2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

IMPROVING THE ROBUSTNESS OF CONVOLUTIONAL NETWORKS TO APPEARANCE VARIABILITY IN BIOMEDICAL IMAGES

Tolga Tasdizen, Mehdi Sajjadi, Mehran Javanmardi, Nisha Ramesh

Scientific Computing and Imaging Institute, University of Utah

ABSTRACT

While convolutional neural networks (CNN) produce state-of-the-art results in many applications including biomedical image analysis, they are not robust to variability in the data that is not well represented by the training set. An important source of variability in biomedical images is the appearance of objects such as contrast and texture due to different imaging settings. We introduce the neighborhood similarity layer (NSL) which can be used in a CNN to improve robustness to changes in the appearance of objects that are not well represented by the training data. The proposed NSL transforms its input feature map at a given pixel by computing its similarity to the surrounding neighborhood. This transformation is spatially varying, hence not a convolution. It is differentiable; therefore, networks including the proposed layer can be trained in an end-to-end manner. We demonstrate the advantages of the NSL for the vasculature segmentation and cell detection problems.

Index Terms— Convolutional networks, domain shift, domain adaptation, segmentation, detection

1. INTRODUCTION

With the increasing availability of computational resources and large datasets, deep learning based methods provide state-of-the-art solutions in applications such as image classification [1], semantic segmentation [2] and object detection [3]. CNNs have been shown to be most successful in the setting of supervised training. This success requires that i) a large enough labeled image set is available for training which captures the data distribution in the presence of potential sources of variability, i.e. pose and appearance of objects, and ii) that the training data distribution is representative of the images that the model is used to make predictions for. Both of these requirements pose significant challenges for biomedical applications. First, large datasets annotated by an expert are costly to generate. This is especially true for image segmentation models which requires fine grained, pixel level ground truth. Second, in biomedical applications, shifts from the training data distribution is likely when the learned model is deployed to make predictions for images collected at a different site or at the same site with different settings. In this paper, we propose a novel neighborhood similarity layer (NSL) that can be used as a layer in CNNs to overcome problems posed by variability due to changes in the appearance of objects such as contrast and texture. NSL is a parameter-free layer that is motivated by grouping with respect to similarity. It computes normalized inner products of feature vectors of the previous layer between a central pixel which acts as a frame of reference and the spatial neighborhood of that pixel. We demonstrate that NSL improves accuracy in regular training and domain adaptation scenarios for cell detection and eye vasculature segmentation.

2. RELATED WORK

2.1. Segmentation and detection with CNNs

Earlier applications of CNNs to biomedical image segmentation use a patch based approach where each pixel is classified one at a time using their surrounding patches fed to a CNN [4]. Patch based approaches can produce a highly localized output, but they are computationally inefficient due to redundant computations on overlapping patches. A more recent approach is the fully convolutional network (FCN) which does not have any fully connected layers and produces an image of class probability vectors that is the same size as the input image [2]. FCNs avoid redundant computations; however, this comes at the cost of localization accuracy. The U-net [5] architecture, a popular FCN for biomedical applications, uses skip connections between the encoder and decoder part of the network to preserve localization accuracy. Fakhry et al. [6] use residual connections between the encoder and decoder parts of a network to segment neurons in electron microscopy images. State-of-the-art results for vasculature segmentation are obtained with a convolutional network that includes specialized side layers [7]. CNNs and FCNs have also been used to detect cell centers [8] and count cells by predicting cell density [9] in microscopy images.

2.2. Learning invariant CNNs

Invariances to a predetermined set of transformations can be forced during training to decrease the undesirable impact of sources of variability on model performance. Tangent prop [10] penalizes the derivative of the network's output with respect to directions corresponding to the desired invariances

978-1-5386-3636-7/18/$31.00

such as rotation and scale change. Data augmentation [11] can be seen as a non-infinitesimal version of tangent prop where new labeled samples are generated with the chosen transformations and used in directly in supervised training. Type-invariant pooling [12] creates a more compact decision layer than data augmentation by creating transformation invariant features rather than capturing all possible transformations. Data augmentation has also been applied to unlabeled samples by penalizing the differences of the network's output to transformed versions of an unlabeled sample [13]. Scattering convolution networks [14] compute translation invariant descriptors that are also stable under deformations. Maxout networks [15] can also learn invariances present in the training data. However, these methods can not adapt to variations that are not predefined or well represented in the training data whereas the approach proposed here can adapt to unforeseen appearance changes that are not present in the training dataset.

2.3. Unsupervised domain adaptation

Domain adaptation methods are used to adapt classifiers to a new dataset whose characteristics might be different from the training set. In early work, co-training [16] generates pseudo-labels for high-confidence data in the target domain. More recent methods use an unsupervised reconstruction task in the target domain with a supervised classification task in the source domain [17, 18]. Residual transfer networks ease the assumption of a shared classifier between the source and target domains [19]. Domain adversarial networks [20] use a gradient reversal layer to extract domain invariant features for classification. In [21], the maximization of the domain classification loss between source and train domain features is replaced with the minimization of the maximum mean discrepancy. In a recent work [22], domain invariance is not forced on the feature representation, rather it is recommended to adapt the source images to the target domain using generative adversarial networks [23] and perform classification in the target domain only. While domain adaption for classification has been widely studied as discussed above, its use for image segmentation is a new area of research [24]. Our work differs from these approaches because it is an invariant learning scheme rather than domain adaptation. We do not use the target data for learning in any way which makes our approach applicable to a wider range of scenarios. For instance, a neuroscientist who wants to test automated cell detection methods in microscopy images could use a pre-trained model with NSL as is without retraining with domain adaptation which would require significant computational resources to be done on a feasible manner. Furthermore, the target dataset might consist of a small number of images, e.g. a single microscopy image, which might not be sufficient for domain adaptation.

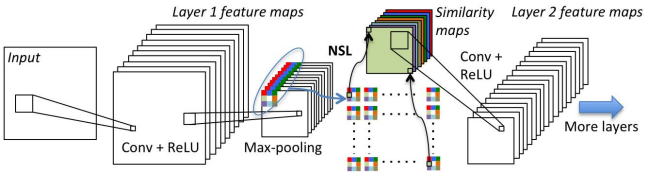

Fig. 1. NSL placed between the first 2 convolutional layers. At any location $\mathbf{x}$ (top left corner shown), the NSL computes a neighborhood of similarities to $\phi(\mathbf{x})$. This creates an image of two dimensional neighborhood similarities which are flattened into one dimensional vectors creating the output of the NSL. The relationship between a 3×3 $\mathcal{N}$ and 8 similarity maps are shown with the color coding.

3. NEIGHBORHOOD SIMILARITY LAYER

The NSL is motivated by grouping with respect to similarity. NSL improves the robustness of a convolutional network by focusing on the similarity of feature vectors and discarding absolute appearances. It transforms its input feature map using the feature vector at each pixel as a frame of reference, i.e. center of attention, for its surrounding neighborhood (Fig. 1). We define an image feature map as a n-dimensional vector valued function $\phi : \Omega \rightarrow \mathbb{R}^n$ where Ω is the discrete d-dimensional image grid and $\mathbb{R}^n$ is the n-dimensional Euclidean space. Let $\mathcal{N}$ be the set of m d-dimensional, non-zero offset vectors $\mathbf{v} \in \Omega$ defining a neighborhood structure. The neighborhood of a pixel $\mathbf{x} \in \Omega$ then follows as the ordered set $\mathcal{N}(\mathbf{x}) = \{\mathbf{x} + \mathbf{v} : \mathbf{v} \in \mathcal{N}\}$. In this paper, $\mathcal{N}(\mathbf{x})$ is taken as a square patch around $\mathbf{x}$ excluding the center pixel $\mathbf{x}$ (Fig. 1). Given a feature map $\phi : \Omega \rightarrow \mathbb{R}^n$ and a neighborhood structure $\mathcal{N}$, the neighborhood similarity layer $\psi : \Omega \rightarrow \mathbb{R}^m$ is the vector map of normalized inner products

$$\psi(\mathbf{x}, \phi, \mathcal{N}) = \left[\frac{\left\langle \phi(\mathbf{x}+\mathbf{v}) - \bar{\phi}, \phi(\mathbf{x}) - \bar{\phi} \right\rangle}{\| \phi(\mathbf{x}+\mathbf{v}) - \bar{\phi} \|^2 \| \phi(\mathbf{x}) - \bar{\phi} \|^2} \right]_{\mathbf{v} \in \mathcal{N}} \quad (1)$$

where $\bar{\phi} = (1/|\Omega|) \sum_{\mathbf{x} \in \Omega} \phi(\mathbf{x})$ is the mean feature vector.

We emphasize the following properties of NSL: (i) it is a parameter free layer and hence independent of training data, (ii) it is not a convolution, but a spatially varying operation that uses the feature vectors at each pixel as a frame of reference, (iii) the output feature map ψ has the same dimensionality as the number of pixels in $\mathcal{N}$ and (iv) when $\mathcal{N}$ is a square patch, the vector $\psi(\mathbf{x})$ corresponds to a square patch of similarities around and excluding $\mathbf{x}$. A NSL can be placed after any feature map producing layer. Note that, a convolutional layer following a NSL operates on a map of feature vectors which correspond to square patches of similarity vectors. Any number of NSL may be used as different layers of a network. In this paper, our experiments are focused on using a single NSL after the first convolutional layer of networks as transformation that improves robustness to appearance variability. Even though NSL (1) is a parameter-free layer, we need to compute its gradient to enable the backpropagation of errors through a

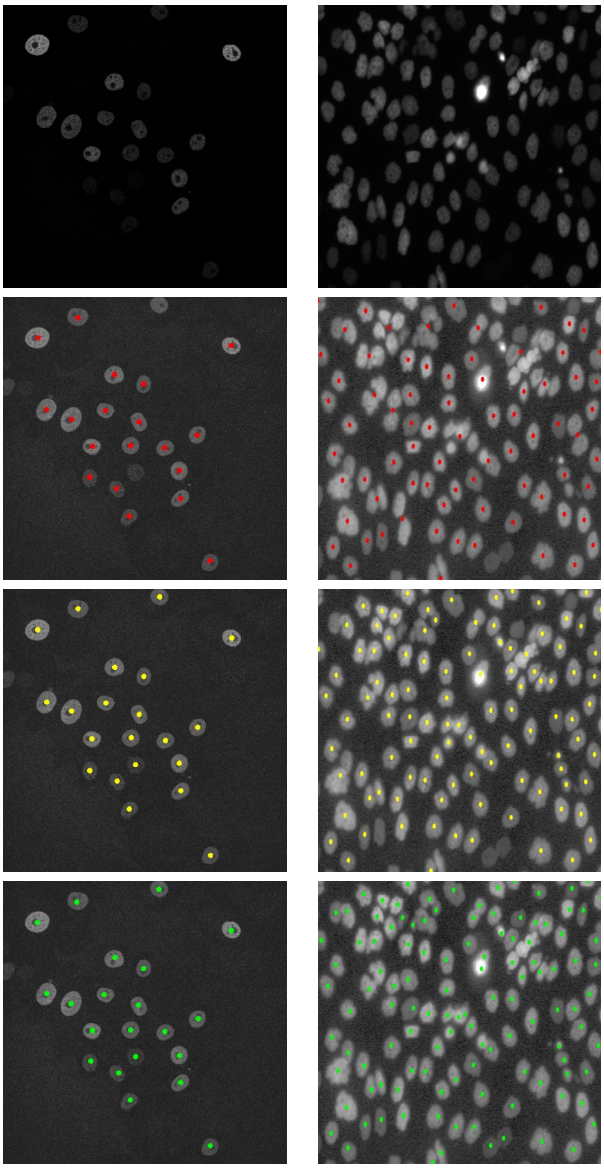

Fig. 2. Left: Fluo-N2DH-GOWT1-1; Right: Fluo-N2DL-HeLa-2 [25]. Top to bottom: Test image, predicted centroids without and with NSL, and ground truth centroids overlaid on contrast enhanced images (for display only).

NSL when it is used in a network. Letting $\tilde{\phi} = \phi - \bar{\phi}$, $\partial\psi/\partial\phi$ is found from (1) and $\partial\psi/\partial\phi = \partial\tilde{\phi}/\partial\phi \times \partial\psi/\partial\tilde{\phi}$ as

$$\left(1 - \frac{1}{|\Omega|}\right)\left[\frac{\tilde{\phi}(\mathbf{x}+\mathbf{v})}{||\tilde{\phi}(\mathbf{x}+\mathbf{v})|| \cdot ||\tilde{\phi}(\mathbf{x})||} - \frac{\left\langle \tilde{\phi}(\mathbf{x}+\mathbf{v}), \tilde{\phi}(\mathbf{x}) \right\rangle \tilde{\phi}(\mathbf{x})}{||\tilde{\phi}(\mathbf{x}+\mathbf{v})|| \cdot ||\tilde{\phi}(\mathbf{x})||^3}\right]$$

4. EXPERIMENTS

4.1. Cell Detection

We used two datasets from the cell tracking challenge [25] to demonstrate the effect of NSL in a supervised learning scenario. Both datasets have two sequences each, one for training and one for testing. The ground truth for these datasets consists of dot annotations at approximately the centers of cells. Our goal is to detect cells, i.e. cell centers; therefore, we place small Gaussian masks centered at the dot annotations to create the target values for training. We train the U-net architecture [5] and a modified U-net where we add a NSL after the first convolution layer. We experimented with 3×3 to 13×13 sizes for $\mathcal{N}$ and accuracy was observed to peak at 9×9. All results in this section are for 9×9 $\mathcal{N}$. Cells are detected as locations of the local maxima of the output of the U-net. Visual results for cell detection on both the datasets are seen in Fig. 2. With the addition of NSL, we are able to identify cells which have low contrast in the Fluo-N2DH-GOWT1 dataset. In the case of the Fluo-N2DL-HeLa dataset, we observe that some of the cells that have been detected as clumps are individually identified with the NSL. For quantitative evaluation, we use the Hungarian algorithm [26] to match ground truth dot annotations with local maxima of the network output and compute precision, recall, and F-score. We have also compared our detection results with state-of-the-art results from [27] and evaluated the U-net without NSL using locally contrast enhanced images. The quantitative results from our evaluations are reported in Table 1. We note that the improvement with NSL significantly exceeds the improvement obtained with contrast enhancement of the input images and provides state-of-the-art accuracy in terms of F-score.

Table 1. Cell detection testing precision, recall and F-score.

	Prec.	Rec.	F-score
Data set	Fluo-N2DH-GOWT1-1		
ECLIP [27]	**100.0**	31.0	47.33
U-net	94.91	82.71	88.39
U-net + contrast enhanced	94.66	86.16	90.21
U-net + 9×9 NSL	96.95	**94.59**	**95.75**
Data set	Fluo-N2DL-HeLa-2		
ECLIP [27]	78.0	**86.0**	81.80
U-net	85.74	69.94	77.03
U-net + contrast enhanced	85.86	71.97	78.30
U-net + 9×9 NSL	**89.60**	79.49	**84.24**

4.2. Cross-domain Vasculature Segmentation

We use the DRIVE [28] and STARE [29] eye vasculature datasets to demonstrate the effect of NSL in a domain adaptation scenario. For the baseline, we use U-net architecture [5] to label pixels as vasculature or background. We also modify this network by adding a NSL after the first convolution layer. The DRIVE [28] and STARE [29] datasets contain 40 and

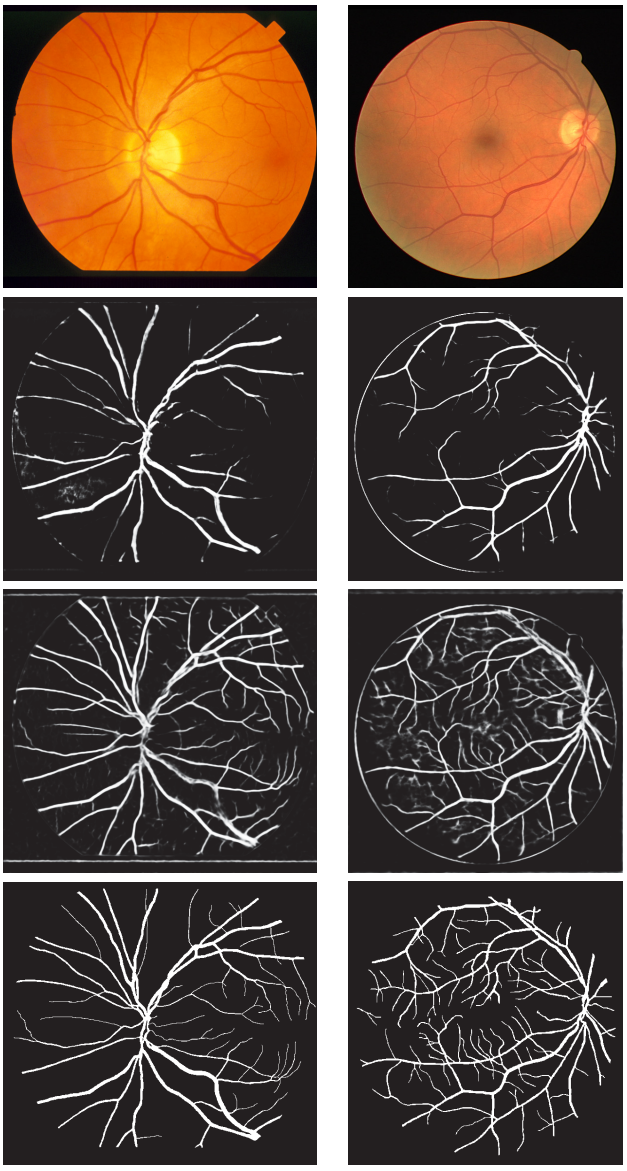

Fig. 3. Left: DRIVE $\rightarrow$ STARE; Right: STARE $\rightarrow$ DRIVE. Top to bottom: Fundus image from test set, without NSL, with 11×11 NSL, expert ground truth.

20 images, respectively, with manually annotated vasculature. We divide the datasets into 50% training and 50% testing for our experiments. Table 2 shows testing F-scores for different combinations of source and target datasets as well as different sizes of $\mathcal{N}$ for NSL. First, we observe that the inclusion of NSL increases the accuracy of segmentation for STARE $\rightarrow$ STARE while moderately decreasing it for DRIVE $\rightarrow$ DRIVE. More importantly, the inclusion of NSL significantly increases the accuracy of the segmentation when the target and source domain differ. Figure 3 illustrates that the recall of vasculature pixels is significantly improved in cross domain experiments with the inclusion of the NSL. We emphasize that no domain adaptation method is used for these experiments, i.e. the training does not use the target domain images in any way. A larger neighborhood size gives more accurate results for DRIVE $\rightarrow$ STARE, whereas the results for STARE $\rightarrow$ DRIVE do not significantly differ between 5×5 and 11×11 $\mathcal{N}$. Remarkably, the results for DRIVE $\rightarrow$ STARE with a 11×11 $\mathcal{N}$ surpass the results for STARE $\rightarrow$ STARE without NSL. To demonstrate that the same improvements in accuracy can not be obtained by histogram matching between domains, we preprocessed both source and target domain images by converting to hue, saturation and intensity space, performing histogram equalization on the intensity values, and converting back to RGB. The network without NSL was found to have F-scores of 79.10% and 59.91% for DRIVE $\rightarrow$ DRIVE and DRIVE $\rightarrow$ STARE, respectively. Finally, we note that our goal in this section is to provide proof of concept of domain adaptation with NSL. Our NSL layer can be incorporated into state-of-the-art networks for eye vasculature segmentation [7] to improve their performance and cross-domain applicability further.

Table 2. Test set F-score of models trained on source and tested on target (source$\rightarrow$ target).

	w/o NSL	5×5 $\mathcal{N}$	11×11 $\mathcal{N}$
DRIVE $\rightarrow$ DRIVE	80.34	80.41	78.83
STARE $\rightarrow$ DRIVE	62.22	69.93	69.92
STARE $\rightarrow$ STARE	73.13	78.08	78.78
DRIVE $\rightarrow$ STARE	65.20	70.49	74.00

5. CONCLUSION

We proposed a NSL that transforms its input feature map using the feature vectors at each pixel as a frame of reference for its surrounding neighborhood. NSL is differentiable; therefore, networks including a NSL can be trained in an end-to-end manner. NSL was shown to induce robustness to appearance variability in CNNs. Supervised training of a CNN can benefit from the inclusion of NSL because nuisance sources of variability, e.g. variable contrast in the cell images, are discarded by NSL. Furthermore, CNNs that include the NSL benefit from a feature representation that is robust to appearance changes between domains, e.g. the appearance of vasculature and background in the eye fundus images from different datasets, resulting in better cross domain accuracy.

6. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105, 2012.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, pp. 3431–3440, 2015.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, pp. 91–99, 2015.

[4] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *NIPS*, pp. 2843–2851, 2012.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[6] A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE Transaction on Medical Imaging*, vol. 36, no. 2, pp. 447–456, 2017.

[7] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Deep retinal image understanding," in *MICCAI*, 2016.

[8] Y. Xue and N. Ray, "Cell detection with deep convolutional neural network and compressed sensing,," *arXiv preprint arXiv:1708.03307*, 2017.

[9] W. Xie, A. Noble, and A. Zisserman, "Microscopy cell counting with fully convolutional regression networks," in *MICCAI Workshop on Deep Learning*, 2015.

[10] P. Simard, B. Victorri, Y. LeCun, and J. Denker, "Tangent prop - a formalism for specifying selected invariances in an adaptive network," in *NIPS*, 1991.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] D. Laptev, N. Savinov, J. Buhmann, and M. Pollefeys, "TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks," in *CVPR*, 2016.

[13] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *NIPS*, pp. 1163–1171, 2016.

[14] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[15] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *ICML*, 2013.

[16] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, ACM, 1998.

[17] M. Ghifary, W. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *ECCV*, 2016.

[18] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, 2016.

[19] M. Long, H. Zhu, J. Wang, and M. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *NIPS*, 2016.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, pp. 1–35, 2016.

[21] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, vol. 37, pp. 97–105, 2015.

[22] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Unsupervised pixel-level domain adaptation with generative adversarial networks." arXiv preprint arXiv:1612.05424, 2016.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adverserial nets," in *NIPS*, 2014.

[24] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[25] C. Solorzano et al., "ISBI cell tracking challenge, 2014. [online]." http://www.codesolorzano.com/celltrackingchallenge/.

[26] H. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[27] E. Türetken et al., "Network flow integer programming to track elliptical cells in time-lapse sequences," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 942–951, April 2017.

[28] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[29] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, 2000.