AAAS
BME Frontiers
Volume 2022, Article ID 9854084, 14 pages
https://doi.org/10.34133/2022/9854084

BMEF
A SCIENCE PARTNER JOURNAL

Research Article

Jointly Optimized Spatial Histogram UNET Architecture (JOSHUA) for Adipose Tissue Segmentation

Joshua K. Peeples,[1] Julie F. Jameson,[2] Nisha M. Kotta,[3] Jonathan M. Grasman,[4] Whitney L. Stoppel ,[2,3] and Alina Zare [1]

[1]*Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA*
[2]*Department of Chemical Engineering, University of Florida, Gainesville, FL 32611, USA*
[3]*J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA*
[4]*Department of Biomedical Engineering, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA*

Correspondence should be addressed to Whitney L. Stoppel; whitney.stoppel@ufl.edu and Alina Zare; azare@ece.ufl.edu

Received 19 November 2021; Accepted 31 March 2022; Published 25 June 2022

Objective. We aim to develop a machine learning algorithm to quantify adipose tissue deposition at surgical sites as a function of biomaterial implantation. *Impact Statement*. To our knowledge, this study is the first investigation to apply convolutional neural network (CNN) models to identify and segment adipose tissue in histological images from silk fibroin biomaterial implants. *Introduction*. When designing biomaterials for the treatment of various soft tissue injuries and diseases, one must consider the extent of adipose tissue deposition. In this work, we analyzed adipose tissue accumulation in histological images of sectioned silk fibroin-based biomaterials excised from rodents following subcutaneous implantation for 1, 2, 4, or 8 weeks. Current strategies for quantifying adipose tissue after biomaterial implantation are often tedious and prone to human bias during analysis. *Methods*. We used CNN models with novel spatial histogram layer(s) that can more accurately identify and segment regions of adipose tissue in hematoxylin and eosin (H&E) and Masson's trichrome stained images, allowing for determination of the optimal biomaterial formulation. We compared the method, Jointly Optimized Spatial Histogram UNET Architecture (JOSHUA), to the baseline UNET model and an extension of the baseline model, attention UNET, as well as to versions of the models with a supplemental attention-inspired mechanism (JOSHUA+ and UNET+). *Results*. The inclusion of histogram layer(s) in our models shows improved performance through qualitative and quantitative evaluation. *Conclusion*. Our results demonstrate that the proposed methods, JOSHUA and JOSHUA+, are highly beneficial for adipose tissue identification and localization. The new histological dataset and code used in our experiments are publicly available.

1. Introduction

1.1. Background. The main goals within biomaterial research for soft tissue regeneration are to develop materials that (a) actively promote cellular infiltration into the scaffold and (b) degrade at similar rates to *de novo* tissue formation. Biomaterials available for surgeons that are applicable for soft tissue injuries include combinations of natural and synthetic materials (e.g., Dacron®, GORE ACUSEAL®, GORE-TEX®, and GORE DUALMESH® patches) [1, 2]. Silk fibroin, a protein extracted from *Bombyx mori* silkworm cocoons, is a natural biomaterial that has shown promising clinical translation [3]. Silk fibroin has mechanical, structural, and chemical parameters that can be tuned to allow for a wide scope of final biomaterial formulations [4, 5]. As a result, silk fibroin biomaterials have been used for a variety of different applications. However, silk fibroin biomaterials have been fabricated without concern for adipose tissue deposition within and around the biomaterial. In the treatment of skeletal muscle disorders, an implanted biomaterial that results in excessive adipose tissue deposition would be considered undesirable. As skeletal muscle atrophies, increases in adipose tissue occur, leading to deformation, paralysis, and even death. On the other hand, adipose tissue functions as a protective layer for organs and maintains body contours. Meaning applications for biomaterials that promote adipose tissue deposition exist. Quantifying adipose tissue accumulation for different biomaterial formulations is necessary to

engineer optimal biomaterials for particular clinical applications. One method to quantify adipose tissue accumulation is to hand label adipocyte area in histological stained images. Manually labeling adipocytes is time-consuming, tedious, and prone to error. Specifically, inconsistency in labeling (e.g., different annotators and fatigue) can also introduce bias when identifying regions of interest.

1.2. Deep Learning for Histological Images. To mitigate human bias and increase efficiency, we can train machine learning models to autonomously identify and quantify the adipose tissue in histological sections. Deep learning, a subarea of machine learning, has been applied to histological images for different tasks including classification, regression, and segmentation [6, 7]. Segmentation tasks include identifying regions of interest pertaining to cells, nuclei, glands, tissue, and tumors [6]. Current state-of-the-art approaches for semantic segmentation are encoder-decoder models [8–12]. UNET is an encoder-decoder model that has been used for biomedical (and other) domains [9]. One biomedical application of UNET is for histological image segmentation [6, 13, 14]. UNET has copy and crop paths (i.e., skip connections) that fuse information from the beginning (i.e., encoder) to the end of the network (i.e., decoder) to promote the identification and localization of regions of interest [6, 9]. The copy and crop connections also facilitate improved learning (i.e., mitigate the effect of vanishing gradient) since these models are typically trained via backpropagation [15]. Despite these advantages, the UNET model can lead to increased computational cost by producing repetitive and unnecessary features [16]. Additionally, high-quality datasets are needed to train and evaluate deep learning approaches for histological image segmentation [6].

1.3. Attention Mechanisms. In machine learning, attention mechanisms have been introduced to improve the performance of models for different applications including natural language processing and semantic segmentation [8, 17, 18]. The motivation behind these approaches is to better train the model to focus on the most relevant and important features in the data [8, 17, 19]. Attention mechanisms have been integrated into UNET to improve segmentation performance (e.g., attention UNET) [16]. Generally, attention mechanisms learn how to weight (i.e., place importance) the input features of the data to encode information such as positions in a sequence [19]. However, attention mechanisms can increase computational costs by increasing the number of learnable parameters in the model. We hypothesize that a simpler (i.e., fewer parameters) attention-inspired fusion approach will lead to improved performance while reducing computational constraints (more details in Section 4.3).

1.4. Problem Statement. Determining adipose tissue accumulation around and within degrading biomaterial-based implants allows engineers and scientists to create biomaterials for specific clinical applications. Previous methods to identify adipose tissue in histological images are tedious, laborious, and susceptible to human bias and error. UNET is a rapid, efficient, and high-throughput method that can be used to quantify adipose tissue. As currently constructed, UNET does not directly encode statistical texture information which has improved performance for semantic segmentation [20–22]. To improve the representation of information throughout the convolutional neural network (CNN), statistical texture features can be used to better characterize the spatial distributions of the data, potentially leading to improved performance as shown in other works that incorporate statistical texture information [7, 20, 21, 23].

In this work, acellular silk fibroin-lyophilized sponges were implanted into the lateral pockets of Sprague Dawley rats. After 1-, 2-, 4-, and 8-week postsurgery, the silk sponges and overlaying tissue were excised. The samples were prepped and stained with hematoxylin and eosin (H&E) and Masson's trichrome. To improve the identification and quantification of adipose tissue, we proposed to extract statistical texture features through the use of histogram layer(s) [24] integrated into the baseline architecture, termed Jointly Optimized Spatial Histogram UNET Architecture (JOSHUA). Additionally, we hypothesized that statistical texture features could also be used as an attention approach to weight important information in the network in order to improve the context learned by each model during training, termed JOSHUA+. The overall framework of this study is shown in Figure 1. Our contributions to the field include the following:

(i) First application of convolutional models to distinctly identify and segment adipose tissue from silk fibroin biomaterial implants in histological images

(ii) Novel incorporation of histogram layers to extract statistical texture and spatial information to improve segmentation performance

(iii) New dataset for the community to evaluate histological images for adipose tissue segmentation

(iv) Simple yet efficient attention-inspired approach to reduce total model parameters and computational costs while achieving comparable and/or better performance

(v) JOSHUA+ accurately identifies differences in adipose tissue accumulation in biomaterial formulations over time

(vi) JOSHUA+ decreases the time to find adipose tissue area by 200% compared to manually annotating

2. Results and Discussion

2.1. Data Selection Using UNET. For machine learning approaches, it is necessary to understand how splitting the data into training and validation sets influence the results of the model using standard segmentation assessment metrics as shown in Figure 2 and Supplemental Table S1. To do this, we evaluated the impacts of time and biomaterial conditions to determine how the data influenced the performance of the model. Ideally, we need a model that can be applied to analyze data from any time point and any silk biomaterial formulation. We investigated five

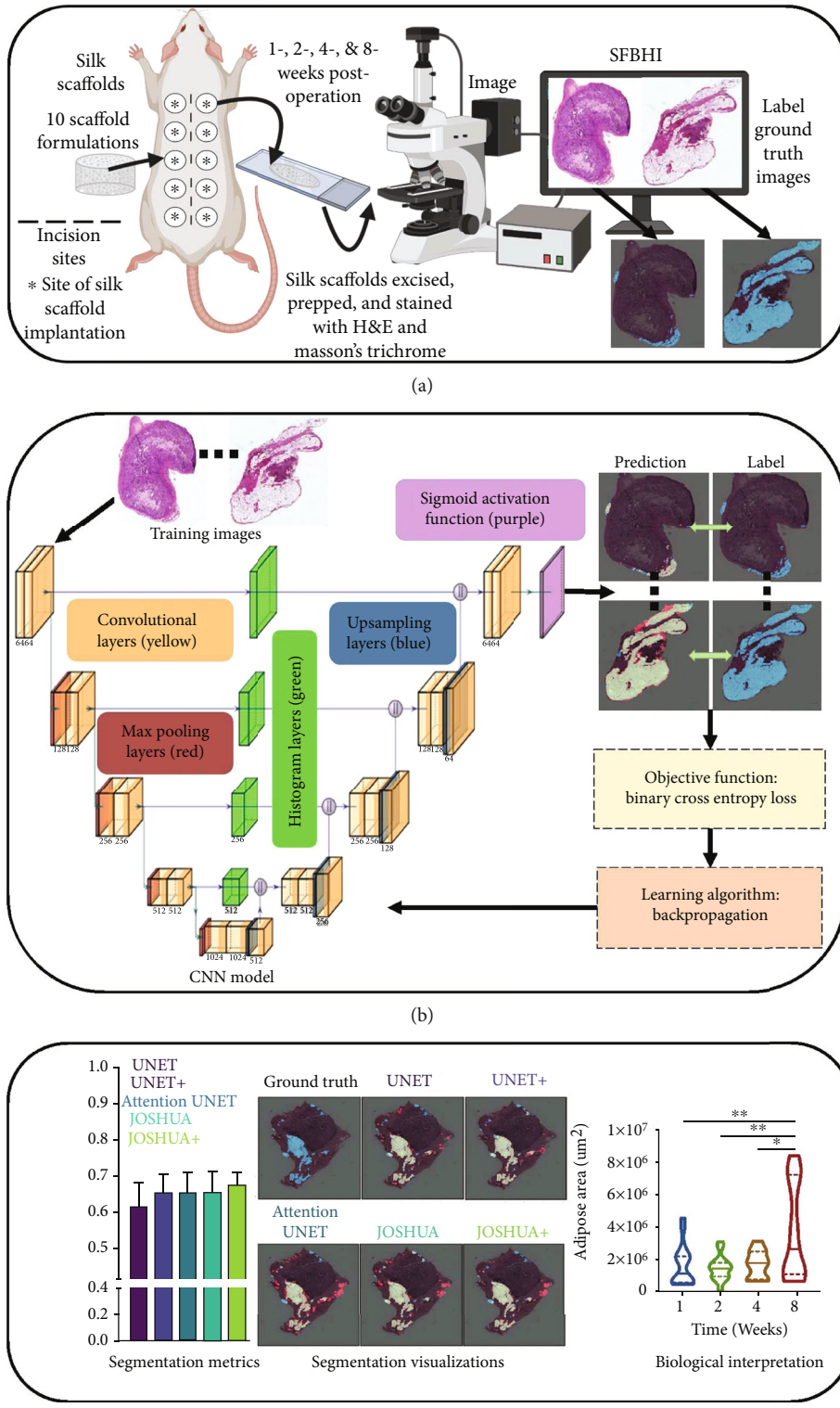

FIGURE 1: Our overall process has three components: data collection, processing, and analysis as shown in Figures 1(a) through 1(c), respectively. (a) For data collection, acellular silk fibroin-lyophilized sponges of varying formulations were subcutaneously inserted into the lateral pockets of Sprague Dawley rats. After 1-, 2-, 4-, and 8-week postsurgery, the silk sponges and overlaying tissue were excised. The samples were prepped and stained with H&E and Masson's trichrome. (b) The next step (i.e., data processing) trains our proposed model through k-fold cross validation. (c) Once training is completed, we can use our machine learning models to quickly segment and quantify the adipose tissue for new samples. We then perform data analysis to connect the model outputs to meaningful biological interpretations.

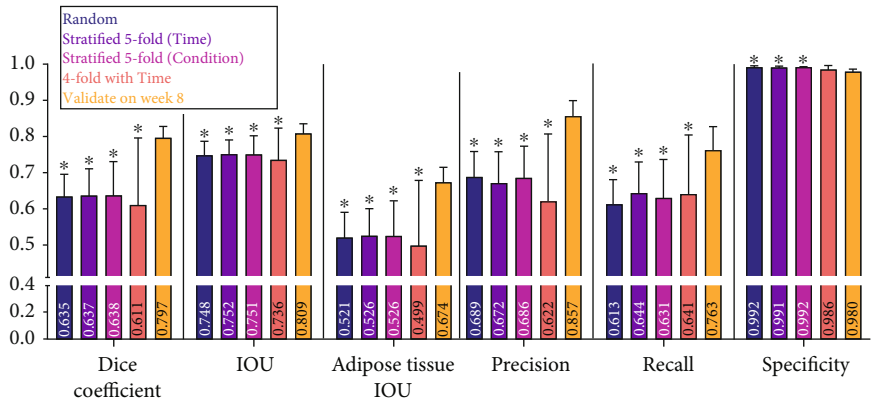

FIGURE 2: Dice coefficient, IOU, adipose tissue IOU, precision, recall, and specificity metrics for data splits, random, stratified 5-fold (time), stratified 5-fold (condition), 4-fold with time, and validate on week 8, on validation images in SFBHI dataset trained on UNET with weighted binary cross entropy. Data are shown as mean ± SD. A one-way analysis of variance followed by Dunnett's multiple comparison test was computed. The asterisks (∗) indicate significant differences as compared to the *validate on week 8* data split ($p < 0.05$).

training and validation splits of our novel *Silk Fibroin Biomaterial Histology Images* (SFBHI) dataset with the baseline UNET model: random, stratified 5-fold (time), stratified 5-fold (condition), 4-fold with time, and validate on week 8. For our SFBHI dataset, we developed *ground truth* for each image by labeling each pixel as either *adipose tissue* or *background*. The ground truth is then used to train the machine learning model to produce predictions based on the input images. We used six metrics to assess segmentation performance on the validation images: Sørensen-Dice coefficient, referred to as dice coefficient, overall intersection over union (IOU), adipose tissue IOU, precision, recall, and specificity. Each metric ranges between 0 and 1. A value of 0 means there is no overlap between ground truth and model prediction, while a value of 1 means the ground truth and model prediction overlap perfectly (details relating to the computation of the segmentation metrics are in Section 4.1.

The five training and validation splits of our novel SFBHI dataset impacted the performance of the baseline UNET model (Figure 2). As detailed in Section 4.1, we used k-fold cross validation. Supplemental Tables S2–S7 detail the percentages of each condition and week used for every data split (random, stratified 5-fold (time), stratified 5-fold (condition), 4-fold with time, and validate on week 8). We found that validating the UNET model on week 8 SFBHI images showed a statistically significant improvement in performance for all metrics except for specificity (Figure 2). Week 8 has the most accumulation of adipose tissue, and as a result, these images are easier for the model to segment and improve metrics such as intersection-over-union (IOU). These models may possibly fail to generalize and perform well when presented with more difficult images (e.g., sparse areas of fat accumulation). This point is further validated by the drop in performance when performing 4-fold cross validation with time. For the three of the four folds, week 8 images are used as training data. However, the overall average is significantly less than validating on week 8 only. Since the model used the "easier" training samples, UNET did not perform as well on the more difficult images with small regions of adipose tissue.

Another interesting observation is that the results of random and stratified (i.e., time and condition) 5-fold cross validation are comparable to one another (Figure 2). The stratified 5-fold cross validation approach considered the biologically meaningful labels, while the random 5-fold cross validation did not. The dataset is relatively balanced across the number of images for each time or condition. In the case of possible data imbalance for future applications, it may be important to leverage biological information (e.g., time and condition) to properly train and validate the model. When comparing the results of stratified 5-fold with time and condition, the model is robust to the use of either biological label. As a result, the model can learn a meaningful mapping between the input and output to identify adipose tissue in either case of labeling.

2.2. Model Comparisons. In practice, we want to deploy the most robust model that will generalize well to both dense and sparse accumulations of adipose tissue. Therefore, we selected the random 5-fold cross validation split to compare the baseline UNET model to our proposed JOSHUA, attention-inspired variants (JOSHUA+ and UNET+), and attention UNET [16] models. The results of the model comparisons are shown in Figure 3 and Table S8. Example qualitative results of segmenting the validation images using our trained models are shown in Figure 3(a). Silk-collagen I (s-c), silk+VEGF$_S$ (s+VEGF$_S$), and silk biomaterials were selected as "easy" data samples with robust adipose presence, whereas additional training was also done on adipose-poor samples such as silk-heparin +VEGF$_S$ (s-h+VEGF$_S$) and s-c-h-VEGF. All models in this work performed well on histological images with large areas of adipose tissue. We observed that the histogram-based models are able to better capture small areas of accumulated fat tissue in comparison to the UNET model as shown on the image from week 8 and condition s-c (Figure 3(a), third row). JOSHUA and JOSHUA+ also

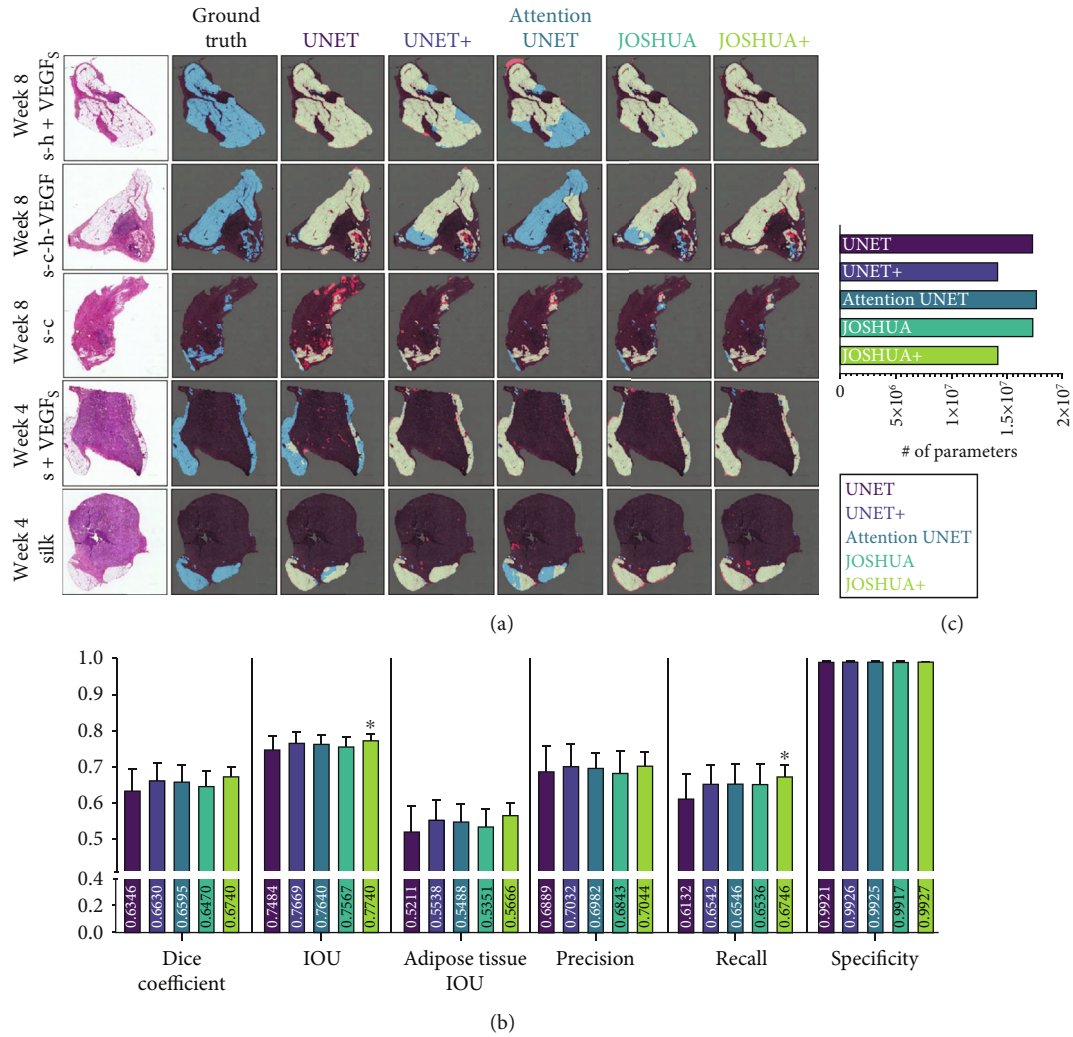

FIGURE 3: (a) Example segmentation results from each model on the SFBHI dataset. The first column displays the input image, and the corresponding ground truth label for the input image is shown in the second column. The remaining columns display the output from each model in comparison to the ground truth. Blue pixels correspond to the ground truth and red pixels correspond to the predicted output. Light green pixels indicate that both the ground truth and predicted output agree. (b) Dice coefficient, IOU, adipose tissue IOU, precision, recall, and specificity metrics for each model for SFBHI dataset. As shown here, the histogram models (JOSHUA/JOSHUA+) and UNET+ improve segmentation results compared to the baseline UNET and attention UNET [16] models. Metrics are shown as mean ± SD. A one-way analysis of variance followed by Dunnett's multiple comparison test was computed. The asterisks (∗) indicate significant differences as compared to UNET ($p < 0.05$). (c) We also evaluated the number of learnable parameters for each CNN model. The attention-inspired variants have approximately 18% fewer parameters than their counterparts.

reduce the number of false positives in comparison to attention UNET for the image from week 8 and condition s-h+VEGF$_S$ (Figure 3(a), first row). The histogram layers integrated into the models demonstrate the effectiveness of statistical texture features for biomedical applications as shown throughout the literature [7, 21, 23].

For our work, we note that each of the models achieved similar performance for the different metrics such as precision ($p > 0.85$). Interestingly, the recall metric for JOSHUA+ statistically improved over UNET alone ($p < 0.02$), along with the value for intersection over union (IOU, $p < 0.04$). For medical image segmentation, particularly for imbalanced datasets (i.e., the class of interest is in low abundance such as in our case), models usually produce high precision and low recall [25–27]. This observed difference in recall further verifies our hypothesis that statistical texture features are beneficial for capturing adipose tissue accumulation and have the potential of improving the precision-recall tradeoff. The average precision and specificity of JOSHUA are slightly lower than the baseline UNET model, indicating a possible overestimation of adipose tissue due to increased sensitivity (i.e., recall) with the addition of histogram layers. JOSHUA could be further improved in later work by increasing the local window size of the histogram layers to improve the discrimination of adipose tissue and background. By increasing the window size, the model can capture more pixel and/or feature values to estimate distinct distributions between the background and adipose tissue.

We found improved performance with the attention-inspired variants of each model, UNET+ and JOSHUA+. Our new method of fusing features from the encoder and decoder components of the architecture encourages joint agreement between features at multiple levels in the network. By encouraging this joint agreement, we improved the mean value for each metric in comparison to the baseline UNET model. This result demonstrates that our fusion approach is beneficial for both the positive (i.e., adipose tissue) and negative (i.e., background) classes. JOSHUA+ is the only model to achieve a statistically significant difference for two metrics, IOU and recall. For overall IOU, JOSHUA+ improves the identification and localization of not only the adipose tissue but also the background information. JOSHUA+ also shows that statistical texture features are more informative as an attention mechanism in comparison to the convolutional feature maps of UNET+. UNET+ and JOSHUA+ have a higher average than attention UNET for most metrics. Our proposed attention-inspired models have fewer learnable parameters than attention UNET as shown in Figure 3(c). As a result, the storage requirements for the trained attention models will be reduced and the fusion operation (i.e., elementwise multiplication) is quicker to compute than the concatenation approach of UNET, attention UNET, and JOSHUA.

Example qualitative results of segmenting the validation images using our trained models are shown in Figure 3(a). All models perform well on large areas of adipose tissue in the histological images. However, all models presented in this study struggled with adipose-poor images (i.e., images with less than 1% of adipose pixels), and this led to a decrease in the overall segmentation metrics. In our SFBHI dataset, approximately 17% of the dataset are adipose-poor samples (20 images). We show example fail cases in Supplemental Figure S4. For adipose-rich samples, the histogram-based models are able to better capture small areas of accumulated fat tissue in comparison to the UNET model as shown in the image from week 8 and condition s-c (Figure 3(a), third row). JOSHUA and JOSHUA+ also reduce the number of false positives in comparison to attention UNET for the image from week 8 and condition s-h+VEGF$_s$ (Figure 3(a), first row). The visualizations are shown to demonstrate the ability of statistical texture features to refine the identification and segmentation of adipose tissue. To provide more insight into the incorporation of statistical textures, we computed the color histograms of the SFBHI dataset in Supplemental Figure S2. The distributions of the adipose tissue and background pixels are different from one another. As a result, all of the models in this study will be able to correctly classify most pixels in the images, which led to limited statistical differences for most metrics in Figure 3(b). However, there is some overlap with the pixel intensity values. In these instances, the statistical texture features are needed to improve performance on these more difficult samples in the data as shown in the segmentation results in Figure 3(a).

2.3. Application of Models to Benchmark Histological Dataset. To demonstrate the effectiveness of the proposed model to be generalized to related segmentation tasks, we evaluated each model with a benchmark dataset, Gland Segmentation in Colon Histology Images (GlaS), for cancerous histological segmentation. The results of the experiments are shown in Figure 4 and Table S9.

All models in this work perform comparatively with the baseline UNET architecture with no significant differences (Figure 4(b)). JOSHUA achieves the highest average for most metrics except for specificity. JOSHUA, JOSHUA+, and UNET+ improved the sensitivity (i.e., recall) over the UNET model. There is a large improvement in each metric (except specificity) for every model in comparison to the SFBHI dataset. The proposed approaches (JOSHUA/JOSHUA+) also have less variability (i.e., small standard deviations) than the UNET model across the metrics.

Overall, JOSHUA achieves the highest average for most metrics except for specificity. The GlaS dataset has a "zoomed in" view of the histological images resulting in reduced background from the histological slides and higher percentage of positive pixels in the images than the SFBHI dataset. Therefore, we see a large improvement in each metric (except specificity) for every model in comparison to the SFBHI dataset due a reduction in false positives as a result of the slide background. An important note here is that JOSHUA, JOSHUA+, and UNET+ improved the sensitivity (i.e., recall) over the UNET model. This result further validates that our histogram layers and fusion approaches better identify the class of interest (i.e., cancerous or adipose tissue). Our proposed approaches also have less variability (i.e., small standard deviations) than the UNET model across the metrics. We also calculated the color histograms of the GlaS dataset in Supplemental Figure S3. The distributions of the cancerous tissue and background pixels are more similar in comparison to the distributions of adipose tissue and background pixels of the SFBHI dataset. For the GlaS dataset, the concatenation approach of JOSHUA was more effective than the elementwise multiplication of JOSHUA+ in terms of achieving better segmentation metrics. The performance of JOSHUA+ and UNET+ is not different by a considerable margin as there is no statistically significant difference between the DICE scores ($p = 0.0629$). For the GlaS dataset, JOSHUA performed better for the segmentation metrics. This result indicated that if the class of interest (i.e., adipose or cancerous tissues) has a more similar distribution of intensity values to the background, JOSHUA will be more effective than JOSHUA+.

Qualitatively, we see similar results to the SFBHI dataset. Each model performs well for large cancerous regions of the image. However, the UNET+ and attention UNET model detect more falsely identified cancerous glands for images in the third and fourth row of Figure 4(a)). For the second row, all models (except for UNET+) have difficulty with detecting the central area of the cancerous gland on the bottom right-hand side of the image. In comparison to the other images, this gland has the most abrupt changes in terms of statistical texture and color information. As stated in the SFBHI section, the window size of the histogram layer(s) can be modified to better characterize the distribution of cancerous and

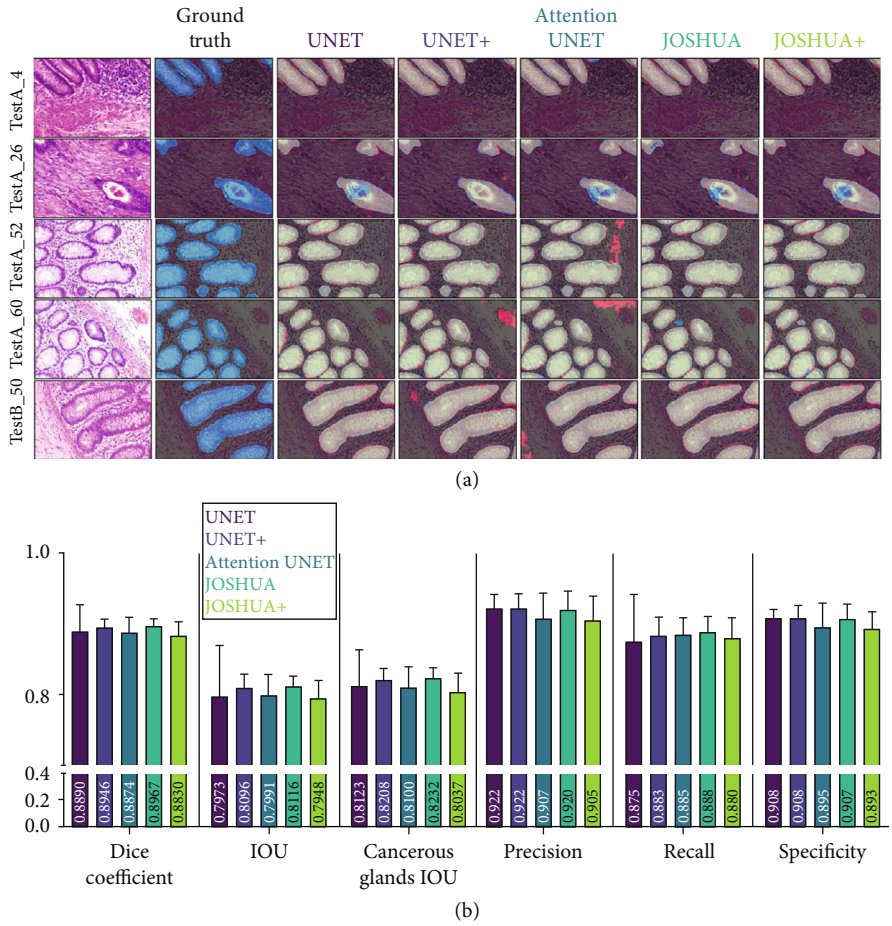

FIGURE 4: (a) Example segmentation results from each model on the GlaS dataset. The first column displays the input image, and the ground truth label for the input image is shown in the second column. The remaining columns display the output from each model in comparison to the ground truth. Blue pixels correspond to the ground truth, and red pixels correspond to the predicted output. Light green pixels indicate that both the ground truth and predicted output agree. As shown here, all models perform comparatively well for this dataset. In some instances, the models with the histogram layers (i.e., JOSHUA and JOSHUA+) reduce the false positives identified by the other models. (b) Dice coefficient, IOU, cancerous glands tissue IOU, precision, recall, and specificity metrics for each model for GlaS dataset. Metrics are shown as mean ± SD. A one-way analysis of variance followed by Dunnett's multiple comparison test was computed. The asterisks (∗) indicate significant differences as compared to UNET ($p < 0.05$).

noncancerous glands to account for images with similar characteristics. The UNET+ model has the fewest number of parameters. This can lead to improved generalization ability in distinct cases such as the "TestA_26" image.

2.4. Applying Model on New Images. To investigate the use of JOSHUA+ to identify the differences in adipose tissue accumulation based on varying biomaterial formulations over time, we took the best JOSHUA+ model (i.e., random initialization three and fold three) to plot the adipose area estimation while considering condition and time (Figure 5). In addition to the original 117 images, we applied our model to 465 new images that were not labelled. Figure 5(a) shows segmentation results from both H&E and Masson's trichrome stained hold-out images. We observed that JOSHUA+ falsely identified adipose tissue for some images (i.e., week 2 H&E and week 1 Masson's trichrome). We expected this would occur because lyophilized silk fibroin sponges can have spherical pores that can be mistaken for spherical adipocytes. JOSHUA+ could be used as a "pre-processing" step to initially identify adipose tissue (or other regions of interest) in tandem with human annotators to ease the burden of labeling images. Figure 5(b) is a graph of adipose area over time for the representative images shown in Figure 5(a). The area of adipose tissue at week 8 was significantly higher than the adipose area at weeks 1, 2, or 4 for the s-c+VEGF$_S$ biomaterial formulation. Figure 5(c) shows the difference in adipose area across 10 biomaterial formulations for week 4. The s-c condition is the only biomaterial formulation to show a significant increase in adipose area at week 4 compared to silk biomaterials. These are important findings as we can produce statistical differences in adipose area over time with varying biomaterial formulations. Future work is needed to determine adipose tissue area with respect to section and scaffold areas to identify biological interpretations of adipose tissue as a function of time and biomaterial formulation.

Figure 5(d) shows the time in seconds to quantify the adipose area for one image for both manually labeling and

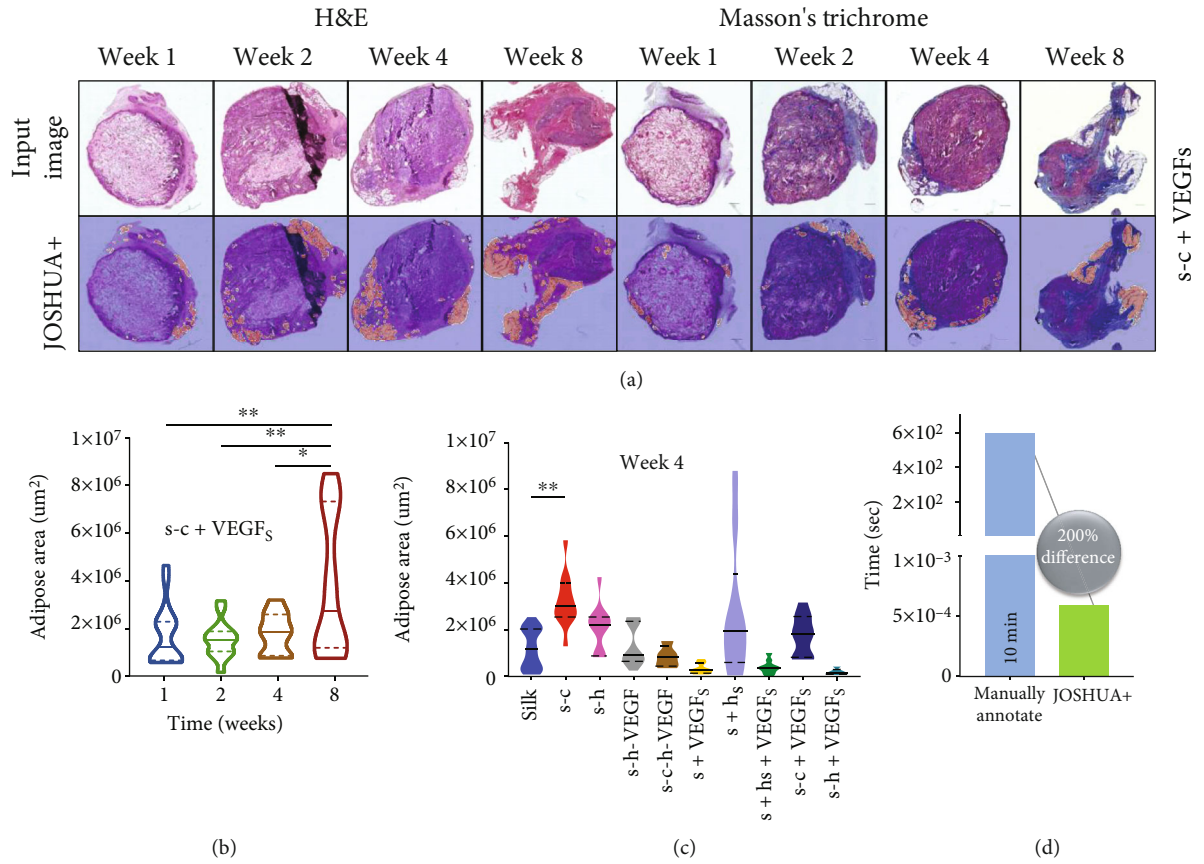

FIGURE 5: (a) SFHBI holdout images and their segmentation results using JOSHUA+. H&E images and their corresponding segmentation results are shown in columns 1-4. Masson's trichrome images and their corresponding segmentation results are shown in columns 5-8. (b) Quantification of adipose area over time for s-c+VEGF $_S$ biomaterial formulation. (c) Adipose area across 10 biomaterial formulations for week 4. (d) The time to manually annotate the adipose area or use JOSHUA+ for one image. There is a 200% difference in time to use JOSHUA+ rather than manually labeling and quantifying adipose area. For Figures 5(b) and 5(c), data are shown with a solid line representing the median and the two dashed lines representing the 25th and 75th percentile quartiles. The Kruskal-Wallis test followed by Dunn's multiple comparison test was computed. The asterisks (∗) indicate significant differences ($p < 0.05$).

using JOSHUA+. This new strategy enables the processing of much larger datasets compared to the existing methods with significantly faster computation speed by leveraging recent advances in deep learning. A limitation of our current work is that the dataset is relatively small for machine learning approaches. As a result, additional images would be beneficial to train and verify the model for improved performance. However, this dataset is relatively large for biomaterial scientists and engineers. Another constraint is that the model was trained using ground truth labels from one annotator. These labels will have some bias; therefore, our method will also benefit by using labels from multiple expert annotators to further improve the robustness and accuracy of the models. In summary, we successfully trained and validated the model with a subset of the data (117 images) and then used JOSHUA+ to analyze a full dataset of 465 images.

3. Conclusion

We present the first investigation into using CNNs to analyze adipose tissue accumulation over time around and within degrading silk-based biomaterials following implantation. We quantified adipose tissue deposition as a function of biomaterial composition and the time postimplantation. Acellular lyophilized silk fibroin sponges were implanted subcutaneously into Sprague Dawley rats. After 1-, 2-, 4-, and 8-week postsurgery, the silk sponges and surrounding tissue were excised at each time point. The samples were fixed, wax embedded, sectioned, and finally stained with H&E or Masson's trichrome. We used CNN models with novel spatial histogram layer(s) to identify and segment regions of adipose tissue in the histological images, allowing for evaluation of the impact biomaterial formulation has on local adipose tissue accumulation. We compared our model, JOSHUA, to the baseline UNET model and an extension of the baseline model, attention UNET, as well as to versions of the models with a supplemental attention-inspired mechanism (JOSHUA+ and UNET+). Our results demonstrate that the developed models, JOSHUA and JOSHUA+, are highly beneficial for adipose tissue identification and localization within and around silk fibroin-lyophilized sponges. Taking inspiration from UNET++ [28] strategies, future work that fuses features from multiple levels in the network within JOSHUA may be a useful addition in situations

where features of interest are in low abundance as a function of total image area and where the clinical application requires that the result have zero margins for error, given the critical implications for the clinical outcome. In this situation, adipose tissue is not cancerous or implicated in patient mortality; adipose tissue in low levels would be normal in muscle tissue—our goal is to identify biomaterial formulation impact on adipose tissue deposition. The main outcome of this work is that we can independently identify the adipose tissue from the biomaterial scaffold with reduced computational load and we are able to draw conclusions about biological mechanisms and results in on-going work. To our knowledge, this study is the first investigation to apply CNN models to identify and segment adipose tissue in histological images from silk fibroin biomaterial implants. It is also the first step in developing automated analysis to guide the formulation of "adaptable" biomaterials that are able to meet the needs of clinicians, surgeons, and patients. Future work was aimed at determining the components of the biomaterial formulation that drive adipose tissue deposition following surgical implantation. We will also explore methods for further optimization by tuning hyperparameters (e.g., window size and the number of bins) and applying weakly supervised learning approaches to reduce the cost of acquiring precise, pixel-level ground truth labels.

4. Materials and Methods

4.1. Experimental and Technical Design

4.1.1. Description of Datasets. In the SFBHI and GlaS datasets, there are 117 and 165 images, respectively. Example images are shown for the SFBHI and GlaS datasets in Figures S1a and S1b. The SFBHI dataset has images at various scales and illuminations, as shown in Figure S1a. Ground truth labels were created for our SFBHI dataset using Hasty [29], an image annotation software. Regions of adipose tissue accumulation in 117 histological images of various silk scaffold conditions were labeled by hand using the software's semantic segmentation feature. An extension of our SFBHI dataset had a total of 465 images that were unlabeled and had two stainings: H&E and Masson's trichrome. These new images were used to evaluate the trained model (i.e., JOSHUA+) for quantifying and analyzing adipose tissue across time and biomaterial formulation.

4.1.2. Evaluation and Training Protocol. For both datasets, we adopt a fivefold cross validation strategy. There were a total of 94 (93 for folds three through five) training images and 23 (94 for folds three through five) validation images for SFBHI in each fold. The training (67 images for folds one through four, 72 for fold five) and validation (18 images for folds one through four, 13 for fold five) data splits from Rony et al. [30] were used for the GlaS dataset. For each dataset, we ran three runs of random initialization for a total of 15 experimental runs. The results shown for SFBHI are averaged across the validation folds for each model, while the GlaS results are average performance metrics on the holdout test evaluated by each model trained and validated on each fold. Convolutional models are dependent on scale [31], so we used bilinear interpolation to resize the SFBHI images to be the same resolution (256×256). The GlaS images did not need to be resized since all images were the same resolution (775×522). We compare a total of five models: baseline UNET [9], UNET+, attention UNET [16], JOSHUA, and JOSHUA+.

To improve the performance and robustness of each model, we followed a similar data augmentation procedure to Rony et al. [30]. The images are randomly flipped horizontally ($p = 0.5$) and rotated by 90° increments (i.e., 0°, 18 0°, and 270°). Random color jittering (i.e., changes in brightness, contrast, and saturation at $p = 0.5$ and hue at $p = 0.05$) is also incorporated in the training images. The images are also iterated over eight times within each minibatch to further artificially increase the size of the training dataset. Random crops of 416×416 are extracted during training for the GlaS dataset. The data splits and random initialization (three random seeds) of each model were fixed, so we could do a fair comparison across the different architectures. Each model was trained for 150 epochs while using Adam [32] optimization with an initial learning rate of 0.001, weight decay 1×10^{-8}, and the gradient values were clipped at 0.1. Early stopping was also implemented to prevent overfitting on the training data. The early stopping epochs were set to 10 for SFBHI and 20 for GlaS. Batch sizes of four and eight were used for GlaS and SFBHI, respectively. Experiments were run on two NVIDIA GeForce RTX 2080TI GPUs. We used binary cross entropy as the loss function to train the models. The SFBHI dataset is heavily imbalanced as some images do not contain much adipose tissue. To counter this, we used weighted binary cross entropy and set the weight in the loss function to three (determined empirically).

In our work, we looked at using stratified and nonstratified cross validation (Section 2.1). For stratification, the division of the data is based on maintaining a similar distribution of labels in both training and validation folds. For our histological images, we have two associated labels: time (1-, 2-, 4-, and 8 -week postsurgery) and condition (silk, silk-collagen I, silk-heparin, silk-heparin-VEGF, silk-collagen I-heparin-VEGF, silk+VEGF_S, silk+heparin_S, silk+heparin_S+VEGF_S, silk-collagen I+VEGF_S, and silk-heparin+VEGF_S). We first used a data split based on a random partition of 5-fold cross validation that did not consider the global labels (i.e., time and condition) associated with each image as in the stratified examples. The distribution of global labels are shown in Supplemental Tables S2 and S3. Using stratified cross validation, we had two data splits: stratified 5-fold (time) and stratified 5-fold (condition) (see Supplemental Tables S4 and S5, respectively). The final two data splits used the time information to divide the data based on the weeks after surgery. In this instance, each fold represents a specific week (e.g., train on 1-, 2-, and 4-weeks, validate on week 8). Since there are only four distinct time points, 4-fold cross validation was used for this data split (i.e., 4-fold with time) (Supplemental Table S6). Lastly, we also experimented with only validating on week 8 images (Supplemental Table S7). By using this data split, we wanted to evaluate if the model(s) could learn sequentially (i.e., train on earlier weeks and predict later weeks).

4.1.3. Evaluation Metrics. There are several common metrics to assess semantic segmentation performance. The first is pixel accuracy, which measures how well the model labels each pixel to the corresponding class. For our segmentation work, we have a two possible classes: positive (i.e., adipose and cancerous tissue) and negative (i.e., background). With these two classes, we have four possible outcomes: true positive (i.e., positive samples correctly identified as positive), false positive (i.e., negative samples incorrectly identified as positive), true negatives (i.e., negative samples correctly identified as negative), and false negatives (i.e., positive samples incorrectly identified as negative). Pixel accuracy is defined as the sum of true positives (TP) and negatives (TN) divided by the sum of all pixels in the image as shown in

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{1}$$

where FP is false positive and FN is false negative. However, pixel accuracy is not an optimal metric to use when there is a severe class imbalance (such as with the SFBHI dataset for adipose tissue vs. background). Therefore, we need other metrics to provide more insight into each model's performance. Three other measures are precision, recall, and specificity. Precision quantifies the correctly identified positive samples, recall (i.e., sensitivity) is how well the model can identify positive samples, and specificity is how well the model identifies true negative samples. Precision, recall, and specificity are defined in

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}. \end{aligned} \tag{2}$$

There is a tradeoff between precision and recall (i.e., as one increases, the other will decrease). To find the balance between the two measures, the F1 measure can be used. For binary problems, the F1 score is equivalent to the Dice coefficient [33] and is shown in

$$\text{Dice} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

A related measure to the Dice score is Intersection Over Union (IOU), also known as the Jaccard index. IOU is defined as the intersection of the model prediction, $\widehat{Y}$, and the true label, Y, divided by the union of the two sets:

$$\text{IOU} = \frac{\left|Y \cap \widehat{Y}\right|}{|Y| + \left|\widehat{Y}\right| - \left|Y \cap \widehat{Y}\right|}. \tag{4}$$

For IOU, we report the results on the positive class or regions of interest as well as the overall IOU (background and positive classes).

4.1.4. Quantification of Adipose Tissue. In addition to evaluating segmentation performance, we also wanted to compute the measurements of adipose tissue found by the machine learning models. To train and evaluate the models, we needed to resize the images to a smaller size as stated previously. In order to scale the amount of adipose tissue area (μm^2) in the reduced images, we used the computation in Equation (5) where the FA_{full} is the full resolution fat pixel area, FA_{down} is the downsampled fat pixel area, IA_{full} is the full resolution image pixel area, IA_{down} is the downsampled image pixel area, and RL is the reference length from the Keyence Analysis software (μm/px):

$$\text{FA}_{\text{full}} = \text{FA}_{\text{down}} \times \frac{\text{IA}_{\text{full}}}{\text{IA}_{\text{down}}} \times \text{RL}^2. \tag{5}$$

4.2. Dataset Collection: Silk Fibroin Biomaterial Histology Images (SFBHI)

4.2.1. Silk Fibroin Extraction. As previously described, silk fibroin solution was prepared through isolation from *Bombyx mori* silkworm cocoons [5]. Five grams of silk cocoons was degummed in 2 L of boiling 0.02 M sodium carbonate (Catalog No. 451614, Sigma-Aldrich, St. Louis, MO) for 30 minutes in order to isolate pure silk fibroin protein. After degumming, the silk fibroin protein was left to air dry in the fume hood for at least 48 hours. The dried silk fibroin protein was then solubilized by denaturation in 9.3 M aqueous lithium bromide (Catalog No. 213225, Sigma-Aldrich, St. Louis, MO) at 60°C for 4 hours. Using a 3.5 kDa molecular weight cutoff dialysis membrane tubing (Catalog No. 08-670-5C, Spectrum™ Spectra/Por™ 3 RC Dialysis Membrane Tubing, 3,500 Dalton MWCO, Thermo Fisher Scientific, Waltham, MA), the solution was dialyzed in deionized water to remove the lithium bromide ions. This produces a solubilized silk solution that was centrifuged twice (>2000 g, 20 min, 4°C) to remove insoluble particles. The silk solution concentration was determined by placing a wet volume of the silk solution in the oven at 60°C and massing the final dry weight. This protocol resulted in a final solution between 5 and 7% weight per volume (wt./v, 0.05-0.07 g/mL) silk solution. Silk solutions were stored at 4°C for up to 3 weeks prior to use in making silk sponges.

4.2.2. Type I Collagen Isolation. All animal experiments were executed under protocols approved by both Tufts University's and University of Florida's Institutional Animal Care and Use Committees and in accordance with the Guide for the Care and Use of Laboratory Animals (NIH, Bethesda MD). As previously described [34], ready-to-use reconstituted type I collagen was prepared. Briefly, rat tail type I collagen was isolated from the tails of adult Sprague Dawley rats. Following lyophilization, the resulting collagen sponge-like material was solubilized in 0.02 N acetic acid.

4.2.3. Scaffold Formation. Two methods of silk scaffold formation were pursued: prefabrication method and postfabrication method. The prefabrication method starts with incorporating heparin (Catalog No. H3149-250KU), heparin

sodium salt from porcine intestinal mucosa (250KU, Sigma-Aldrich, St. Louis, MO, 100 g/mL), type I collagen (200 g/mL), and vascular endothelial growth factor (VEGF) (Catalog No. 100-20, Recombinant Human VEGF165, Peprotech, Rocky Hill, NJ, 0.1 g/mL). A mixture of heparin, collagen I, and VEGF was created prior to being added to diluted aqueous silk solution (3% weight per volume (wt./v, 0.03 g/mL)). Following published protocols [4], the silk solution was poured into wells of a 6 well plate and frozen in -80 °C freezer overnight, forming isotropic silk scaffolds. Sponges were lyophilized at -80 °C and 0.185 mbar for 5-7 days (FreeZone 12 Liter -84°C Console Freeze Dryer, Labconco, Kansas City, MO) postfreezing. To make the sponge-like scaffolds water insoluble, the scaffolds were water annealed for 12 hours at room temperature to induce β-sheet formation (silk fibroin protein crystallization) [35, 36]. Sponge-like scaffolds were cut using a 6 mm diameter biopsy punch to a create cylinder 3 mm in height. The post-fabrication method is described by pouring the diluted aqueous silk alone into wells of a 6 well plate and freezing the solution in -80 °C freezer overnight. This is again followed by lyophilizing the silk sponge at -80 °C and 0.185 mbar for 5-7 days (FreeZone 12 Liter -84°C Console Freeze Dryer, Labconco, Kansas City, MO). These silk sponges were also made water insoluble by water annealing for 12 hours at room temperature to induce β-sheet formation (silk fibroin protein crystallization) [35, 36]. Prior to trimming the scaffolds, VEGF and heparin were solubilized where VEGF_S and heparin_S designate soluble VEGF and heparin added postfabrication, respectively. Scaffolds were then soaked in solubilized solutions of VEGF_S and/or heparin_S for 30 minutes. Sponge-like scaffolds were cut using a 6 mm diameter biopsy punch to a create cylinder 3 mm in height.

4.2.4. In Vivo Subcutaneous Procedures. All procedures were conducted under animal protocols approved by Tufts Institutional Animal Care and Use Committee. All animals used in this study were 7-8-week-old Sprague Dawley rats (~ 225 g, Charles River Laboratories, Wilmington, MA). Following general anesthesia of oxygen and isoflurane, the back of the rat was shaved, and the skin was prepped using three sets of betadine disinfectant followed by an alcohol wipe. Sterilized acellular silk scaffolds (6 mm diameter × 3 mm height) were subcutaneously implanted in lateral pockets of each rat through up to three small longitudinal incisions made through the skin. The incisions were then closed using surgical clips. The animals were euthanized at 1-, 2-, 4-, and 8-week postsurgery. The silk scaffolds (along with the overlying tissue) were then excised and collected for histological examination.

4.2.5. Histological Staining of In Vivo Samples. After a series of graded ethanol and xylene incubations, the samples were fixed with 10% phosphate buffered formalin (Thermo Fisher Scientific, Waltham, MA) and embedded in paraffin. The samples were sectioned to 7-15 m thickness and deparaffinized before staining with hematoxylin (Catalog No. SDGHS280, Thermo Fisher Scientific, Waltham, MA) and eosin (Catalog No. SDHT1101128, Thermo Fisher Scientific, Waltham, MA) or Masson's trichrome (Catalog No. HT15, Sigma-Aldrich, St. Louis, MO). Hematoxylin stains cell nuclei in a dark purple and blue color, while eosin stains the extracellular matrix and cytoplasm pink. Masson's trichrome is a multicolor staining technique. Keratin and muscle fibers are stained in red. Collagen and bone are stained in blue or green. Cytoplasm is stained in light red and pink. Cell nuclei are stained in dark brown or black. The samples were embedded in DPX Mountant (Sigma-Aldrich, St. Louis, MO) following staining and dehydration and then imaged using a Keyence BZ-X800 series microscope at 20x magnification.

4.3. Jointly Optimized Spatial Histogram UNET Architecture (JOSHUA). The proposed model, Jointly Optimized Spatial Histogram UNET Architecture (JOSHUA), is shown in Figure 6. The model integrates histogram layers [24] in the architecture along the copy and crop connections. By adding the histogram layers in these locations, we are able to include statistical texture information from the encoder and concatenate these features in the decoder to improve the context provided by the "binned" convolutional feature maps. Given D number of convolutional features maps of height M and width N from the encoder, $X_{\text{encoder}} \in \mathbb{R}^{M\times N\times D}$, the K number of B-bin histogram representations of height R and width C, $H \in \mathbb{R}^{R\times C\times B\times K}$, is computed using Equation (6) with a sliding window of size $S \times T$:

$$H_{rcbk} = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} e^{-\gamma_{bk}^{2}\left(x_{r+s,c+t,k}-\mu_{bk}\right)^{2}}, \tag{6}$$

where μ_{bk} and γbk are the bin center and widths, respectively, of the b^{th} histogram bin for the k^{th} input feature map. After the convolutional features are binned (mapped to the range of [0, 1]), we aggregate the average counts for each bin across the spatial dimensions of the feature maps to compute the local histograms ($S = 2$, $T = 2$). The bin centers and widths of the histogram layers are updated via back-propagation during training in order to improve the representation of the statistical texture features.

The architecture for our proposed model is shown in Figure 6. For the upsampling layers, we used bilinear upsampling instead of transposed convolution. We incorporated a total of four histogram layers along the copy and crop connections in the model. The histogram features maps are then concatenated onto the feature maps from the decoder, $X_{\text{decoder}} \in \mathbb{R}^{R\times C\times D}$, at each level of the network. In order to have an equal number of feature maps as in the baseline UNET model, we added $1 \times 1 \times C$ convolutional layers to reduce the input feature dimensions (i.e., constrain the product of the number of bins, B, and reduce input feature maps, K, to equal the original number of input feature maps, D). For example, in the first level of the network, the number of convolutional feature maps from the encoder is 64. We set each histogram layer to have a total of 16 bins. The input feature maps are then reduced to a total of four to produce 64 histogram feature maps (i.e., four input convolutional maps produce 16 histogram representations each for a total of 64 histogram feature maps).

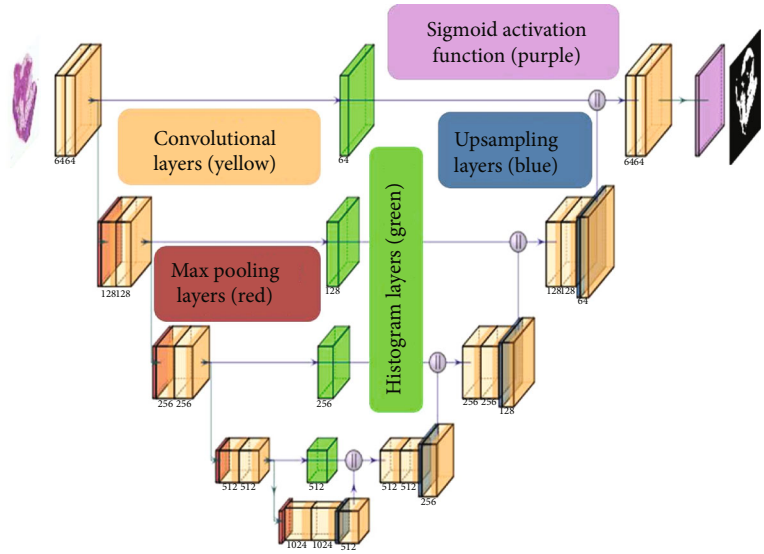

FIGURE 6: The JOSHUA model is comprised of convolutional (with ReLU activations) (yellow), max pooling (red), and upsampling layers (blue) as in the baseline UNET model [9]. Our proposed method integrates histogram layers (green) along the copy and crop connections to transform the input convolutional features maps to their statistical texture representations. The histogram feature maps are then concatenated with the features maps in the decoder . We then apply a sigmoid activation function (purple) on the output layer and threshold the output to distinguish between the positive (i.e., adipose tissue) and negative classes (i.e., background).

4.4. Attention-Inspired Models: JOSHUA+ and UNET+. In addition to the histogram model, we also propose alternative models to JOSHUA and UNET: JOSHUA+ and UNET+. Instead of concatenating the feature maps from the encoder and decoder layers, we perform an elementwise multiplication between two encoder and decoder feature maps. Through this approach, we fuse features in such a way that we have a joint agreement between the features from the beginning and ending (similar to an *AND-ing* operation). By performing this elementwise multiplication, the number of feature maps is reduced by half, leading to a decrease in the number of parameters in the decoder branch of the model. Our approach is inspired from the *Scaled Dot-Product* self-attention method [19]. Attention functions map a query to an output given a set of key-value inputs [19]. A weight for each value based on the score from a compatibility function on the query and a given key is used to compute a weighted sum of values for the output. Given matrices of queries (Q), keys (K), and values (V), self-attention is computed by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)V, \tag{7}$$

where d_k is the dimensionality of each key vector. The compatibility function shown here is the matrix product between the queries and keys normalized by factor of $1/\sqrt{d_k}$. The softmax function maps these values to a range of $[0, 1]$ to indicate the similarity (i.e., compatibility) of the queries and keys.

As noted by Vaswani et al. [19], the dot-product attention mechanism is computationally less expensive than the additive attention mechanism in terms of speed and space efficiency. For JOSHUA+, the compatibility function will be the histogram layer output, H, that maps input feature maps, X_{encoder}, to a range of $[0, 1]$ based on the bin centers and widths that are estimated in the network. The decoder feature maps, X_{decoder}, serve as our *values* (i.e., V) to perform the elementwise multiplication between H and X_{decoder} to compute our attention outputs as shown in

$$\text{Attention}_{\text{JOSHUA}}(H, X_{\text{decoder}}) = HX_{\text{decoder}}. \tag{8}$$

Our approach has an advantage over attention models such as attention UNET [16] in that these models have less learnable parameters and reduced computational complexity. JOSHUA+ also results in a more constrained attention mechanism due to elementwise multiplication between H and X_{decoder}. Through this operation, we encourage the decoder feature maps to be tied more directly to encoder feature values that have larger compatibility scores to a corresponding bin (i.e., enforce statistical similarities in the data). To evaluate the impact of using a normalized compatibility function (*i.e.*, $\text{softmax}(\mathbf{QK}^T/\sqrt{d_k})$ and H), we also perform a similar operation for UNET+ except that we compute the element-wise product between the encoder and decoder feature maps (i.e., compatibility function in this case is the identity function applied to the encoder feature maps).

4.5. Statistical Analysis. For Figures 2–4, data are shown as mean ± SD. A one-way analysis of variance (ANOVA) followed by Dunnett's multiple comparison test was computed. For Figures 5(b) and 5(c), data are shown with a solid line representing the median and the two dashed lines representing the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile quartiles. The Kruskal-Wallis test followed by Dunn's multiple comparison test was

computed because the data was not normally distributed. The asterisks (∗) indicate significant differences ($p < 0.05$).

Data Availability

The histological dataset and code for our experiments are publicly available: https://github.com/GatorSense/Histological_Segmentation. We have full resolution and downsampled (256 × 256 pixels) images available on the code repository.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Authors' Contributions

JK Peeples and JF Jameson conceived the idea. JK Peeples, JF Jameson, WL Stoppel, and A Zare planned and designed the research. WL Stoppel and JM Grasman executed the animal work. JF Jameson and NM Kotta collected and labelled the data. JK Peeples, JF Jameson, and NM Kotta performed the analysis. JK Peeples, JF Jameson, and NM Kotta wrote the manuscript. WL Stoppel and A Zare reviewed and edited the manuscript. All authors read and approved the final manuscript. Joshua K. Peeples and Julie F. Jameson contributed equally to this work.

Acknowledgments

The Stoppel Lab thanks the undergraduate students that helped with silk solution preparation and the initial startup of the lab at the University of Florida and the undergraduate students that started the initial work at the Tufts University, especially Elizabeth C. Bender who is a current PhD student at the University of Texas at Austin. Figure 1 was created with a license to http://BioRender.com/. The authors acknowledge the University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication (URL: http://researchcomputing.ufl.edu). JK Peeples recognizes support from the National Science Foundation Graduate Research Fellowship (DGE-1842473). The authors thank the NIH Tissue Engineering Research Center (P41 EB002520) for support of this work in its initial stages. WLS acknowledges support from the National Institutes of Health Institutional Research and Academic Career Development Awards Program at the Tufts University (K12GM074869, Training in Education and Critical Research Skills (TEACRS)), which supported her prior to the start of her faculty position. JMG would like to acknowledge support from an NIH postdoctoral fellowship (F32-DE026058), which funded him prior to the start of his faculty position. JMG is currently supported by the National Center for Advancing Translation Sciences (NCATS), a component of the National Institutes of Health (NIH) under award number KL2TR003018. JFJ acknowledges support from the University of Florida Herbert Wertheim College of Engineering Graduate School Preeminence Award and Institute for Cell and Tissue Science and Engineering Pittman Fellowship.

Supplementary Materials

Figure S1: example images from each histological dataset for adipose tissue (S1a) and cancer gland segmentation (S1b). Figure S2: color histograms (16 bins, equally spaced) of the normalized intensity values of each image in the SFBHI dataset (excludes test images). Figure S3: color histograms (16 bins, equally spaced) of the normalized intensity values of each image in the GlaS dataset (excludes test images). Figure S4: example segmentation results on adipose-poor samples (i.e., less than 1% of adipose pixels in image) from each model on the SFBHI dataset. Table S1: metrics for each data split on validation images in SFHBI dataset trained on UNET with binary cross entropy. Table S2: global distribution of random data split based on time. Table S3: global distribution of random data split based on conditions. Table S4: global distribution of stratified 5-fold time data split. Table S5: global distribution of stratified 5-fold condition data split. Table S6: global distribution of 4-fold with time data split. Table S7: global distribution of validate on week 8 data split. Table S8: metrics for each model on images in SFHBI dataset Table S9: metrics for each model on images in GlaS dataset trained with binary cross entropy loss. *(Supplementary Materials)*

References

[1] A. K. Gaharwar, I. Singh, and A. Khademhosseini, "Engineered biomaterials for in situ tissue regeneration," *Nature Reviews Materials*, vol. 5, no. 9, pp. 686–705, 2020.

[2] G. S. Hussey, J. L. Dziki, and S. F. Badylak, "Extracellular matrix-based materials for regenerative medicine," *Nature Reviews Materials*, vol. 3, no. 7, pp. 159–173, 2018.

[3] M. Gholipourmalekabadi, S. Sapru, A. Samadikuchaksaraei, R. L. Reis, D. L. Kaplan, and S. C. Kundu, "Silk fibroin for skin injury repair: where do things stand?," *Advanced Drug Delivery Reviews*, vol. 153, pp. 28–53, 2020.

[4] J. Rnjak-Kovacina, L. S. Wray, K. A. Burke et al., "Lyophilized silk sponges: a versatile biomaterial platform for soft tissue engineering," *ACS Biomaterials Science & Engineering*, vol. 1, no. 4, pp. 260–270, 2015.

[5] D. N. Rockwood, R. C. Preda, T. Yucel, X. Wang, M. L. Lovett, and D. L. Kaplan, "Materials fabrication from Bombyx mori silk fibroin," *Nature Protocols*, vol. 6, no. 10, pp. 1612–1631, 2011.

[6] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, article 101813, 2021.

[7] M. K. Ghalati, A. Nunes, H. Ferreira, P. Serranho, and R. Bernardes, "Texture analysis and its applications in biomedical imaging: a survey," *IEEE Reviews in Biomedical Engineering*, vol. 15, pp. 222–246, 2022.

[8] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2021.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pp. 234–241, Springer, 2015.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, 2015.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[12] T. Lei, R. Wang, Y. Wan, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: a survey," 2020, https://arxiv.org/abs/2009.13120.

[13] Z. Swiderska-Chadaj, H. Pinckaers, M. van Rijthoven et al., "Learning to detect lymphocytes in immunohistochemistry with deep learning," *Medical Image Analysis*, vol. 58, pp. 101547–101547, 2019.

[14] T. de Bel, M. Hermsen, B. Smeets, L. Hilbrands, J. van der Laak, and G. Litjens, "Automatic segmentation of histopathological slides of renal tissue using deep learning," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, article 1058112, International Society for Optics and Photonics, 2018.

[15] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021.

[16] O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention u-net: learning where to look for the pancreas," 2018, https://arxiv.org/abs/1804.03999.

[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3640–3649, Las Vegas, NV, USA, 2016.

[18] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, CA, USA, 2019.

[19] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[20] A. Danilov and A. Yurova, "Automated segmentation of abdominal organs from contrast-enhanced computed tomography using analysis of texture features," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 36, no. 4, article e3309, 2020.

[21] B. Hui, Y. Liu, J. Qiu, L. Cao, L. Ji, and Z. He, "Study of texture segmentation and classification for grading small hepatocellular carcinoma based on ct images," *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 199–207, 2021.

[22] L. Zhu, D. Ji, S. Zhu, W. Gan, W. Wu, and J. Yan, "Learning statistical texture for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12537–12546, Nashville, TN, USA, 2021.

[23] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikainen, "From bow to CNN: two decades of texture representation for texture classification," *International Journal of Computer Vision*, vol. 127, no. 1, pp. 74–109, 2019.

[24] J. Peeples, W. Xu, and A. Zare, "Histogram layers for texture analysis," *IEEE Transactions on Artificial Intelligence*, 2021.

[25] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2018.

[26] M. Yeung, E. Sala, C.-B. Schonlieb, and L. Rundo, "Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, article 102026, 2022.

[27] Z. Li, K. Kamnitsas, and B. Glocker, "Analyzing overfitting under class imbalance in neural networks for image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 1065–1077, 2020.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet ++: a nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2018.

[29] "Hasty.ai - a single application for all your vision ai needs," https://hasty.ai/.

[30] J. Rony, S. Belharbi, J. Dolz, I. B. Ayed, L. McCaffrey, and E. Granger, "Deep weakly-supervised learning methods for classification and localization in histology images: a survey," 2019, https://arxiv.org/abs/1909.03354.

[31] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6054–6063, Seoul, South Korea, 2019.

[32] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[33] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced nlp tasks," 2019, https://arxiv.org/abs/1911.02855.

[34] N. Rajan, J. Habermehl, M. F. Coté, C. J. Doillon, and D. Mantovani, "Preparation of ready-to-use, storable and reconstituted type I collagen from rat tail tendon for tissue engineering applications," *Nature Protocols*, vol. 1, no. 6, pp. 2753–2758, 2006.

[35] H.-J. Jin, J. Park, V. Karageorgiou et al., "Water-stable silk films with reduced β-sheet content," *Advanced Functional Materials*, vol. 15, no. 8, pp. 1241–1247, 2005.

[36] X. Hu, K. Shmelev, L. Sun et al., "Regulation of silk material structure by temperature-controlled water vapor annealing," *Biomacromolecules*, vol. 12, no. 5, pp. 1686–1696, 2011.