

# ArCSEM: Artistic Colorization of SEM Images via Gaussian Splatting

Takuma Nishimura<sup>1</sup>, Andreea Dogaru<sup>1</sup>,  
 Martin Oeggerli<sup>2</sup>, and Bernhard Egger<sup>1</sup>

<sup>1</sup> Friedrich-Alexander-Universität Erlangen-Nürnberg  
 {takuma.nishimura, andreea.dogaru, bernhard.egger}@fau.de  
<sup>2</sup> Micronaut  
 info@micronaut.ch

**Abstract.** Scanning Electron Microscopes (SEMs) are widely renowned for their ability to analyze the surface structures of microscopic objects, offering the capability to capture highly detailed, yet only grayscale, images. To create more expressive and realistic illustrations, these images are typically manually colorized by an artist with the support of image editing software. This task becomes highly laborious when multiple images of a scanned object require colorization. We propose facilitating this process by using the underlying 3D structure of the microscopic scene to propagate the color information to all the captured images, from as little as one colorized view. We explore several scene representation techniques and achieve high-quality colorized novel view synthesis of a SEM scene. In contrast to prior work, there is no manual intervention or labelling involved in obtaining the 3D representation. This enables an artist to color a single or few views of a sequence and automatically retrieve a fully colored scene or video. Project page: <https://ronly2460.github.io/ArCSEM>

**Keywords:** Artistic Colorization · Novel View Synthesis · Scanning Electron Microscope · human AI co-creation

## 1 Introduction

Throughout history, different art forms have been used to express ones' creative perspective and share it with the world. As the society and technology evolve, artistic endeavours reach beyond set boundaries in the exploration of imaginative challenges. Visual arts in particular are being revitalized by novel approaches at the intersection of computer graphics, computer vision and human creativity, where artistic expression entangles with academic research.

Following the advancement of deep learning-based methods, applications such as neural style transfer [17, 25, 27, 70], image colorization [8, 22, 67], text-prompted image generation [46, 65] and editing [3, 23] have become easily-accessible for both professional artists and curious creators. Though these methods are capable of impressive results, they are limited to the 2D domain. Another research area in

computer vision that has seen tremendous improvements in the recent years is optimizing 3D scenes from multi-view inputs. Either with the goal of synthesising novel views or reconstructing the underlying 3D surface, the scene representation plays an important role in achieving high-quality results. Popular choices are point clouds [13, 43, 47], neural fields [1, 40, 57], voxels [14], hybrid representations [11, 37], and more recently Gaussian splats [24, 28].

These recent advancements have dramatically expanded artistic expression into three-dimensional space, allowing us to freely change textures [4, 6, 61], geometries [2, 62], illumination [15, 48], and introduce new objects into scenes without disrupting the environment [19, 52]. This progress offers artists unprecedented creative freedom. Our research aims to extend these advanced 3D creation techniques into the microscopic world.

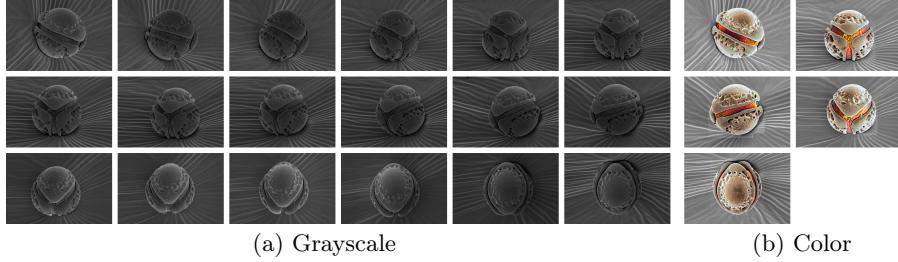
To explore this, we use images captured by a Scanning Electron Microscope (SEM). SEMs are used for the examination and analysis of nanoscale structures. An electron gun generates a beam that thoroughly scans the surface. As the beam interacts with the surface, it emits signals. The microscope’s detectors capture these emitted electrons. The quantity of electrons detected from each point is then converted into corresponding pixel values, resulting in a high-resolution grayscale image that reveals the intricate surface structure. SEM images share similarities with optical images, exhibiting diffuse and specular reflectance and effects similar to optical shadowing. The fundamental distinction lies in the particle flow: in SEM imaging, the particles travel in the opposite direction compared to optical imaging.

We experiment with multi-view grayscale images of a pollen granule captured by tilting the sample while keeping the microscope fixed. However, tilting alters the incident angle between the electron beam and the surface, which causes the emitted electrons to vary across regions of the sample. This induces view-dependent variations in electron emission and scattering, which are perceived as illumination changes in the final SEM images.

Leveraging our captured grayscale dataset, we achieve novel view synthesis (NVS) via a precise 3D representation of the pollen modeled with 2D Gaussian Splatting (2DGS) [24]. To address the aforementioned illumination variations, we apply an image specific affine color transformation (ACT) to the Gaussians, as proposed by [10]. Moreover, based on the 3D representation, we further introduce colors into the scene, guided by artistic intuition. In addition to our grayscale dataset, we incorporate up to five color images created by a professional artist, Martin Oeggerli. These color images are then used for the colorization of the scene, by adapting ideas from [69] to propagate the color information via pseudo-colors and semantic correspondences across views. We showcase the capabilities of our method, ArCSEM, by generating expressive colorized novel views of an SEM scene with artistic guidance.

The key contributions of our work are as follows:

- We obtain a precise and intricate 3D representation of a pollen captured by SEM, enabling novel view synthesis.



**Fig. 1:** Our dataset. (a) A subset of 18 out of 32 grayscale images, arranged left to right in the first two rows, and front to top in the bottom row. (b) All manually colored images shown in the following order: leftmost, center, rightmost, angled, and top view

- We achieve 3D colorization of the grayscale 3D scene using a limited number of manually colored images, enabling colored novel view synthesis.
- We demonstrate the effectiveness of the proposed approach compared to previous methods through comprehensive experiments.

## 2 Related work

**Colorization.** Colorization techniques can be broadly categorized into two main approaches: statistical methods and semantic methods. Statistical methods, pioneered by [45], utilize color statistics of images, focusing on correcting overall color distributions between source and target images. While computationally efficient, these methods often lack semantic understanding, potentially leading to contextually inappropriate results. The advent of deep learning has transformed the field [8], making semantic methods more prevalent. These methods consider content-aware correspondences and leverage architectures based on Convolutional Neural Networks (CNNs) [22, 67] or Transformers [32] to extract semantic features and produce more contextually coherent colorization. Beyond 2D images, 3D colorization methods have also been developed for meshes [66], point clouds [16, 34], and voxels [60]. The few works [20, 55] that consider the colorization of SEM images are limited to the 2D domain and work without specific color guidance. In contrast, we focus on artistic 3D colorization of SEM images, using specific color inputs, which allows artists to guide and control the process.

**3D SEM scenes.** Only one line of works [72, 73] has focused on 3D shape reconstruction of complex objects from SEM images, particularly of a cat flea. However, these approaches employ traditional photogrammetry and complex computer graphics techniques, which demand extensive mathematical calculations and laborious work. In contrast, our pipeline provides a much simpler method for representing 3D scenes, it is straightforward and requires no customization to model grayscale scenes.

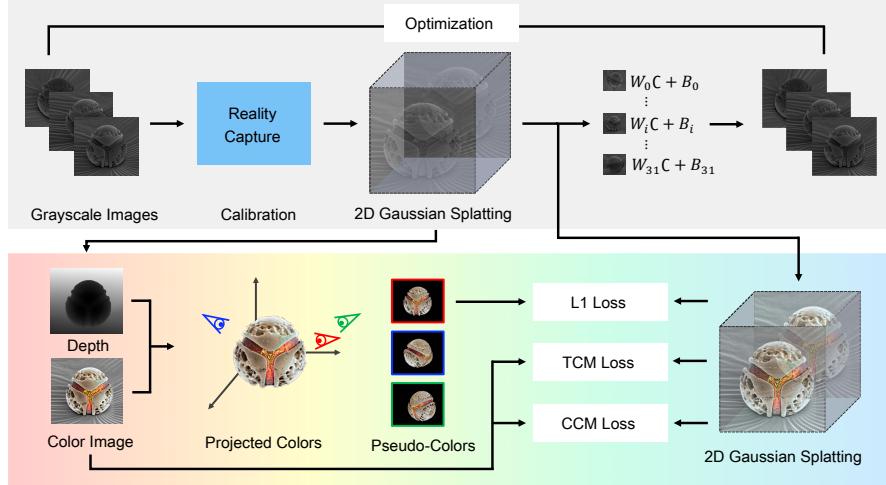
**Appearance editing in NeRF and Gaussian splats.** Unlike 2D editing, 3D editing requires maintaining both geometric and appearance consistency across viewpoints while ensuring natural-looking edits. NVS techniques, such as Neural Radiance Fields (NeRF) [40] and its variants [1, 5, 14], have enabled the photorealistic rendering of arbitrary views. The versatility of these 3D representations enabled the development of various methods for appearance editing with diverse control modalities. Image-guided approaches [9, 26, 35, 41, 64, 69] leverage visual references to control the editing process. Text-based methods [12, 21, 29, 53, 56, 58, 71] employ natural language descriptions to manipulate scene appearances. Other methods [18, 31, 33, 39, 42] allow for manual color specification or tool-based editing. Recently, Gaussian Splatting [24, 28] has emerged as a breakthrough technique, representing scenes explicitly with numerous 3D Gaussians primitives. Providing real-time rendering and competitive quality, the framework has already enabled several editing techniques [7, 36, 63]. Our proposed method belongs to the latter line of works, but, in contrast to existing approaches that mainly focus on stylization or color replacement using extracted color palettes, we build on Ref-NPR [69] which is more suitable for our grayscale-to-color setting.

### 3 Method

Our proposed method, illustrated in Fig. 2, employs a two-stage training process: grayscale 3D scene optimization and colorization. We start with a grayscale scene representation by fitting 2DGS on the SEM image dataset calibrated with RealityCapture [44]. To handle varying illumination, we apply an affine color transformation to the decoded color, using image-specific weights and biases. In the second stage, we use the grayscale model to generate depth maps and project the artist-provided colors into 3D space. For views without color data, we use a nearest-neighbor search to obtain pseudo-colors. Finally, we fine-tune the initial grayscale model using the color images and the computed pseudo-colors by keeping the geometry fixed and optimizing the spherical harmonics coefficients for all degrees using losses inspired from Ref-NPR [69].

#### 3.1 Grayscale Training

**Calibration.** Our grayscale dataset is acquired using SEM. Although SEM utilizes parallel electron beams, resulting in orthographic projection, we approximate it using perspective projection to ensure compatibility with existing NVS methods. Under this setting, an exceptionally large focal length of approximately 50,000 pixels is estimated, with the cameras positioned very far from the scene content. To this end, we experimented with COLMAP [50, 51] and more advanced feature extraction and matching methods, including SuperGlue [49] and LoFTR [54], but found them to struggle with the peculiar images produced by SEM. Instead, we obtained satisfying results using RealityCapture [44] with shared intrinsic parameters across all images, while allowing for image-specific



**Fig. 2:** Overview of our two-stage approach: (a) **Grayscale training**: We fit 2DGS [24] with an image-specific affine color transformation to the grayscale images calibrated with RealityCapture [44]. (b) **Colorization**: 2DGS depth maps are used to project colors from limited manually colorized images into 3D space as pseudo-colors. Together with the input color views, the pseudo-colors guide the colorization of the grayscale model via L1, TCM, and CCM loss functions.

distortion coefficients. This approach yielded the most accurate and highly precise calibration for our dataset.

**Backbone model.** While the original method we built upon, Ref-NPR [69], utilizes Plenoxels [14], we found this backbone to be inappropriate for our setup, often resulting in visual artifacts (see Fig. 4). Alternatively, we considered the more recent 3D Gaussian Splatting [28] method which represents the scene using 3D Gaussian primitives, but encountered similar issues including visible floaters caused by poor geometry fitting. Lastly, the successor method 2DGS [24], which uses 2D oriented Gaussian disks to represent the scene, succeeded in modeling the SEM images, rendering accurate depth maps and grayscale images.

**View-specific effects.** To accommodate varying illumination conditions, we employ an affine color transformation (ACT) as proposed by [10]. Each Gaussian holds spherical harmonics coefficients, which are subsequently converted to output intensity values. Prior to rasterization, we apply an image-dependent transformation on the decoded illumination,  $\mathbf{L}$ . During the subsequent colorization stage, we found that applying this affine transformation led to a degradation in output quality, so we omit it in the second stage. The transformation uses three weights  $\mathbf{W} = \{w_1, w_2, w_3\}$  and biases  $\mathbf{b} = \{b_1, b_2, b_3\}$  and is defined as  $\mathbf{L}' = \mathbf{W} \cdot \mathbf{L} + \mathbf{b}$ . When rendering novel views, we average the weights and biases

of the training views, resulting in plausible and consistent illumination, as can be seen in Fig. 3.

### 3.2 Colorization

Our colorization method is based on Ref-NPR [69], which was introduced for 3D stylization. We made several key modifications and enhancements to adapt it to our dataset and colorization needs. This section details our approach, highlighting the differences and improvements. We rely on three components: pseudo-color supervision for views lacking color information, a Template-based Correspondence Module for propagating colors via the grayscale feature space, and a Coarse Color-Matching Loss to ensure global color consistency.

**Pseudo-color supervision.** For color transfer, we first utilize the depth information of the grayscale model to unproject the input pixels of the colored views into the 3D space. Then, for each pixel in the other views we compute a pseudo-color as the color of the closest colored point to its unprojected location. This color is then used as a supervision signal via the loss defined in Eq. 1, where  $\hat{\mathbf{C}}_{pc}$  denotes the pseudo-color, and  $\hat{\mathbf{C}}_{\hat{x}}$  refers to the color rendered by the model. If there is no colored point within a given radius, we exclude the respective pixel from the loss calculation.

$$\mathcal{L}_{\text{pseudo-color}} = \frac{1}{N_{pc}} \left\| \hat{\mathbf{C}}_{pc} - \hat{\mathbf{C}}_{\hat{x}} \right\|_1. \quad (1)$$

Ref-NPR employs Reference Ray Registration with a grid system, where colors are assigned to grids and a single color is selected for each pixel. However, this approach does not consider the distance between points during the selection phase; it only filters out pixels that exceed a threshold in the final image, making it difficult to create precise pseudo-colors and potentially resulting in artifacts in the final output.

Our approach is designed to accommodate high-resolution datasets without encountering memory constraints. The grid-based method in Ref-NPR becomes computationally infeasible for our data, as the higher number of required grids to match our cinematic resolution would exhaust available memory resources.

While Ref-NPR also considers the cosine similarity of ray directions in the final image, we found that this factor had minimal impact on quality for our dataset. Therefore, we simplified our approach by concentrating exclusively on spatial proximity.

Pseudo-colors for the background regions often introduce noise and inconsistencies across images, as the number of background pixels varies significantly between images, potentially skewing the loss calculation in Eq. 1. To address these issues, we employ Segment Anything Model [30] to generate precise masks of the pollen granule. This segmentation allows us to effectively extract only the pollen and eliminate the interference of the background elements.

**Template-based Correspondence Module.** We employ TCM proposed by Ref-NPR as a loss function to propagate colors to the areas that do not have a ground truth color assigned in the colorized input views, by using matches in the feature space of the grayscale images. This loss minimizes the cosine distance between the features  $F_{I_g}$  of the rendered color image and a constructed guidance feature  $\hat{F}_{I_g}$  of the view  $I_g$ . The grayscale image  $I_g$  is fed into a VGG network [38] to extract the feature map  $F_{I_g}$ . The feature maps of the reference colorized images  $S_k$  and their grayscale version  $I_k$  are extracted as  $F_{S_k}$  and  $F_{I_k}$  respectively. For each location  $(i, j)$  in the guidance feature map  $\hat{F}_{I_g}^{(i,j)}$ , we consider the grayscale feature  $F_{I_g}^{(i,j)}$  and search for the nearest grayscale feature across reference views  $F_{I_k}^{(i^*,j^*)}$  and take the corresponding feature of the colorized image  $F_{S_k}^{(i^*,j^*)}$ , as defined in Eq. 2. Further details on TCM can be found in [69].

$$\hat{F}_{I_g}^{(i,j)} = F_{S_k}^{(i^*,j^*)}, \quad \text{where } (i^*, j^*), k = \arg \min_{(i', j'), k'} \text{dist}\left(F_{I_g}^{(i,j)}, F_{I_{k'}}^{(i',j')}\right), \quad (2)$$

$$\mathcal{L}_{\text{TCM}} = \text{dist}(F_{I_g}, \hat{F}_{I_g}) \quad (3)$$

**Coarse Color-Matching Loss.** Although TCM helps estimate color for occluded regions, it can result in global color inconsistencies and mismatches. To address this limitation, we also consider a coarse color-matching loss [69] that operates at the patch level to minimize color differences, as defined in Eq. 4. Using the index  $(i^*, j^*)$  obtained in Eq. 2, let  $\bar{C}$  denote the average color of a patch, and  $C_{S_k}$  and  $C_{I_g}$  refer to the patches in the input color image and rendered image respectively. See [69] for details.

$$\mathcal{L}_{\text{coarse-color}} = \frac{1}{N} \sum_{i,j} \|\bar{C}_{I_g}^{(i,j)} - \bar{C}_{S_k}^{(i^*,j^*)}\|_2^2. \quad (4)$$

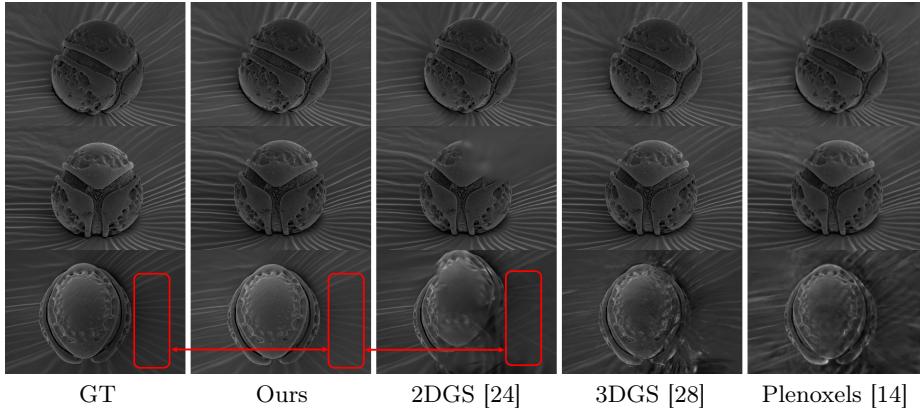
**Optimization.** For views with available colorized image, we directly optimize the L1 loss between rendered and input images. For the other views, our final loss function is defined in Eq. 5.  $\lambda$ s are respective weights for each loss term.

$$\mathcal{L} = \lambda_{\text{pc}} \mathcal{L}_{\text{pseudo-color}} + \lambda_{\text{TCM}} \mathcal{L}_{\text{TCM}} + \lambda_{\text{cc}} \mathcal{L}_{\text{coarse-color}} \quad (5)$$

## 4 Experiments

### 4.1 Dataset

Our dataset comprises 32 high-resolution ( $3072 \times 2048$ ) SEM images of a pollen, captured along two primary axes. The horizontal axis consists of 20 images spanning from left to right, providing a comprehensive lateral view, while the vertical axis includes 12 images, with the camera moving in an arc from the frontal view to the top view of the pollen. We illustrate 18 of these grayscale images in Fig. 1a, and use the entire set of 32 images as our training dataset. Fig. 1b shows colorized versions of 5 of the grayscale images, manually colorized by an artist known for his work on microscopic subjects, Martin Oeggerli.



**Fig. 3:** Grayscale novel views. The first two rows show novel views in the lateral trajectory and the bottom row indicates the view from the top. All models are trained at  $3072 \times 2048$ . Red squares highlight areas with illumination differences. ACT in Ours effectively normalizes the differences across views.

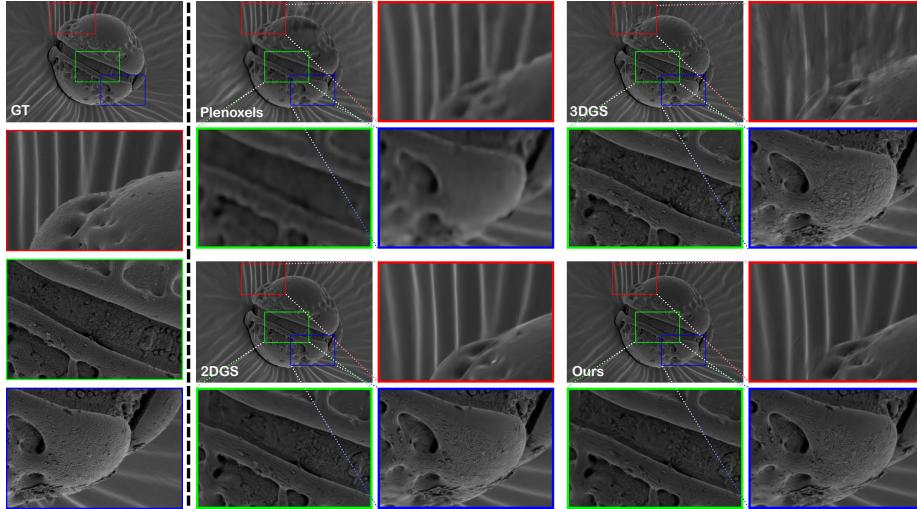
#### 4.2 Implementation details

Plenoxels [14] uses an initial grid resolution of  $128 \times 128 \times 128$ , which is increased up to  $304 \times 304 \times 128$ . Larger grid resolution led to training collapse, likely due to variations in scene illumination. The model was trained for 38,400 iterations at each grid resolution, with a total of 115,200 iterations. Using Plenoxels trained on grayscale, we train Ref-NPR for 10 epochs in the colorization phase. To address the computational and memory demands of the Template-based Correspondence Module (TCM) and coarse color-matching (CCM) loss, we compute them on images scaled down by a factor of 4. The grayscale 3D Gaussian Splatting (3DGS) [28] and 2D Gaussian Splatting (2DGS) [24] models are trained for 60,000 epochs, with 20,000 additional epochs for the colorization process. We initialize the affine color transformation (ACT) by adding small random perturbations to the identity transformation (weights  $w_i = 1$  and biases  $b_i = 0$ , for  $i = 1, 2, 3$ ) and optimize the parameters using a learning rate of 0.0001.

#### 4.3 Grayscale

We qualitatively compare the backbones on grayscale novel view synthesis and quantitatively on rendering the training views by evaluating common image quality metrics: PSNR, SSIM [59], and LPIPS [68]. All models were trained at the full  $3072 \times 2048$  resolution. Although Plenoxel's grid resolution was insufficient to fully represent this high resolution, we prioritized consistent training conditions to ensure a fair comparison of the best possible rendered quality across all methods especially since rendering at cinematic resolution is our goal.

The rendered novel views are presented in Fig. 3, with detailed close-ups shown in Fig. 4. We observe in Fig. 3 that our method demonstrates the ability to generate high-quality novel views with notable fidelity compared to other



**Fig. 4:** Grayscale novel views with closeups. The left side displays the nearest Ground Truth (GT) image along with its closeups. This novel view is precisely between two adjacent GTs. The right side shows the generated novel views and their corresponding closeups. Our method (2DGS+ACT) exhibits superior quality in synthesising novel views.

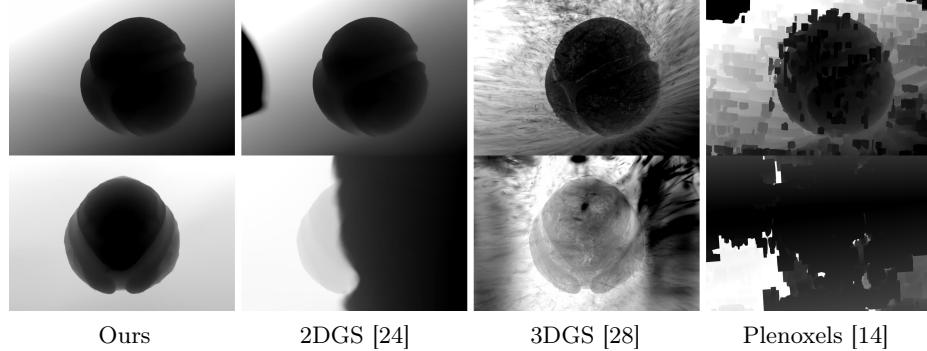
methods. ACT enables 2DGS to achieve scene representation without the prominent floaters that often appear in other approaches, especially in the view from the top. Additionally, in the bottom view (red squares), our method produces a slightly brighter image compared to the ground truth and 2DGS. This observation indicates that ACT effectively normalizes illumination across different viewpoints, resulting in more consistent lighting conditions, as well as improved rendering quality. As depicted in Fig. 4, while Plenoxels lacks fine details, the other methods model the scene with high precision. Still, 3DGS exhibits various artifacts, which are not observed in either 2DGS or 2DGS+ACT.

The rendered depth maps presented in Fig. 5 are consistent with the grayscale predictions: all other approaches face challenges in accurately estimating depths from top viewpoints. This is due to significant illumination variations in the top views (see the bottom row of Fig. 1a). Notably, 3DGS exhibits significant artifacts, not only in top views, but also in lateral perspectives. In contrast, our approach generates remarkably clean and consistent depth maps for the entire scene, demonstrating superior performance across various viewpoints.

Although the quantitative evaluation presented in Tab. 1 indicates that 3DGS outperforms in terms of rendering quality of the training views, 2DGS actually produces superior results in terms of perceived qualitative novel view quality (see Fig. 4). Moreover, the integration of ACT improves the results of 2DGS across all considered metrics. We do not quantitatively evaluate the methods for NVS in order to avoid further reducing the already limited training set.

Model	PSNR ( $\uparrow$ )	SSIM ( $\uparrow$ )	LPIPS ( $\downarrow$ )
Plenoxels [14]	30.94	0.855	0.556
3DGS [28]	<b>37.50</b>	<b>0.902</b>	<b>0.461</b>
2DGS [24]	35.25	0.867	0.511
2DGS + ACT (Ours)	36.32	0.890	0.489

**Table 1:** Evaluation of rendered training images. 3DGS and our method achieve the best quantitative scores for reproducing the training views, demonstrating the ability of modelling SEM images.



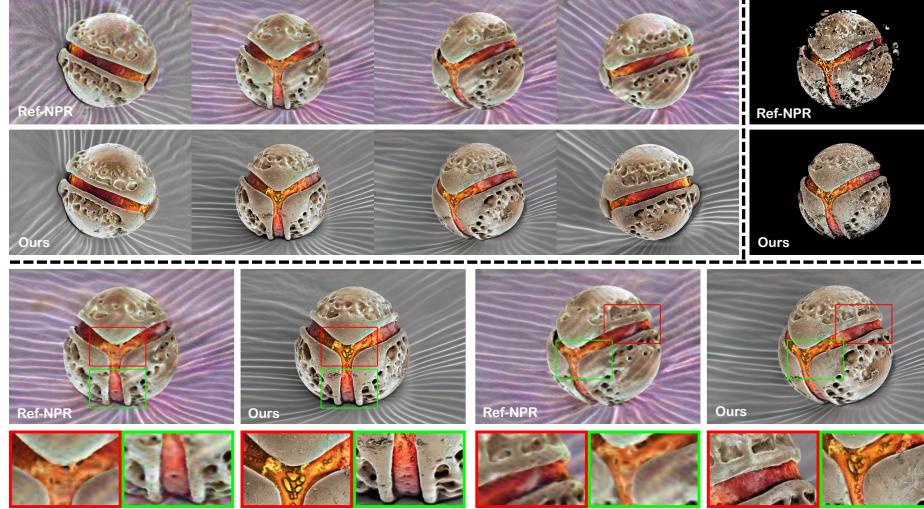
**Fig. 5:** Predicted depth maps of novel views. The view in the first row is sampled from the horizontal trajectory, while the view in the second is sampled from the vertical trajectory. All other methods failed to predict depths in the vertical trajectory. The depth map accuracy is essential for correctly projecting colors into the 3D space.

#### 4.4 Colorization

**Qualitative evaluation.** We compare our method against Ref-NPR [69] for colorization using five color images as input and present the results in Fig. 6. For fairness, both methods were trained at a lower resolution ( $768 \times 512$ ), due to Plenoxel’s limitation on grid resolution. Note that, at this grid resolution, the grayscale rendering is of similar quality at low and high image resolution, as depicted in Fig. 3. The rendered views by Ref-NPR exhibit noticeable purple artifacts in the background. To address this, the method uses CCM loss. However, our experiments revealed that excessive reliance on this loss affected the fine details. Showcased here is the best result achieved in our experiments with Ref-NPR. Our method, in contrast, successfully colorizes even the very fine details, particularly in the complex geometry and appearance of the pollen’s center. Regarding pseudo-colors, our method generates nearly perfect pseudo-colors, whereas Ref-NPR’s grid-based color selection method results in voxel-like artifacts. Although our method also shows some inconsistencies at the pseudo-color level, the final results exhibit a smooth transition on the surface color.

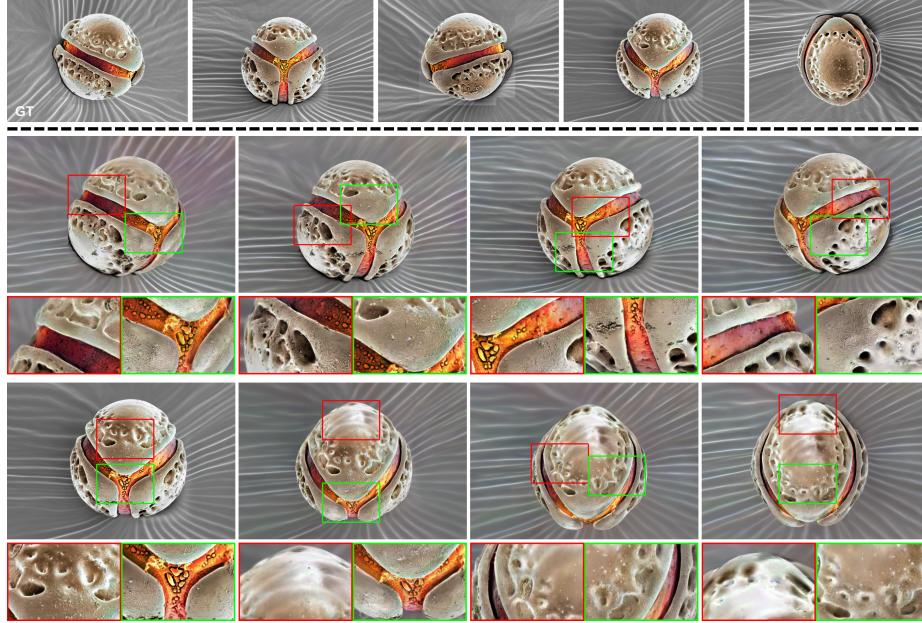
**High-resolution colorization.** To obtain the final result, we trained our model with five color images at the original resolution ( $3072 \times 2048$ ). The generated novel views and corresponding close-ups are shown in Fig. 7. Our method accurately

captures fine details such as black outlines of small circular protrusions at the center and the gradual transition of red hues from center to periphery. The boundary of the pollen is clear, and color changes on the gray surface are also captured. Overall, our method achieved uniform and high-quality colorization across the entire scene.



**Fig. 6:** Qualitative comparison between Ref-NPR [69] and our method, both trained on  $768 \times 512$  resolution images with five colored images. Top left: Synthesized novel views. Top right: Pseudo-colors. Bottom: closeups of novel views.

**Different number of color inputs.** Our final results utilize five color images, but we also investigated the impact of using fewer artist colored images to guide the colorization of the scene. We examined scenarios with one, two, three, and four color images. The four-image case omits the top view (rightmost in Fig. 7). The three-image case uses only lateral views. The two-image case employs the frontal and rightmost lateral views. The one-image case relies only on the frontal view. Fig. 8 shows the pseudo-colors and the corresponding colorization results with varying numbers of color inputs. In the single-image case, which only uses the frontal view, it tends to produce darker red on the sides. Strong reflections appear on the sides and top. This effect originates from the pseudo-colors, as seen in the first row of the single image case. Nevertheless, even with just one image, our method achieves reasonably accurate colorization. In the two-image case, which lacks the leftmost view, the predicted colors are also darker on the sides, though not as strong as in the single-image case. In the three- and four-color image cases, which lack the top views, slight color inconsistencies are noticeable on the top compared to the five-image case. As the area covered by color inputs increases, the colorization accuracy improves.

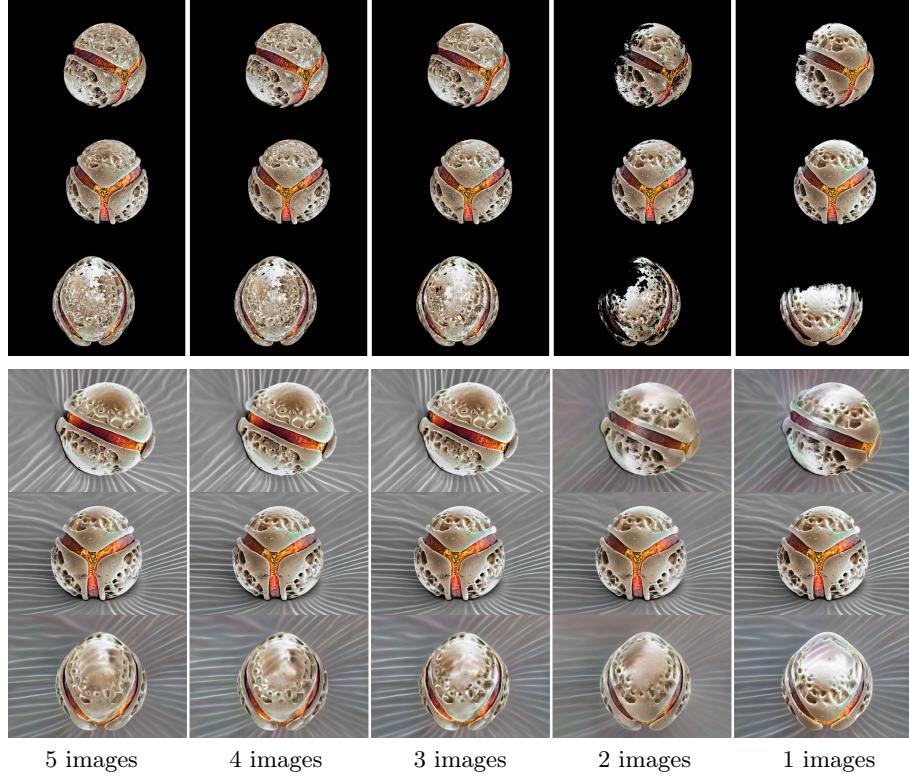


**Fig. 7:** Novel views and closeups generated by our method. Top: All input color images. Bottom: Synthesized novel views at 3072×2048 resolution, with corresponding closeups.

#### 4.5 Ablation study

To evaluate the contribution of each component in our proposed model, we conducted an ablation study. We compared the full model against versions without TCM, without CCM, without ACT, and without all components.

Fig. 9 illustrates the generated novel views from this study. The case (w/o TCM) shows noticeable color inconsistencies on the surface, likely due to variations in the pseudo-colors. This highlights the effectiveness of TCM in estimating colors, although it introduces subtle background color shifts, with slight reddish (second row) or bluish tints (fourth row) in the Full model case. Still, the CCM ablation shows the role of the loss in reducing global color inconsistencies, as these background color artifacts are mitigated in the full model. The case without ACT leads to more pronounced artifacts (third row), originating from the grayscale representation (see Fig. 3). Additionally, reddish background color shifts are also seen in this case, and the color on the sides (fourth row) is much darker than in other cases. This phenomenon is likely due to the presence of these floaters, which adversely impacted the effectiveness of CCM, leading to its failure. Moreover, compared to the case without any components, the color prediction in our full model performs remarkably well. Consequently, each component contributes to our final results.

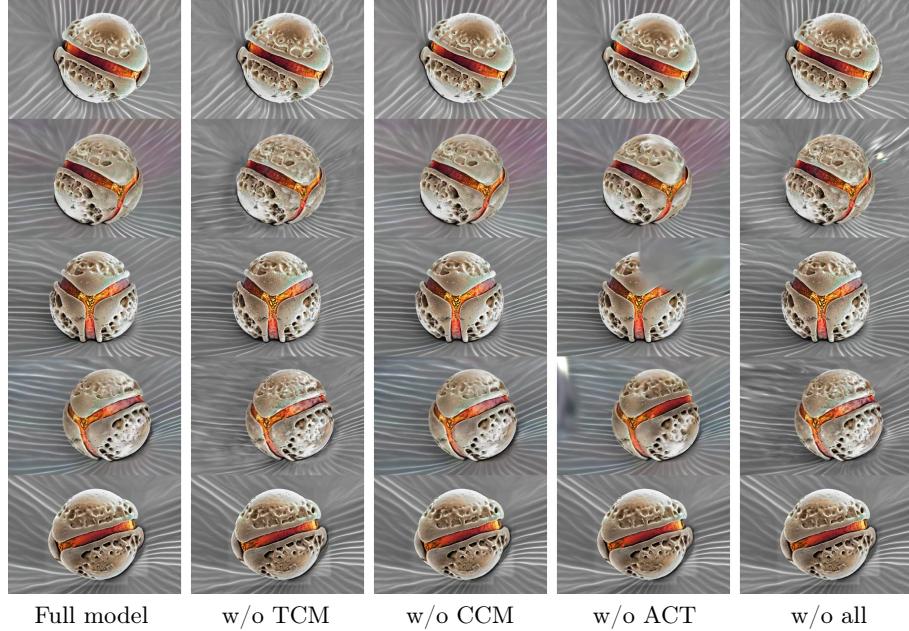


**Fig. 8:** Novel views with varying amounts of manually colored input images. Whilst for a limited range of viewpoints we reach excellent results even from a single colored image, more images improve the overall colorization quality visible (especially when animated).

## 5 Limitations

Our method requires substantial processing time on high-end hardware. Generating pseudo-colors takes approximately 5 hours on an NVIDIA A100 GPU due to the need for each pixel to identify the nearest color from among all pixels across the five color images of pollen. Additionally, the colorization process itself takes around 3 hours. This time requirements are induced by TCM and the CCM loss.

While our method produces high-quality results, some challenges remain. As illustrated in Fig. 3, accurately modeling very fine surface patterns in the grayscale stage has proven challenging. Moreover, as depicted in Fig. 3, artifacts persist in the colorization results. These include a subtle reddish tint in the background, minor green discoloration near the central raised area and along the sides, line-shaped artifacts in the red regions on both sides, and a blurred area in the apical region.



**Fig. 9:** Ablation studies. This figure illustrates the impact of different model components on colorization quality. From left to right: Our full model, without Template-based Correspondence Module (TCM), without coarse color-matching (CCM) loss, without affine color transformation (ACT), without all modules.

As illustrated in Fig. 8, the current single-view case is limited to the colors visible in the image. However, this limitation could potentially be overcome by employing segmentation or feature-based approaches, or even a diffusion-based method to estimate the colors of the unseen parts of the object.

## 6 Conclusion

We achieved cinematic colorization of pollen images captured by a Scanning Electron Microscope. Our approach, which incorporates color projection onto 3D space, affine color transformation, a Template-based Correspondence Module, and a Coarse Color-Matching loss, demonstrated superior performance on our dataset compared to existing methods. We effectively validated the necessity and efficacy of each component in achieving our results. From an artistic perspective, we are confident in the value of our work; introducing color to a monochrome realm offers a visually arresting and mesmerizing experience. Moreover, our method particularly enables us to reduce the number of viewpoints artists need to color manually. By eliminating the manual annotations through our novel view synthesis process, our approach not only enhances efficiency but also opens new creative possibilities for artists.

## Acknowledgments

We would like to thank Maximilian Weiherer for valuable discussions and support with the camera calibration. Andreea Dogaru was funded by the German Federal Ministry of Education and Research (BMBF), FKZ: 01IS22082 (IRRW). The authors are responsible for the content of this publication. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b112dc IRRW. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. This project was supported by the special fund for scientific works at the Friedrich-Alexander-Universität Erlangen-Nürnberg.

## References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
2. Binniger, A., Hertz, A., Sorkine-Hornung, O., Cohen-Or, D., Giry, R.: Sens: Part-aware sketch-based implicit neural shape modeling. In: Computer Graphics Forum. vol. 43. Wiley Online Library (2024)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
4. Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In: ICCV (2023)
5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensororf: Tensorial radiance fields. In: European conference on computer vision. pp. 333–350. Springer (2022)
6. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. In: ICCV (2023)
7. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting (2023)
8. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: Proceedings of the IEEE international conference on computer vision. pp. 415–423 (2015)
9. Chiang, P.Z., Tsai, M.S., Tseng, H.Y., Lai, W.S., Chiu, W.C.: Stylizing 3d scene via implicit representation and hypernetwork. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1475–1484 (2022)
10. Darmon, F., Porzi, L., Rota-Bulò, S., Kortschieder, P.: Robust gaussian splatting. arXiv preprint arXiv:2404.04211 (2024)
11. Dogaru, A., Ardelean, A.T., Ignatyev, S., Zakharov, E., Burnaev, E.: Sphere-guided training of neural implicit surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20844–20853 (2023)
12. Dong, J., Wang, Y.X.: Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. Advances in Neural Information Processing Systems **36** (2024)

13. Franke, L., Rückert, D., Fink, L., Stamminger, M.: Trips: Trilinear point splatting for real-time radiance field rendering. In: Computer Graphics Forum. p. e15012. Wiley Online Library (2024)
14. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxtels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5501–5510 (2022)
15. Gao, J., Gu, C., Lin, Y., Zhu, H., Cao, X., Zhang, L., Yao, Y.: Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. arXiv preprint arXiv:2311.16043 (2023)
16. Gao, R., Xiang, T.Z., Lei, C., Park, J., Chen, Q.: Scene-level point cloud colorization with semantics-and-geometry-aware networks. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2818–2824. IEEE (2023)
17. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
18. Gong, B., Wang, Y., Han, X., Dou, Q.: Recolornerf: Layer decomposed radiance fields for efficient color editing of 3d scenes. arXiv preprint arXiv:2301.07958 (2023)
19. Gordon, O., Avrahami, O., Lischinski, D.: Blended-nerf: Zero-shot object generation and blending in existing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2941–2951 (2023)
20. Goytom, I., Wang, Q., Yu, T., Dai, K., Sankaran, K., Zhou, X., Lin, D.: Nanoscale microscopy images colorization using neural networks. arXiv preprint arXiv:1912.07964 (2019)
21. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19740–19750 (2023)
22. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics (TOG) **37**(4), 1–16 (2018)
23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control.(2022). URL <https://arxiv.org/abs/2208.01626> (2022)
24. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
25. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
26. Huang, Y.H., He, Y., Yuan, Y.J., Lai, Y.K., Gao, L.: 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. pp. 18342–18352 (2022)
27. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
29. Kim, H., Lee, G., Choi, Y., Kim, J.H., Zhu, J.Y.: 3d-aware blending with generative nerfs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22906–22918 (2023)

30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
31. Kuang, Z., Luan, F., Bi, S., Shu, Z., Wetzstein, G., Sunkavalli, K.: Paletten-erf: Palette-based appearance editing of neural radiance fields. arXiv preprint arXiv:2212.10699 (2022)
32. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=5NA1Pin1GFu>
33. Lee, J.H., Kim, D.S.: Ice-nerf: Interactive color editing of nerfs via decomposition-aware weight optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3491–3501 (2023)
34. Liu, J., Dai, S., Li, X.: Pccn: Point cloud colorization network. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3716–3720. IEEE (2019)
35. Liu, K., Zhan, F., Chen, Y., Zhang, J., Yu, Y., El Saddik, A., Lu, S., Xing, E.P.: Stylerf: Zero-shot 3d style transfer of neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8338–8348 (2023)
36. Liu, K., Zhan, F., Xu, M., Theobalt, C., Shao, L., Lu, S.: Stylegaussian: Instant 3d style transfer with gaussian splatting. arXiv preprint arXiv:2403.07807 (2024)
37. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)
38. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 730–734 (2015). <https://doi.org/10.1109/ACPR.2015.7486599>
39. Mazzucchelli, A., Garcia-Garcia, A., Garces, E., Rivas-Manzaneque, F., Moreno-Noguer, F., Penate-Sanchez, A.: Irene: Instant recoloring of neural radiance fields (2024), <https://arxiv.org/abs/2405.19876>
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
41. Nguyen-Phuoc, T., Liu, F., Xiao, L.: Snerf: stylized neural implicit representations for 3d scenes. arXiv preprint arXiv:2207.02363 (2022)
42. Radl, L., Steiner, M., Kurz, A., Steinberger, M.: LAENeRF: Local Appearance Editing of Neural Radiance Fields. In: CVPR (2024)
43. Rakhimov, R., Ardelean, A.T., Lempitsky, V., Burnaev, E.: Npbg++: Accelerating neural point-based graphics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15969–15979 (2022)
44. Realitycapture. <http://www.capturingreality.com>
45. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21**(5), 34–41 (2001). <https://doi.org/10.1109/38.946629>
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
47. Rückert, D., Franke, L., Stamminger, M.: Adop: Approximate differentiable one-pixel point rendering. ACM Transactions on Graphics (ToG) **41**(4), 1–14 (2022)
48. Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V., Theobalt, C.: Nerf for outdoor scene relighting. In: ECCV (2022)

49. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
50. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
51. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14. pp. 501–518. Springer (2016)
52. Shahbazi, M., Claessens, L., Niemeyer, M., Collins, E., Tonioni, A., Van Gool, L., Tomboli, F.: Inserf: Text-driven generative object insertion in neural 3d scenes. arXiv preprint arXiv:2401.05335 (2024)
53. Song, L., Cao, L., Gu, J., Jiang, Y., Yuan, J., Tang, H.: Efficient-nerf2nerf: Streamlining text-driven 3d editing with multiview correspondence-enhanced diffusion models. arXiv preprint arXiv:2312.08563 (2023)
54. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
55. Venema, E.: Colorizing Scanning Electron Microscopy Images With Diffusion Models. Master’s thesis, The University of Bergen (2023)
56. Wang, C., Jiang, R., Chai, M., He, M., Chen, D., Liao, J.: Nerf-art: Text-driven neural radiance fields stylization. arXiv preprint arXiv:2212.08070 (2022)
57. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NeurIPS (2021)
58. Wang, Y., Cheng, J.S., Feng, Q., Tao, W.Y., Lai, Y.K., Li, K.: Tsnerf: Text-driven stylized neural radiance fields via semantic contrastive learning. Computers & Graphics **116**, 102–114 (2023)
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)
60. Yang, Z., Liu, L., Huang, Q.: Learning generative neural networks for 3d colorization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
61. Yeh, Y.Y., Huang, J.B., Kim, C., Xiao, L., Nguyen-Phuoc, T., Khan, N., Zhang, C., Chandraker, M., Marshall, C.S., Dong, Z., et al.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In: CVPR (2024)
62. Yuan, Y.J., Sun, Y.T., Lai, Y.K., Ma, Y., Jia, R., Gao, L.: Nerf-editing: Geometry editing of neural radiance fields. In: CVPR (2022)
63. Zhang, D., Chen, Z., Yuan, Y.J., Zhang, F.L., He, Z., Shan, S., Gao, L.: Stylizedgs: Controllable stylization for 3d gaussian splatting (2024), <https://arxiv.org/abs/2404.05220>
64. Zhang, K., Kolkin, N., Bi, S., Luan, F., Xu, Z., Shechtman, E., Snavely, N.: Arf: Artistic radiance fields. In: European Conference on Computer Vision. pp. 717–733. Springer (2022)
65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
66. Zhang, M., Liao, J., Yu, J.: Deep exemplar-based color transfer for 3d model. IEEE transactions on visualization and computer graphics **28**(8), 2926–2937 (2020)

67. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14. pp. 649–666. Springer (2016)
68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
69. Zhang, Y., He, Z., Xing, J., Yao, X., Jia, J.: 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. pp. 4242–4251 (2023)
70. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
71. Zhuang, J., Wang, C., Liu, L., Lin, L., Li, G.: Dreameditor: Text-driven 3d scene editing with neural fields. arXiv preprint arXiv:2306.13455 (2023)
72. Zivanov, J.: Reconstruction of intricate surfaces from scanning electron microscopy. Ph.D. thesis, University\_of\_Basel (2017)
73. Zivanov, J., Vetter, T.: Multiview reconstruction of complex organic shapes. In: BMVC. pp. 157–1 (2015)