CNT-NeRF: Carbon Nanotube Forest Depth Layer Decomposition in SEM Imagery using Generative Adversarial Networks

Nguyen P. Nguyen1, Ramakrishna Surya2, Prasad Calyam1, Kannappan Palaniappan1, Matthew Maschmann2, Filiz Bunyak1*
1Department of Electrical Engineering and Computer Science
2Department of Mechanical and Aerospace Engineering
University of Missouri-Columbia, MO, USA
Email: {npntz3, rst7b}@mail.missouri.edu
{ma8pz, calyamp, pal, maschmannm, bunyak}@missouri.edu

Abstract

Carbon nanotube (CNT) forests are imaged using scanning electron microscopes (SEMs) that project their multi-layered 3D structure into a single 2D image. Image analytics, particularly instance segmentation is needed to quantify structural characteristics and to predict correlations between structural morphology and physical properties. The inherent complexity of individual CNT structures is further increased in CNT forests due to density of CNTs, interactions between CNTs, occlusions, and lack of 3D information to resolve correspondences when multiple CNTs from different depths appear to cross in 2D. In this paper, we propose CNT-NeRF, a generative adversarial network (GAN) for simultaneous depth layer decomposition and segmentation of CNT forests in SEM images. The proposed network is trained using a multi-layer, photo-realistic synthetic dataset obtained by transferring the style of real CNT images to physics-based simulation data. Experiments show promising depth layer decomposition and accurate CNT segmentation results not only for the front layer but also for the partially occluded middle and back layers. This achievement is a significant step towards automated, image-based CNT forest structure characterization and physical property prediction.

1. Introduction

Carbon nanotubes (CNTs) [17] are tubular nanostructures with remarkable mechanical, thermal, electrical, and chemical properties [9, 18]. Single-walled CNTs (SWNTs) may exhibit metallic or semiconducting traits. Multi-walled CNTs (MWNTs) can be transformed into yarns [41] that are not only electrically conductive and strong, but also retain their flexibility and can even be tied into a knot. The properties of CNTs make them an appealing material for a wide range of applications, including dry spinning conductive, high-strength fibers [19, 41], piezoresistive sensing [23, 24, 30], electrochemical energy storage [5, 8], and thermal interface materials [6, 7].

Production of individual CNTs in isolation remains a challenge and is not practical for device-level integration. CNT forests, with dense populations synthesized on a support substrate, offer a solution by forcing the CNTs to grow vertically. Persistent van der Waals bonds, created by interactions between contacting CNTs, resist mechanical loads and result in an open-cell foam-like morphology. Despite their fascinating properties, CNT forests’ characteristics are significantly diminished compared to individual CNTs. For instance, while an individual CNT has an elastic modulus of around 1 TPa, a CNT forest’s compressive elastic modulus are frequently on the order of 1-10 MPa [26], akin to natural rubber. Variations in CNT forest morphology created during cooperative synthesis [1,22,31] are thought to be the root cause of the wide range of deformation mechanisms observed in compressed CNT forests [2,3,16,25–27,34,40].

Testing CNT forests’ physical properties often necessitates their destruction, limiting further data collection. Imaging and image analysis offer a nondestructive opportunity to indirectly predict CNT forests’ properties. CNT forests are imaged using scanning electron microscopy (SEM) or transmission electron microscopy (TEM). Figure 1 shows a CNT forest micropillar imaged at different zoom levels using SEM. Image analytics seeks to quantify structural characteristics such as diameter, orientation, curvature, tortuosity, density, spatial layout etc. to predict correlations between structural morphology and physical properties. This characterization requires robust segmentation of
Figure 1. The scanning electron microscope (SEM) was used to capture images of a carbon nanotube (CNT) pillar. The images depict (a) a full view of the pillar, and (b) a magnified side view of the CNT pillar. (c) Various applications of CNTs: semiconductors, energy storage and thermal materials.

CNTs within CNT forests. This is a highly challenging task since the inherent complexity of individual CNT structures is further increased due to density of CNTs in CNT forests, front to back occlusion, lack of 3D information in SEM images to resolve correspondences when multiple CNTs from different depths appear to cross in 2D etc.

Earlier attempts to analyze CNT images mostly relied on traditional image processing techniques. For instance, thresholding was used in [12] to create partial CNT masks that were used to determine the diameters of CNTs, while class-entropy maximization was used in [38] to segment CNT images captured at modest magnification levels (800X-4000X). More recently neural networks-based approaches started to be used for CNT image analysis. In [35], a hybrid approach, combining thresholding and classification using multi-layer neural networks, was used to segment sparse, non-overlapping CNTs in small image patches. In [13, 14], deep learning networks were used to analyze synthetic CNT forest images generated from physics-based simulations to predict mechanical properties. In [29], a self-supervised learning network CNTSegNet was introduced for semantic segmentation of CNTs based on weak labels and orientation histograms computed in Fourier space.

In this paper, we present a neural radiance field (NeRF) [21, 36] inspired deep learning network that aims to reduce image complexity in order to ensure accurate segmentation/tracing of individual CNTs. The proposed generative deep learning network decomposes CNT forest images into their depth layers and generates preliminary segmentation for each layer. The network converts a single 2D image into K 2D images corresponding to different depth layers. This 2.5D representation resolves front-versus-back relationships and occlusions, reduces image complexity for instance segmentation, lowers potential id-switches from one CNT to another during CNT tracing, and ultimately allows for a more comprehensive understanding of the spatial arrangement and intricacies of CNTs in the analyzed images.

2. Methods

2.1. Neural radiance field (NeRF)

Neural radiance field (NeRF) [28] is a recent neural networks-based approach for synthesizing novel views of complex 3D scenes using sparse set of views. The process starts with a virtual camera that casts rays into the scene to sample the 3D coordinates \((x, y, z)\) in the scene with viewing angles \((\theta, \phi)\) in order to project output colors and densities (probability of visibility/transparency). In the next step, a 3D scene dataset is utilized to train the neural network, enabling it to learn the relationships between the 5D input data (3D points and viewing directions) and their corresponding colors and densities [28]. Once trained, the neural network can take in new 3D points and viewing directions as input, and infer a set of colors and densities corresponding to each point in the scene. These inferred properties can then be used to render more realistic and accurate 3D scenes using techniques such as volume rendering or ray tracing.

Use of neural networks as black-box models for inferring properties of 3D scenes represents a significant breakthrough in the field of computer graphics, for applications from scientific imaging to video game/education design. As this technology advances, it has the potential to revolutionize the way we visualize and interact with 3D data, including complex microscopy data.

2.2. NeRF inspired CNT layer synthesis

Scanning electron microscope (SEM) projects the multi-layered 3D structure of CNT forests into a single 2D image. The planar neural radiance field proposed in [21, 36] uses planes instead of rays to represent the camera frustum. This method can be used for single-view image rendering as well as depth estimation. In our application, by sampling a num-
ber of planes within the camera frustum at different depth levels, we can efficiently capture the structural properties of complex 3D CNT forests. Our NeRF inspired CNT layer synthesis solution is designed for segmentation and depth estimation based solely on pixel (grayscale) intensity values. We intentionally ignore the camera matrix parameters and rely solely on depth layer information and single-view CNT forest images as inputs. This approach enables us to achieve accurate segmentation and depth estimation results while reducing computational complexity and potential errors caused by incorporating more parameters.

2.3. Network architecture

We assume that the segmentation mask for a CNT forest image consists of $K$ layers partitioning the CNTs according to their depths indicated by similar grayscale intensity levels in the input image. With this assumption, we designed a novel framework for CNT forest depth layer decomposition and segmentation using a generative adversarial network (GAN). Figure 2 illustrates our network architecture and its training approach. The first block of this framework is the generator, which includes an encoder and a decoder. Our network utilizes a ResNet-34 [15] encoder to extract features from the input images. The decoder (using the architecture of Monodepth2 [11]) takes the extracted features from the latent space, combines them with the prior information from the disparity vector, and generates prediction outputs. The disparity vector [21, 33] is a positional encoding, similar to the ones used in transformer models [4, 37]. It maps the layer depth information into a $2 \times L$ dimension embedding space. Given a layer with depth $z_i$, its encoded disparity vector is computed as [21]:

$$\nu(z_i) = [\sin(2^0 \pi z_i), \cos(2^0 \pi z_i), \sin(2^1 \pi z_i), \cos(2^1 \pi z_i), ... , \sin(2^{(L-1)} \pi z_i), \cos(2^{(L-1)} \pi z_i)]$$

where the frequency parameter $L$ was set to 10.

The ResNet-34 [15] encoder is used to extract robust and
informative features from the input images, which are then used to generate high-quality predictions in the decoder. The disparity vector provides additional information about the depth of each layer, allowing our network to better distinguish between the adjacent layers and their orders.

For each output layer \( i \), our segmentation network generates two channels: (1) a mask channel \( M_i \); and (2) corresponding sigma channel \( \sigma_i \), which controls visibility. Pixel-wise multiplication of the masks and corresponding sigma channels yields the final segmentation mask. This setting enables effective handling of occlusion issues where retrieval of CNTs in the back layers that may be obstructed by those in the front layers.

The second block of our CNT segmentation network, also using a ResNet-34 model [15], is a discriminator responsible for classifying the output and ground truth masks as fake or real after they are applied to the input images. This process enables the discriminator to determine authenticity of the generated masks and provides feedback to the generator on how to improve its performance.

By integrating a discriminator into our network architecture, we can improve quality and accuracy of the generated masks. The discriminator provides an additional level of feedback and supervision, which can help the generator to learn the nuances and subtleties of the data more effectively. Below we introduce the discriminator and generator objective functions used to optimize the proposed GAN model.

### 2.4. Discriminator loss function

Assuming a GAN model with a generator \( G \) and a discriminator \( D \), we feed the input image \( I \) and the disparity vector \( v_d \) into this GAN model to generate a prediction mask \( M_{\text{pred}} \). The discriminator’s goal is to classify both ground truth \( M_{\text{GT}} \) and prediction \( M_{\text{pred}} \) masks. To achieve this, the discriminator loss \( L_D \) is calculated using cross-entropy loss and is minimized during the training process.

\[
L_D = \log D(M_{\text{GT}} \circ I) + \log (1 - D(M_{\text{pred}} \circ I)) = \log D(M_{\text{GT}} \circ I) + \log (1 - D(G(I, v_d) \circ I))
\]

where \( \circ \) refers to pixel-wise multiplication operation.

### 2.5. Generator loss function

If the generator is capable of generating perfectly realistic data, the discriminator should classify this data as real with high confidence. To achieve this, the generator’s training objective is to minimize the difference between the predicted confidence value and the ground truth value, which represents the true classification of the generated data. The generator loss for classification is defined as:

\[
L_{G_{\text{class}}} = \log (1 - D(M_{\text{pred}} \circ I)) = \log (1 - D(G(I, v_d) \circ I))
\]

Segmentation is a challenging task that requires a strong loss function to ensure that the predicted mask closely matches the ground truth mask. In addition to the classification loss above, we utilized two other loss functions, dice loss and scale invariant loss [10].

Given a prediction mask \( M_{\text{pred}} \) and a ground truth mask \( M_{\text{GT}} \), dice loss is computed using the following equation:

\[
L_{G_{\text{Dice}}} (M_{\text{pred}}, M_{\text{GT}}) = 1 - 2 \times \frac{|M_{\text{pred}} \cap M_{\text{GT}}|}{|M_{\text{pred}}| + |M_{\text{GT}}|} = 1 - 2 \times \frac{|G(I, v_d) \cap M_{\text{GT}}|}{|G(I, v_d)| + |M_{\text{GT}}|}
\]

It is common to use dice loss to match the prediction map and its ground truth mask. We have taken this approach a step further by incorporating a scale-invariant loss component, inspired by the method proposed in [10] for depth estimation. This addition aims to enhance the alignment of relative differences between each pair of pixels in the prediction map and the corresponding pair of pixels in the ground truth mask.

Suppose a ground truth mask and associated prediction map contains \( n \) pixels and \( y_i, y_i^* \) denote intensity values at the \( i^{th} \) pixel of the associated prediction map and the ground truth mask respectively. We define \( d_i \) as the difference (in log scale) between the \( i^{th} \) pixel in the prediction map and the corresponding pixel in the ground truth mask:

\[
d_i = \log(y_i) - \log(y_i^*)
\]

Using \( d_i \), the scale-invariant loss computes the relative difference between each pair of pixels \( i \) and \( j \) in the prediction map and in the ground truth mask as following to minimize their distance.

\[
L_{G_{\text{DiS}}} (M_{\text{pred}}, M_{\text{GT}}) = \frac{1}{2n^2} \sum_{i,j} ((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*))^2 \\
= \frac{1}{n} \sum_i d_i^2 - \frac{1}{n^2} \sum_{i,j} d_i d_j = \frac{1}{n} \sum_i d_i^2 - \frac{1}{n^2} \left( \sum_i d_i \right)^2
\]

The overall generator loss is computed as the weighted sum of the three loss components (classification, dice, and scale invariance) described above:

\[
L_G = k_1 \times L_{G_{\text{class}}} + k_2 \times L_{G_{\text{Dice}}} + k_3 \times L_{G_{\text{DiS}}}
\]

where weights \( k_1, k_2, k_3 \) were empirically set to 0.1, 0.3, and 0.6 respectively.
3. Experimental results

3.1. Datasets

**Synthetic images of CNT forests:** In order to thoroughly evaluate the performance of the proposed layer segmentation approach, we generated an SEM-style, realistic-looking, synthetic CNT forest image dataset with associated ground truth masks. This synthetic dataset was generated by fusing multiple layers of 2D binary synthetic images obtained using the physics-based simulation technique described in [13]. Prior to fusion using a pixel-wise max operation, the individual layers (as shown in Figure 3a-c) were first smoothed by a Gaussian filter, then multiplied with a global weight according to their depth order in the combined image. The fused images (as shown in Figure 3d) were further improved by style transfer from a real SEM image (as shown in Figure 3f) using the Fourier Domain Adaptation (FDA) method described in [39]. FDA transfers the low-frequency features in Fourier space from the reference image to the target image, resulting in synthetic images with desired realistic styles (as shown in Figure 3e). The synthetic dataset comprised of 133 images of size 512 × 512 pixels with two versions with and without FDA style transfer. The synthetic dataset was partitioned into a training set of 106 and a test set of 27 images.

**SEM images of CNT forests:** The carbon nanotube (CNT) forests used in this study were produced using a chemical vapor deposition (CVD) technique as described in [20]. CNT forest images were acquired by utilizing a FEI Quanta scanning electron microscope (SEM) at a pixel dwell time of 10 µs and magnification of 50,000X. In total, 94 image patches of size 512 × 512 pixels were used to transfer SEM image style to synthetic images.

3.2. Experimental results on synthetic images

To assess the performance of our approach, we trained our network on these two different datasets: multi-layer synthetic images without FDA style transfer, and multi-layer synthetic images with FDA style transfer. Initially, we trained the generator in 125 epochs using both the dice loss and scale-invariant loss. Following this, we proceeded to fine-tune the generator by training it alongside the discriminator for an additional 50 epochs. We conducted a comparison between CNT-NeRF and U-net [32], a widely used architecture for segmentation. Table 1 presents segmentation dice scores for the proposed CNT-NeRF and U-net networks. We used a Resnet-34 [15] backbone encoder in the U-net network and trained it with the same training dataset as the proposed CNT-NeRF network.

As demonstrated in Table 1, CNT-NeRF outperformed
U-net by achieving higher average dice scores for both datasets, with and without FDA-style transfer. CNT-NeRF has proven to be effective in segmenting the front and middle layers, and particularly accurate in extracting the back layer, even in the presence of occlusions. CNT-NeRF’s performance surpassed that of U-net on Layer 1 by 1-to-9%, and on Layer 2 by 2-to-13%. Notably, for the back layer (Layer 3), CNT-NeRF’s dice scores consistently surpassed U-net’s by a substantial 7-to-17% margin for both training and test sets with and without style transfer. CNT-NeRF’s performance remained less affected as the depth of the layers increased, in contrast to U-net, which exhibited a significant decline in layer 3.

Figure 4 (for the synthetic image without style transfer), Figure 5 (for the synthetic images with style transfer) show the depth layer decomposition, along with their corresponding segmentation outputs from synthetic images. In these figures, rows (a) present the synthetic images and their segmentation ground truth masks for the three layers. Row (b) and (d) depict the segmentation results for CNT-NeRF and U-net, respectively, starting from the front (layer 1) and extending to the back (layer 3). Row (c) and (e) highlight the segmentation outcomes (in green) of CNT-NeRF and U-net, overlaid on the segmentation masks (in red) and the synthetic image (in grayscale). The yellow areas signify the intersections where the segmentation outcomes align with the ground truth masks. The residual red regions correspond to false positives of the segmentation tasks. By visual inspection, row (c) of CNT-NeRF displays fewer red/green regions compared to row (e) of U-net. This indicates that CNT-NeRF exhibits a lower rate of mis-prediction (false negatives and false positives) than U-net for synthetic images, both with and without style-transfer.

Additionally, CNT-NeRF exhibited better performance against overfitting, as evidenced by the minimal difference between training and test scores, ranging from 0-to-3%. In contrast, U-net experienced a performance drop of 4% and 10% when transitioning from the training set to the test set. This underscores CNT-NeRF’s ability to generalize and maintain its effectiveness across different datasets.

4. Conclusions

In this paper, we proposed CNT-NeRF, a generative adversarial network for simultaneous depth layer decomposition and segmentation of CNT forests in SEM images. CNT-NeRF converts a single 2D image into $K \times 2D$ images. This 2.5D representation aims to reduce image complexity and resolve front-versus-back relationships and occlusions for a robust instance segmentation performance. Training of the network was done using our photorealistic multi-layer synthetic images as input and associated physics-based binary synthetic layers as target labels. Promising depth layer decomposition and 7-to-15% improved CNT segmentation results were obtained compared to U-net segmentation network. The proposed depth layer decomposition and segmentation process is an important step towards automated and non-destructive characterization of CNT forest physical properties and our ultimate goal of human out of the loop material discovery.

Acknowledgement. This work was partially supported by the National Science Foundation under award number CMMI-2026847. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References


<table>
<thead>
<tr>
<th></th>
<th>TRAINING SET</th>
<th>TEST SET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Layer 1 (Front)</td>
<td>Layer 2 (Middle)</td>
</tr>
<tr>
<td>CNT-NeRF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synthetic</td>
<td>0.86</td>
<td>0.84</td>
</tr>
<tr>
<td>Synthetic + style transfer</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>U-net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synthetic</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>Synthetic + style transfer</td>
<td>0.86</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Figure 4. Depth layer decomposition and CNT segmentation results on multi-layer, physics-based, synthetic images (without FDA style transfer). Row (a) exhibits the synthetic image, and its segmentation ground truth masks for the three layers, extending from the front (layer 1) to the back (layer 3). Row (b) and (d) illustrate the segmentation outcomes for CNT-NeRF and U-net, respectively. Row (c) and (e) highlight the segmentation results (in green) of CNT-NeRF and U-net superimposed on the segmentation masks (in red) and the synthetic image (in grayscale). The yellow regions represent the intersections between the segmentation outcomes and the segmentation ground truth masks.
Figure 5. Depth layer decomposition and CNT segmentation results on multi-layer, physics-based, synthetic images (with style transfer from a real SEM image of CNT forest). Row (a) shows the synthetic image where the background is derived from a real image, and its segmentation masks for the three layers, spanning from the front (layer 1) to the back (layer 3). The segmentation outcomes for CNT-NeRF and U-net are depicted in row (b) and (d), respectively. Row (c) and (e) showcase the segmentation outcomes (in green) of CNT-NeRF and U-net overlaid on the segmentation masks (in red) and the synthetic image (in grayscale). The yellow areas illustrate the intersections between the segmentation outcomes and the segmentation ground truth masks.


