ChromaDistill: Colorizing Monochrome Radiance Fields with Knowledge Distillation

Ankit Dhiman[1,2] R Srinath[1] Srinjay Sarkar[1] Lokesh R Boregowda[2] R Venkatesh Babu[1]
[1]Vision and AI Lab, IISc Bangalore [2]Samsung R & D Institute India - Bangalore

Abstract

Colorization is a well-explored problem in the domains of image and video processing. However, extending colorization to 3D scenes presents significant challenges. Recent Neural Radiance Field (NeRF) and Gaussian-Splatting (3DGS) methods enable high-quality novel-view synthesis for multi-view images. However, the question arises: How can we colorize these 3D representations? This work presents a method for synthesizing colorized novel views from input grayscale multi-view images. Using image or video colorization methods to colorize novel views from these 3D representations naively will yield output with severe inconsistencies. We introduce a novel method to use powerful image colorization models for colorizing 3D representations. We propose a distillation-based method that transfers color from these networks trained on natural images to the target 3D representation. Notably, this strategy does not add any additional weights or computational overhead to the original representation during inference. Extensive experiments demonstrate that our method produces high-quality colorized views for indoor and outdoor scenes, showcasing significant cross-view consistency advantages over baseline approaches. Our method is agnostic to the underlying 3D representation and easily generalizable to NeRF and 3DGS methods. Further, we validate the efficacy of our approach in several diverse applications: 1.) Infra-Red (IR) multi-view images and 2.) Legacy grayscale multi-view image sequences. Project Webpage: https://val.cds.iisc.ac.in/chroma-distill.github.io/

1. Introduction

Adding color to a monochromatic signal is a longstanding problem [6, 16, 22, 22, 56] in computer vision and graphics. This monochromatic signal can be obtained from special sensors such as infra-red (IR) sensors or legacy content such as old movies. A range of methods have been proposed to colorize images/videos [21, 26, 42]; however, colorization of 3D scenes is challenging as it needs to maintain 3D consistency for realistic colorization. Recent exploration of 3D representations (e.g., NeRF [28] and 3DGS [18]) has enabled effective modeling of complex real-world 3D scenes given multi-view images. Leveraging this, we formulate the problem of colorization of 3D representations given input multi-view grayscale images of a scene. To solve this effectively, we raise the following question: *Can we leverage rich knowledge learned from existing image colorization approaches to colorize these 3D representations?*

This practical setting for colorizing 3D representations has many applications: ***a)*** generating colorized novel views from legacy images/videos, ***b)*** generating colorized novel views from monochromatic signals such as IR and ***c)*** enhancing the performance of discriminative models (e.g., object detection) [50] on monochromatic signals by applying colorization prior to inference. A straightforward approach to colorize a 3D representation involves applying image colorization methods [19, 56] to the input views before training the 3D representation. However, this simplistic method leads to 3D inconsistencies (Fig. 1) across views since each view is independently colorized, resulting in inconsistent color assignments to the same 3D point. Another promising approach is to use video colorization methods on the generated novel-view sequence. This approach ensures temporal consistency (Fig. 1) but fails to guarantee 3D consistent colorization, as it is not grounded in 3D representation.

In the context of 3D colorization, earlier works [36, 49] tried to colorize point clouds. Another direction is to add texture to a predefined mesh [4, 52]. However, these methods are limited to simple synthetic objects. Recent 3D representations (e.g., NeRF, 3DGS) effectively capture high-quality geometry of real-world scenes and can propagate losses from 2D images due to their differentiable nature. This underscores a need for novel techniques to colorize these representations and enable realistic colorization of complex real scenes.

To colorize these 3D representations, a recent method [44] lifts the encoder features of any 2D vision model to the 3D representation. The features are rendered in 2D and passed through the vision model's decoder to generate a consistent novel view and obtain the final RGB image. However, the features are encoded at a very low resolution and may lead

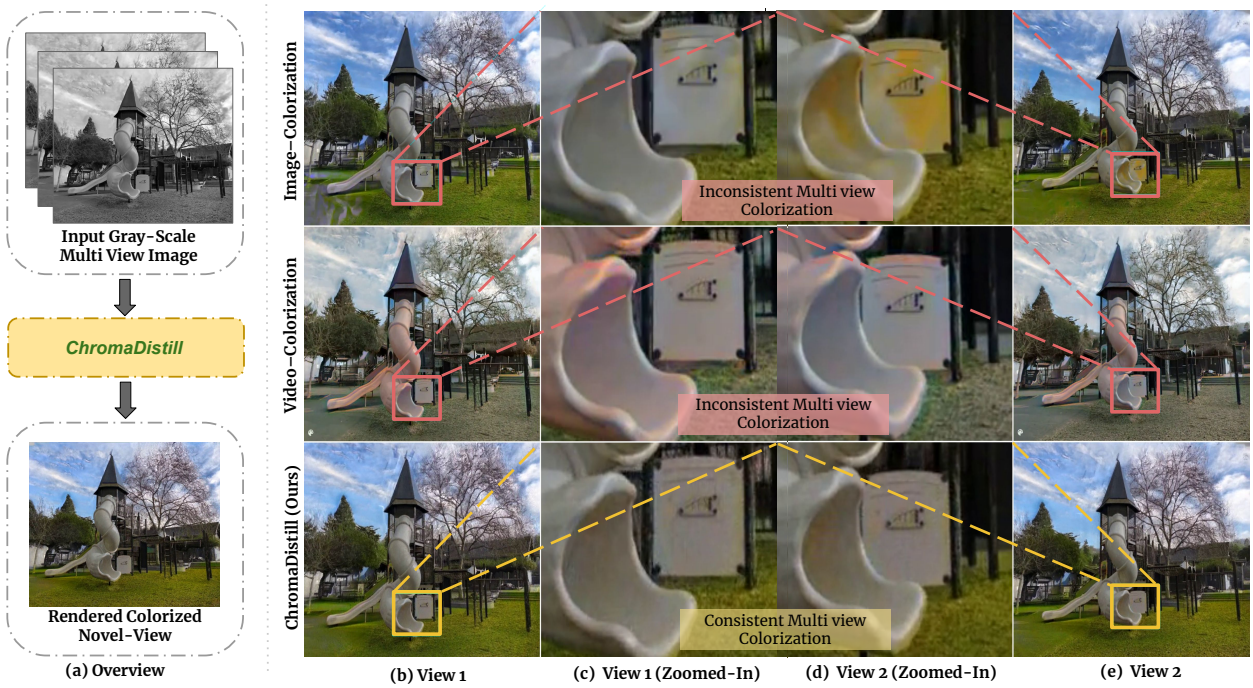

Figure 1. (a) Overview of our method. Given input multi-view gray-scale views, the proposed approach "ChromaDistill" is able to generate colorized views which are 3D consistent. Two colorized novel-views (b) and (e) by I. Image-colorization baseline, II. Video-colorization baseline, and III. our approach on "playground" scene from LLFF [27] dataset. State-of-the-art colorization baselines generate 3D inconsistent novel-views as shown in zoomed-in regions in (c) and (d).

to inconsistent 3D colorization due to independent decoding with the image space decoder. We hypothesize that there are two crucial requirements for high-quality 3D colorization: ***i)*** 3D consistency and ***ii)*** accurate colorization with minimal bleeding artifacts.

In this work, we propose a novel framework for accurate colorization of 3D representations by distilling knowledge from state-of-the-art image colorization methods. We introduce a two-stage process for effective colorization. In the first stage, we train a luminance radiance field to learn the geometry from grayscale images effectively. In the second stage, we freeze the geometry and distill the chroma component from the pre-trained image colorization network. This two-stage training effectively decouples the geometry and colorization of the 3D scenes, leading to high-quality colorization outputs with refined geometry. Notably, this strategy incurs no additional cost for training a separate colorization module for the radiance field networks. Further, we propose a novel *multi-scale self-regularization* technique to mitigate the desaturation (washed-out color) effects when distilling from the colorization network.

We demonstrate the effectiveness of our approach in colorizing both front-facing and unbounded 3D scenes from widely used 3D datasets [20, 27, 47]. Our method significantly outperforms all the baselines in multi-view consistency and realism of colorization. We also compare favorably to state-of-the-art stylization approaches. Further, we show results on two downstream tasks: **1)** Colorizing multi-view infrared (IR) images and **2)** Colorizing legacy grayscale content. Notably, when used for the downstream object detection task, the colorized IR images significantly improve the detection scores. Our primary contributions are:

- We introduce a novel approach *ChromaDistill* for colorizing radiance field networks to produce 3D consistent colorized novel views from input grayscale multi-view images.
- We propose a multi-scale self-regularization to mitigate de-saturation in the distilled color.
- We show that the proposed colorization approach is generalizable to any 3D representation e.g 3DGS [18] and NeRF [28].
- We demonstrate our approach on two real-world applications for novel view synthesis: input multi-view IR images and input grayscale legacy content.

2. Related Work

Image Colorization. One of the earliest deep-learning based methods [16] used a CNN to estimate color for the

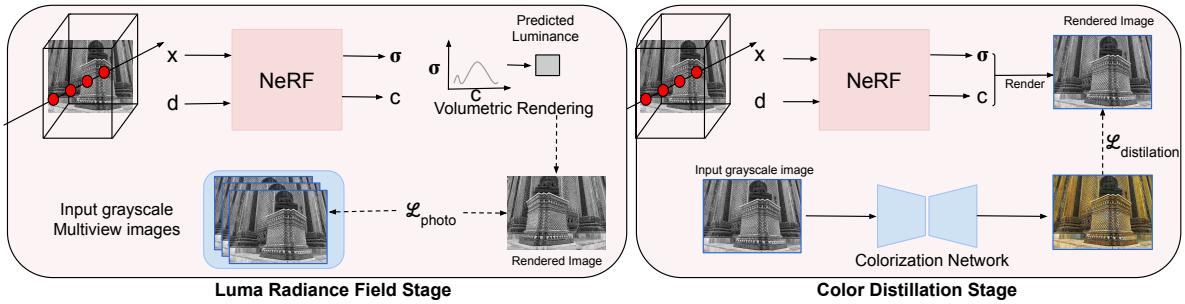

Figure 2. Overall architecture of our method. First, we train a radiance field network from input multi-view grayscale images in the "Luma Radiance Field Stage". Next, we distill knowledge from a teacher colorization network trained on natural images to the radiance field network trained in the previous stage.

grayscale images by jointly learning global and local features. Larsson et al. [22] train the model to predict per-pixel color histograms by leveraging pre-trained networks for high and low-level semantics. Zhang et al. [57] also colorize a grayscale image using a CNN network. GANs have also been used for the image colorization task. [41] uses a generator to produce the chroma component of an image from a given grayscale image, which is conditioned on semantic cues. GAN methods exhibit strong generalization to new images. Recently, diffusion-based methods [7,53] have shown superior performance on this task.

Many methods [9, 16, 22, 56] colorize images only with a grayscale. As there can be multiple plausible colorized images, [8,19,25,48] explores generating diverse colorization. Some of these methods use generative priors for diverse colorization. These methods [37,41,59] use semantic information for better plausible colorization.

Video Colorization. Compared to image colorization, video colorization is more challenging as it has to color an entire sequence while maintaining temporal consistency along with spatial consistency. [23] introduces an automatic approach for video colorization with self-regularization and diversity without using any label data. [54] presents an exemplar-based method that is temporally consistent and remains similar to the reference image. They use a recurrent framework using semantic correspondence and color propagation from the previous step.

3D Representations. NeRF [28] has become a popular choice of 3D representation for novel-view synthesis tasks. Representations like InstantNGP [29], Plenoxels [11], DVGO [38], Mip-NeRF360 [2], Zip-NeRF [3] have enhanced the original NeRF [28] by reducing aliasing, training and rendering time. Recently, Gaussian-Splatting [18], which uses rasterization of splats instead of volumetric rendering, was introduced, accelerating training and achieving real-time rendering. Further, these representations have become popular for solving other tasks such as dynamic scenes [24, 31, 51], hierarchical scenes [10], text-to-3D generation [39,46], and large-scale scenes [33].

Knowledge Distillation. Hinton et al. [13] distilled the soft targets generated by a larger network to a smaller network. Some common approaches include distillation based on the activations of hidden layers in the network [12], distillation based on the intermediate representations generated by the network [1], and distillation using an adversarial loss function to match the distributions of activations and intermediate representations of the two networks [45].

3. Method

3.1. Preliminaries

NeRF. NeRF [28] represents the implicit 3D geometry of a scene by learning a continuous function f whose input is 3D location x and a viewing direction d and outputs are color c and volume density σ, which is parameterized by a multi-layer perceptron (MLP) network. During rendering, a ray r is cast from the camera center along the viewing direction d and is sampled at different intervals. Then, NeRF estimates the color of a pixel by weighted-averaging of the colors of sampled 3D points using volumetric rendering [28]. The MLP is learned by optimizing the squared error between the rendered pixels $f(r)$ and the ground truth pixels $I(p)$ from multiple input views:

$$L_{photo} = ||I(p) - f(r)||_2^2 \quad (1)$$

Hybrid Representations. Recently, hybrid representations like InstantNGP [29], Plenoxels [11], and DVGO [38] have become popular as they use grid-based representation, which is much faster than the traditional NeRF representations. We develop upon Plenoxels [11], which represents a 3D scene with sparse voxel grids, and learn spherical harmonics and density for each voxel grid. Spherical harmonics are estimated for each of the color channels. For any arbitrary 3D location, density and spherical harmonics are trilinearly interpolated from the nearby voxels. Plenoxels also use the photometric loss described in NeRF [28] (Eq. 1). Additionally, they also use total variation (TV) regulariza-

tion on the voxel grid. The final loss is as follows:

$$L_{rendering} = L_{photo} + \lambda_{TV} L_{TV} \quad (2)$$

3.2. Overview

Given a set of multi-view grayscale images of a scene $X = \{X_1, ..., X_n\}$ and corresponding camera poses $P = \{P_1, ..., P_n\}$, we learn a radiance field network f_θ which predicts density σ and color c along a camera ray r. To achieve this, we propose a two-stage learning framework. Even though the input to the radiance field network is multi-view grayscale images, we can still learn the underlying geometry and luminance of the scene. This is the first stage in our pipeline: *"Luma Radiance Field Stage"* which learns the geometry and luminance of the 3D scene. Next, we distill the knowledge from a pre-trained colorization network trained on natural images to the learned radiance field network in the previous stage. This is *"Color Distillation Stage"* in our method. Fig. 2 illustrates the overall pipeline of our method. We discuss *"Luma Radiance Field Stage"* in Section 3.3 and *"Color Distillation Stage"* in Section 3.4.

3.3. Luma Radiance Field Stage

We train a neural radiance field network using Plenoxels [11] f_θ to learn the implicit 3D function of the scene. As our method does not have access to the color image, we take photometric loss w.r.t to the ground-truth grayscale image following Eq. 1. We observe that it has no issues in learning the grayscale images, both qualitatively and quantitatively. (Appendix C.1 in the supplementary material)

3.4. Color Distillation Stage

From the previous stage, we have a trained radiance field f_θ, which has learned the implicit 3D function of the scene but generates grayscale novel views. Directly updating color information in the learned implicit 3D function is not possible. To update the implicit 3D representation, we must compute the loss on the rendered view. Therefore, the optimal strategy for colorizing a radiance field network is to distill knowledge from pre-trained colorization networks trained on a large dataset of natural images.

We propose a color distillation strategy that transfers color details to a 3D scene parameterized by f_θ from any image colorization network $\mathcal{T}$ trained on natural images. More precisely, given a set of multi-view grayscale images of a scene $X = \{X_1, ..., X_n\}$, we pass them through the colorization network $\mathcal{T}$ to obtain a set of colorized images $I^C = \{I_1^C, I_2^C, ..., I_n^C\}$. Corresponding to the camera poses of these images, we obtain rendered images $I^R = \{I_1^R, I_2^R, ..., I_n^R\}$ from f_θ trained in the previous stage on X. We convert both I_i^C and I_i^R to *Lab* color space and distill knowledge from the color network $\mathcal{T}$. Then, our distillation loss, $\mathcal{L}_D$, can be written as :

Algorithm 1: Color Distillation With Multi-Scale Regularization (Appendix B.5 for notations)

Input: Trained NeRF model f_θ on multi-view grayscale images, colorization teacher network $\mathcal{T}$
Output: Colorized NeRF model

function LOOP(for each image i=1,2.....N do)
  $\mathcal{L}_i \leftarrow \phi$
  $I_i^C \leftarrow \mathcal{T}(X_i)$.
  $\mathcal{P}_a \leftarrow \phi$
  $\mathcal{P}_b \leftarrow \phi$
  function LOOP(for each scale s=K,...,1,0 do)
  ${}^sI_i^C \leftarrow downsample(I_i^C, 2^s)$.
  ${}^sI_i^R \leftarrow f_\theta(P_i, s)$
  $\mathcal{L}_i \leftarrow \mathcal{L}_i + \mathcal{L}_{distill}({}^sI_i^C, {}^sI_i^R)$.
    function IF(s != K)
      $\mathcal{L}_i \leftarrow \mathcal{L}_i + ||\mathcal{P}_a - {}^sa_i^R|| + ||\mathcal{P}_b - {}^sb_i^R||$
  $\mathcal{P}_a \leftarrow upsample({}^sa_i^R, 2)$
  $\mathcal{P}_b \leftarrow upsample({}^sb_i^R, 2)$
  Update f_θ

$$\mathcal{L}_D(I_i^C, I_i^R) = ||L_i^C - L_i^R||^2 + ||a_i^C - a_i^R|| + ||b_i^C - b_i^R|| \quad (3)$$

To summarize, we minimize MSE loss between the luma channel and use L1 loss for a and b channels. MSE loss preserves the content of the original grayscale images and L1 loss on the chroma channels distills the color from the colorization network. We briefly summarize this color distillation in Appendix B.4 in the supplementary material.

Multi-scale regularization. Sometimes, colorized views appear to desaturated or washed out. To mitigate this, we introduce multi-scale regularization in the colorized views. In multi-scale regularization, we analyze an image at different scales by constructing image pyramids that correspond to different scales of an image. The lowest level of the pyramid contains the image structure and dominant color, while the finer level, as the name indicates, contains finer features. We create an image pyramid by progressively sub-sampling an image. Then, we start color distillation at the coarsest scale, as discussed in the previous section. For subsequent scales, we regularize the predicted chroma channels with the prediction from the previous scale. We provide details about this regularization in Algorithm 1. $\mathcal{P}_a$ and $\mathcal{P}_b$ are placeholders to keep the interpolated predicted chroma channels from the previous scale. We use bilinear interpolation to upsample the chroma channels. Distilling color from coarsest-to-finest levels ensures prominent colors are learned during optimization, which mitigates the desaturation in the colorized views. We provide ablation in Appendix C.2 in the supplementary material.

Implementation Details We use Plenoxel as our radiance field network representation (Section 3.1). We use the loss

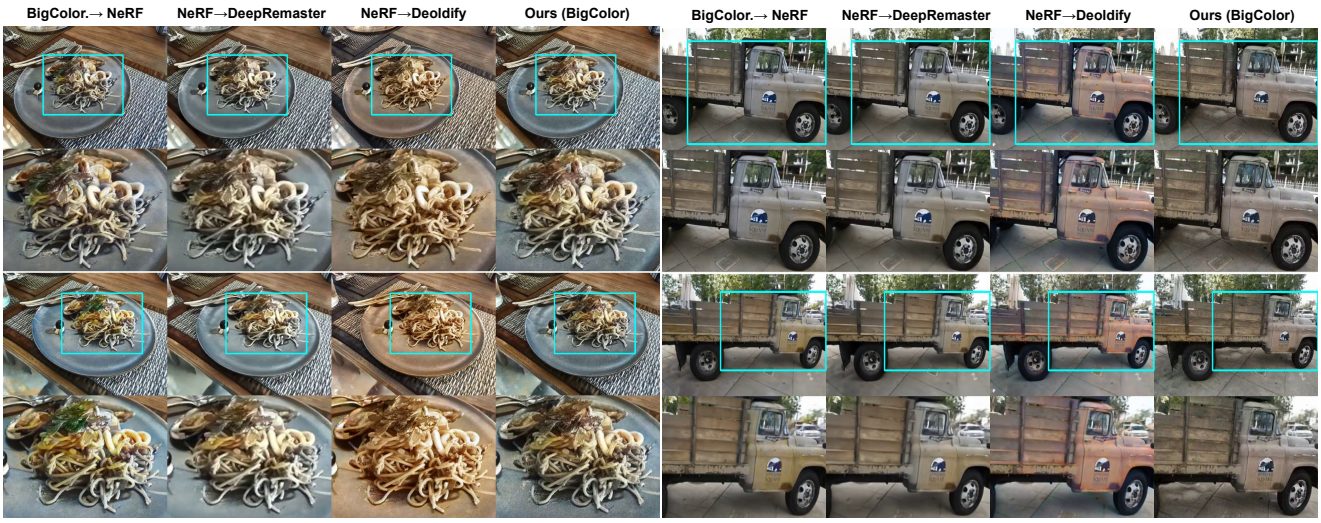

Figure 3. **Qualitative results of our method on baselines for "Pasta" and "Truck" scene.** We display two novel views rendered from different viewpoints, with rows 1 and 3 at the original resolution and rows 2 and 4 zoomed in on the highlighted regions. Even the video-based baselines (columns 2 and 3) exhibit inconsistencies. Note the color change in highlighted regions in "Truck" scene.

described in Eq. 2 for the datasets used in our experiments. During the Color Distillation stage, we estimate the loss in *Lab* color space as described in Eq. 3. We train Color Distillation stage only for 10 epochs.

4. Experiments

This section presents qualitative (Section 4.1) and quantitative (Section 4.2) experiments to evaluate our method. Our method's effectiveness is demonstrated with two image colorization teacher networks [57] and [19]. We compare our approach with two trivial baselines: 1.) colorize input multi-view grayscale images and then train a radiance field network, and 2.) colorize the generated novel-view grayscale image sequence using a video colorization method. To quantitatively evaluate, we use a cross-view consistency metric using a state-of-the-art optical flow network RAFT [40] discussed in [14, 30]. Further, we also compare our method with a concurrent work and show that the stylization methods are unsuitable for the colorization task in 3D.

In addition, we conducted a user study to qualitatively evaluate the colorization results. We also present ablation for multi-scale regularization in Appendix C.2 in the supplementary material. Finally, we show results on two real-world downstream applications - colorization of radiance field networks trained on 1.) Infra-Red (IR) and 2.)In-the-wild grayscale images. Our experiments show that our approach outperforms the baseline methods, producing colorized novel views while maintaining 3D consistency. We encourage readers to watch the supplementary video to better assess our work.

Datasets. We conduct experiments on two types of real-scenes: i) forward-facing real scenes LLFF [27] and Shiny [47] dataset; and ii) 360° unbounded real-scenes Tanks & Temples (TnT) [20] dataset. LLFF [27] dataset provides 24 scenes captured using a handheld cellphone, and each scene has $20-30$ images. Shiny [47] has 8 scenes with multi-view images. Tanks & Temples (TnT) [20] also has 8 scenes that are captured in realistic settings with an industry-quality laser scanner for capturing the ground truth. These datasets have a variety in terms of objects, lighting, and scenarios. For experimentation purposes, we convert the images in the dataset to grayscale.

Baselines. We compare with the following baselines:

1. **Image Colorization → Novel View Synthesis.** : Train Plenoxels [11] on colorized images using image colorization method [19, 56].

2. **Novel View Synthesis → Video Colorization**: Train Plenoxels [11] on grayscale images and obtain colorized novel-views by applying state-video colorization methods [15, 34] to the rendered images.

For baseline 1, we use [57] and [19] for colorizing the input views, thus creating two versions for this baseline. Similarly, for baseline 2, we create two versions using DeepRemaster [15] and DeOldify [34]. Further, we compare our results with stylization works and a contemporary work Color-NeRF [5].

4.1. Qualitative Results

Image Colorization → Novel View Synthesis. We compare our method with both versions of this baseline in Fig. 11. We generate novel views from two different viewpoints to facilitate a better comparison of the 3D consis-

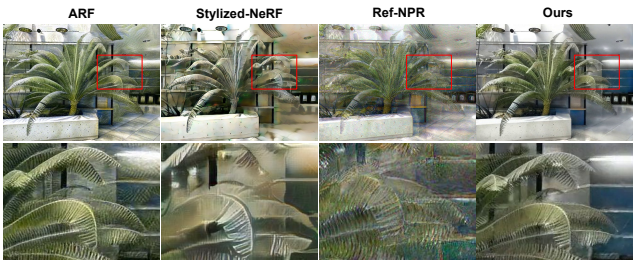

Figure 4. (left-to-right) Results from ARF [55], Stylized-NeRF [14], Ref-NPR [58] and Our method. (Bottom Row) Zoomed-in region of the highlighted region. Check the artifacts from results in stylization works

tency. The baselines exhibit significant color variation in the "Cake" scene, while our method produces results without color variation. Similarly, for "Leaves" and "Pasta" scenes, color variations can be observed in the highlighted region. We also observe similar 3D consistency in the TnT [20] dataset, as shown in Fig. 11 in the bottom two sets. Our method visually demonstrates better 3D consistency in the generated novel views.

Novel View Synthesis → Video Colorization. We compare with the video-colorization-based baseline in Fig. 3 for the "Pasta" scene from LLFF [27] dataset and the "Truck" scene from TnT [20] dataset. The video-based baseline shows better consistency than the image-based baseline but still produces inconsistent colorization. Our method preserves consistency due to explicit modeling in 3D. We can observe a color change in the plate and truck body from Deoldify [34] baseline version. Our method preserves color consistency on the truck body and plate across two views.

Comparison with NeRF-Stylization methods. We also compare our method with NeRF-stylization methods ARF [55], Stylized-NeRF [14] and Ref-NPR [58] by giving a color image as a style image. We observe artifacts in results from these stylization methods in Fig. 4. Stylization involves transferring the overall style of one image to another image or video, focusing on overall texture differences using loss functions like LPIPS. In contrast, colorization emphasizes achieving plausible colors by accurately representing local color values. Therefore, stylization techniques are unsuitable for the colorization of radiance fields.

Comaprison with Color-NeRF [5] We compared our method against a concurrent work Color-NeRF. Fig. 5 shows qualitative results for a forward-facing scene from LLFF. We observe that the novel-views from Color-NeRF are not consistent. Notice the color change on the plate. In comparison, results from our method are consistent. Further, it takes nearly 10 hours end-to-end on RTX A6000 for a forward-facing scene for Color-NeRF. Compared to this, our method takes only 1 hour with Plenoxel backbone and 30 minutes with 3DGS backbone. We produce more comparison results in Appendix C.3.

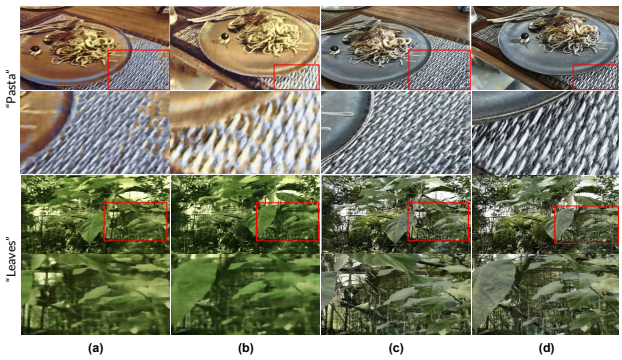

Figure 5. (a) & (b) Novel-views from Color-NeRF [5] and (c) & (d) Novel-views from our method. Bottom row of each scene illustrates zoomed-in regions. Notice the inconsistency in Color-NeRF.

Figure 6. (First column) Grayscale novel-view. Colorized novel-views from our method with Gaussian-Splatting [18] backbone for "Train"(Top) and "Truck"(Bottom) scenes. These results demonstrate that our method maintains multi-view consistency and extends seamlessly to rasterization-based 3D representation.

Table 1. Quantitative results for cross-view short-term and long-term consistency on LLFF dataset.

	Short-Term Consistency ↓				Long-Term Consistency ↓			
	Cake	Pasta	Buddha	Leaves	Cake	Pasta	Buddha	Leaves
BigColor → NeRF	0.037	0.030	0.022	0.015	0.060	0.039	0.033	0.024
NeRF → DeepRemaster	0.018	0.015	0.015	0.015	0.032	0.023	0.023	0.021
NeRF → DeOldify	0.023	0.034	0.017	0.032	0.033	0.049	0.022	0.040
Ours(Colorful Image Colorization)	**0.009**	**0.009**	**0.008**	0.009	**0.013**	**0.017**	**0.012**	0.015
Ours(BigColor)	0.019	0.015	0.015	**0.008**	0.033	0.025	0.023	**0.013**

Results with Gaussian-Splatting [18] backbone Our method can be extended to Gaussian-Splatting representation, which uses rasterization of Gaussian splats for rendering. We provide details for training with Gaussian-Splatting as backbone in Appendix B.6 of the supplementary material. Fig. 6 provides qualitative results on two scenes: "Train" and "Truck" from TnT [20] dataset. Similar to the backbone with Plenoxels, we observe that the colorized novel-views are multi-view consistent.

4.2. Quantitative Results.

Measurement of 3D consistency. To evaluate the 3D consistency across generated novel views, we adopt a strategy proposed by [21], which is also used by various NeRF-based stylization methods [14, 30]. First, we render novel views from the colorized radiance field. We need optical flow and occlusion masks between two views to compute the metric. The occlusion mask, denoted as M, represents occluded regions, out of bounds, or with motion gradients.

Table 2. Quantitative results for cross-view long-term consistency on Tanks & Temples dataset.

Long-Term Consistency ↓	Horse	M60	Train	Truck
Colorful Image Colorization → NeRF	0.018	0.017	0.028	0.035
BigColor→ NeRF	0.022	0.027	0.034	0.038
NeRF → DeepRemaster	**0.017**	0.022	**0.024**	0.032
NeRF → DeOldify	0.032	0.031	0.025	0.031
Ours(Colorful Image Colorization)	0.018	**0.015**	0.026	**0.020**
Ours(BigColor)	0.020	0.021	0.031	0.028

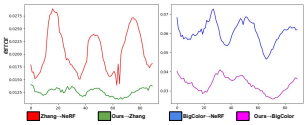

Figure 7. Metrics distribution for (Left) [56] and (Right) BigColor [19] for "cake" scene. We observe that variation from our method has less variance compared to both versions of the image-colorization-based baseline.

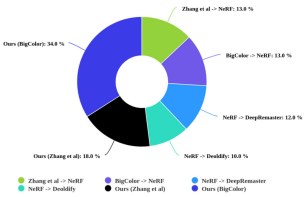

Figure 8. User Study. Our result maintains view consistency after colorization and performs better than the baselines.

We use RAFT [40] to predict the optical flow between two views. Then, we warp a rendered view I_i to obtain a warped view $\hat{I_{i+\Delta}}$; where Δ is the frame-index offset. Consistency error is defined as:

$$E_{\text{consistency}}\left(I_{i+\Delta}, \hat{I}_{i+\Delta}\right) = \frac{1}{|M|}\left\|I_{i+\Delta} - \hat{I}_{i+\Delta}\right\|^2 \quad (4)$$

Similar to [14, 30] we show this metric on short-range and long-range pairs. For color consistency, we measure error only in the chroma channels.

Table 1 and 2 show short-term and long-term consistency metrics for LLFF and TnT datasets, respectively. We observe that our qualitative findings align with these quantitative results. For our method with [56] and BigColor, we observe that short-term and long-term consistency improves for different scenes. For "Pasta" scene in Tab. 1 we see 70% and 56.41% reduction when compared with image-based baseline. We observe significant improvement in metrics when compared with different baselines. Further, our approach generates better cross-view consistent novel-views, regardless of the pre-trained colorization teacher. Additionally, our method produces more consistent novel views than video-based baselines.

Fig. 7 shows the distribution of metrics for the entire rendering sequence for both pre-trained models. The error curve from our method is consistently lower and smoother than the baselines, validating our claim of consistency in novel views obtained from our distillation method.

User Study. We provide users with 12 colorized sequences from LLFF [27], Shiny [47], Shiny Extended [47] and Tanks & Temples (TnT) [20] to compare our method with baseline techniques. Users were asked to select the scene with the best view consistency, vivid color, and no color spill into neighboring regions. We invited 30 participants and asked them to select the best video satisfying these criteria. Fig. 8 shows that our method was preferred 52% of the time.

Figure 9. **Results on In-the-wild grayscale-sequences.** First column represents the input grayscale scene. Columns 2-3 illustrate the colorized novel-view sequence from our method. (Top Row) "Cleveland in 1920s - House". (Bottom Row) " Mountain - Cinematic Video". Our method generates consistent colorized views.

Figure 10. (Column 1) Input multi-view IR Sequence. (Columns 2 and 3) Colorized multi-views from Our method. Our approach yields consistent novel-views for a different input modality.

5. Applications

Multi-View IR images. Our method is highly significant for modalities that do not capture color information. One such popular modality is IR images. For this experiment, we obtain data from [32]. This dataset is generated from a custom rig consisting of IR and multi-spectral (MS) sensor and RGB camera. This dataset contains 16 scenes and 30 views per modality. We show novel views in Fig. 10. We observe that a teacher trained on natural images works well for colorizing the scene. We also discuss the benefits of colorization for the object-detection task in Appendix C.5 in the supplementary material.

In-the-wild grayscale images. We demonstrate our approach's capability to colorize real-world old videos. We extract an image sequence from an old video: "Cleveland in 1920s" and pass them through COLMAP [35] to extract camera poses. Then, we use our framework to generate the color novel views from this grayscale legacy content input. Similarly, we generate novel views for "Mountain" sequence. We can observe in Fig. 9 that our method can get 3D consistent novel views for such sequences. This is useful in the restoration of the old legacy content.

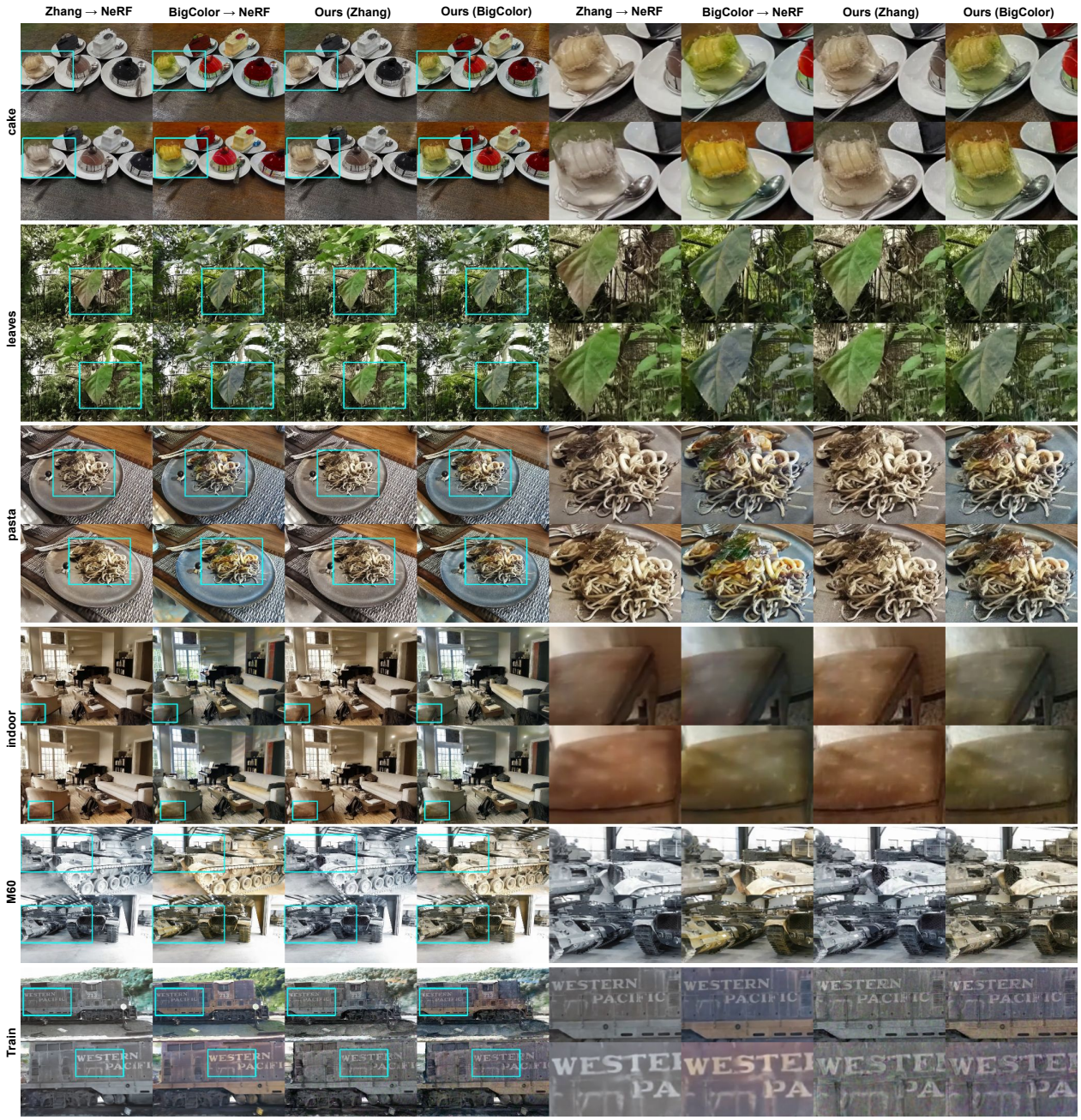

Figure 11. **Qualitative results of our method with image-colorization baselines.** We display two rows of each scene, each rendered from a different viewpoint. The first four columns depict the original resolution results, while the last four columns show zoomed-in regions of the highlighted areas in the first four columns. The image-based baselines have color inconsistencies in their results, whereas our distillation strategy (columns 3, 4, 7, 8) maintains color consistency across different views.

6. Conclusion

We present ChromaDistill, a novel method for colorizing radiance field networks trained on multi-view grayscale images. We use a distillation framework that leverages pre-trained colorization networks on natural images, ensuring superior 3D consistency compared to baseline methods. Multi-scale self-regularization prevents color desaturation during distillation. Our experiments demonstrate robustness to variations in teacher networks. Generated novel views from our method exhibit greater 3D consistency than baselines. Additionally, our method extends seamlessly to rasterization-based representations. A user study showed a preference for our approach. We also demonstrate applications to multi-view IR sensors and legacy image sequences.

Contents

A. Introduction

We present additional results and other details related to our proposed method: ChromaDistill. We present training details in Appendix B.1. We explain the downstream applications in Appendix B.2 and B.3. We present additional experimental results in Appendix C.

B. Implementation Details

B.1. Training Details

We use Plenoxels [11] as neural radiance field representation in our experiments. This representation uses a sparse 3D grid based representation with spherical harmonic (SH) coefficients. For the first stage, luma radiance field, we use the default Plenoxel grid recommended for the type of dataset. We use batch-size of 5000 with RMSProp as optimizer. In the first stage, we use both photometric losses and total-variation (TV) loss proposed in the plenoxels [11]. In the distillation stage, first we get the colorized images from the teacher network. In our experiments, we present result with two image-colorization teachers : 1.) Zhang et al. [56] and 2.) Bigcolor [19]. These colorized images are then used in the distillation stage. When distilling color, we convert the colorized image to "Lab" color space.

Algorithm 2: Color Distillation Algorithm

Input: Trained Nerf Model on Multi-view Grayscale images f_θ, colorization teacher network $\mathcal{T}$

Output: Colorized radiance field network $\mathcal{T}$.

function LOOP(for each image i=1,2.....N do)

$\mathcal{L}_i \leftarrow \phi$

$I_i^C \leftarrow \mathcal{T}(X_i)$.

$I_i^R \leftarrow f_\theta(P_i)$

$\mathcal{L}_i \leftarrow \mathcal{L}_i + \mathcal{L}_{distill}(I_i^C, I_i^R)$

Update f_θ

B.2. Infra-Red Muli-Views

Multi-spectral or Infra-red (IR) sensors are more sensitive to the fine details available in the scene than RGB sensors. Poggi et al. [32] proposed Cross-spectral NeRF (X-NeRF) to model a scene using different spectral sensors. They built a custom rig with a high-resolution RGB camera and two low-resolution IR and MS cameras and captured 16 forward-facing scenes for their experiments. We extracted IR multi-view images and camera poses from the proposed dataset. We naively normalize the IR view between 0 and 1; thus treating it as a grayscale multi-view input sequence. We then apply our method to colorize this view. Our method is effective in colorizing views from different modalities.

B.3. In-the-wild GrayScale Multi-Views

Other than different multi-spectral sensors, there exist lot of in-the-wild grayscale content either in the form of legacy old videos or monochromatic cameras. We extract these multi-view image sequences and then pass these images through COLMAP [35] to extract camera poses. For legacy grayscale image sequences, as there are lot of unnecessary artefacts which affects the perfomrance of COLMAP [35], we pass this sequnce through the video restoration method proposed in [43]. We use the extracted camera-pose and grayscale multi-view image sequence as input for the propsoed method and obtain 3D consistent color-views. This downstream task has a lot of application in Augmented-reality(AR)/Virtual Reality (VR).

B.4. Overview of the Color Distillation Algorithm

Algorithm 2 gives an overview of the color distillation algorithm, For each camera pose, we render a view from the radiance field network trained in grayscale images f_θ. To distill the loss, we colorize the gray-scale teacher using a teacher colorization network

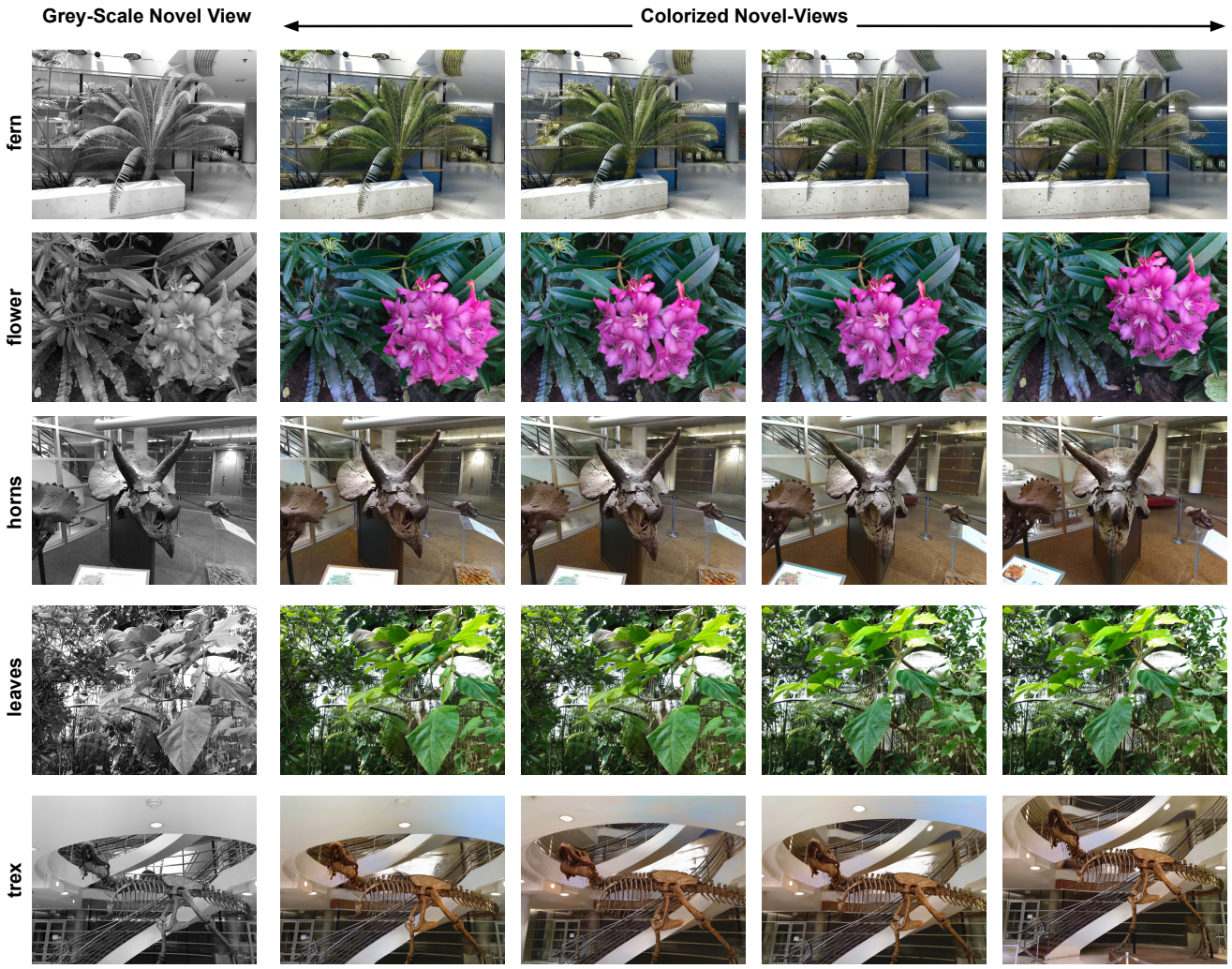

Figure 12. **Colorized Novel-views from Gausssian Splatting 3D representation,** We show results for LLFF scenes : fern, flower, horn, leaves and trex. First column is a grayscale novel-view followed by colorized novel-views using our strategy.

B.5. Notation for "Color Distillation With Multi-Scale Regularization"

- f_θ : NeRF model trained in stage 1 on multi-view grayscale images
- $\mathcal{L}_i$: Loss for i^{th} image in training-set
- $\mathcal{P}_a, \mathcal{P}_a$: Placeholder to save chroma a and b channels from previous scale
- ${}^sI_i^C \leftarrow downsample(I_i^C, 2^s)$: Downsample the image from pre-trained colorization at original resolution by a factor 2^s
- ${}^sI_i^R \leftarrow f_\theta(P_i, s)$: Render an image with the corresponding pose at scale s i.e output width and height be downscaled by a factor 2^s
- $\mathcal{P}_a \leftarrow upsample({}^sa_i^R, 2)$: upsamples the chroma a and b channels for next scale by a factor of 2
- Our method starts from the coarsest scale K i.e image resolution is downscaled by a factor of 2^K

B.6. Colorization Pipeline using Gaussian Splatting [18]

Our proposed knowledge distillation method can be further applied to alternative 3D representations such as Gaussian Splatting [18], which uses rasterization rather than ray-tracing for rendering. We adhere to the default hyperparameters suggested in the original study. Training is conducted only with the luma component up to $15k$ iterations, as Gaussian densification and pruning occur only up to this point. After that, we distill the "a" and "b" channels from

Table 3. Quantitative analysis of GrayScale views

	cake	pasta	buddha	leaves
PSNR	27.772	21.951	23.206	22.146
SSIM	0.855	0.785	0.804	0.784
LPIPS	0.242	0.305	0.347	0.210

Figure 13. Novel views generated from the input grayscale images for *playground* scene in Tanks & Temples [20] dataset.

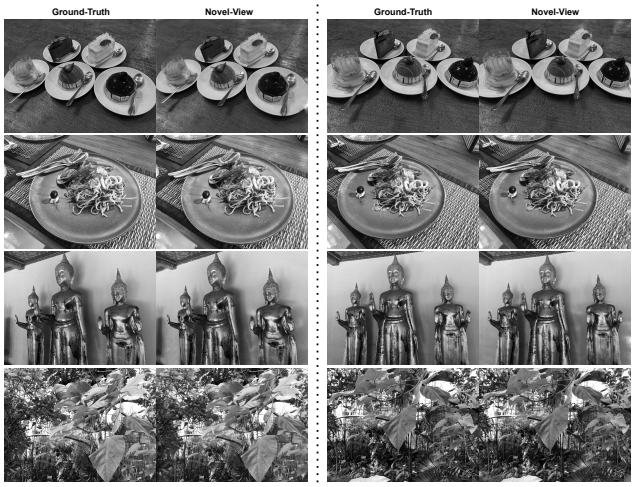

Figure 14. (Top to Bottom) : Comparison of ground-truth and novel-view for grayscale inputs for cake, pasta, buddha and leaves scene.

the teacher colorization network until $30k$ iterations. We present more qualitative results in Fig. 12. Further, we use a different teacher network for colorization model DdColor [17].

C. Experimental Results

C.1. Grayscale Novel Views

We present quantitative results for generated grayscale novel views from "Luma Radiance Field Stage" (Stage 1) in Table 3. We also compare the generated novel-views with the ground-truth grayscale views in Fig. 13 and 14. We observe that generated novel-views are of good quality. This shows that learning monochromatic signal using a radiance field representation is achievable.

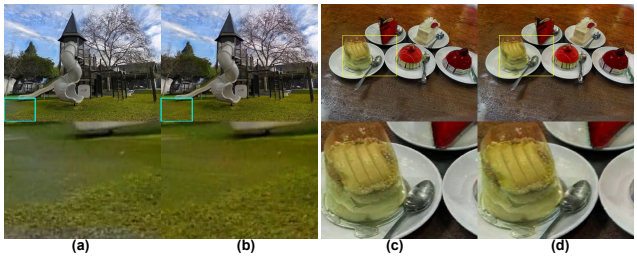

Figure 15. The effect of applying multi-scale regularization on the "playground"((a) and (b)) and "Cake" ((c) and (d)) scene. The highlighted region in the playground (b) and cake (d) had better color in the multi-scale regularization image (than the one w/o multi-scale regularization. Colors in w/o multi-scale regularization are slightly desaturated.

Table 4. Characteristic comparison of Our method with Color-NeRF [5]

Method	Extra Parameters	Inference Speed	Supports other 3D representation
Color-Nerf [5]	Yes	High	No
Ours	No	Low	Yes

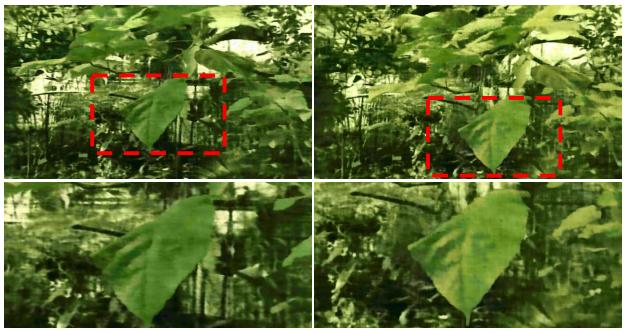

Figure 16. (Top row) Novel-views for "leaves" scene from [5]. (Bottom row) Zoomed in region of the highlighted region. Notice the color change in the leaf.

C.2. Impact of multi-scale regularization

We performed ablation studies on the impact of multi-scale regularization. When distilling color at the original resolution, some areas appeared de-saturated, as seen in the highlighted regions in Fig. 15 (a) & (c). To overcome this issue, we employed multi-scale regularization, which mitigated the color de-saturation during the distillation process. This is evident in the improved color on the grass in playground and on top of the cake, as seen in Fig. 15 (b) & (d). One can observe that a bluish patch is not there with the proposed multi-scale technique. These results demonstrate that our regularization method effectively addresses the color de-saturation problem in the generated views.

C.3. Comparison with Color-NeRF [5]

Color-NeRF [5] is a contemporary work that also solves a similar task. We show additional qualitative results from

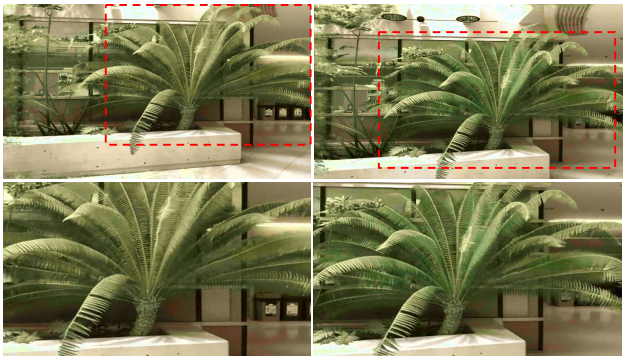

Figure 17. (Top row) Novel-views for "fern" scene from [5]. (Bottom row) Zoomed in region of the highlighted region. Notice that the shade of the fern change from light green to a darker shade of green.

Table 5. Quantitative comparison of Our method with Color-NeRF [5]. Our method outperforms Color-NeRF for cross-view consistency.

		pasta	**fern**
Short-Term Consistency (↓)	Color-NeRF [5]	0.077	0.021
	Ours	**0.009**	**0.010**
Long-Term Consistency (↓)	Color-NeRF [5]	0.129	0.029
	Ours	**0.017**	**0.011**

Color-NeRF in Fig. 16 and 17. We observe that cross-view consistency is not maintained by their method. Further, we compare with the cross-view consistency metrics described in the main paper. Tab. 5 shows that our method performs better short-term and long-term consistency when compared with Color-NeRF. We also draw a comparison of their methodology with ours in Tab. 4. We observe that whereas our method doesnot require any extra parameters to learn color. Color-NeRF requires a separate MLP to learn the color representation. Further, their method is too specific to NeRF architecture. Whereas ours can be easily used with any 3D representation.

C.4. Additional Results

We present additional qualitative results in Fig. 18, Fig. 21 and Fig. 22. We observe that our approach yields 3D consistent color views than the baseline methods. We also present quantitative results in Table 7 and 8. Our method achieves better cross-view consistency compared with the baselines.

Figure 18. Novel views from the "Different Room", "Fern", and "Ninja bike" scenes are shown in the top, middle, and bottom rows, respectively. Note the consistency across views. To better appreciate these results, please refer to the supplementary video.

Table 6. Ablation results show that using the distillation strategy in the "Lab" color space leads to superior cross-view consistency performance across various scenes.

	Cake	**Pasta**	**Three Buddha**	**Leaves**
Ours(RGB)	0.034	0.027	0.023	0.021
Ours(Lab)	**0.033**	**0.025**	**0.023**	**0.019**

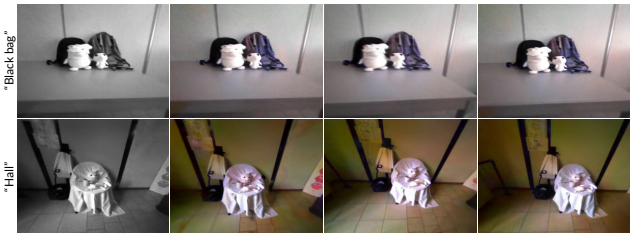

Figure 19. **More IR samples.** (Column 1) Input multi-view IR Sequence. (Columns 2, 3 & 4) Colorized multi-views from Our method. Our approach yields consistent novel-views for a different input modality.

C.5. Demonstration on Downstream task.

We show downstream results in Fig. 20. We observe that objects are consistently detected in the colorized novel-views. This downstream task is very useful to enable downstream tasks such as detection for IR sensors.

C.6. Ablation on color-space

We show ablation on color space in Tab. 6, We clearly see better cross-view consistency achieved with "Lab" color space.

D. Discussion

D.1. Impact of Colorization Teacher Networks.

The proposed method is compatible with any colorization technique. The quality of colorization depends on the selected teacher colorization network. By utilizing BigColor [19] and [56], our approach ensures multi-view consistency irrespective of the chosen teacher network. For instance, in the "cake" scene, [56]. produce dull colors for various objects. In contrast, BigColor generates vivid and sharp colors for different objects. Likewise, we employ DDColor as the teacher colorization network in our infrared experiments.

D.2. Video Colorization Baselines.

Video-colorization methods can generate different colorized outputs for differently rendered trajectories. For example, if we render N videos from N trajectories: $T_1, T_2, ...T_n$ and feed them independently to a video colorization method, this can lead to different outputs even when the same reference images are given. Hence, even though feed-forward video colorization methods can generate temporally consistent views they do not guarantee 3D consistency. Compared to these baselines, our method ensures 3D consistency.

Table 7. Quantitative results for short-term consistency

Scene	BigColor [19] → NeRF	NeRF → DeepRemaster [15]	NeRF → DeOldify [34]	Ours(BigColor [19])
pond	0.022	0.013	0.025	**0.010**
benchflower	0.025	0.013	0.022	**0.010**
chesstable	0.021	0.015	0.022	**0.012**
colorspout	0.025	0.013	0.031	**0.011**
lemontree	0.026	0.015	0.022	**0.014**
stove	0.014	0.010	0.019	**0.008**
piano	0.016	0.010	0.015	**0.009**
redplant	0.029	0.015	0.033	**0.014**
succulents	0.025	0.016	0.027	**0.015**
ninja	0.015	0.011	0.021	**0.007**

Table 8. Quantitative results for long-term consistency

Scene	BigColor [19] → NeRF	NeRF → DeepRemaster [15]	NeRF → DeOldify [34]	Ours(BigColor [19])
pond	0.035	0.017	0.028	**0.015**
benchflower	0.043	0.019	0.030	**0.016**
chesstable	0.033	0.023	0.028	**0.021**
colorspout	0.040	**0.020**	0.051	**0.020**
lemontree	0.041	**0.020**	0.027	0.021
stove	0.018	0.015	0.024	**0.012**
piano	0.026	0.014	0.019	**0.013**
redplant	0.041	0.021	0.041	**0.020**
succulents	0.040	**0.024**	0.032	0.026
ninja	0.021	0.015	0.027	**0.012**

Figure 20. To demonstrate the effectiveness of colorization, we conducted an object detection task on both original infrared (IR) views and their corresponding colorized counterparts. Notably, in (a), no objects are detected in the IR view, and only one out of two objects is detected in (c). However, objects are consistently detected in the colorized views, showcasing the enhanced performance achieved through colorization.

Figure 21. **Qualitative results of our method with baselines.** We display two rows of each scene, each rendered from a different viewpoint. The first four columns depict the original resolution results, while the last four columns show zoomed-in regions of the highlighted areas in the first four columns. The baselines have color inconsistencies in their results, whereas our distillation strategy (columns 4 & 8) maintains color consistency across different views. (Top to bottom) Order of scenes : pond, benchflower, chesstable, colorspout, lemontree

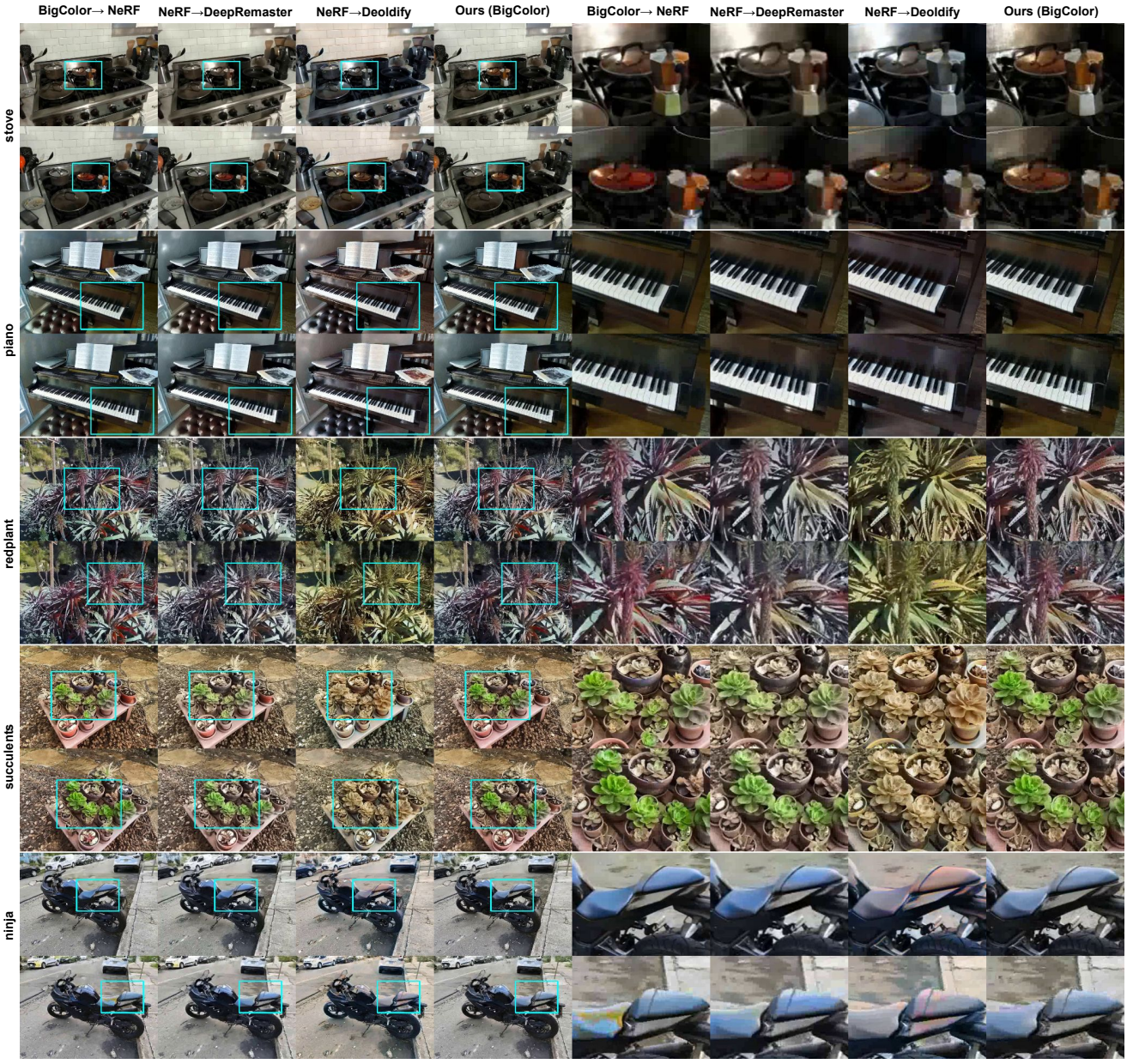

Figure 22. **Qualitative results of our method with baselines.** We display two rows of each scene, each rendered from a different viewpoint. The first four columns depict the original resolution results, while the last four columns show zoomed-in regions of the highlighted areas in the first four columns. The baselines have color inconsistencies in their results, whereas our distillation strategy (columns 4 & 8) maintains color consistency across different views. (Top to bottom) Order of scenes : stove, piano, redplant, succulents, ninja

References

[1] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020. 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields supplemental material. 2023. 3

[4] Alexey Bokhovkin, Shubham Tulsiani, and Angela Dai. Mesh2tex: Generating mesh textures from image queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8918–8928, 2023. 1

[5] Yean Cheng, Renjie Wan, Shuchen Weng, Chengxuan Zhu, Yakun Chang, and Boxin Shi. Colorizing monochromatic radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1317–1325, 2024. 5, 6, 9, 11, 12

[6] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, 2015. 1

[7] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei. Automatic controllable colorization via imagination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2609–2619, 2024. 3

[8] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[9] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE international conference on computer vision*, 2015. 3

[10] Ankit Dhiman, R Srinath, Harsh Rangwani, Rishubh Parihar, Lokesh R Boregowda, Srinath Sridhar, and R Venkatesh Babu. Strata-nerf: Neural radiance fields for stratified scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17603–17614, 2023. 3

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 4, 5, 9

[12] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019. 3

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[14] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5, 6, 7

[15] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6), 2019. 5, 14

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4), 2016. 1, 2, 3

[17] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 328–338, 2023. 11

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 1, 2, 3, 6, 9, 10

[19] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. Bigcolor: Colorization using a generative color prior for natural images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 2022. 1, 3, 5, 7, 9, 13, 14

[20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4), 2017. 2, 5, 6, 7, 11

[21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 1, 6

[22] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016. 1, 3

[23] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 3

[24] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 3

[25] Safa Messaoud, David Forsyth, and Alexander G Schwing. Structural consistency and controllability for diverse colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[26] Simone Meyer, Victor Cornillère, Abdelaziz Djelouah, Christopher Schroers, and Markus Gross. Deep video color propagation. *arXiv preprint arXiv:1808.03232*, 2018. 1

[27] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 2, 5, 6, 7

[28] 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. *Communications of the ACM*, 65(1), 2021. 1, 2, 3

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 3

[30] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 5, 6, 7

[31] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3

[32] Matteo Poggi, Pierluigi Zama Ramirez, Fabio Tosi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Cross-spectral neural radiance fields. In *2022 International Conference on 3D Vision (3DV)*, pages 606–616. IEEE, 2022. 7, 9

[33] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 3

[34] Antoine Salmona, Lucía Bouza, and Julie Delon. Deoldify: A review and implementation of an automatic colorization method. *Image Processing On Line*, 12, 2022. 5, 6, 14

[35] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 7, 9

[36] Takayuki Shinohara, Haoyi Xiu, and Masashi Matsuoka. Point2color: 3d point cloud colorization using a conditional generative network and differentiable rendering for airborne lidar. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1062–1071, 2021. 1

[37] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[38] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 3

[39] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020. 5, 7

[41] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 3

[42] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 1

[43] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. *CVPR*, 2022. 9

[44] Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, Ravi Ramamoorthi, et al. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21367–21377, 2024. 1

[45] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018. 3

[46] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[47] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. *CoRR*, abs/2103.05606, 2021. 2, 5, 7

[48] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 3

[49] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8929–8939, October 2023. 1

[50] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23320–23330, 2023. 1

[51] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 3

[52] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, and Xiaojuan Qi. Texture generation on 3d meshes with point-uv diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4206–4216, 2023. 1

[53] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[54] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 3

[55] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*. Springer, 2022. 6

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016. 1, 3, 5, 7, 9, 13

[57] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 3, 5

[58] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4242–4251, 2023. 6

[59] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *International Journal of Computer Vision*, 128, 2020. 3