https://doi.org/10.2352/EI.2022.34.13.CVAA-169

Improving semantic segmentation of fine art images using photographs rendered in a style learned from artworks

Thomas Heitzinger; TU Wien, Computer Vision Lab; Vienna, Austria
David Stork; Consultant; Portola Valley, CA 94028 USA

Abstract

Our central goal was to create automatic methods for semantic segmentation of human figures in images of fine art paintings. This is a difficult problem because the visual properties and statistics of artwork differ markedly from the natural photographs widely used in research in automatic segmentation. We used a deep neural network to transfer artistic style from paintings across several centuries to modern natural photographs in order to create a large data set of surrogate art images. *We then used this data set to train a separate deep network for semantic image segmentation of genuine art images. Such data augmentation led to great improvement in the segmentation of difficult genuine artworks, revealed both qualitatively and quantitatively. Our unique technique of creating surrogate artworks should find wide use in many tasks in the growing field of computational analysis of fine art.*

Introduction and background

Semantic segmentation is an important step in many tasks in the analysis of natural photographs (e.g., captioning, question answering, search), medical images (identifying tumors or cancerous lesions), remote sensing (identifying land and water forms, foliage and vegetation), and autonomous driving (lane following, obstacle identification and avoidance). Currently the most accurate segmentation methods rely on deep neural networks trained with large corpora of photographs or stills from videos—up to hundreds of thousands of examples.[12, 2]

Segmentation is also a key early step in many techniques of automatic art analysis, including the analysis of compositional styles, identifying figures or "actors" as a step to inferring the story, moral, or meaning expressed by an artwork,[17] and others. For a number of reasons, semantic image segmentation of art images has proven difficult for machine learning methods based on deep networks trained with photographs.[3] First, art images in the Western canon vary widely in style—color palettes, marks, contours, and so forth. Second, art images often include imaginary or non-existent objects, such as dragons, halos, satyrs, and nymphs. Third, even realist art often includes objects that violate physical constraints, such as flying putti and dripping pocket watches. None of these attributes have counterparts in the natural photographs generally used for training deep neural networks. As a result, existing deep networks for segmentation perform quite poorly on stylized fine art images, as we shall confirm in the following sections.

Given the abilities of deep networks to learn complex functions and relationships from adequately large training sets, the central problem in the development of such networks for art analysis would appear to be that there are too few art images for training such networks for tasks such as semantic segmentation. The largest image datasets used in deep learning use millions—or in private datasets up to hundreds of millions—of natural photographs.[5, 18] By contrast, the total number of relevant fine art images is several orders of magnitude smaller. Estimates of completed painting by leading artists—including the most prolific artists—are much lower: Picasso (13500), Pierre-Auguste Renoir (4000), Jean-Baptiste-Camille Corot (3000), Vincent van Gogh (900), Paul Cézanne (900), Johannes Vermeer (34), and Leonardo da Vinci (18). Even a collection of the artworks of thousands of artists yields a dataset far smaller than the minimum corpora of photographs used in accurate deep network systems.

Overall approach and compute environment

Our approach to this problem is to create a large corpus of new images by transferring a learned art style to natural photographs.[7, 8, 6] The images in the resulting dataset will be referred to as *surrogate artworks* and might include an image of a jumbo jet rendered in the style of Monet. We then use this corpus of surrogate artworks to refine or *transfer train* an existing deep network for segmentation to better reflect the properties of target genuine artworks. We then apply this transfer-trained deep network to the problem of segmenting genuine artworks, ones not in the training set. As we shall see, this two-step approach leads to great improvement in the segmentation of particularly difficult, non-realistic genuine artworks as we specify both qualitatively and quantitatively, below.

In general, the task of transferring domain knowledge from a source domain to a target domain is referred to as *domain adaptation* [15, 10]. Recent work [1] shows that adaptation techniques based on neural style transfer can outperform state-of-the-art GAN-based image translation approaches [11, 14] on semantic segmentation and object recognition tasks. To the best of our knowledge, this work is the first to demonstrate the use of neural style transfer to generate a large-scale surrogate dataset and to demonstrate its effectiveness in training and applying semantic segmentation models to the analysis of artworks.

Our base dataset of photographs[19] and our Deeplab v3 segmentation network[4] have been described elsewhere, have been used in several studies, and are well understood. The backbone of our segmentation network was ResNet50.[9] We used the well-tested VGG19 network for feature extraction[16] in the style transfer process.[7] In all cases, we used the network architectures as described in the references; it is our overall system architecture and creation of surrogate database for transfer training that are novel.

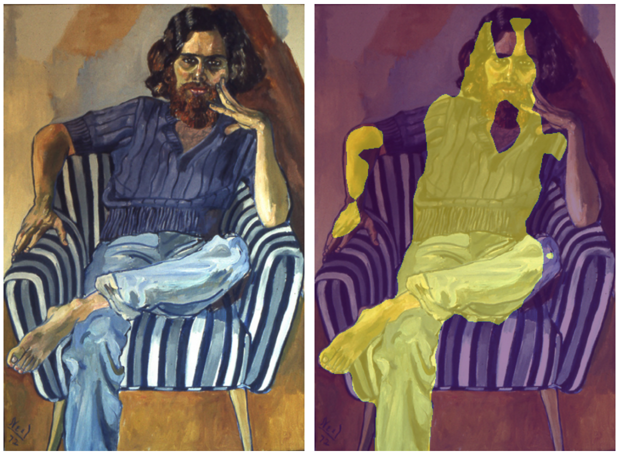

Figure 1. *Example of baseline segmentation performance on Alice Neel's portrait* Dana Gordon*, oil on canvas (1972). This is a particularly challenging segmentation problem because of the artist's style avoids strong lighting clues, both color and line design in the shirt and chair are similar and employ somewhat unnatural colors. Notice the segmentation omission errors at of the arms, neck, and forehead, and the commission errors beneath the subject's foot. Additional baseline examples appear in the middle row of Fig. 3.*

Baseline performance: Semantic segmentation of art images by networks trained on natural photographs

Throughout our work, the corpus of art images consists of 105 full-color paintings from the Western canon, scraped from online databases and scaled to 320 × 320 pixels, a convention that speeds batch processing of images in our work. We use a Deeplab v3 [4] segmentation network with a ResNet [9] backbone with 50 layers, where the baseline segmentation network was trained on natural photographs. The network is pretrained on the *COCO train2017* [12] dataset which features 21 classes, including a person class. As we are only interested in the segmentation of persons the detection head is reduced to a single output channel for our binary segmentation problem. The baseline network is then finetuned on a subset of the Baidu People Segmentation Dataset[19] consisting of 5209 half-body and full-body shots of people with varied backgrounds. Annotation is provided in the form of accurate binary segmentation masks. We split the dataset into a training set and a validation set in a roughly 4:1 ratio.

Figure 1 shows sample baseline segmentation of a modern portrait by the network trained with only natural photographs. Notice the numerous segmentation errors. As we shall see in the results section, such errors arise in part because the styles (colors, textures, and so on) of the artworks differs from that of the natural photographs used for training.

The segmentation performance can be quantified by the standard IoU metric:

$$U = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IoU}(T_i, P_i). \qquad (1)$$

where $N = 21$ is the number of independent art images used for testing, T_i and P_i are the sets of pixels of target and predicted segmentation masks, respectively, and IoU is the intersection over union metric defined as $\mathrm{IoU}(A,B) = \frac{|A \cap B|}{|A \cup B|}$.[13] Of course $0 \leq U \leq 1$, where the higher the value of U, the higher the quality of segmentation. We find that for this baseline case $U_b = 62.1\%$.

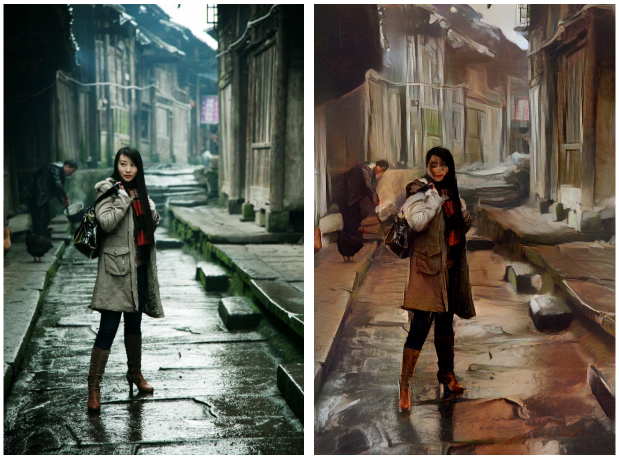

Figure 2. *Natural photograph before and after application of style transfer from our art data set. Notice the somewhat unnatural ochre and brown tones in the surrogate artwork.*

Neural transfer of artistic style to natural photographs

We created our surrogate training set by style transfer of 60 paintings randomly selected from our corpus to the baseline training set of 4146 natural photographs, using a VGG-19 network for feature extraction and the style-transfer protocols to combine the content of natural photographs with the style of artworks.[7] A surrogate validation set of 1063 images was generated in analogous fashion. Both content and style images used to generate the surrogate training and validation sets are disjoint. Furthermore, the final test set of real artworks consists of 21 additional unseen artworks. Figure 2 shows typical style transfer results. Notice, informally, that the transferred style is not a single one from the history of art, but some form of mixture of many styles, as is appropriate given we will be applying our system to art from throughout the full period, and hence a variety of styles.

We took the base segmentation network that had been finetuned on 4146 natural photographs and transfer trained it with the same number of surrogate paintings to maximal IoU score on the surrogate validation set of 1063 images. The same learning protocols as in the base network was used.

We then applied the resulting network to the task of segmenting 21 genuine artworks—none of which were used to transfer style to the surrogate database. As is proper protocol, images in our test set were not used in any way in the training of our overall system.

All models were run on a system with an Intel Xeon E5-2697v3 CPU (14 cores), 128GB RAM and 4 Nvidia GeForce GTX Titan X GPUs, using *PyTorch 1.6.0.*

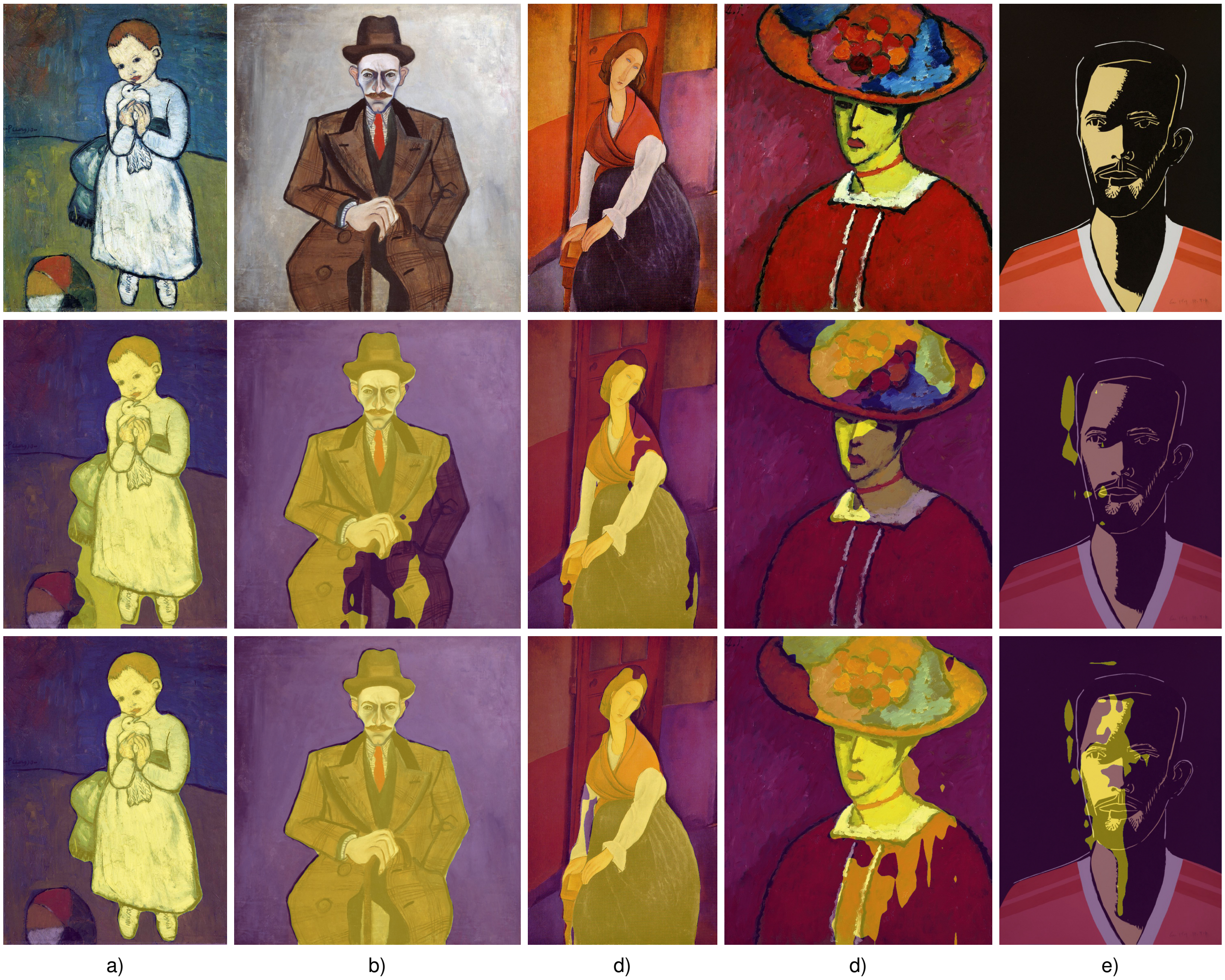

Figure 3. *Top row: paintings, middle row, segmentation with baseline network, bottom row, segmentation after network transfer trained with surrogate art images. a) Pablo Picasso's* Child with dove, *b) Alice Neel's* Kenneth Dolittle, *c) Amedeo Modigliani's* Jeanne Hebuterne (aka In Front of a Door), *d) Alexej von Jawlensky's* Schokko with Wide-Brimmed Hat, *and e) Alex Katz's* Daniel. *In all these challenging cases, the network transfer trained with surrogate art images performs better than the baseline case based on just natural photographs.*

Results: Segmentation after transfer training with photographs styled by artworks

Figure 3 shows typical segmentation results for difficult artworks, both before and after transfer training by the surrogate art data set. Our base network was trained only on natural photographs. Nevertheless, for Pablo Picasso's ”Child with Pigeon” (see column a), it generates a largely accurate segmentation mask with an IoU score of 88.7% and small false positive region at the child's feet. In comparison, our improved network generates a nearly perfect mask at an IoU score of 97.4%. Generally notice that qualitatively the segmentations are superior to the base case, in some cases significantly so, most notably Alice Neel's *Kenneth Dolittle* (IoU improvement of 27%). Informally, numerous obvious segmentation errors—of both omission and commission—arising in the base system are eliminated by training with the surrogate artworks. The improvements in segmentation quality achieved with our method are most evident in Alexej von Jawlensky's *Schokko with Wide-Brimmed Hat* shown in column d. The baseline network only succeeds in capturing small parts of the face of the woman, which is rendered by the artist in an unnatural yellow skin tone. Our enhanced network captures the entire face and neck area, leaving only the chest largely unrecognized. The last example of Alex Katz's *Daniel* represents a special case. None of Katz's artworks were used in the generation of our surrogate dataset, i.e., neither the baseline network nor our improved network was trained on styles found in Katz's artworks. What is remarkable about this example is its reduced style and overall flat appearance – compared to the previous examples and artworks used in the creation of our surrogate dataset, fine structures are much less pronounced. Nonetheless, and although the segmentation errors in the final Daniel are quite severe, the segmentation is still far better than in the base case, as shown in column e. Finally, Figure 4 represents a case where neural style transfer produces unconvincing surrogate artworks. The floral and fabric patterns of

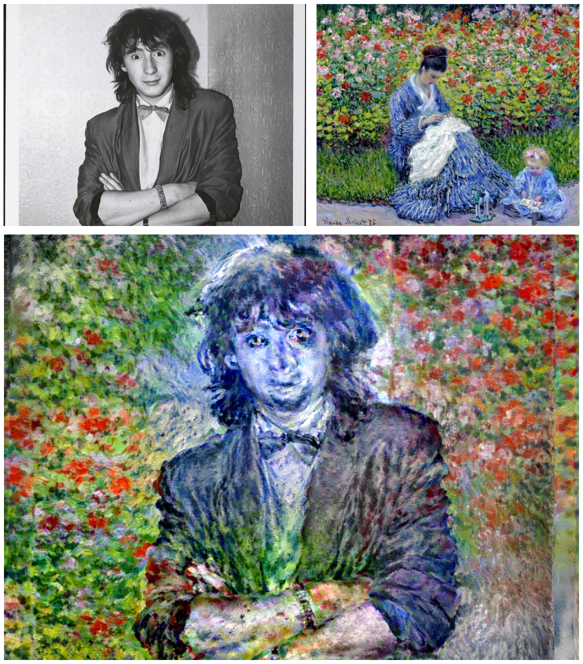

Figure 4. An example of poor fusion of content (top left) and style images (top right) into a surrogate artwork (bottom).

Claude Monet's *Camille Monet and a child in the artist's garden in Argenteuil*, which serve as the style component, are incorrectly transferred to the subject's face and arms. These effects can be caused by unexpected image features, e.g. if the image is black and white, as is the case in this example, but they can also be due to a lack of semantic understanding of the network used for feature extraction in the style transfer process.

There is a quantitative benefit from the use of our surrogate training set, as expressed by the performance measure in Eq. 1. The final system has a performance of $U_t = 74.9\%$, which is of course superior to the baseline $U_b = 62.1\%$, as measured on 21 independent artworks.

Conclusions and future directions

We have demonstrated that deep network transfer of artistic style to a large corpus of readily available natural photographs can produce large data sets of surrogate art images. A baseline deep network architecture for semantic segmentation, transfer trained with this large data set of surrogate art images, leads to significantly greater accuracy on segmentation of difficult art images. Our work strongly suggests that it is the *style*, not the diversity of shapes, is the key factor in segmentation of art images.

Future work could focus on training segmentation networks with larger sets of both genuine and surrogate artworks, and alterations of network architecture and training protocols tailored to art problems. Similarly, we can extend this work to non-binary segmentation, that is, to semantic segmentation of numerous classes relevant to art analysis. Likewise, we expect that the imposition of priors and focused data sets relevant to a specific task improves results. Thus for instance if our ultimate task is to automatically segment images of *Impressionist* artworks, we should transfer style from just Impressionist artworks to create a surrogate data set. Likewise for other periods, such as *Mannerist*, *Expressionist* and so forth. In some cases, this technique might apply restricted to artists with sufficiently large oeuvre and distinctive style, possibly Vincent van Gogh or Georges Seurat.

We believe the methods presented in this paper should find use in computational approaches to art scholarship, particularly higher-level interpretation tasks such as extracting simple meanings from artworks.[17] Our technique of training neural networks with surrogate photographs is not limited to the task of semantic segmentation and could facilitate the transfer of image analysis tools that already exist for natural photographs to the domain fine art paintings. Conceivable examples include art conservation and analysis through digital inpainting of paintings based on deep networks trained with surrogate artworks from the relevant style. Similarly, visual search tools could facilitate indexing and search of large catalogue records of collections.

ACKNOWLEDGEMENTS

This work is an extension of a project begun in *Computer vision and image analysis of art* taught by the second author at the Technical University of Vienna, January–February 2021. We thank Professor Robert Sablatnig for support of both that course and this current research. This work has been partly funded by the Austrian security research program KIRAS of the Federal Ministry for Transport, Innovation and Technology (BMVIT) under Grant 873495.

References

[1] D. Aysegul, L. Ming-Yu, W. Ting-Chun, Z. John, and K. Jan. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Applications*, 39(12):2481–2495, 2017.

[3] A. Bourached, G. H. Cann, R. Griffiths, and D. G. Stork. Recovery of underdrawings and ghost-paintings via style transfer by deep convolutional neural networks: A digital tool for art scholars. In D. G. Stork and K. Heumiller, editors, *Computer vision and analysis of art*. IS&T, 2021.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[6] Z. Ding, N. M. Nasrabadi, and Y. Fu. Task-driven deep transfer learning for image classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2414–2418, 2016.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[8] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3985–3993, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for im-

age recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] R. Ievgen, M. Emilie, H. Amaury, S. Marc, and B. Younès. *Advances in domain adaptation theory*. Elsevier, 2019.

[11] H. Judy, T. Eric, P. Taesung, Z. Jun-Yan, I. Phillip, S. Kate, E. Alexei, and D. Trevor. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in COntext. In *European conference on computer vision*, pages 740–755, 2014.

[13] F. C. Monteiro and A. C. Campilho. Distance measures for image segmentation evaluation. *AIP Conf. Proc.: Numerical Analysis and Applied Mathematics ICNAAM*, 1479:794–797, 2012.

[14] I. Naoto, F. Ryosuke, Y. Toshihiko, and A. Kiyoharu. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.

[15] B.-D. Shai, B. John, C. Koby, K. Alex, P. Fernando, and V. J. Wortman. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. `arXiv:1409.15565v6`.

[17] D. G. Stork, A. Bourached, G. H. Cann, and R. Griffiths. Computational identification of significant figures in paintings through symbols and attributes. In D. G. Stork and K. Heumiller, editors, *Computer vision and analysis of art*. IS&T, 2021.

[18] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[19] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan. Early hierarchical contexts learned by convolutional networks for image segmentation. In *2014 22nd International Conference on Pattern Recognition*, pages 1538–1543. IEEE, 2014.

Author Biography

Thomas Heitzinger holds an MSc in Applied Mathematics and an MSc in Logic and Computation. He is currently a PhD student at the Computer Vision Lab (CVL) at TU Wien. His main research interests are semantic segmentation for sepcialized target applications and 3D scene understanding on resource constrained hardware and human behavior analysis using depth sensors and other non-RGB-based visual sensors, as well as computer vision without strong texture information.

David G. Stork holds degrees in physics from the Massachusetts Institute of Technology and the University of Maryland at College Park. He has made technical contributions in theoretical mechanics, computational imaging, computer vision, optics, machine learning, pattern classification, computational data acquisition, concurrency theory, cryptography, visual psychophysics and perception, statistics, combinatorics, and other areas. He studied art history at Wellesley College and was an artist-in-residence at the New York State Council of the Arts.