MULTI-VIEW COLOR CORRECTION WITH MULTI-OBJECTIVE GEOMETRY-GUIDED VOXEL CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

The modern image processing pipeline is often designed for one single image without any spatial and temporal information. Each component in the ISP can introduce slight view dependent differences that cause larger visual inconsistencies, thereby impacting virtual reality experience and affecting downstream tasks like texture mapping and some 3D reconstruction tasks. We formulate the color consistency as a multi-view geometry-guided optimization problem. Utilizing depth and camera poses to back-project image into 3D space to jointly optimize all correspondence camera views in a global manner through voxelization. We optimizing a trainable per-view bilateral grid as our color transformation matrix, applied indirectly in 3D voxel space to achieve multi-view color consistency. To further improve the color consistency while preserving the visual fidelity of the image, we also introduce a robust multi-objective color consensus method for color correction instead of operating on sRGB space directly. The proposed method shows near SOTA spatial color consistency and better performance in complex cases.

1 INTRODUCTION

High quality 3D representation of real world scenes are essential for a wide range of applications, including fields like virtual reality, 3D visualization and reconstruction, smart city and robotics. A good reconstruction model should capture both an accurate physical geometry structure and its visual appearance. Not only will it enhance the visual experience for end users, but it also gives advantages for downstream tasks. A common approach for such data acquisition involves one or more cameras to capture a scene from multiple viewpoints to cover both geometry and appearance information as much as possible (Li et al., 2026; Chang et al., 2017). For a typical multi-view image acquisition pipeline, the images are often captured under different conditions, resulting in differences in color, brightness, contrast and exposures even when they are observing the exact same object. With such data discrepancies, they can produce visual artifacts in immersive visualization, especially for immersive applications where appearance discontinuities can greatly reduce the realism of the scene (Serrano et al., 2019), and further degrade the performance of any downstream 3D reconstruction tasks (Hong et al., 2023). Therefore achieving image color consistency across each view is an essential preprocess step.

Image correction and visual consistency correction has been an active topic in many multi-view related areas. In traditional domain such as image stitching (Xia et al., 2017) and panoramic image generation, color correction and color transfer has already been developed to reduce the visible seams between overlapping regions. Early approaches often estimate a global correction gain (Brown & Lowe, 2007) or transform matrix between overlapping images (HaCohen et al., 2013; Finlayson et al., 2016). While more modern methods formulate this as a global optimization problem for more robust global consistency (Moulon et al., 2013). Similarly, color transfer and color harmonization has been developed for a more complex appearance transformation from one image to another (Shin et al., 2025). In the 3D reconstruction domain, color correction and appearance embedding (Martin-Brualla et al., 2020) has also been researched to minimize the color discrepancies across multi-view images.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, there are some limitations among existing methods. In 2D image color correction topic, they often struggle to handle complex color discrepancies (Xia et al., 2019), and they relay on good initial matching points for the color correspondences generated by structure from motion (SfM). Some existing methods estimate a single global gain (Brown & Lowe, 2007) or color transformation for each image. While they can compensate for simple global exposure or color shift, they can not handle color discrepancies that vary within a single image. Another major challenge is the lack of ground truth color consistent data (Cheng et al., 2015).

Motivated by these limitations. We propose a geometry guided color correction approach. Instead of appearance feature matching, we project observations into a common world frame using depth and pose, quantized into voxels and forming color correspondences per voxels as the raw input for optimization. This adds supervision even on textureless surfaces. For each shared voxel, we compute a robust consensus chromaticity from the original observations using Gaussian color confidence, we also add a saturation-preserving regularization that aligns observations to the consensus hue direction while allowing colors to retain their saturation, reducing desaturation artifacts, which yields better result than operating on sRGB space. We use bilateral grid (Gharbi et al., 2017) as our color correction model as it is expressive enough to model complex color transformation while maintaining a relatively small number of parameters and it can be performed in low resolution. We tackle the color consistency correction as an optimization process without the need for ground truth data, which yields a near SOTA and sometimes SOTA results compare to existing methods.

2 Related Works

Multi-view Color Correction Multi-view color correction has been a long-standing area of study in the context of 2D image stitching and multi-view reconstruction. In early image stitching application, several blending techniques are used for reducing the inconsistent color/texture artifacts in overlapping region (Agarwala, 2007; Szeliski et al., 2011). However blending is insufficient for large appearance discrepancies in non-overlapping regions. Later approaches such as Fecker et al. (2008) and Qian et al. (2013) utilize the correspondences found in the overlapped area to build a transfer function or mapping table that can be applied to all images. More recent methods (Yu et al., 2024; Li et al., 2022) use sparse matching points as their correspondences generated by SfM (structure from motion). To minimize the color differences between images, several nonparametric methods are developed (Pitie et al., 2005; Su et al., 2012). Reinhard et al. (2001) is the first to introduce a parametric global color transfer method based on image statistics. Hong et al. (2023) expand upon this by operating in YCbCr and treats luminance and chrominance separately. Closest to our setting, Madeira et al. (2024) exploit RGB-D geometry with NeRF (Mildenhall et al., 2020), and a small per-image MLP is then trained to ensure cross-view color consistency.

Appearance Embedding and Color Transfer In the field of 3D reconstruction, per-image appearance embedding is used with NeRF (Mildenhall et al., 2020) and 3DGS (Kerbl et al., 2023). NeRF-W (Martin-Brualla et al., 2020) use GLO (Bojanowski et al., 2019) to jointly model appearance and 3D geometry. In the work of HDRNet (Gharbi et al., 2017), they shows that bilateral grid (Chen et al., 2007) is an expressive, edge preserving model for image enhancement with low computation requirement. Following HDRNet, recent work Wang et al. (2024) adopts bilateral grids as a lightweight, generalized representation of ISP processing for each camera view. Shin et al. (2025) decouple appearance harmonization from scene reconstruction by predicting a per-view 3D bilateral grid with a cross-view Transformer, aligning all views to a reference frame and training in a self-supervised manner. Wang et al. (2026) instead unify appearance embedding and bilateral grids in a multi-scale fashion trained end-to-end with the Gaussian-splatting objective, improving reconstruction robustness. Niemeyer et al. (2025) specifically tackle the exposure inconsistency problem by introducing an additional NeRF to model the 3d exposure representation. However, these methods are primarily designed to decompose the per-view disagreement away from shared 3D representation for more robust reconstruction rather than to output a color-consistent image set directly. The reference alignment strategy can introduce a bias when the reference view can not covers all parts of a scene.

3 METHODS

3.1 OVERVIEW

Our central idea is to formulate the multi-view color consistency problem as a geometry-guided global optimization process. Instead of solely relying on sparse feature matching (Tian & Cohen, 2016), we incorporate depth and camera poses information to get better semi-dense correspondence. Given a set of images together with their depth and camera poses, all valid observations are projected into a shared 3D space and voxelized. In this common 3D space, we group all the related pixels together through voxelization to form the input basis for our optimization. For a pixel p in image i, let $c_i(p) \in \mathbb{R}^3$ denote its RGB color. We want a camera dependent transformation

$$\hat{c}_i(p) = F_i(c_i(p), p) \tag{1}$$

where F_i is a 3D bilateral grids matrix that maps the original observation to its corrected color. Rather than estimate F_i independently for each view, we build constraints between views through their common observations. The resulting objective therefore couples different views according to their actual geometric co-visibility. During the optimization step, directly using all observations within all shared voxels substantially increases the complexity and memory of the optimization. Instead we adopt a sampling strategy utilizing mini batches SGD (Bottou, 1991) to optimize the 3D bilateral grids.

3.2 GEOMETRY-BASED VOXEL CORRESPONDENCE

Let s be the unit voxel grid length and x_o an origin for the scene grid. Every valid world point x_i projected from pixel p_i and image i is assigned the integer voxel coordinate $q(x)$, and the observation pixel set associated with voxel $\mathcal{V}_k$ is

$$q(x) = \left\lfloor \frac{x_i - x_o}{s} \right\rfloor \qquad \mathcal{V}_k = \big\{(i,p) \;\big|\; q\big(x_i(p)\big) = k\big\} \tag{2}$$

Let M_k denote the set of distinct cameras represented in $\mathcal{V}_k$. A voxel is considered eligible only when $|M_k| \geq 2$.

This construction treats each voxel set as a **weak geometric correspondence**. Points in the same voxel are not assumed to be exact projections of the same 3D point. We assume those points belong to a relatively small local surface neighborhood whose intrinsic color should be similar across views. Only valid pixels with reliable depth are considered. Under and over exposed pixels are excluded from correspondence construction since their chromaticity is noisy and unstable. This geometry correspondence construction provides an advantage that even on textureless surfaces, it yields reliable matching points.

3.3 HIERARCHICAL CORRESPONDENCE SAMPLING

Simple voxelization will produce multiple observations from the same camera within a single voxel grid. Directly aggregating all observations will results in false color and blurry texture by mixing color from different surfaces. Thus we add a hierarchical sampling process to refine the correspondences inside each voxel grid. We instead create a voxel consensus to represents the voxel grid. This allowing us to optimize all cameras observing the same region jointly. Avoiding error accumulation when propagating through each image pair.

For voxel $\mathcal{V}_k$, one camera and one of its observations are sampled as a pivot, with world position x_{pivot}. For every other camera $i \in M_k$ sharing the same voxel grid, the representative observation is selected by nearest 3D position:

$$p^*_{ki} = \arg \min_{p:(i,p)\in\mathcal{V}_k} \|x_i(p) - x_{\text{pivot}}\|_2^2 \tag{3}$$

The resulting representative set is

$$\mathcal{R}_k = \{(i, p^*_{ki}) \mid i \in M_k\} \tag{4}$$

The pivot changes over repeated sampling. As a result, different local points inside a voxel will be involved in the supervision of the transformation over time rather than fixing the voxel to one arbitrary representative. This stochastic construction approximates dense correspondence coverage without going through every cross-camera pixel pair.

3.4 Robust Color Consensus

Standard sRGB images are gamma-encoded for display, so their values are not proportional to light intensity. Performing color transformation or averaging colors directly in sRGB would mix non-linear values produce undesirable behavior (Larry Gritz, 2026). The images are first converted from sRGB to linear RGB before optimization and the corrected result is converted back to sRGB. For each shared voxel grid, the objective is to minimize the disagreement among the representative observations after correction.

Inspired by the previous works (Ran et al., 2019; Hong et al., 2023), we decoupled the plain RGB into different objectives for optimization. More specifically, we divide the disagreement into chromaticity, exposure, and hue/saturation. Let $c_{ki} \in \mathbb{R}^3_+$ be the linear RGB color selected from camera i in voxel $\mathcal{V}_k$. Its normalized chromaticity χ is

$$\chi(c_{ki}) = \frac{c_{ki}}{\mathbf{1}^\mathsf{T} c_{ki}} \tag{5}$$

Chromaticity removes overall intensity and retains the relative color information. This allows views with large brightness disagreement to constrain color without requiring equal exposure. Each observation in the representative set $\mathcal{R}_k$ receives a confidence based on its distance from the set mean $\bar{\chi}_k$. Ensuring that observations close to the dominant color contribute more. The robust voxel chromaticity target t_k is:

$$w_{ki} = \exp\left(-\frac{\|\chi(c_{ki}) - \bar{\chi}_k\|_2^2}{2\sigma_c^2}\right) \qquad t_k = \frac{\sum_{i \in M_k} w_{ki}\chi(c_{ki})}{\sum_{i \in M_k} w_{ki}} \tag{6}$$

Directly moving every observation to the same chromaticity can reduce saturation when a voxel contains different colors caused by noise. The vividness should be preserved where the original is genuinely saturated. A saturation-preserving target separates chromaticity into a direction away from neutral gray and a magnitude along that direction. Let the neutral chromaticity be $n = [1/3 \quad 1/3 \quad 1/3]^\mathsf{T}$. For observation i in voxel $\mathcal{V}_k$, define

$$v_{ki} = \chi(c_{ki}) - n \qquad s_{ki} = \|v_{ki}\|_2 \tag{7}$$

where v_{ki}/s_{ki} represents hue direction and s_{ki} represents saturation in the chromaticity simplex. For chromaticity consensus target t_k, we define its voxel consensus hue direction d_k and saturation $\bar{s}_k$ as

$$d_k = \frac{t_k - n}{\|t_k - n\|_2} \qquad \bar{s}_k = \frac{\sum_{i \in M_k} w_{ki} s_{ki}}{\sum_{i \in M_k} w_{ki}} \tag{8}$$

To avoid desaturation, the target saturation for each camera blends its original saturation with the voxel consensus saturation:

$$\rho_{ki} = \text{clip}\left[\left(\frac{s_{ki}}{s_{\max}}\right)^\gamma, 0, 1\right] \tag{9}$$

$$s^*_{ki} = \rho_{ki} s_{ki} + (1 - \rho_{ki})\bar{s}_k \tag{10}$$

where $s_{\max} = \sqrt{2/3} \approx 0.816$ is the maximum distance from neutral gray to a vertex of the RGB chromaticity simplex, and γ controls the saturation-preservation curve. The camera-specific saturation-preserving target is

$$t_{ki}^{\text{hue}} = \text{Norm}\left(\text{clip}_{[0,1]}\left[n + d_k s_{ki}^*\right]\right) \tag{11}$$

where Norm normalizes the vector to sum to one. All observations share the robust consensus hue direction d_k, but strong saturated colors preserve more of their original saturation value. While weak saturated observations are pulled more toward the voxel consensus saturation. This avoids the washed-out appearance caused by enforcing one averaged chromaticity on every camera.

Finally, we add a weighted exposure consensus in log space following the same logic as chromaticity to address brightness discrepancies. By separating the plain RGB into several properties, chromaticity and exposure are calculated in their own axis, which in return allowing us to achieve better color consistency while maintaining better color vividness.

3.5 Per-camera Bilateral-grid

A single global gain or matrix per camera assumes that the color discrepancy is approximately uniform throughout the image, which is often not the case. A transformation at every pixel would be highly susceptible to noise and require heavy computations. We use a bilateral grid as an intermediate solution. Every camera i has a low-resolution coefficient grid

$$G_i \in \mathbb{R}^{12 \times D \times H_g \times W_g} \tag{12}$$

indexed by normalized luminance l, vertical image coordinate v, and horizontal coordinate u. Each grid vertex stores 12 coefficients representing a 3×4 RGB transformation. At observation (i, p), we get the transformation from trilinear interpolation at

$$T_i(p) = \text{TriInterp}\big(G_i, u(p), v(p), l_i(p)\big) \tag{13}$$

For a pixel p from image i in the voxel k, its corrected linear RGB color is

$$\hat{c}_i(p) = \hat{c}_{ki} = T_i(p) c_i(p) \tag{14}$$

Because the coefficients are defined on a low-resolution grid, the resulting full resolution transformation remains smooth rather than behaving as an unconstrained per-pixel correction.

3.6 Optimization Objective

Let $\Theta = \{G_i\}_{i=1}^N$ denote all camera grids. The objective combines cross-view agreement with photometric and spatial regularization:

$$\mathcal{L}(\Theta) = \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{expo}} \mathcal{L}_{\text{expo}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}} \tag{15}$$

Multi-view chromaticity consistency For every representative observation, the corrected chromaticity should agree on the voxel consensus:

$$\mathcal{L}_{\text{cons}} = \frac{1}{\sum_k |M_k|} \sum_k \sum_{i \in M_k} \left\| \chi\big(\hat{c}_{ki}\big) - t_{ki}^{\text{hue}} \right\|_2^2 \tag{16}$$

Exposure consistency Exposure disagreement is measured in log space. Using the same confidence weight logic as chromaticity but instead using only the intensity,

$$L(c) = 0.2126\, c_R + 0.7152\, c_G + 0.0722\, c_B \qquad \ell_k^* = \frac{\sum_{i \in M_k} w_{ki}\, \log\big(L(c_{ki})\big)}{\sum_{i \in M_k} w_{ki}} \tag{17}$$

The exposure term is therefore

$$\mathcal{L}_{\text{expo}} = \frac{1}{\sum_k |M_k|} \sum_k \sum_{i \in M_k} \Big(\log\big(L(\hat{c}_{ki})\big) - \ell_k^* \Big)^2 \tag{18}$$

Bilateral-grid smoothness Often the correspondence constraints do not uniformly cover every image location or luminance. To propagate reliable corrections into weakly constrained regions while preventing abrupt changes, we impose smoothness directly on the bilateral-grid coefficients. The regularizer is

$$\mathcal{L}_{\text{tv}} = \mathcal{L}_l + \mathcal{L}_v + \mathcal{L}_u \tag{19}$$

3.7 Implementation Details

We parameterize our color correction model as a set of trainable bilateral grids. The bilateral grids are initialized as identity matrix since color consistent multi-view images should remain the same after the transformation. Each view is assigned a grid with resolution $32 \times 16 \times 8$ (horizontal, vertical and luminance respectively). Correspondences are built on the voxel grid with size $s = 0.15\,\text{m}$, and only pixels with valid depth are back-projected. A voxel is considered valid only when it is observed by at least two cameras. We optimize the per-view grids with Adam for 30 epochs with mini batch of 8192 sampled voxel correspondences. The learning rate is set to 0.01. For the loss function, λ_{cons} is set to 1.0, $\lambda_{\text{expo}} = 0.1$ and $\lambda_{\text{tv}} = 10.0$. For saturation preservation, the hyperparameter $\gamma = 2.0$.

4 Experiments

4.1 Dataset

We evaluate on Realsee3D dataset (Li et al., 2026) with a collection of 1000 real-world multi-view scenes. Each scene provides a set of equirectangular sRGB images together with per-view depth maps with size 800×1600, camera poses, and valid masks. Every scene contains a semi color-consistent ground-truth capture and an augmented version in which per-view color, exposure, and white-balance shifts are injected to emulate the inconsistencies produced by a per-image ISP. All methods are given only the augmented views and their geometry. Reported numbers are means over the 1000 scenes, and only the front/back/left/right cube images converted from the panorama image are used. We also evaluate on the Matterport3D dataset (Chang et al., 2017), another large-scale indoor scene dataset. It contains 90 indoor scenes captures, and each scene has RGB images together with depth information and camera poses. However, compared with Realsee3D dataset, this dataset contains many color discrepancies, which make it well suited for our color consistency correction evaluation.

4.2 Baseline

We compare against two methods from the multi-view color correction that can output corrected images. Shin et al. (2025) is a Transformer-based, feed-forward harmonizer that predicts a per-frame 3D bilateral grid correction matrix. Madeira et al. (2024) is a two-stage depth-guided NeRF model that trains a per-view mlp to predict corrected color for each pixel. We also present two variants of our own approach, one operating on plain RGB and one using our multi-objective color consensus. Every method is evaluated on the same augmented views, the two ablation variants of our approach share all geometry and hyperparameters. The augmented uncorrected views serve as the lower bound for each methods.

4.3 Evaluation Metrics

We measure color consistency directly in the shared 3D space using a *Voxel-SSIM* agreement score, and also report the affine aligned PSNR and SSIM scores against GT and a saturation statistic. All metrics are measured in sRGB.

Voxel-SSIM. Following the philosophy of SSIM (Wang et al., 2004). We propose a Voxel SSIM like metrics that measure the level, spread and structure. For every voxel observed by at least two cameras, each contributing camera i is reduced to a per-channel mean $\mu_i \in \mathbb{R}^3$, a per-channel standard deviation $\sigma_i \in \mathbb{R}^3$. A weighted consensus $(\mu_{\text{cons}}, \sigma_{\text{cons}})$ is formed across the cameras of the voxel. Following the three-factor structure of SSIM, each camera is scored against the consensus with

$$l_i = \frac{1}{3}\sum_c \frac{2\mu_i^c \mu_{\text{cons}}^c + c_1}{(\mu_i^c)^2 + (\mu_{\text{cons}}^c)^2 + c_1} \qquad s_i = \frac{1}{3}\sum_c \frac{2\sigma_i^c \sigma_{\text{cons}}^c + c_2}{(\sigma_i^c)^2 + (\sigma_{\text{cons}}^c)^2 + c_2} \tag{20}$$

$$r_i = \tfrac{1}{2}\left(1 + \frac{\langle \mu_i - \bar{\mu}_i,\ \mu_{\text{cons}} - \bar{\mu}_{\text{cons}} \rangle}{\|\mu_i - \bar{\mu}_i\|_2 \, \|\mu_{\text{cons}} - \bar{\mu}_{\text{cons}}\|_2}\right) \tag{21}$$

where $\bar{\mu}$ denotes the average RGB vector. The *level* term l measures per-channel mean agreement (color cast/exposure), the *spread* term s measures per-channel standard-deviation agreement (contrast/texture energy), and the *structure* term r is the Pearson correlation (Pearson, 1895) of the mean-subtracted RGB vectors, i.e. agreement of hue direction independent of brightness.

Affine-aligned PSNR/SSIM. To measure fidelity to the ground truth or original input, we report affine aligned PSNR and SSIM against ground truth following the previous evaluation protocol (Barron et al., 2022). For equirectangular images, these metrics are evaluated on the four horizontal cube faces (front/right/back/left) rendered from each panorama.

4.4 Results

Table 2 reports the mean of each metric over the dataset. All correction methods substantially improve the multi-view consistency compared to the augmented input. Our proposed multi-objective attains the highest Voxel-SSIM (0.915), improving Voxel-SSIM by $+0.113$ over the input and exceeding both the plain-RGB baseline (0.902) and the feedforward methods (0.902) in the Realsee3D dataset. While Neura has a significantly lower score in affine-aligned PSNR and SSIM, even though it achieves the second best Voxel-SSIM.

For affine-aligned PSNR/SSIM evaluation against the semi ground truth. The Chroma achieves the best result. This is expected as Chroma is trained on massive dataset not as a purely color correction model but also an image recovery model. It can recover lost details during color shift or exposure data augmentation, thus making its outputs align more closely with the ground truth. Our multi-objective and plain RGB methods trade a little reference fidelity (≈ 30.4 dB) for stronger multi-view agreement, since maximizing mutual consistency and minimizing distance to a single reference are not the same objective. And Neura shows the opposite behavior from Chroma as it trades the image details for color consistency.

Table 1: Saturation analysis over Realsee3D dataset

Method	Sat
Original input	0.1841
Augmented input	0.3627
RGB baseline	0.2946
Multi-obj	0.3342

For application such as seamless immersive viewing and downstream photometric 3D tasks, mutual consistency is the more relevant criterion, and the proposed objective is best on it. Figure 1 shows a representative cube face rendered from two scenes under each method. However, consistency alone is an insufficient objective, it can be trivially inflated by desaturation, which Neura exhibits. For the two variants of our approach, the saturation comparison Table 1 explains why the decomposed objective is better in practice. Enforcing a single averaged color in plain RGB space drives saturation down significantly. Our multi-objective retains much more saturation while achieving the best consistency, confirming that the chromaticity/exposure and hue/saturation split preserves vivid colors that plain-RGB averaging removes. Meanwhile, Chroma has a reference selection step, in some complex cases, there is no good reference image that can cover all views in a scene, resulting in lost of color contrast.

Table 2: Metrics real-world multi-view scenes for Realsee3D and Matterport3D dataset. Voxel-SSIM and its level, spread, and structure factors measure cross-view color agreement in the shared 3D space (higher is better). Augmented is the uncorrected input. Best in bold black, second best in bold blue.

	Vox-SSIM↑	Level↑	Spread↑	Struct↑	Affine PSNR↑	Affine SSIM↑
Realsee3D						
Ground truth	0.8953	0.978	0.863	0.924	—	—
Augmented (input)	0.8022	0.941	0.858	0.792	—	—
Neura[1]	**0.9073**	**0.988**	**0.881**	**0.963**	**15.33**	**0.7299**
Chroma	0.9016	0.965	0.856	0.951	**31.87**	**0.9629**
RGB baseline (Ours)	0.9019	0.967	**0.863**	0.945	30.37	0.9474
Multi-obj (Ours)	**0.9146**	**0.968**	0.861	**0.973**	**30.38**	**0.9474**
Matterport3D						
Raw input	0.8280	0.936	0.843	0.842	—	—
Chroma	0.8410	**0.943**	**0.853**	0.862	**34.00**	**0.9621**
RGB baseline (Ours)	**0.8624**	0.941	**0.847**	**0.900**	30.25	0.9250
Multi-obj (Ours)	**0.8748**	**0.953**	0.844	**0.940**	**33.47**	**0.9443**



Figure 1: Qualitative comparison on two scenes in Realsee3D. Our multi-objective method removes cross-view color and exposure differences while retaining neutral saturation. Neura has better color consistency scores but has lost a significant amount of details. Chroma shows strong color shift due to the bias introduced by the selected reference frame. While our plain RGB exhibit slight false color.

5 Conclusions and Limitations

We presented a geometry-guided framework for multi-view color consistency correction. By using depth and camera poses, we establish geometric relationships between observations across different views. After voxelization, a carefully designed correspondence construction step is followed to create a robust chromaticity based consensus that is decomposed into chromaticity, hue-preserving saturation, and log-space exposure terms. Then a per-view bilateral grid is trained upon these correspondences to minimize the final corrected observation. Overall, the proposed method achieves near state-of-the-art performance and better result in complex cases without any ground truth supervision and massive datasets.

[1]Due to the poor performance and long per-scene training procedure, we only report the average score of the first 10 scenes as we deem it is a sufficient representation for this method.

Limitations. Our proposed approach requires high quality of depth map and camera poses. Significant depth or pose errors can cause unwanted mixing of observations, resulting in degradation for corrected images. Second, our work is a per-scene optimization, unlike massively trained model that can process hundreds of image inputs in a few seconds. We still need extra time to minimize the color differences online.

Future Work. Although our method needs extra geometry information to work. But it can cooperate with other 3D foundation models, to produce color consistent image set to expand and refine existing dataset. This self-supervised formulation also has the potential to be combined with 3D reconstruction model to further improve not only the quality of the reconstruction but correct the image at the same time in one single step, rather than multi-step correction.

AI USE STATEMENT

We used generative AI tools only as general-purpose assistants for writing and coding such as polishing the wording of the paper, and assisting with general structure coding. Mainly for scripts for loading and batch processing raw data, and write structured logging files. We have not used generative AI to design the method or its objective or any experimental data or results. We have reviewed and manual edited all AI-assisted text, and all AI-assisted code was checked and verified. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Aseem Agarwala. Efficient gradient-domain compositing using quadtrees. *ACM SIGGRAPH 2007 papers*, 2007. URL https://api.semanticscholar.org/CorpusID:519241.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields, 2022. URL https://arxiv.org/abs/2111.12077.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks, 2019. URL https://arxiv.org/abs/1707.05776.

Léon Bottou. Stochastic gradient learning in neural networks. 1991. URL https://api.semanticscholar.org/CorpusID:12410481.

Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. *ACM Trans. Graph.*, 26(3):103–es, July 2007. ISSN 0730-0301. doi: 10.1145/1276377.1276506. URL https://doi.org/10.1145/1276377.1276506.

Dongliang Cheng, Brian Price, Scott Cohen, and Michael S. Brown. Beyond white: Ground truth colors for color constancy correction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 298–306, 2015. doi: 10.1109/ICCV.2015.42.

Ulrich Fecker, Marcus Barkowsky, and AndrÉ Kaup. Histogram-based prefiltering for luminance and chrominance compensation of multiview video. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1258–1267, 2008. doi: 10.1109/TCSVT.2008.926997.

Graham D. Finlayson, Han Gong, and Robert B. Fisher. Color homography color correction, 2016. URL https://arxiv.org/abs/1607.05947.

Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):1–12, 2017. ISSN 1557-7368. doi: 10.1145/3072959.3073592. URL http://dx.doi.org/10.1145/3072959.3073592.

Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Optimizing color consistency in photo collections. *ACM Trans. Graph.*, 32(4), July 2013. ISSN 0730-0301. doi: 10.1145/2461912.2461997. URL https://doi.org/10.1145/2461912.2461997.

Zhonghua Hong, Changyou Xu, Xiaohua Tong, Shijie Liu, Ruyan Zhou, Haiyan Pan, Yun Zhang, Yanling Han, Jing Wang, and Shuhu Yang. Efficient global color, luminance, and contrast consistency optimization for multiple remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:622–637, 2023. doi: 10.1109/JSTARS.2022.3229392.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

Eugene d'Eon Larry Gritz. The importance of being linear, 2026. URL https://developer.nvidia.com/gpugems/gpugems3/part-iv-image-effects/chapter-24-importance-being-linear.

Linyuan Li, Yan Wu, Xi Li, Lingli Wang, Tong Rao, Jie Zhou, Cihui Pan, and Xinchen Hui. Realsee3d: A large-scale multi-view rgb-d dataset of indoor scenes (version 1.1), 2026. URL https://doi.org/10.5281/zenodo.17826242.

Yunmeng Li, Yinxuan Li, Jian Yao, Ye Gong, and Li Li. Global color consistency correction for large-scale images in 3-d reconstruction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3074–3088, 2022. doi: 10.1109/JSTARS.2022.3167264.

Tiago Madeira, Miguel Oliveira, and Paulo Dias. Neural colour correction for indoor 3d reconstruction using rgb-d data. *Sensors (Basel, Switzerland)*, 24, 2024. URL https://api.semanticscholar.org/CorpusID:270784422.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *CoRR*, abs/2008.02268, 2020. URL https://arxiv.org/abs/2008.02268.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020. URL https://arxiv.org/abs/2003.08934.

Pierre Moulon, Bruno Duisit, and Pascal Monasse. Global multiple-view color consistency. *Proceedings of Conference on Visual Media Production*, 11 2013.

Michael Niemeyer, Fabian Manhardt, Marie-Julie Rakotosaona, Michael Oechsle, Christina Tsalicoglou, Keisuke Tateno, Jonathan T. Barron, and Federico Tombari. Learning neural exposure fields for view synthesis, 2025. URL https://arxiv.org/abs/2510.08279.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. ISSN 03701662. URL http://www.jstor.org/stable/115794.

F. Pitie, A.C. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pp. 1434–1439 Vol. 2, 2005. doi: 10.1109/ICCV.2005.166.

Yuntao Qian, Danping Liao, and J. Zhou. Manifold alignment based color transfer for multiview image stitching. *2013 IEEE International Conference on Image Processing*, pp. 1341–1345, 2013. URL https://api.semanticscholar.org/CorpusID:14700634.

Qing Ran, Wenjing Zhao, and Jieqing Feng. Intrinsic color correction for stereo matching. *Computers Graphics*, 82:22–31, 2019. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2019.05.008. URL https://www.sciencedirect.com/science/article/pii/S0097849319300639.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, July/August 2001. doi: 10.1109/38.946629.

Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutiérrez, Aaron Hertzmann, and Belen Masia. Motion parallax for 360° rgbd video. *IEEE Transactions on Visualization and Computer Graphics*, 25:1817–1827, 03 2019. doi: 10.1109/TVCG.2019.2898757.

Jisu Shin, Richard Shaw, Seunghyun Shin, Zhensong Zhang, Hae-Gon Jeon, and Eduardo Perez-Pellitero. Chroma: Consistent harmonization of multi-view appearance via bilateral grid prediction, 2025. URL https://arxiv.org/abs/2507.15748.

Zhuohan Su, Daiguo Deng, Xue Yang, and Xiaonan Luo. Color transfer based on multiscale gradient-aware decomposition and color distribution mapping. *Proceedings of the 20th ACM international conference on Multimedia*, 2012. URL https://api.semanticscholar.org/CorpusID:9537391.

Richard Szeliski, Matthew Uyttendaele, and Drew Steedly. Fast poisson blending using multi-splines. In *2011 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–8, 2011. doi: 10.1109/ICCPHOT.2011.5753119.

Qi-Chong Tian and Laurent D. Cohen. Color correction in image stitching using histogram specification and global mapping. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, 2016. doi: 10.1109/IPTA.2016.7821034.

Nan Wang, Yuantao Chen, Lixing Xiao, Weiqing Xiao, Bohan Li, Zhaoxi Chen, Chongjie Ye, Shaocong Xu, Saining Zhang, Ziyang Yan, Pierre Merriaux, Lei Lei, Tianfan Xue, and Hao Zhao. Unifying appearance codes and bilateral grids for driving scene gaussian splatting, 2026. URL https://arxiv.org/abs/2506.05280.

Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing, 2024. URL https://arxiv.org/abs/2406.00448.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Menghan Xia, Jian Yao, Renping Xie, Mi Zhang, and Jinsheng Xiao. Color consistency correction based on remapping optimization for image stitching. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2977–2984, 2017. doi: 10.1109/ICCVW.2017.351.

Menghan Xia, Jian Yao, and Gao Zhi. A closed-form solution for multi-view color correction with gradient preservation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157:188–200, 09 2019. doi: 10.1016/j.isprsjprs.2019.09.004.

Dinglong Yu, Junxing Yang, Tianjiao Wang, Qiaoye Yu, He Huang, and Fei Deng. Image color consistency correction method based on iterative strategy and quadratic convex optimization. *Sensors and Materials*, 36(12):5427–5443, 2024. doi: 10.18494/SAM5078. URL https://sensors.myu-group.co.jp/article.php?ss=5078.