https://doi.org/10.2352/ISSN.2470-1173.2020.14.COIMG-377
© 2020, Society for Imaging Science and Technology

Efficient Multilevel Architecture for Depth Estimation from a Single Image

Bruno Artacho, Nilesh Pandey, and Andreas Savakis; Rochester Institute of Technology; Rochester, NY, USA.

Abstract

Monocular depth estimation is an important task in scene understanding with applications to pose, segmentation and autonomous navigation. Deep Learning methods relying on multilevel features are currently used for extracting local information that is used to infer depth from a single RGB image. We present an efficient architecture that utilizes the features from multiple levels with fewer connections compared to previous networks. Our model achieves comparable scores for monocular depth estimation with better efficiency on the memory requirements and computational burden.

Introduction

Estimating depth from a single image is an important problem in computer vision. The goal of monocular depth estimation algorithms is to obtain a depth value for every pixel in an image. Obtaining a precise depth map directly benefits several applications such as semantic segmentation [2], pose estimation [26], object detection and tracking [27]. Recent approaches, based on deep learning, utilize an encoder-decoder framework in an attempt to fuse features of the image for extracting depth information [6]. However, memory requirements of recent methods are high and it would be beneficial to reduce their footprint.

The objective of this work is to achieve monocular depth estimation results that are comparable to state-of-the-art, while reducing the size of the network. We propose a framework called Structure Aware Waterfall for depth estimation (SAWdepth) network. Our SAWdepth method consists of an encoder-decoder structure that projects a higher amount of scales inside the network, compared to state-of-the-art methods [6], while having a smaller number of feature maps, resulting in a significantly smaller and consequently easier to train and faster network. Examples of depth estimation obtained with our method are shown in Figure 1.

An important component of our architecture is the use of the Waterfall Atrous Spatial Pooling (WASP) module [1], which combines the cascaded approach for atrous convolution with the larger FOV obtained from parallel configuration from the Atrous Spatial Pyramid Pooling (ASPP) module [3]. Our SAWdepth architecture, based on the Waterfall module for atrous spatial pooling and an encoder-decoder structure, which significantly reduces the size of the network.

Related Work

In current research, the vast majority of depth estimation methods rely in the use of CNNs for the inference of depth. Initial developments of monocular methods for depth estimation were conducted by [7] using a multi-scale approach. The use of ResNet [12] backbone for feature extraction generated significant improvement in depth estimation accuracy in [20]. Other networks utilized for the task of depth estimation include DenseNet [16] and the Squeeze-and-Excitation network (SENet) [14].

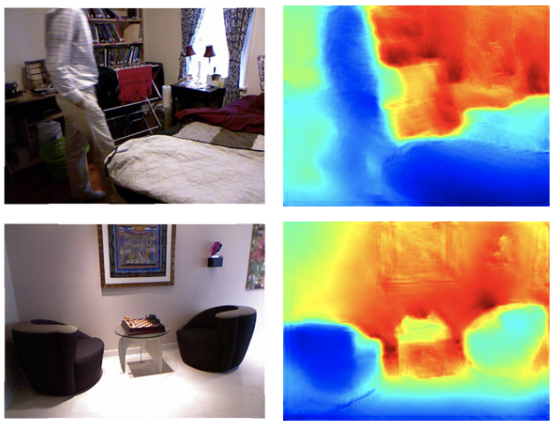


Figure 1. *Depth estimation examples obtained with our SAWdepth method.*

A complication resulting from the lack of pooling layers is a reduction of spatial invariance. Thus, additional techniques are used to recover spatial definition, namely, Conditional Random Fields (CRF) and atrous Convolutions. The implementation of postprocessing CRF by Li et al. [21] improved the efficiency of networks for the depth estimation of small objects that were previously hard to identify due to loss of resolution from pooling. Aiming for better delineation of objects in the image, [30] combines CNN and CRF in a single network to incorporate the probabilistic method of Gaussian pairwise potentials during inference. A limitation of architectures using CRF is that CRF has a greater difficulty capturing boundaries, as these regions have low confidence in the unary term of the CRF energy function.

The work of Ronneberger et al. [25] introduced the U-Net architecture, consisting of a "U" shape network for the encoder and decoder stages of processing. The U-Net approach can be applied for the depth estimation task that presents similar complexity to semantic segmentation and in addition requires a contextual interpretation of the image.

An important challenge with pixel-wise tasks incorporating CNN layers is the significant reduction of resolution caused by pooling. Semantic segmentation with Fully Convolutional Networks (FCN) [22] addressed the resolution reduction problem by deploying upsampling strategies across deconvolution layers. These attempt to reverse the convolution operation and increase the feature map size back to the dimensions of the original image.

Multi-scale approaches became popular for overcoming the loss of pooling [8]. Hao et al. [11] initially made use of atrous

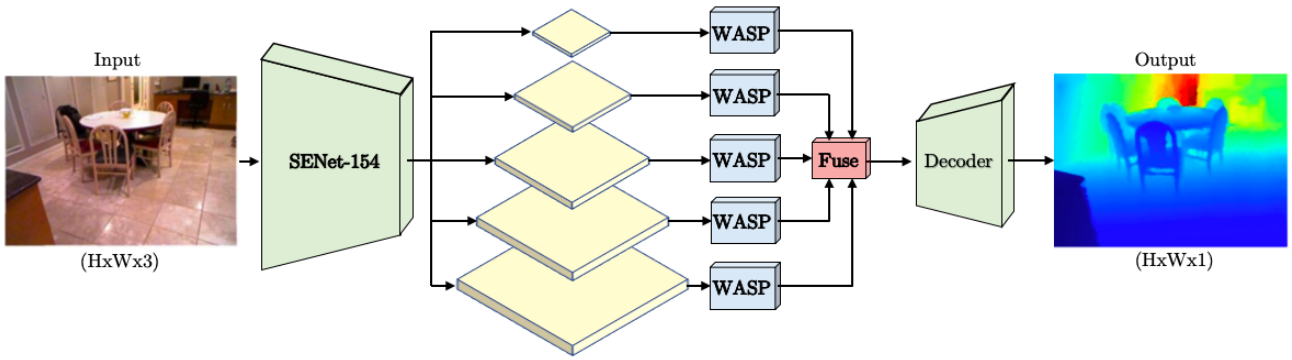


Figure 2. *SAWdepth architecture for depth estimation. The input color image of dimensions (HxW) is fed through the SENet backbone and WASP modules to obtain 256 feature channels at each scale with 4 dilation rates each. The decoder module generates the final estimation for depth at the original resolution.*

convolutions to access multiple scales for depth. Similarly, [15] implements a multi-scale approach with improved results by fusing feature scales, although it still lacks in precision for more complex objects. Other methods that use multi-scales include [6], as well as [28] which combines the multi-scale with CRF.

With the goal of reducing the complexity of depth estimation, the approach in [8] redefines depth estimation as an ordinal regression task by implementing a spacing-increasing discretization (SID) strategy. The use of ordinary regression loss applies a multi-scale approach in order to avoid additional unpooling layers and obtain a larger number of scales in the network.

Approaches to estimate depth without supervision were introduced by unsupervised [10] and semi-supervised methods [19]. The methods obtain the losses of the network by comparing the differences between the left and right side of the estimation map. A similar approach was used by [5] using a pair-wise information ranking system to propose and determine the estimation of depth in the image. The work by [24] aims to estimate depth using a geometric neural network. The method combines the geometric relation of the depth and normal surfaces.

Networks for depth estimation and semantic segmentation exhibit significant commonalities. A multitude of networks employ similar structures and occasionally perform both tasks simultaneously. Several networks rely on leveraging information from the backbone to perform both tasks in multi-scale approaches [7], [29], and [17].

A popular technique for maintaining the original resolution is the use of dilated or atrous convolutions [3]. Atrous convolutions aim to increase the size of the receptive fields in the network, avoid downsampling, and generate a multi-scale framework for processing. In the simpler case of one-dimensional convolution, the output of the signal is defined as follows:

$$y[i] = \sum_{l=1}^{L} x[i + rl] \cdot w[l] \tag{1}$$

where r is the rate of dilation, $\omega[l]$ is the filter of length L, $x[i]$ is the input, and $y[i]$ is the output. A rate value of one results in a regular convolution operation.

Motivated by the success of the Spatial Pyramids applied on pooling operations [13], the ASPP architecture was successfully incorporated in DeepLab [3] for semantic segmentation. The ASPP approach assembles atrous convolutions in four parallel branches with different rates, that are combined by fast bilinear interpolation with an additional factor of eight. This configuration recovers the feature maps in the original image resolution. The increase in resolution and FOV in the ASPP network can be beneficial for contextual segmentation as well as depth estimation.

The Res2Net module [9] is another promising multi-scale backbone architecture that achieves improved representations. The WASP module, recently introduced by [1], allows the application of atrous convolutions in a hybrid configuration between parallel and cascade assembling, leveraging both the increased FOV and reduced size of the network. We leverage this capability of the WASP method in the SAWdepth framework.

Methodology

We propose an efficient architecture for depth estimation that makes use of the large FOV generated by the WASP module combined with an encoder-decoder structure to fuse the multiple scales of representation extracted through our network.

The processing pipeline is shown in Figure 2. The input image is initially fed into a deep CNN, namely a SENet-154 architecture, following approaches by [15] and [6]. The resultant score maps from five different levels are fed into five different WASP modules for further extraction of features across scales. Our decoder extracts the final depth estimation as a combination of all fused scales obtained from the WASP modules.

WASP Module

Introduced by [1], the WASP module generates an efficient multi-scale representation that helps the network to increase the number of scales obtained without significantly increasing the size of the network. The WASP architecture, shown in Figure 3, is designed to leverage both the larger FOV of the ASPP configuration and the reduced size of the cascade approach to obtain multi-scale representations. The WASP module combines the benefits of the ASPP [3], Cascade [4], and Res2Net [9] modules.

The WASP module utilizes atrous convolutions, which are fundamental to ASPP, to maintain a large FOV. It also performs a cascade of atrous convolutions at increasing rates to gain efficiency. Furthermore, WASP incorporates multi-scale features. In contrast to ASPP and Res2Net, WASP does not immediately par-

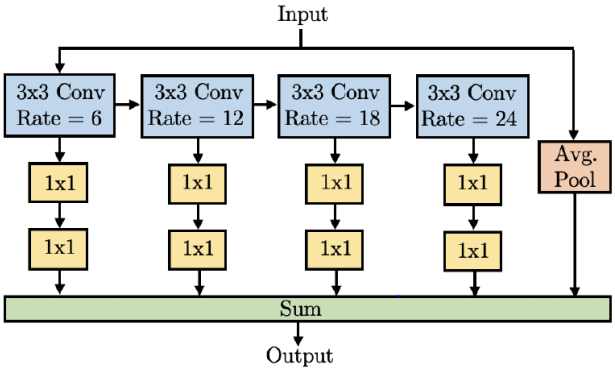


Figure 3. *Waterfall architecture in the WASP module [1].*

allelize the input stream. Instead, it creates a waterfall flow by first processing through a filter and then creating a new branch. WASP also goes beyond the cascade approach by combining the streams from all its branches and average pooling of the original input to achieve a multi-scale representation.

WASP aims to reduce the number of parameters in order to deal with memory constraints and overcome the main limitation of atrous convolutions. The four branches in WASP have different FOV and are arranged in a waterfall-like fashion. The atrous convolutions in WASP start with a small rate of 6, which consistently increases in subsequent branches (rates of 6,12,18,24). This configuration gains efficiency due to the smaller filter sizes, and creates multi-scale features with each branch that are combined to obtain a richer representation. The WASP module is utilized in the SAWdepth architecture of Figure 2 for depth estimation.

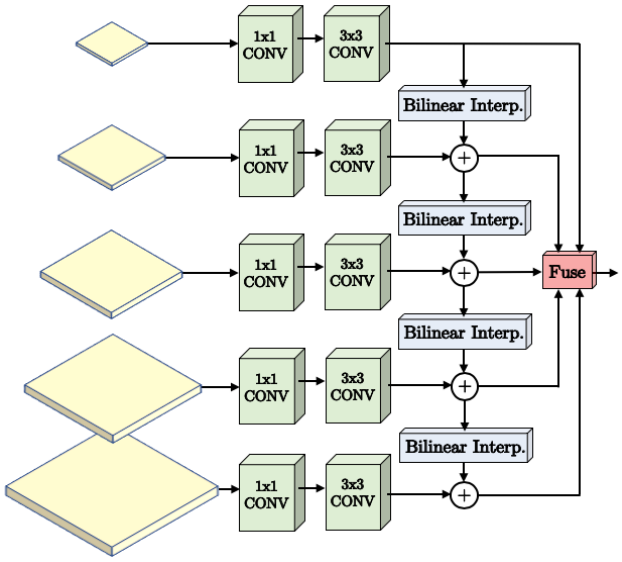


Figure 4. *Decoder module used in the SAWdepth pipeline. The inputs to the decoder are 256 channels for each scale after the WASP module. Each scale of the decoder is added to the next scale. The output of the decoder is the fuse of all maps and scales.*

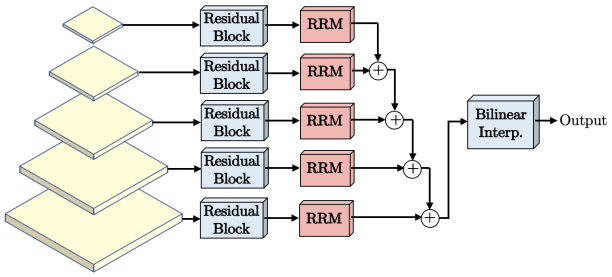


Figure 5. *Residual Block based encoder and decoder used for experiments.*

Decoder

Our decoder module estimates the final depth from multiple scales generated from five WASP modules, using higher resolutions to refine coarser scales. Figure 4 shows the structure of our decoder. Differently then previous works using a multi-scale encoder [6], we used a larger amount of maps from lower level features and a lower amount of maps from the pyramid feature stage (WASP module). By re-configuring the decoder in this fashion, we are able to further reduce the size of the network while maintaining the larger amount of map scales.

The depth maps resulting from lower resolutions are intended to obtain a more general overall depth in the image, while higher resolution maps extract the details in the image depth. We predict the depth from each image after two convolutional layers that are added to the higher resolution branch after bilinear interpolation. By computing the mean square error at every level of resolution in our decoder, we refine the depth estimation at various levels of detail, extracting the final depth estimation from the combined higher resolution layer combined with all other lower resolution representations.

Structure Aware Residual Network

In addition to our proposed SAWdepth method, we also experimented we other approaches for depth estimation while reducing the size of the network. We considered the Residual Block, obtained from the ResNet architecture, at every scale of our backbone, resulting in a Structure Aware Residual Block (SARB). The output of each residual block is processed through a Residual Refinement Module (RRM) introduced by [6]. After fusing all scales, the output consists of the bilinear interpolation of the scales to the original resolution of the image. Figure 5 shows the stages used after the backbone for the estimation of depth in this experimental method.

Experiments and Results

Dataset

We performed training and testing of SAWdepth and SARB based on the NYU v2 depth dataset [23]. NYU v2 is an indoor segmentation and depth dataset, consisting of RGBD images obtained with the Microsoft Kinect sensor. The dataset is composed of 1,449 densely labelled indoor images, paired with their depth images. We utilized data augmentation on the NYU v2 samples to obtain a training set of over 50,000 images. The data augmentation for the training dataset consists of horizontal flipping, rotation of up to 5 degrees, cropping, light variation, color normalization,

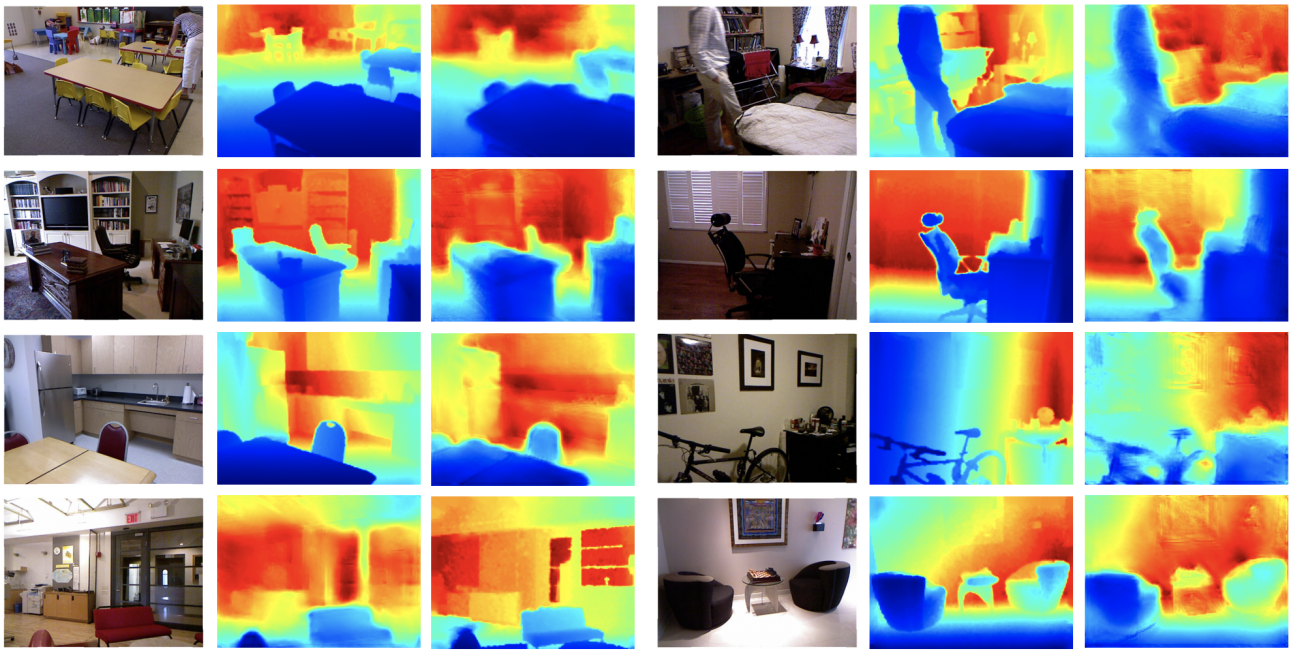


Figure 6. *Results sample for NYUD v2 dataset [23].*

and color jitter including brightness, contrast, and saturation, following similar procedures adopted by [15] and [6].

Similarly to previous studies, we evaluate our methods by assessing the Root Mean Square Error (RMSE). We used a Mean Square Error (MSE) during training in a Stochastic Gradient Descent (SGD) optimizer. We input the native resolution of the input image without resizing, in order to train the network with the most detail possible. We adopted a starting learning rate of 10^{-5} that is regulated through an Adam optimizer [18]. All experiments were performed using PyTorch 1.0 running on Ubuntu 16.04. The workstation has an Intel i5-2650 2.20GHz CPU with 16 GB of RAM and an NVIDIA Tesla V100 GPU.

Results

Following training, the SAWdepth and SARB methods were compared with SARPN in terms of accuracy and network size. The results for the NYUD v2 dataset are presented in Table 1. An RMSE loss of mIOU of 0.561 was achieved in only 4 epochs with SAWdepth, in contrast to 20 epochs used for SARPN [6]. Our SAWdepth network reduced the memory required for SARPN by 41.9%, from an original size of 6.33 GB to 3.68 GB, for a batch size of 1. We also tested the residual block network, and obtained a size reduction of 41.5% and a RMSE loss of 0.589. A comparison between SAWdepth and SARB shows that SAWdepth outperforms SARB in both accuracy and size, due to the use of the WASP module.

Table 2 shows a comparison of SAWdepth with other state-of-the-art methods, based on RMSE, for depth estimation on the NYUD v2 dataset. The results demonstrate the SAWdepth is competitive with respect to performance while significantly reducing the network size, based on the results of Table 1. Examples of depth estimation with SAWdepth for the NYUD v2 dataset are shown in Figure 6. It is noticeable from these examples that our method estimates the depth of main objects in the image with good accuracy. Challenging conditions for estimation include objects and walls located further away from the camera, having a less define reference frame.

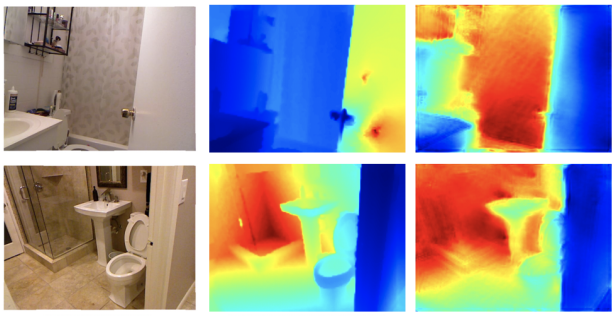
A significant source of error occurred from the presence of transparent objects such as glass and windows, since they do not present a solid surface for interpretation. Since most images in the dataset contains an upper middle section as the furthest away from the camera, this resulted in a bias to assign larger depth for regions near the image center. Representative examples of fail

Table 1: Depth estimation results and size comparison of SAWdepth with SARPN [6] for the NYUD v2 dataset [23].

Method	RMSE	Network Size	Size Reduction
SARPN [6]	0.514	6.33 GB	-
SAWdepth (ours)	0.561	3.68 GB	41.9%
SARB (ours)	0.589	3.70 GB	41.5%

Table 2: Results and comparison with other state-of-the-art methods for the NYUD v2 dataset [23].

Method	RMSE
SARPN [6]	0.514
Hu et al. [15]	0.530
SAWdepth (ours)	0.561
Geonet [24]	0.569
Xu et al. [28]	0.586
Li et al. [21]	0.821



(a) Original (b) Ground-Truth (c) SAW

Figure 7. *Occurrence of fail cases containing mirrors, glass, and light variation.*

cases are shown in Figure 7 for the presence of glass and mirrors in the image, as well as the bias, from training on this dataset, to locate larger depth in the center of the image.

Conclusions

We presented SAWdepth, a multi-scale architecture based on the WASP module for efficient depth estimation that drastically decreases the size of the network compared to other methods. The smaller size of our architecture results in faster training and easier implementation that improves its usefulness in applications, such as pose estimation, autonomous driving, and scene analysis.

Acknowledgments

This research was funded in part by National Science Foundation grant ♯1749376.

References

[1] Bruno Artacho and Andreas Savakis, "Waterfall Atrous Spatial Pooling Architecture for Efficient Semantic Segmentation," Sensors 19.24, p. 5361, 2019.

[2] Yuanzhouhan Cao, Chunhua Shen, and Heng Tao Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," IEEE Transactions on Image Processing 26.2, pp. 836–846, 2016.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Allan L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution and Fully Connected CFRs," IEEE Transactions on Pattern Analysis and Machine Intelligence 40.4, pp. 834–845, 2018.

[4] Liang-Chieh Chen, Goerge Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," CoRR abs/1706.05587, URL: http://arxiv.org/abs/1602.06541, 2017.

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng, "Single-image depth perception in the wild," Advances in Neural Information Processing Systems, pp. 730–738, 2016.

[6] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha, "Structure-aware residual pyramid network for monocular depth estimation," arXiv preprint arXiv:1907.06023, URL: https://arxiv.org/abs/1907.06023, 2019.

[7] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," Advances in Neural Information Processing Systems, pp. 2366–2374, 2014.

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep ordinal regression network for monocular depth estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2002–2011, 2018.

[9] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr, "Res2Net: A New Multi-scale Backbone Architecture," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[10] Ravi Garg, B. G. Vijay Kumar and Gustavo Carneiro, and Ian Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," European Conference on Computer Vision, Springer, pp. 740–756, 2016.

[11] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu, "Detail Preserving Depth Estimation from a Single Image Using Attention Guided Networks," International Conference on 3D Vision (3DV), pp. 304–313, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 37.9, pp. 1904–1916, 2015.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.

[15] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.1043–1051, 2019.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.

[17] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention driven loss," Proceedings of the European Conference on Computer Vision (ECCV), pp. 53–69, 2018.

[18] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization". In: arXiv preprint arXiv:1412.6980, 2014.

[19] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe, "Semi-supervised deep learning for monocular depth map prediction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6647–6655, 2017.

[20] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," Fourth International Conference on 3D vision (3DV), pp. 239–248, 2016.

[21] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1119–1127, 2015.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrel, "Fully Convolutional Networks for Semantic Segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

2015.

[23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor Segmentation and Support Inference from RGBD Images," Proceedings of the European Conference on Computer Vision (ECCV), 2012.

[24] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 283–291, 2018.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," International Conference on Medical image computing and computer-assisted intervention, pp. 234-241, 2015.

[26] J. Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and others, "Efficient human pose estimation from single depth images," IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12, pp. 2821–2840, 2012.

[27] Shuran Song and Jianxiong Xiao, "Deep sliding shapes for amodal 3d object detection in RGB-D images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 808–816, 2016.

[28] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, Nicu Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5354–5362, 2017.

[29] Dan Xu, Wanli Ouyang, Xiaogang Wang, Nicu Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 675–684, 2018.

[30] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, Elisa Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3917–3925, 2018.

Author Biography

Bruno Artacho received his B.Eng from Sao Paulo State University (2015) and his M.Eng. from Memorial University of Newfoundland (2017), both in Electrical Engineering. He worked at Transport Canada (Ottawa, Canada) as part of the Unmanned Aerial System Task Force to assess risk and update the Canadian Air Traffic Policy. He is currently pursuing his Ph.D. in Engineering at the Rochester Institute of Technology in Rochester, NY.

Nilesh Pandey completed his Bachelor in Electrical Engineering at Ramrao Adik Institute of Technology (2016) and received his Master in Computer Engineering from Rochester Institute of Technology (2019). His research interests include Generative Adversarial Networks (GAN), and applications for tracking and pose estimation architectures.

Andreas Savakis is Professor of Computer Engineering and Director of the Center for Human-aware Artificial Intelligence (CHAI) at Rochester Institute of Technology (RIT). He received his Ph.D. in Electrical and Computer Engineering from North Carolina State University. Prior to joining RIT, he was Senior Research Scientist at Kodak Research Labs. His research interests include computer vision, deep learning, machine learning, domain adaptation, object tracking, human pose estimation, and scene analysis. Dr.Savakis has coauthored over 120 publications and is co-inventor on 12 U.S. patents.

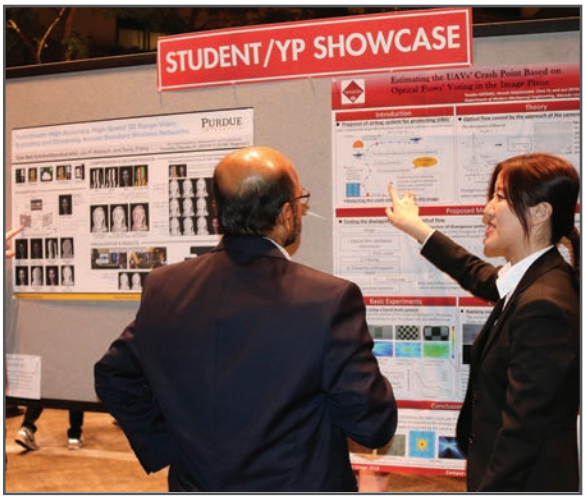
JOIN US AT THE NEXT EI!

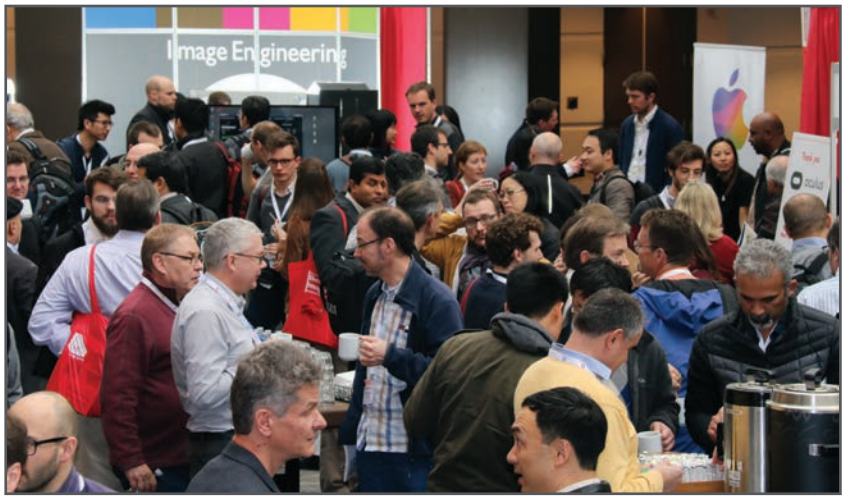
IS&T International Symposium on

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