https://doi.org/10.2352/ISSN.2470-1173.2021.11.HVEI-112

Controllable Medical Image Generation via Generative Adversarial Networks

Zhihang Ren, Stella X. Yu, David Whitney; UC Berkeley / ICSI; Berkeley, California, USA

Abstract

Radiologists and pathologists frequently make highly consequential perceptual decisions. For example, visually searching for a tumor and recognizing whether it is malignant can have a life-changing impact on a patient. Unfortunately, all human perceivers—even radiologists—have perceptual biases. Because human perceivers (medical doctors) will, for the foreseeable future, be the final judges of whether a tumor is malignant, understanding and mitigating human perceptual biases is important. While there has been research on perceptual biases in medical image perception tasks, the stimuli used for these studies were highly artificial and often critiqued. Realistic stimuli have not been used because it has not been possible to generate or control them for psychophysical experiments. Here, we propose to use Generative Adversarial Networks (GAN) to create vivid and realistic medical image stimuli that can be used in psychophysical and computer vision studies of medical image perception. Our model can generate tumor-like stimuli with specified shapes and realistic textures in a controlled manner. Various experiments showed the authenticity of our GAN-generated stimuli and the controllability of our model.

Introduction

Because of its significant impact on health and well-being, medical image perception has been the focus of a great deal of research [1, 2, 3], from studies on its limits to studies on how to improve it. However, researchers often encounter a relative paucity of data and resources needed to pursue further investigation. While there are many publicly available medical imaging datasets, these are often limited, inadequately annotated, or outdated, e.g., The Digital Database for Screening Mammography (DDSM[19]). Moreover, the public datasets (e.g.,[5]) are not sufficiently large to support certain research questions.

Therefore, many researchers resort to using their own data from hospitals. Although this approach can ensure sufficient data is collected, it is often extracted from small geographic areas that are not representative of the broader population. Another issue with this method is the tedious and time-consuming process it requires to sort, categorize, and de-identify the data, and make it public. Moreover, it requires experts to perform meticulous annotations that are costly and time intensive [4]. Finally, and most importantly for the present purposes, these types of medical images are specific to each individual patient, which allows almost no room for researchers to manipulate them in order to meet desired experimental configurations.

To tackle this problem in psychophysical experiments, artificial medical stimuli have been employed [6, 7]. On one hand, artificial stimuli can be easily generated and manipulated, such as shape morphing and background replacement. On the other hand, those stimuli are obviously fake and completely unlike those that doctors routinely examine. Consequently, expert radiologists rightly worry that these psychophysical experiments do not accurately represent their daily diagnostic tasks.

Thus, generating authentic and easily controllable medical image stimuli is critical for medical image perception studies. Current research in computer vision provides us with a promising approach, using Generative Adversarial Networks (GAN). Generative Adversarial Networks have been utilized for generating authentic materials [14, 13, 8] for years, such as faces, cars, landscapes, and so on. Trained on real image datasets, GAN can generate various authentic samples that have similar semantic context to that of real images. Besides this, the generation is easily conditioned [8], which means manipulating generated samples is possible through the design and the input of the GAN.

Inspired by Generative Adversarial Networks, we utilized this computational model to generate medical image stimuli. We then tested our model on mammogram stimuli generation. Furthermore, we generated tumor-like stimuli with specified shapes and realistic textures using our GAN model, which effectively controls the similarity of the generated stimuli. Various experiments showed the authenticity of our GAN-generated stimuli and the controllability of our model.

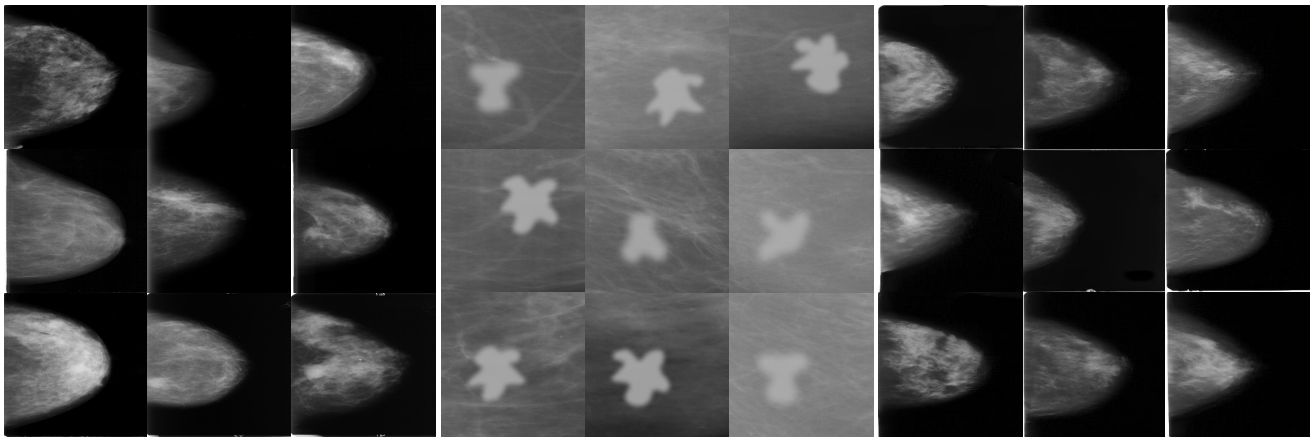
Generative Adversarial Network

A Generative Adversarial Network (GAN) is a powerful deep learning model with two networks, i.e., a generator network and a discriminator network [9]. The two networks learn from each other in an adversarial way. In summary, the generator produces authentic images from random noise vectors and tries to fool the discriminator, while the discriminator tries to distinguish the fake samples (generated from generator) from the real samples. The whole process can be conceptually described as a min-max game shown in Equation 1, where G represents the generator, D represents the discriminator, $p_{data}(x)$ indicates the real data distribution, and $p_z(z)$ indicates the noise vector distribution.

$$\min_G \max_D E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

Ideally, through the adversarial training, GAN can approximate the real data distribution (manifold) parameterized by the generator network. Similar samples between two specific samples can be picked along the path between the corresponding points on the manifold. This can be done by interpolating the corresponding latent vectors and forwarding them through the generator.

Originally, the training of the Generative Adversarial Network (GAN) was highly unstable and many strategies have been proposed to tackle this problem [10, 11, 12, 13]. In this paper, we adopt the structure from StyleGAN [14], where the latent space



(a) Real Mammograms (b) Previous Stimuli [6] (C) GAN Generated Mammograms

Figure 1. *Comparison of (a) real mammograms, (b)stimuli used in previous studies [6], and (c)our GAN generated mammograms. A particular example from previous study consists of naive morphed shapes (tumors) and healthy mammogram backgrounds, which are obviously fake and do not represent realistic stimuli for radiologists. It is clear that our mammogram generation can mimic the texture for both tumor and non-tumor regions and they have reasonable shapes compared to real mammograms.*

$\mathscr{Z}$ is first mapped into the $\mathscr{W}$ space through a non-linear mapping network (an 8-layer MLP) and then merged into the synthesis network via adaptive instance normalization (AdaIN) at each convolutional layer [15, 16, 17]. Gaussian noise is added after each convolutional layer, before the AdaIN layer.

StyleGAN has the ability to generate high-resolution realistic images of faces, cats, bedrooms, and cars. It can control the image details from coarse to fine by changing the AdaIN parameters and the input noises at different network levels. Using StyleGAN, we can easily generate various authentic medical images by changing the input noise vectors. However, the generation is unconditional; this means, in order to get the desire input (e.g., the mammogram having a specific shape or texture.), we need to pick samples from a large number of generated images because we have no control over the attributes of the images (e.g., the shape and the texture of the mammogram). It is tedious and time-consuming. Moreover, we do not have the latent codes for real data if we want to generate similar medical images between two real ones. An intuitive idea is that we can encode the latent vector z from the desired medical image. Then the similar medical images can be generated by interpolating the corresponding encoded vectors.

Method

To train then utilize the encoder, first, the discriminator and generator of StyleGAN [14] are pretrained. Then the generator of styleGAN is fixed. While training the encoder network, traditional methods [20, 21] regularize the encoder on the latent space, encouraging the encoder to encode the same latent codes for the corresponding generated images regardless of the reconstructed images. This method can deteriorate the reconstruction quality. Instead, we adopt the idea from In-domain GAN inversion [18], where the regularization of the encoder is on the image space. In detail, the encoded vector is fed into the generator again and the L2 reconstruction loss regularizes the encoder on the image space. The training is conducted on the real data and the adversarial loss helps the reconstructed image to be more realistic. Moreover, perceptual loss [22] is utilized. The whole process can be summarized as follows

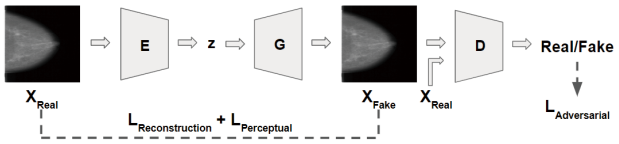


Figure 2. *Overview of proposed method. The generator (G) and discriminator (D) are from StyleGAN[14]. The training has two phases. In the first phase, the generator and discriminator will be trained first without the encoder (E) via adversarial loss $L_{adversarial}$. In the second phase, the generator (G) will be fixed. The encoder (E) and discriminator (D) will be trained adversarially via the reconstruction loss $L_{reconstruction}$, the perceptual loss $L_{perceptual}$, and the adversarial loss $L_{adversarial}$. The dashed lines indicate how to compute the corresponding loss metrics.*

$$\min_E ||x - G(E(x))||_2 + \lambda_1 ||F(x) - F(G(E(x)))||_2 - \lambda_2 E_{x\sim p_{data}(x)}[\log D(G(E(x)))] \tag{2}$$

$$\min_D E_{x\sim p_{data}(x)}[\log D(G(E(x)))] - E_{x\sim p_{data}(x)}[\log D(x)] + \frac{\gamma}{2} E_{x\sim p_{data}(x)}[||\nabla_x D(x)||_2^2] \tag{3}$$

where $p_{data}(x)$ indicates the real data distribution, x is the real image, E represents the encoder, F indicates the VGG feature extraction [23], and λ_1, λ_2 and γ are weights for the perceptual loss, the adversarial loss, and the gradient penalty. An overview of the training process can be find in Figure 2.

Since inverse mapping will never be perfect, additional optimization is required for a better reconstruction for each image. Starting with the output code from the encoder, the optimization updates the encoded vectors based on the reconstruction loss and

the perceptual loss but is still regularized by the encoder. The optimization process can be described as below.

$$z^{inv} = \min_{z} ||x - G(E(x))||_2 + \lambda_3 ||F(x) - F(G(z))||_2 + \lambda_4 ||z - E(G(z))||_2 \quad (4)$$

where z^{inv} is the optimized inverse code, λ_3 and λ_4 are weights for the perceptual loss, and the code reconstruction loss (i.e., the encoder regularization).

We test our proposed method on the mammogram generation task. Similar experiments can be conducted with different medical imaging modalities.

Perceptual loss

Perceptual loss has been utilized as a similarity metric in many computer vision tasks, such as style transfer [24, 25], and image super-resolution [22], both of which are ill-posed problems. For style transfer, there is no ground truth to act as a reference. For image super-resolution, many high-resolution images can be sampled to generate the same low-resolution image. In order to achieve the tasks, semantic information of the input images should be maintained. Thus, per-pixel loss is no longer suitable since it cannot capture the semantic difference between the output and ground truth. For example, in style transfer, there are usually drastic changes in color and texture compared to the input images.

Perceptual loss is computed as the difference between high-level features from a pretrained loss network which is usually a feature extractor of an image classification network. Compared to the per-pixel loss, which depends only on low-level pixel information, perceptual loss is more robust in image similarity measurement during training.

Implementation details

We conduct our experiment on Digital Database for Screening Mammography (DDSM) [19] dataset which consists of 2,620 cases of normal, benign, and malignant cases. During training, we only use the mammograms which have tumors inside, i.e., the benign and malignant cases. The GAN part is pretrained based on StyleGAN. The encoder consists of one initial convolution, 8 residual blocks, and one dense layer. And the batch normalization is utilized for all modules in the encoder. While training the encoder, the generator is fixed. Only the encoder and discriminator are updated according to the loss function shown in Equation 2 and Equation 3 respectively. For the perceptual loss, we use *conv*4_3 feature layer in VGG [23] as illustrated in [18]. Hyperparameters are set as $\lambda_1 = 5e^{-5}$, $\lambda_2 = 0.1$, $\lambda_3 = 5e^{-5}$, $\lambda_4 = 2$, and $\gamma = 10$. And we use the Adam optimizer [27] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is set to 0.0001

Mammogram generation

The pretrained StyleGAN[14] approximates the real mammogram distribution (manifold) which is parameterized by the generator. Then, the authentic mammograms can be sampled from the learned manifold. To do so, we sample latent codes from a normal distribution and use the generator to map the latent codes onto the learned manifold.

Mammogram interpolation

Mammogram interpolation is utilized to generate similar stimuli between two desired mammograms. Since the GAN generator already approximates the real data manifold, similar mammograms can be found between any two mammograms on the path that links them on the manifold. Given two latent codes from either unconditional generation or encoded from real mammograms, we interpolate the latent codes and the generator can help to find one of the linking paths by mapping the interpolated latent codes onto the manifold.

Controllable mammogram generation

Controllable mammogram generation is utilized to generate similar desired mammograms (i.e. the end points for interpolation). With the generator and the encoder network, we can achieve the controllable mammogram generation, where we can combine the desired tumor texture with the given shape template.

First, we crop the tumor texture region and paste it onto the shape template. Then, we use the encoder to obtain a latent code for this combined raw image. Because the regularization of the encoder is on the image space, the latent code can already carry certain semantic information from both the shape template mammogram and the tumor texture mammogram. Finally, we apply the masked optimization, only using the tumor texture region to compute the reconstruction loss.

Human evaluation

To verify the authenticity of our GAN generated mammograms, we designed a judgement test where participants were randomly presented 100 mammograms with the same amount of real and generated (fake) samples. In this task, they were asked to classify each mammogram as real or fake as well as rate their level of confidence with their selection. In total, 6 participants were involved.

To make sure participants were paying attention and not guessing randomly, we asked a subset of participants to do the judgment test a second time. Participants were not told about the second judgment task prior to the first judgment, eliminating any chance that they purposely remember their first responses. At last, we compute the test-retest similarity in term of the Sokal-Michene metric[26].

Results

In this section, we will show the mammogram stimuli generation quality, the corresponding human evaluation result, the interpolation result, and the results when generating mammograms given specific shapes and textures.

Mammogram generation

Examples of the GAN generated samples are shown in Figure 1 (c). A particular example from previous study [6] consists of naive morphed shapes (tumors) and healthy mammogram backgrounds, which are obviously fake and do not represent realistic stimuli for radiologists. It is clear that our mammogram generation can mimic the texture for both tumor and non-tumor regions and has reasonable shapes compared to real mammograms.

Mammogram interpolation

The interpolation results are shown in Figure 3. Through the interpolation, the mammograms change gradually from one to the other and they are similar to the neighboring images. Moreover, through the interpolation, we can generate a similar stimuli loop where the stimuli gradually changing from image A to image B, then to image C, and finally back to image A, as shown in Figure

3.

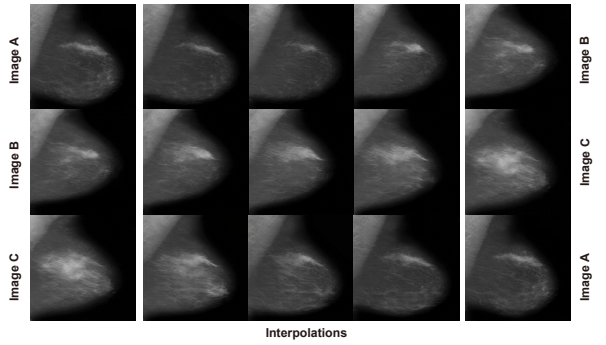


Figure 3. *Mammogram interpolation result. Here we evenly pick three mammograms along the interpolation just for illustration. We show that we can generate a similar stimuli loop where the mammogram can gradually change into other mammograms, and finally change back.*

Controllable mammogram generation

The final results compared with the tumor texture mammogram, shape template mammograms, directly stitching results, traditional image blending results, and the results without final optimization (i.e., directly generated from the encoder output), are shown shown in Figure 4. For the traditional image blending results, though it blends the tumor texture region into the shape template mammogram, the surrounding region of the tumor is not realistic compared to our GAN generated results. The results directly generated from the encoder outputs do not have the same tumor texture as given. It is clear that only results obtained after optimization have the same tumor texture and shape as given.

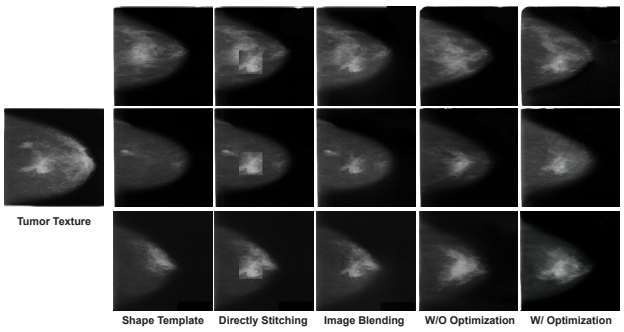


Figure 4. *Controllable mammogram generation results. We compared our final results with the traditional image blending method and the results without final optimization (i.e., directly generated from the encoder output). It is clear that only the results after final optimization have the same tumor texture and shapes as given.*

Human evaluation

The performance of the judgement test participants in terms of the Receiver Operating Characteristic (ROC) curve is shown in Figure 5. The Receiver Operating Characteristic (ROC) curve is a plot which can indicate the ability of a binary classifier as its discrimination threshold is changing. A ROC curve that is close to the diagonal indicates performance at chance level. A ROC curve that is close to upper left corner indicates stronger discrimination power. For all the participants, their performance curves are near the diagonal, which is near chance discrimination performance, and the d' is 0.02 on average, which indicates that the generated mammograms appeared authentic.

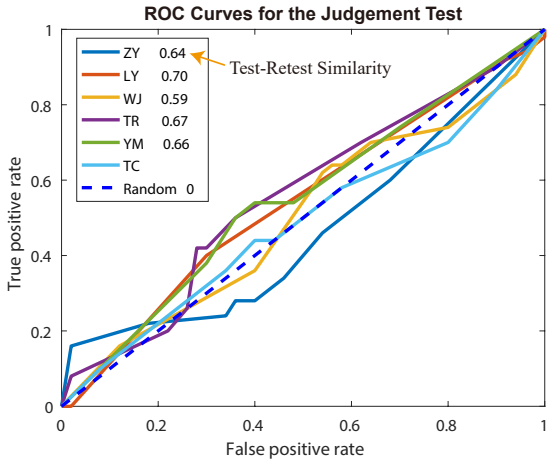


Figure 5. *Human judgement test for our mammogram generation quality. All the Receiver Operating Characteristic (ROC) curves lie near the diagonal, indicating that our generated mammogram successfully fooled participants. The high test-retest similarities indicate that participants did not just randomly guess the result.*

The test-retest similarity in term of the Sokal-Michene metric[26] is shown after each participant initial in Figure 5. The high similarities indicate that participants did not just randomly guess the mammogram label.

Discussion

While other methods [28] can only achieve unconditional mammogram generation, our method provides control over the shape and texture of the generated mammograms. Since the stimuli used in psychophysical experiments of medical image perception often need to be similar and controllable, we can manually combine the desired tumor texture with the similar shape templates to create the required stimuli. Therefore, this work largely benefits psychophysical experiments by establishing the ability to manipulate and control life-like medical images.

Summary

We proposed to use the Generative Adversarial Network to generate medical image stimuli for studies of medical image perception. Similar medical image stimuli can be generated through the interpolation of the corresponding latent codes. Desired stimuli can be manually combined with desired attributes, e.g. object shape and tumor texture, in a controllable manner. We tested our method on the mammogram stimuli generation task. Empirically, we proved the authenticity of our synthesized mammograms with a psychophysical discrimination task.

Acknowledgments

This work has been supported by National Cancer Institute (NCI) under grant # 1R01CA236793-01.

References

[1] Degnan, Andrew J., et al. Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. Academic radiol-

ogy 26.6: 833-845 (2019).
[2] Sunday, Mackenzie A., Edwin Donnelly, and Isabel Gauthier. Individual differences in perceptual abilities in medical imaging: the Vanderbilt Chest Radiograph Test. Cognitive Research: Principles and Implications 2.1: 1-10 (2017).
[3] Itri, Jason N., and Sohil H. Patel. Heuristics and cognitive error in medical imaging. American Journal of Roentgenology 210.5: 1097-1105 (2018).
[4] Willemink, Martin J., et al. Preparing medical imaging data for machine learning. Radiology 295.1: 4-15 (2020).
[5] F. W. Prior et al., TCIA: An information resource to enable open science, 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1282-1285. (2013)
[6] Manassi M, Kristjansson A , Whitney D. Serial dependence determines object classification in visual search. Journal of Vision, 17(10):221–221 (2017)
[7] Manassi M, Kristjansson A , Whitney D. Serial dependence in a simulated clinical visual search task. Sci Rep 9, 19937 (2019)
[8] Park, Taesung, et al. Semantic image synthesis with spatially-adaptive normalization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems pp. 2672-2680 (2014)
[10] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C. Improved training of wasserstein gans. In Advances in neural information processing systems pp. 5767-5777 (2017)
[11] Arjovsky, M., Chintala, S., Bottou, L. Wasserstein gan. arXiv preprint arXiv:1701.07875. (2017)
[12] Brock, A., Donahue, J., Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations. (2018)
[13] Karras, T., Aila, T., Laine, S., Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196. (2017)
[14] Karras, T., Laine, S., Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4401-4410 (2019)
[15] Dumoulin, V., Shlens, J., Kudlur, M. A learned representation for artistic style. arXiv preprint arXiv:1610.07629. (2016)
[16] 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)
[17] Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H. D., Courville, A., Bengio, Y. Feature-wise transformations. Distill, 3(7), e11. (2018)
[18] Zhu, J., Shen, Y., Zhao, D., Zhou, B. In-domain gan inversion for real image editing. In Proceedings of European Conference on Computer Vision (ECCV). (2020)
[19] Bowyer, K., Kopans, D., Kegelmeyer, W. P., Moore, R., Sallam, M., Chang, K., Woods, K. The digital database for screening mammography. In Third international workshop on digital mammography. Vol. 58, p. 27 (1996)
[20] Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A. Inverting layers of a large generator. In: ICLR Workshop (2019)
[21] Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A. Seeing what a gan cannot generate. In: ICCV (2019)
[22] Johnson, J., Alahi, A., Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Springer, Cham. (2016)
[23] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
[24] Gatys, L. A., Ecker, A. S., Bethge, M. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576. (2015)
[25] Gatys, L., Ecker, A. S., Bethge, M. Texture synthesis using convolutional neural networks. Advances in neural information processing systems, 28, 262-270. (2015)
[26] Zhang, B., Srihari, S. N. Properties of binary vector dissimilarity measures. In Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing (Vol. 1). (2003)
[27] Kingma, D. P., Ba, J. Adam: A method for stochastic optimization. In: ICLR (2015)
[28] Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B. High-resolution mammogram synthesis using progressive generative adversarial networks. arXiv preprint arXiv:1807.03401. (2018)

Author Biography

Zhihang Ren received his BS in Electrical and Electronic Engineering from University of Electronic Science and Technology of China and his MS in Electrical and Computer Engineering from UC San Diego. He is now a PhD candidate in Vision Science at UC Berkeley.

Stella Yu received her Ph.D. from the School of Computer Science at Carnegie Mellon University. After a postdoctoral fellowship at UC Berkeley in computer vision, she became a Clare Booth Luce Professor at Boston College. She is now the Vision Group Director at the International Computer Science Institute, a Senior Fellow at the Berkeley Institute for Data Science, and a faculty member of Computer Science, Vision Science, Cognitive and Brain Sciences at UC Berkeley.

David Whitney received his BA in Psychology, Economics, and Philosophy from Boston University and his MA and PhD in Psychology from Harvard University. He served as a Post-doctoral Fellow at the University of Western Ontario and as an Associate Professor at UC Davis. He is now a professor in the Department of Psychology and the Helen Wills Neuroscience Institute at UC Berkeley. In addition, he serves as the Director of Cognitive Sciences at UC Berkeley.

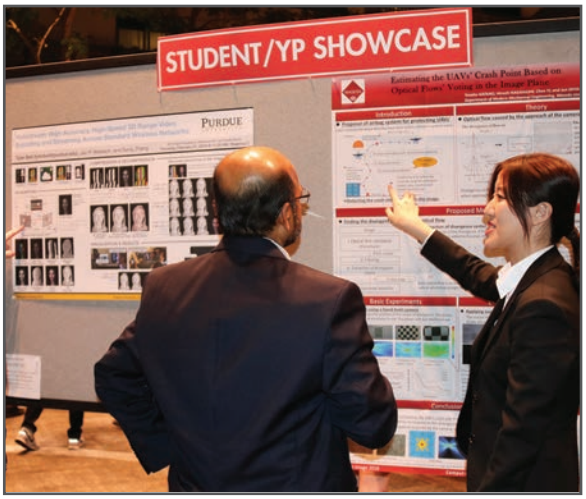
JOIN US AT THE NEXT EI!

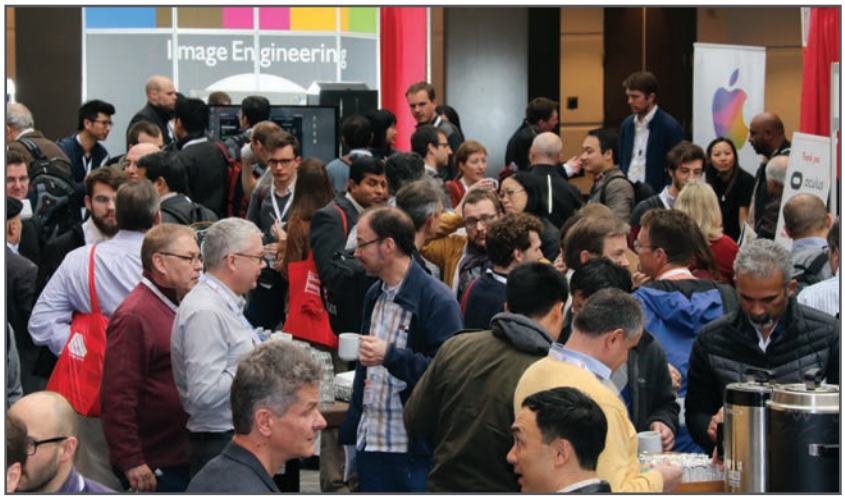
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org