

CoRF : Colorizing Radiance Fields using Knowledge Distillation

Ankit Dhiman^{1,2} R Srinath¹ Srinjay Sarkar¹ Lokesh R Boregowda² R Venkatesh Babu¹

¹Vision and AI Lab, IISc Bangalore ²Samsung R & D Institute India - Bangalore

Abstract

Neural radiance field (NeRF) based methods enable high-quality novel-view synthesis for multi-view images. This work presents a method for synthesizing colorized novel views from input grey-scale multi-view images. When we apply image or video-based colorization methods on the generated grey-scale novel views, we observe artifacts due to inconsistency across views. Training a radiance field network on the colorized grey-scale image sequence also does not solve the 3D consistency issue. We propose a distillation based method to transfer color knowledge from the colorization networks trained on natural images to the radiance field network. Specifically, our method uses the radiance field network as a 3D representation and transfers knowledge from existing 2D colorization methods. The experimental results demonstrate that the proposed method produces superior colorized novel views for indoor and outdoor scenes while maintaining cross-view consistency than baselines. Further, we show the efficacy of our method on applications like colorization of radiance field network trained from 1.) Infra-Red (IR) multi-view images and 2.) Old grey-scale multi-view image sequences.

1. Introduction

Colorization is an important and well-studied problem [17, 2, 15, 42] in computer graphics where the objective is to add color to a monochromatic signal. This monochromatic signal can either be obtained from special sensors such as IR sensor or it can be in the form of legacy content. Recently, NeRF-based methods have become popular to generate novel views of a scene while learning the underlying geometry of the 3D scene implicitly using multi-view input images. Our research focuses on a precise scenario: generating colorized novel views in a 3D consistent manner from monochromatic input multi-view images. Fig. 1 illustrates our approach.

Colorization is a well-studied problem in the image [17, 2, 15, 42] and video domain [14, 19, 34]. However, it is not well addressed for the novel view synthesis task. Solving this problem is essential because it requires the radiance field to generate colorized novel views with limited re-

sources i.e., only grey-scale views are available. Colorizing grey-scale multi-view image sequences holds tremendous potential in augmented reality (AR) and virtual reality (VR) applications, especially in restoring legacy content. Also, the proposed approach has applications in other modalities, such as infra-red sensors, which capture shapes and objects in scenes but do not capture color information.

Colorization is an ill-posed problem. Recovering the true color from a grey-scale observation is not trivial. For example, given a grey-scale image of a flower, predicting if the flower is red or blue, or pink is impossible. Hence, given a grey-scale observation, there can be multiple possibilities of color. The objective here is to find a color which looks natural and aesthetically pleasing. Another problem is that the entire image should be colorized consistently maintaining spatial consistency. The color assigned to an object in a scene should not leak into its surrounding. Similarly, the radiance field colorization should be 3D consistent i.e. the color assigned to an object or a region should not change drastically with the change in camera movement. Image and video colorization methods fail to model this aspect during colorization as shown in Fig. 1.

Colorizing monochromatic signals such as black-and-white images have been thoroughly investigated in the literature [17, 2, 42, 15]. Traditional methods solved an objective function to colorize the images using sparse inputs such as scribble [17, 25]. Recently, deep learning methods [10, 45, 2, 42] have been used to solve the colorization task in videos and images and are proven to be very effective. This is because colorization requires a rich understanding of the content of the video, such as the objects, their temporal and spatial relationships, and global temporal context. Deep Learning methods are well-known to have this understanding and learn it for large-scale real-world video datasets.

We can apply image colorization methods to the input grey-scale images and train a radiance field network, but the generated novel views will not be 3D consistent. Similarly, we can apply video colorization methods on the generated novel-view sequence, which may be temporally consistent but does not guarantee 3D consistency as shown in Fig. 1. Another approach is to use generative capability for 3D aware colorized view synthesis using techniques

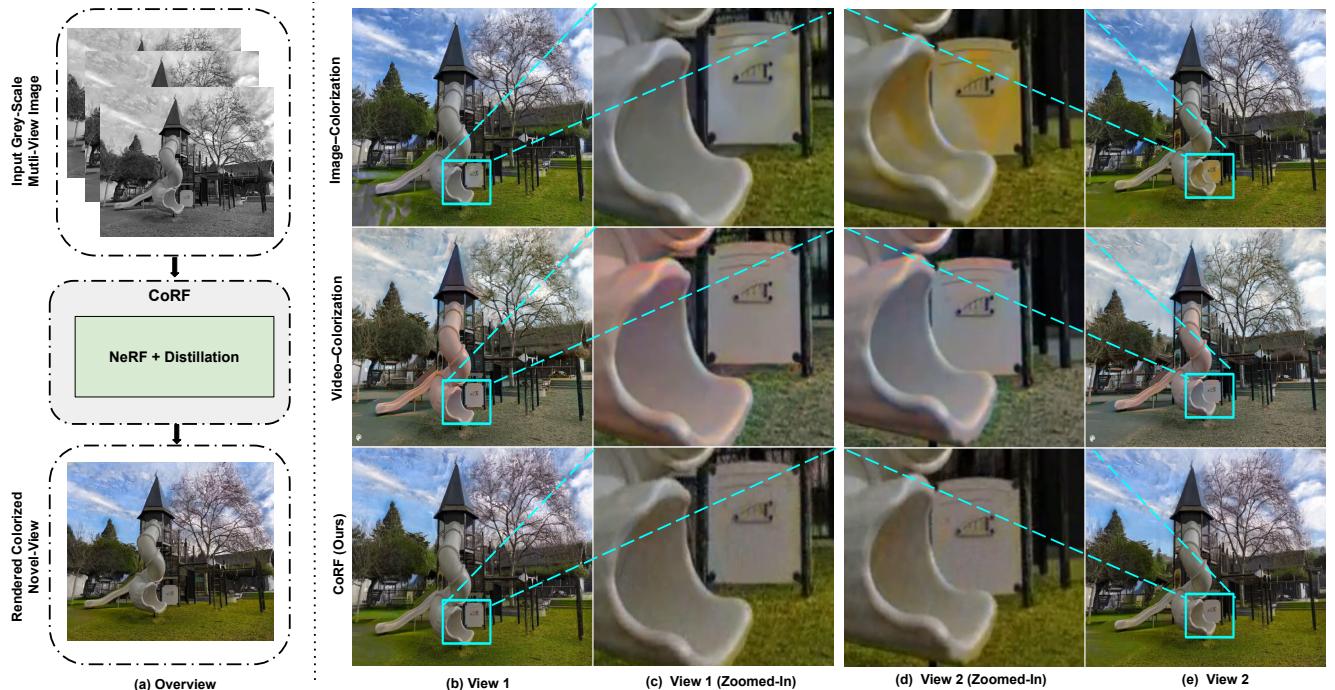


Figure 1. (a) Overview of our metod. Given input multi-view grey-scale views, the proposed approach “CoRF” is able to generate colorized views which are 3D consistent. Two colorized novel-views (b) and (e) by I. Image-colorization baseline, II. Video-colorization baseline, and III. our approach on “playground” scene from LLFF [20] dataset. State-of-the-art colorization baselines generate 3D inconsistent novel-views as shown in zoomed-in regions in (c) and (d).

such as GSN [5], GRAF [29]. These methods suffer from low-quality novel view synthesis and are category specific. Hence, it’s impractical to train these methods on multiple scenes for the colorization task as it loses the capability of generating photo-realistic novel views for a single scene.

We propose a distillation-based method based to leverage the existing deep image colorization methods. This strategy incurs no additional cost for training a separate colorization module for the radiance field networks. We divide our training process into two stages. In stage 1, we train a radiance field network on input grey-scale multi-view images. In stage 2, we distill knowledge from a teacher colorization network into the trained radiance field network in stage 1. We also regularize the model using a multi-scale self-regularization technique to mitigate any spatial color inconsistency. We show the effectiveness of our approach on various grey-scale image sequences generated from the existing datasets such as LLFF [20] and Shiny [37]. We also show results on two downstream tasks: 1.) Colorizing multi-view IR images and 2.) Colorizing In-the-wild grey-scale content. Our main contributions are:

- We propose a novel approach *CoRF* for colorizing radiance field networks to produce 3D consistent colorized novel views from input grey-scale multi-view images.
- We propose a multi-scale self-regularization to reduce

spatial inconsistencies.

- We demonstrate our approach on two real-world applications for novel view synthesis: input multi-view IR images and input grey-scale legacy content.

2. Related Work

Image Colorization. One of the earliest deep-learning-based methods was proposed by [11] which estimates the color of the grey-scale images by jointly learning global and local features through a CNN. [15] trains the model to predict per-pixel color histograms by leveraging pre-trained networks for high and low-level semantics. [43] also colorizes a grey-scale image using a CNN network. GANs have also been used for the image colorization task. [33] uses a generator to produce the chromaticity of an image from a given grey-scale image which is conditioned on semantic cues. GAN methods have good generalization on new images.

Many methods [4, 15, 42, 11] colorize the image automatically i.e. just with a grey-scale input. As there can be multiple plausible colorized images for a grey-scale input, [3, 18, 38, 12] look into generating diverse colorization. Some of these methods use generative priors for diverse colorization. These methods [33, 30, 44] use semantic information for better plausible colorization which is semantically consistent.

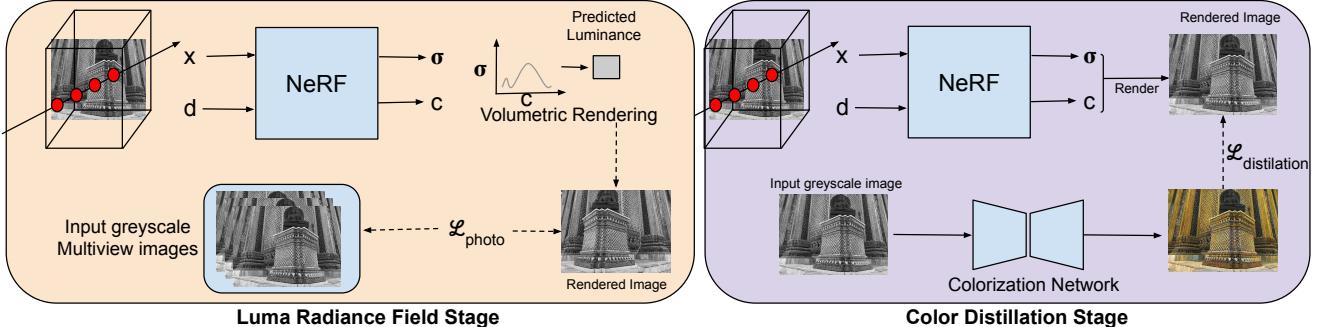


Figure 2. Overall architecture of our method. First, we train a radiance field network from input multi-view grey-scale images in the “Luma Radiance Field Stage”. Next, we distill knowledge from a teacher colorization network trained on natural images to the radiance field network trained in the previous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

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. [16] introduces an automatic approach for video colorization with self-regularization and diversity without using any label data. [39] 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.

Knowledge Distillation. [8] imitated the soft targets generated by a larger network to a smaller network. Since then a lot of work has been done in this area. Some common approaches include distillation based on the activations of hidden layers in the network [7], 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 in the two networks [36]. This knowledge transfer mitigates the problem of large-scale datasets in real-world problems.

3. Method

3.1. Preliminaries

NeRF. NeRF [21] represents the implicit 3D geometry of a scene by learning a continuous function 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 is cast from the camera center along the viewing direction d and is sampled at different intervals. Then, the color of the pixel is determined by performing a weighted average of the color at each of the sampled 3D points using volumetric rendering [21] with f . Finally, the MLP is learned by optimizing the squared error between the rendered pixels and the ground truth pixels from multiple input views:

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

Hybrid Representations. Recently, hybrid representa-

tions like InstantNGP [22], Plenoxels [6], DVGO [31] have become popular as they use grid-based representation which is much faster than the traditional NeRF representations. We develop upon Plenoxels [6] which represents a 3D scene with sparse voxel grids and learns spherical harmonics and density for each of the 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 [21] (Eq. 1). Additionally, they also use total variation (TV) regularization on the voxel grid. Final loss function is described as :

$$L_{tot} = L_{recon} + \lambda_{TV} L_{TV} \quad (2)$$

3.2. Overview

Given a set of multi-view grey-scale images of a scene $X = \{X_1, \dots, X_n\}$ and corresponding camera poses $P = \{P_1, \dots, P_n\}$, 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 grey-scale images, we can still learn the underlying geometry and luminance of the scene. This is “Luma Radiance Field Stage” in our method. Next, we distill the knowledge from a 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 [6] 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 greyscale image following Eq. 1. We show that the radiance field network has no issues in learning the grey-scale images, both qualitatively and quantitatively in Appendix C.1 in the sup-

Algorithm 1: Color Distillation Algorithm

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

Output: Colorized radiance field network

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  $f_\theta$ 
```

plementary 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 grey-scale novel views. However, image colorization is a generative task; which requires a large amount of diverse training images to produce photo-realistic color images. This is difficult to do in the case of radiance field networks because often there are fewer training images per scene. Hence, we strongly believe that the best strategy for colorizing a radiance field network is to distill knowledge from already trained colorization networks trained on a large number 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 grey-scale images of a scene $X = \{X_1, \dots, X_n\}$, we pass them through the colorization network \mathcal{T} to obtain set of colorized images $I^C = \{I_1^C, I_2^C, \dots, I_n^C\}$. Corresponding to the camera poses of these images, we obtain rendered images $I^R = \{I_1^R, I_2^R, \dots, I_n^R\}$ from the radiance field network 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 can be written as :

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

To summarize, we minimize MSE loss between the luma channel and use L1 loss for a and b channels. MSE loss between luma channels preserves the content of the original grey-scale images and L1 loss on the chroma channels distills information from the colorization network.

Multi-scale regularization. As image colorization is done individually on each ground-truth grey-scale image. It often leads to different colorization across multiple views. Hence, we further introduce losses to regularize this inconsistency. In multi-scale regularization, we analyze an image at different scales by constructing image pyramids that cor-

Algorithm 2: Color Distillation With Multi-Scale Regularization

Input: Trained NeRF model f_θ on multi-view greyscale images

Output: Colorized NeRF model

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

```
 $\mathcal{L}_i \leftarrow \phi$ 
 $\mathcal{P}_a \leftarrow \phi$ 
 $\mathcal{P}_b \leftarrow \phi$ 
function LOOP(for each scale s=1,2,...,K do)
 ${}^s I_i^C \leftarrow downsample(I_i^C, s)$ .
 ${}^s I_i^R \leftarrow f_\theta(P_i, s)$ 
 $\mathcal{L}_i \leftarrow \mathcal{L}_i + \mathcal{L}_{distill}({}^s I_i^C, {}^s I_i^R)$ .
function IF(s != K)
 $\mathcal{L}_i \leftarrow \mathcal{L}_i + \|\mathcal{P}_a - {}^s a_i^R\| + \|\mathcal{P}_b - {}^s b_i^R\|$ 
 $\mathcal{P}_a \leftarrow interpolate({}^s a_i^R, 2s)$ 
 $\mathcal{P}_b \leftarrow interpolate({}^s b_i^R, 2s)$ 
Update  $f_\theta$ 
```

respond to different scales of an image. The lowest level of the pyramid contains the image structure and dominant features while the finer level as the name indicates contains finer features like texture, etc. 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 of this algorithm in Algorithm 2. \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.

3.5. Implementation Details

As described in Section 3.3, we use Plenoxel [6] as our radiance field network representation. We use the suggested setting for the datasets used in our experiments. During the Color Distillation stage, we estimate the loss in *Lab* color space. We use the deferred back propagation technique proposed by ARF [40] to backpropagate the loss. In this stage, we train only for 10 epochs.

4. Experiments

In this section, we present quantitative (Section 4.1) and qualitative (Section 4.2) experiments to evaluate our method. Our methods effectiveness was demonstrated with two image colorization teacher networks [43] and [12]. To summarize, our method takes a set of grey-scale posed images of a given scene and learns to generate colorized novel views. We compare our approach with two trivial baselines: 1.) colorize input multi-view grey-scale images



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.

and then train a radiance field network, and 2.) colorize the generated novel-view grey-scale 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 [32] used in SNeRF [23] and Stylized-NeRF [9]. Additionally, we conduct a user study to qualitatively evaluate the colorization results. We also present ablations on the critical design choices in our proposed approach in Appendix C.3 in the supplementary material. Finally, we show the effectiveness of our approach on two real-world downstream applications - colorization of radiance field networks trained on 1.) Infra-Red (IR) and 2.) In-the-wild Grey-Scale images. Our experiments show that our distillation approach outperforms the baseline methods, producing colorized novel views while maintaining 3D consistency. Our distillation strategy can be used to achieve 3D consistent colorization of NeRFs by incorporating advancements in image colorization networks. We encourage readers to watch the supplementary video to assess our work better.

Datasets. We conduct experiments on two types of real-scenes: i) forward-facing real scenes LLFF [20] and Shiny [37] dataset; and ii) 360° unbounded real-scenes Tanks & Temples (TnT) [13] dataset. LLFF [20] dataset provides 24 scenes captured using a handheld cellphone, and each scene has 20 – 30 images. The camera poses are extracted through COLMAP [28]. Shiny [37] has 8 scenes with multi-view images. Tanks & Temples (TnT) [13] also has 8 scenes which 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. The supplementary material contains more details about the dataset. For experimentation

purposes, we convert the images in the dataset to grey scale using a well-known image-format converter. We use the resolution size per the recommended configuration files in Plenoxel [6].

Baselines. We compare CoRF with the following baselines:

1. **Image Colorization → Novel View Synthesis.** : Train Plenoxels [6] on colorized images using state-of-the-art image colorization method [42, 12] on input grey-scale images.
2. **Novel View Synthesis → Video Colorization:** Obtain colorized novel-views by applying state-of-the-art video colorization methods [10, 26] on the novel-view image sequence obtained from the Plenoxel [6] trained on grey-scale multi-view images.

All baselines use the same radiance field representation: Plenoxel [6]. For baseline 1, we use [43] and [12] for colorizing the input views, thus creating two versions for this baseline. Similarly, for baseline 2, we create two versions using DeepRemaster [10] and DeOldify [26]. We did not use image colorization techniques on the rendered grey-scale views because they do not consider temporal and multi-view consistency. Similarly, we did not apply video-colorization techniques to the multi-view grey-scale images because different input views could lead to different sequences for the video-colorization network.

4.1. Qualitative Results

Image Colorization → Novel View Synthesis. We compare our method with both versions of this baseline in

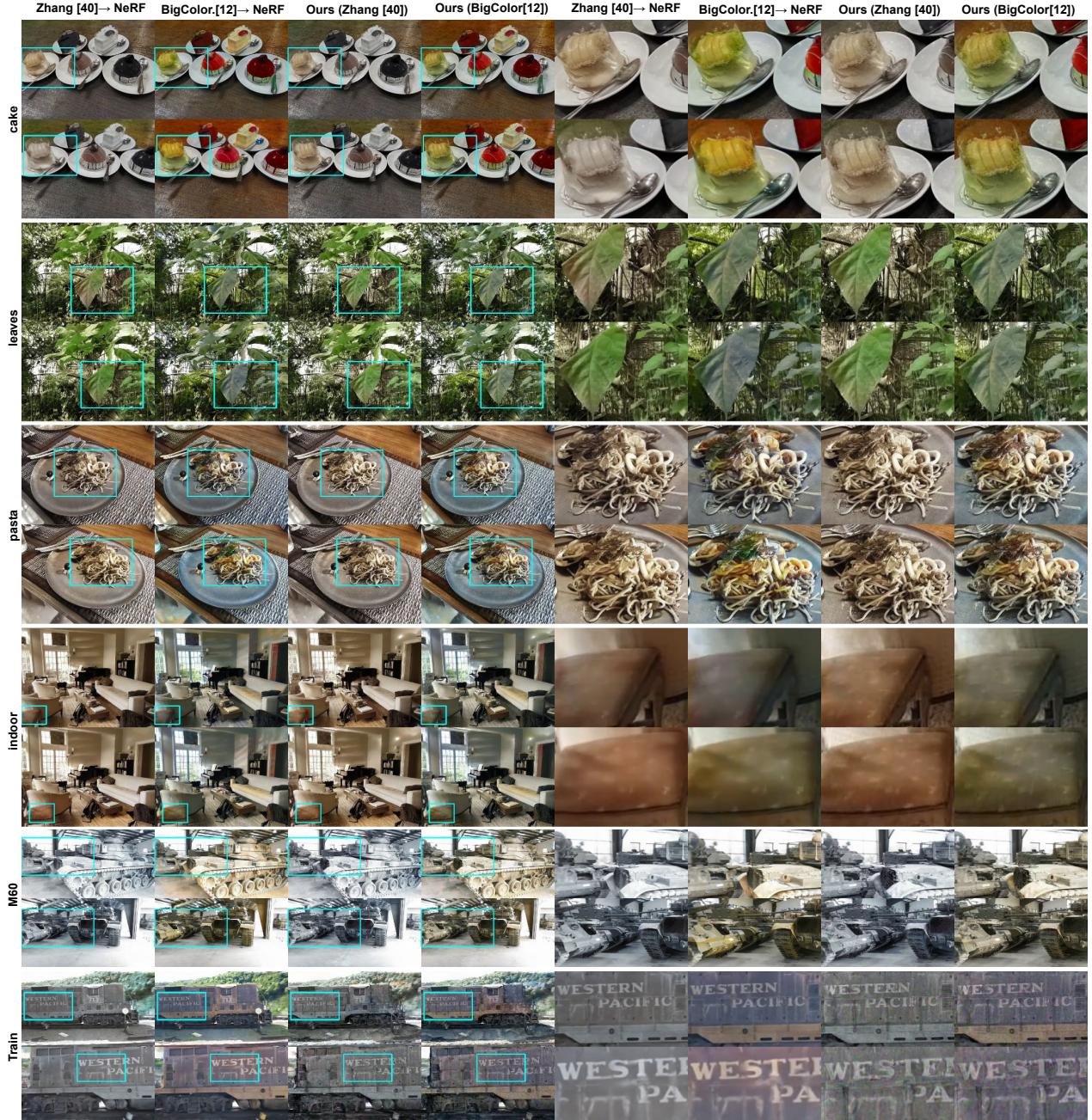


Figure 4. **Qualitative results of our method with image-colorization baselines.** We display two rows of each scene, each rendered from a different viewpoint. The first four columns depict the original resolution results, while the last four columns show zoomed-in regions of the highlighted areas in 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.

Fig. 4. We generate novel views from two different viewpoints to facilitate a better comparison of the 3D consistency. The baselines exhibit color variation in the “Cake” scene, while our strategy produces results without color variation. Similarly, in the “Leaves” and “Pasta” scenes, color variations can be observed in the highlighted leaf and pasta. We also observe similar 3D consistency in the

TnT [13] dataset, as shown in Fig. 4 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 [20] dataset and the “Truck” scene from TnT [13] dataset. The video-based



Figure 5. Results from (a) ARF [41] and (c) Our method. (b) Zoomed-in region of (a) and (d) Zoomed-in region of (c). Check the perceptual artifacts from results in color-ARF.

Table 1. Quantitative results for cross-view short-term and long-term consistency on LLFF [20] dataset.

Short-Term Consistency ↓	Cake	Pasta	Three Buddha	Leaves
[42] → NeRF	0.014	0.014	0.010	0.014
BigColor [12] → NeRF	0.037	0.030	0.022	0.015
NeRF → DeepRemaster [10]	0.018	0.015	0.015	0.015
NeRF → DeOldify [26]	0.023	0.034	0.017	0.032
Ours ([42])	0.009	0.009	0.008	0.009
Ours(BigColor [12])	0.019	0.015	0.015	0.008
Long-Term Consistency ↓				
[42] → NeRF	0.021	0.024	0.015	0.024
BigColor [12] → NeRF	0.060	0.039	0.033	0.024
NeRF → DeepRemaster [10]	0.032	0.023	0.023	0.021
NeRF → DeOldify [26]	0.033	0.049	0.022	0.040
Ours ([42])	0.013	0.017	0.012	0.015
Ours(BigColor [12])	0.033	0.025	0.023	0.013

baseline version results exhibit better consistency than the image-based baseline but still generate inconsistent colorization. Our method preserves consistency due to explicit modeling in 3D. Specifically, we can observe a color change in the plate in the Deoldify [26] baseline version. Similarly, in the “Tanks” scene, we can observe color consistency on the truck body across two views for our method.

Comparison with NeRF-Stylization methods. We also compare our method with a popular NeRF-stylization method ARF [41] by giving a color image as a style image. We show results in Fig. 5 and we clearly observe artifacts in results from ARF. The stylization task involves transferring the overall style of one image to another image or video. For instance, a prominent loss function used in stylization work is LPIPS, which primarily penalizes differences in overall texture rather than local color values. On the other hand, the colorization task prioritizes achieving plausible colors, focusing on accurately representing local color values. Hence, stylization works cannot be utilized for the colorization task for radiance fields.

Novel View Synthesis. We show additional results in Appendix C.2 of the supplementary material. Our method maintains 3D consistency across all views despite challenging lighting conditions and scenes.

4.2. Quantitative Results.

Measurement of 3D consistency. To evaluate the 3D consistency across generated novel views, we adopt a strategy proposed by [14], which is also used by various NeRF-based stylization methods, such as SNeRF [23] and

Table 2. Quantitative results for cross-view long-term consistency on Tanks & Temples [13] dataset.

Long-Term Consistency ↓	Horse	M60	Train	Truck
[42] → NeRF	0.018	0.017	0.028	0.035
BigColor [12] → NeRF	0.022	0.027	0.034	0.038
NeRF → DeepRemaster [10]	0.017	0.022	0.021	0.032
NeRF → DeOldify [26]	0.032	0.031	0.025	0.031
Ours ([42])	0.018	0.015	0.026	0.020
Ours(BigColor [12])	0.020	0.021	0.031	0.028

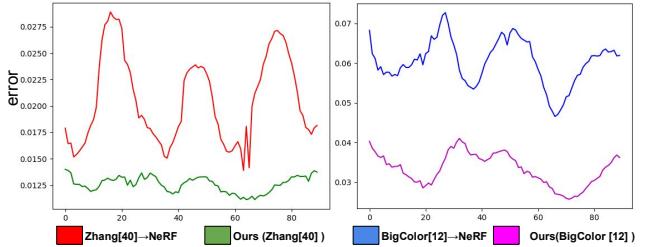


Figure 6. Metrics distribution for (Left) [42] and (Right) Big-Color [12] for “cake” scene. We observe that variation from our method has less variance compared to both versions of the image-colorization-based baseline.

Stylized-NeRF [9]. Firstly, we render novel views from the colorized radiance field. We require optical flow and occlusion masks between two views to compute the metric. We use views from a radiance field network trained on original color images to generate these masks. The occlusion mask, denoted as M represents regions that are occluded, out of bounds, or with motion gradients. We use RAFT [32] 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 :

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

SNeRF [23] and Stylized-NeRF [9] show this metric on short-range pairs and long-range pairs. Note that for color consistency, we measure error only in the chroma channels.

Table 1 and 2 show metrics for short-term and long-term consistency, respectively. The generated novel views in the rendering trajectory have a temporal difference of 10 in short-term consistency and 30 in long-term consistency. Our qualitative findings align with these quantitative results. We compared two versions of image-colorization baselines. For [42], we observe that short-term and long-term consistency improved for different scenes. However, we observed significant improvement in metrics when using Big-Color [12] due to its more colorful views leading to significant variations across multiple perspectives. Nonetheless, 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. For the “Train” scene in Table 2, even though video-based baselines perform better qualitatively our method yields consistent novel-views

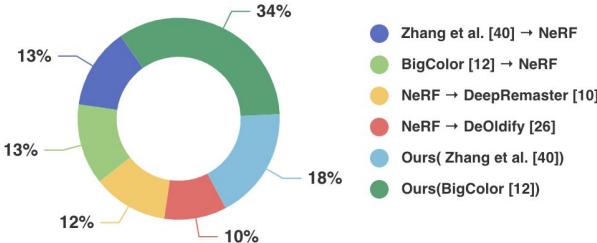


Figure 7. User Study. Our result maintains view consistency after colorization and perform better than the baselines.

Table 3. 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

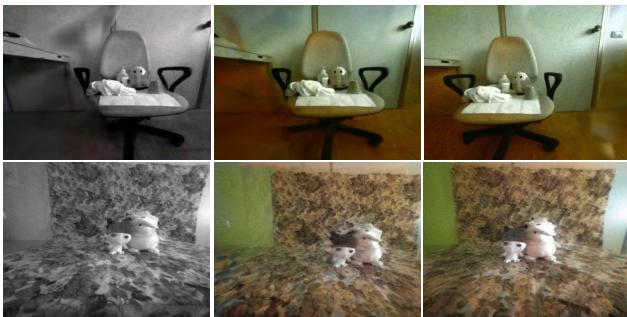


Figure 8. (Column 1) Input multi-view IR Sequence. (Columns 2 and 3) Colorized multi-views from Our method. Our approach yields consistent novel-views for a different input modality.

as shown in Fig. 4. Fig. 6 shows the distribution of metrics for the entire novel-view sequence for both teachers in a scene, with our error curve consistently lower and smoother than the baselines, validating our claim of consistency in novel views obtained from our distillation method.

User Study. To compare our method with baseline techniques, we provided users with 12 colorized sequences from LLFF [20], Shiny [37], Shiny Extended [37] and Tanks & Temples (TnT) [13]. The users were asked to select the scene with the best view consistency and relevant colors without spilling in the neighboring regions. We invited 30 participants and asked them to select the best video satisfying the aforementioned criteria. Fig. 7 shows that the proposed distillation method was preferred 52% of the time indicating the 3D consistency in our method.

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 [24]. This dataset is generated



Figure 9. **Results on In-the-wild grey-scale-sequences.** First column represents the input grey-scale scene. Column 2-4 illustrates 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.

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. 8. We observe that a teacher trained on natural images works well for colorizing the scene. Also, as our approach is invariant with the choice of teacher, we can also use a colorization network that is trained on IR images as a teacher network.

In-the-wild grey-scale images. We show a real-world scenario where our approach can be used to restore old videos by colorization. We extract an image sequence from an old video of “Cleveland in 1920s”. We extracted the frames from the video and pass them through COLMAP [27] to extract camera poses. Then we use our framework to generate the color novel views from this grey-scale legacy content input. Similarly, we generate novel views for “Mountain” sequence. We can observe in Fig. 9 that our method is able to get 3D consistent novel views for such In-the-wild sequences.

6. Conclusion

We present CoRF, a novel method for colorizing radiance field networks trained on input multi-view grey-scale images. A novel distillation framework is proposed, which leverages the pre-trained colorization networks trained on natural images which are more 3D consistent than the baseline methods. We propose a multi-scale self-regularization that prevents de-saturating in color during distillation. Through our experiments, we show that this distillation is invariant of the color teacher network, hence can adapt to advancement in the image colorization domain. Our method outperforms all the baselines both qualitatively and quantitatively. Generated novel views from our approach are more 3D consistent than the baselines. We also conduct a user study in which our method was preferred by the participants. Further, we demonstrate the application of our approach for multi-view IR sensors and legacy image sequences. In future work, we will like to explore real-world applications in more detail.

Appendix

Table of Contents

A Introduction	9
B Implementation Details	9
B.1. Training Details	9
B.2. Infra-Red Multi-Views	9
B.3. In-the-wild Grey-Scale Multi-Views	9
C Experimental Results	9
C.1. Grey-Scale Novel Views	9
C.2. Additional Results	9
C.3. Ablations	10

A. Introduction

We present additional results and other details related to our proposed method : CoRF. 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 [6] 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 [6]. 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.* [42] and 2.) Bigcolor [12]. These colorized images are then used in the distillation stage. When distilling color, we convert the colorized image to “Lab” color space.

B.2. Infra-Red Multi-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.* [24] 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

Table 4. Quantitative analysis of Grey-Scale 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 10. Novel views generated from the input grey-scale images for *playground* scene in Tanks & Temples [13] dataset.

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 grey-scale 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 Grey-Scale Multi-Views

Other than different multi-spectral sensors, there exist lot of in-the-wild grey-scale 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 [27] to extract camera poses. For legacy grey-scale image sequences, as there are lot of unnecessary artefacts which affects the performance of COLMAP [27], we pass this sequence through the video restoration method proposed in [35]. We use the extracted camera-pose and grey-scale multi-view image sequence as input for the proposed method and obtain 3D consistent color-views. This downstream task has a lot of application in Augmented-reality(AR)/Virtual Reality (VR).

C. Experimental Results

C.1. Grey-Scale Novel Views

We present quantitative results for generated grey-scale novel views from “Luma Radiance Field Stage” (Stage 1) in Table 4. We also compare the generated novel-views with the ground-truth grey-scale views in Fig. 10 and 11. We observe that generated novel-views are of good quality. This shows that learning monochromatic signal using a radiance field representation is achievable.

C.2. Additional Results

We present additional qualitative results in Fig. 12, Fig. 14 and Fig. 15. We observe that our approach yields 3D consistent color views than the baseline methods. We also present quantitative results in Table 5 and 6. Our method achieves better cross-view consistency compared with the baselines.

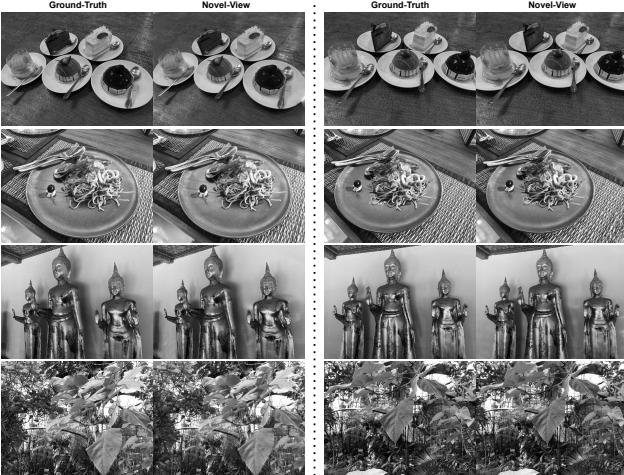


Figure 11. (Top to Bottom) : Comparison of ground-truth and novel-view for grey-scale inputs for cake, pasta, buddha and leaves scene.



Figure 12. 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.

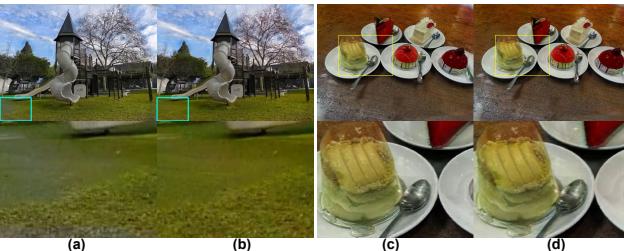


Figure 13. The effect of applying multi-scale regularization on the “playground”((a) and (b)) and “Cake” ((c) and (d)) scene. The highlighted region in the playground (b) and cake (d) had better color in the multi-scale regularization image (than the one w/o multi-scale regularization). Colors in w/o multi-scale regularization are slightly desaturated.

C.3. Ablations

We performed ablation studies on the choice of color space and the impact of multi-scale regularization. How-

Table 5. Quantitative results for short-term consistency

Scene	BigColor [12] → NeRF	NeRF → DeepRemaster [10]	NeRF → DeOldify [26]	Ours(BigColor [12])
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
colorsput	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 6. Quantitative results for long-term consistency

Scene	BigColor [12] → NeRF	NeRF → DeepRemaster [10]	NeRF → DeOldify [26]	Ours(BigColor [12])
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
colorsput	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

ever, when distilling color at the original resolution, some areas appeared de-saturated, as seen in the highlighted regions in Fig. 13(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. 13(b) & 8(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.

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] Zehzhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, 2015. 1
- [3] 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. 2
- [4] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE international conference on computer vision*, 2015. 2
- [5] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [6] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinzhong 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
- [7] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation



Figure 14. 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, colorsput, lemontree

- boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019. 3
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [9] 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, 7

- [10] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6), 2019. 1, 5, 7, 10
- [11] 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 simulta-

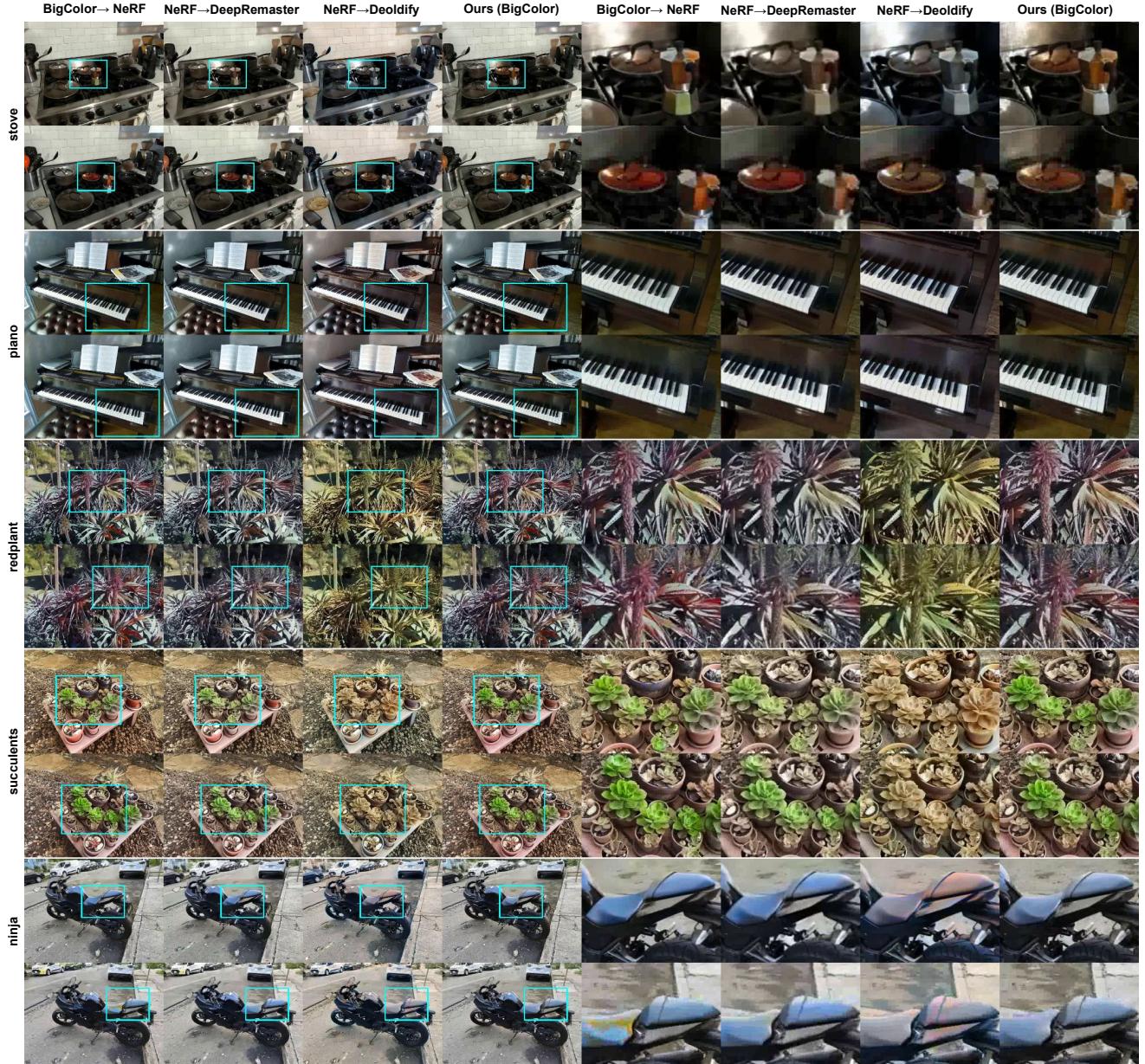


Figure 15. **Qualitative results of our method with baselines.** We display two rows of each scene, each rendered from a different viewpoint. The first four columns depict the original resolution results, while the last four columns show zoomed-in regions of the highlighted areas in the first four columns. 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

- neous classification. *ACM Transactions on Graphics (ToG)*, 35(4), 2016. 2
- [12] 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. 2, 4, 5, 7, 9, 10
- [13] 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. 5, 6, 7, 8, 9
- [14] 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, 7
- [15] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th Eu-*

- ropean Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, 2016. 1, 2
- [16] 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
- [17] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization, 2004. 1
- [18] 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. 2
- [19] Simone Meyer, Victor Cornillère, Abdelaziz Djelouah, Christopher Schroers, and Markus Gross. Deep video color propagation. *arXiv preprint arXiv:1808.03232*, 2018. 1
- [20] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortíz-Cayón, 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, 8
- [21] 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. 3
- [22] 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
- [23] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 5, 7
- [24] Matteo Poggi, Pierluigi Zama Ramirez, Fabio Tosi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Cross-spectral neural radiance fields. *arXiv preprint arXiv:2209.00648*, 2022. 8, 9
- [25] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM Transactions on Graphics (ToG)*, 25(3), 2006. 1
- [26] Antoine Salmona, Lucía Bouza, and Julie Delon. Deoldify: A review and implementation of an automatic colorization method. *Image Processing On Line*, 12, 2022. 5, 7, 10
- [27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 8, 9
- [28] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5
- [29] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33, 2020. 2
- [30] 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. 2
- [31] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 3
- [32] 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
- [33] 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. 2
- [34] 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
- [35] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. *CVPR*, 2022. 9
- [36] 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
- [37] Suttsak Wizadwongsu, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. *CoRR*, abs/2103.05606, 2021. 2, 5, 8
- [38] 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. 2
- [39] 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
- [40] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields, 2022. 4
- [41] 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. 7
- [42] 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, 2, 5, 7, 9
- [43] 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. 2, 4, 5
- [44] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *International Journal of Computer Vision*, 128, 2020. 2
- [45] Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou. Vegan: Video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia*, 2022. 1