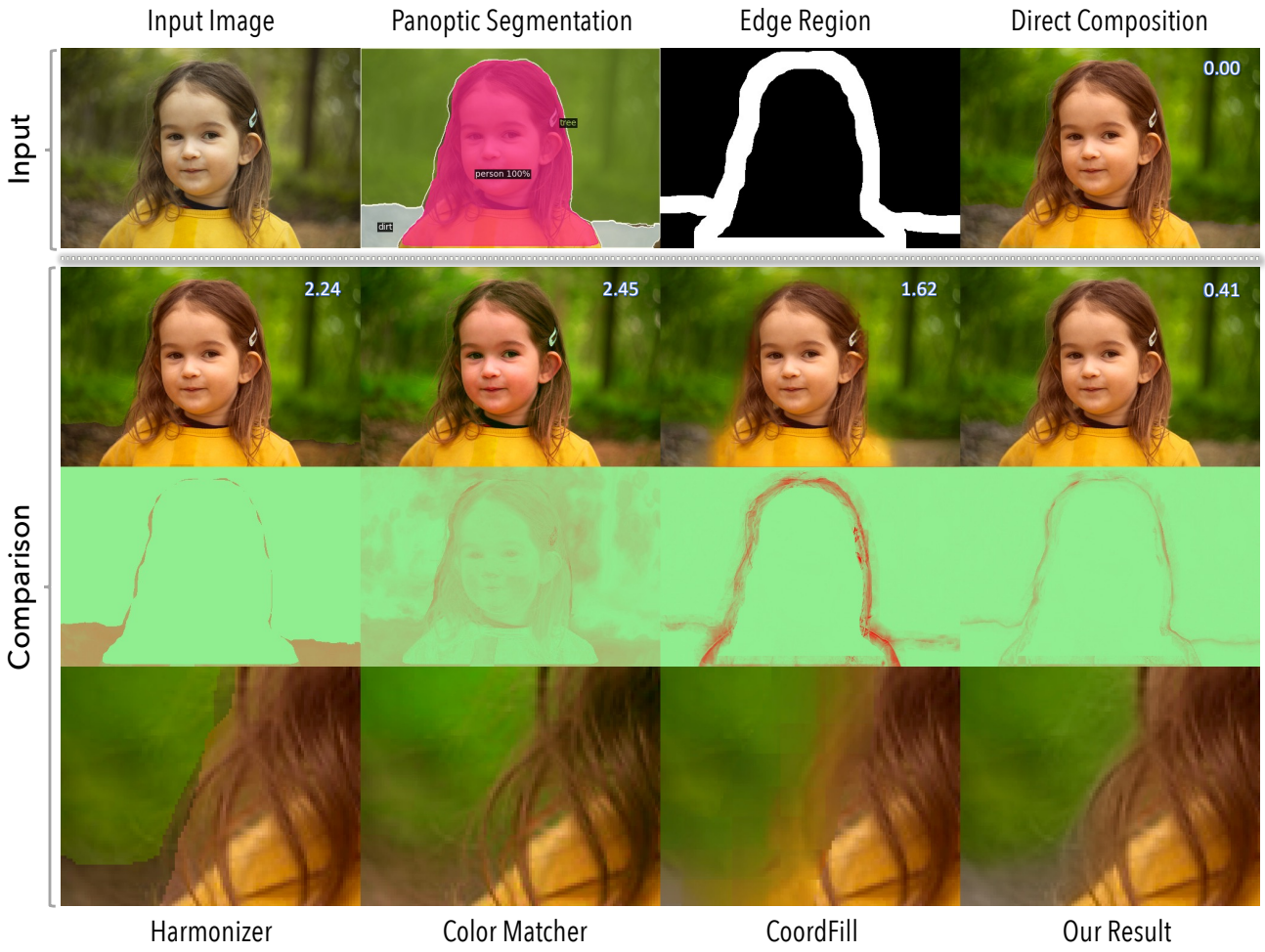


Figure 7. Visual comparisons of multi-object color fusion results obtained by Harmonizer [6], Color Matcher [5], CoordFill [8] and our method. Each quartet of rows follows a similar structure to Figure 6, displaying the original input, the result images, difference maps, and zoomed-in regions, along with W_1 distances for color similarity evaluation.

7.4. Limitation and Failure Cases

As shown in Figure 8, CFNet faces challenges with transparent areas due to inherent complexities and potential training data gaps. Despite producing generally acceptable results, there are color inconsistencies. In addition, as shown in robustness evaluation, if the ratio of colors to be predicted is too high, CFNet's ability of color inpainting is limited. Future endeavors will emphasize dataset diversity, especially transparency, and algorithm refinements.

Figure 8. Failure cases. CFNet's output illustrates color discrepancies within transparent regions.

References

[1] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S. Brown. Image recoloring based on object color distributions. In *Eurographics 2019 - Short Papers*. The Eurographics Association, 2019. 2

[2] Chenjie Cao, Qiaole Dong, and Yanwei Fu. ZITS++: image inpainting by improving the incremental transformer on structural priors. *CoRR*, abs/2210.05950, 2022. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 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, 2021. 3

[5] Christopher Hahne and Amar Aggoun. Plenopticam v1.0: A light-field imaging framework. *IEEE Transactions on Image Processing*, 30:6757–6771, 2021. 2, 13, 17

[6] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W. H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *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*, pages 690–706. Springer, 2022. 2, 10, 13, 17

[7] Jingtang Liang, Xiaodong Cun, Chi-Man Pun, and Jue Wang. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *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*, pages 334–349. Springer, 2022. 2, 10, 13

[8] Weihuang Liu, Xiaodong Cun, Chi-Man Pun, Menghan Xia, Yong Zhang, and Jue Wang. Coordfill: Efficient high-resolution image inpainting via parameterized coordinate querying. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 1746–1754. AAAI Press, 2023. 2, 10, 13, 17

[9] Alexis Van Hurkman. *Color correction handbook: professional techniques for video and cinema*. Pearson Education, 2014. 2