Cell Image Segmentation using Generative Adversarial Networks, Transfer Learning, and Augmentations

Michael Majurski
NIST
Gaithersburg MD 20899
michael.majurski@nist.gov

Petru Manescu
NIST
Gaithersburg MD 20899
peter04xrt@gmail.com

Sarala Padi
NIST
Gaithersburg MD 20899
sarala.padi@nist.gov

Nicholas Schaub
Department of Neurology
University of Michigan
nischaub@med.umich.edu

Nathan Hotaling
National Eye Institute at NIH
Bethesda MD 20892
nathan.hotaling@nih.gov

Carl Simon Jr
NIST
Gaithersburg MD 20899
carl.simon@nist.gov

Peter Bajcsy
NIST
Gaithersburg MD 20899
peter.bajcsy@nist.gov

Abstract

We address the problem of segmenting cell contours from microscopy images of human induced pluripotent Retinal Pigment Epithelial stem cells (iRPE) using Convolutional Neural Networks (CNN). Our goal is to compare the accuracy gains of CNN-based segmentation by using (1) un-annotated images via Generative Adversarial Networks (GAN), (2) annotated out-of-bio-domain images via transfer learning, and (3) a priori knowledge about microscope imaging mapped into geometric augmentations of a small collection of annotated images.

First, the GAN learns an abstract representation of cell objects. Next, this unsupervised learned representation is transferred to the CNN segmentation models which are further fine-tuned on a small number of manually segmented iRPE cell images. Second, transfer learning is applied by pre-training a part of the CNN segmentation model with the COCO dataset containing semantic segmentation labels. The CNN model is then adapted to the iRPE cell domain using a small set of annotated iRPE cell images. Third, augmentations based on geometrical transformations are applied to a small collection of annotated images. All these approaches to training CNN-based segmentation model are compared to a baseline CNN model trained on a small collection of annotated images.

For very small annotation counts, the results show accuracy improvements up to 20% by the best approach in comparison to the accuracy achieved using a baseline U-Net model. For larger annotation counts these approaches asymptotically approach the same accuracy.

1. Introduction

Dry macular degeneration is a loss of rod and cone cells caused by the death of Retinal Pigment Epithelial (RPE) cells. Age related Macular Degeneration (AMD[1]) affects 1 in 50 people by the age of 50. Recently, a novel experimental therapy was developed with induced Pluripotent Stem Cells (iPSCs) [5]. However, evaluating the functions of living iPSCs is challenging. Traditional evaluation approaches include assays testing DNA/RNA expression, immunolabeling, running gels, assessment of secreted proteins or other factors, and physiological tests. Many of these tests are invasive (require lysing or fixing the cells), labor intensive (hours to days to perform the assay), expensive (ELISA kits and gene arrays), and/or increase the likelihood of contaminating the cell population (placing tools into the culture area to measure physiological function).

Nevertheless, a correlation between cell function and the visual structure of RPE monolayers has been reported [3, 24]. Based on visual inspections, healthy iRPE monolayers consist of cells that are tightly packed together and their maturation over time correlates with the amount of

[1] https://nei.nih.gov/eyedata/amd

cell pigment. Thus, one can determine whether an iRPE cell implant is healthy or not based on image analyses of live cells imaged by a bright-field microscope and transformed to absorbance images. By segmenting cell boundaries from absorbance images, estimates of pigment concentration and shape features per cell and per population can be related to implant functional test measurements. The accuracy of the segmentation could have a significant impact on feature-driven modeling and the derived biological conclusions [17]. While fluorescently stained images of iRPE monolayers have been segmented using classical segmentation techniques involving edge detections and morphological operations [19, 23], brightfield absorbance images of living cells have not been segmented since the membranes of the cells are not easily discernible.

In order to automatically segment cells from absorbance images with high accuracy, we use a Convolutional Neural Network (CNN) based segmentation method. The main challenge of this supervised approach is that one must decide how to overcome the gap between millions of coefficients in a CNN model to be optimized and the labor required to sufficiently optimize model coefficients. In this paper, we compare three basic approaches that are less labor demanding than manual annotations of cell images, such as (1) building a Generative Adversarial Network (GAN) from un-annotated images, (2) performing transfer learning from already annotated Common Objects in Context (COCO) [16] dataset for Semantic Segmentation with out-of-cell microscopy domain, and (3) image data augmentation driven by a-priori knowledge about invariances in imaging. Other approaches not explored in this work include weakly and semi-supervised methods [20, 15, 29].

This work presents a segmentation accuracy comparison of six CNN-based segmentation models created by combining these three approaches and by applying the models to absorbance iRPE cell microscopy images. We assume that a small number of manual annotations are available to complete the domain adaptation of models pre-trained by GAN or COCO-based transfer learning. The benchmark segmentation accuracy for this comparative study comes from the most common approach in which the model is trained only on annotated images.

The rest of the paper is organized into four sections as follows: Section 2 presents related work on CNN-based microscopy image segmentation using only a few annotations and highlights the contributions of our work. Section 3 describes the comparative methodology in detail and Section 4 shows preliminary evaluation results. Finally, Section 5 discusses the results and Section 6 concludes this work.

2. Related Work

Recent advances in deep learning have led to novel segmentation techniques based on convolutional neural networks (CNN) [4, 22]. Among many types of CNN models, the U-Net model [22] has been successfully applied to segmenting biological images. To train CNN models, the challenge lies in obtaining a large number of training segments (i.e., annotations) that are usually created manually. This manual effort is costly and as a result, the number of annotations needed for training CNN models with millions of estimated coefficients is not always available. Moreover, when CNN models are applied in the bio-medical domains, microscopy and medical images can only be annotated by subject matter experts. Thus, the annotation creation by experts is limited to a small number of manually prepared samples for training CNN models. To overcome this lack of annotated training samples, approaches like data augmentation [9], transfer learning [7], and representation learning via Generative Adversarial Networks (GAN) [14] have been proposed in the past years.

Data augmentation techniques are based on geometrical and/or spectral transformations of training images while preserving their reference labels. Geometrical transformations rotate, translate, and mirror training images [9]. Spectral transformations alter the intensities of pixel values in the training set [6]. However, a challenge appears in selecting the suitable augmentation models and the range of their parameters to capture the variability of the entire image dataset and future image collections (i.e., both training and test images).

Transfer learning (TL) usually refers to fine-tune models already trained on different tasks and datasets that have plenty of annotations, such as the ImageNet dataset [7], Common Objects in Context (COCO) [16], or the PASCAL Visual Object Classes (VOC) [10]. This TL approach consists of replacing the last layers of the pre-trained models with randomly initialized ones that fit the purpose of the new application. Next, all network weights are optimized with respect of the biomedical training dataset [6].

Another approach to overcoming the lack of training annotations is based on Generative Adversarial Networks (GAN). GANs are unsupervised learning models that are able to generate detailed realistic synthetic images [11]. Thus, the GANs can increase the number of annotated training samples and hence yield improved accuracy of CNN-based classification or localization tasks [1].

GAN models can be considered as a two-player game between a generator, which learns how to generate samples resembling real data, and a discriminator, which learns how to discriminate between real and generated data. Both the generator and the discriminator cost functions are minimized simultaneously. The iterative minimization of cost functions eventually leads to a Nash equilibrium where neither can further unilaterally minimize its cost function. In the end, the GAN discriminator provides an abstract unsupervised representation of the input images. In a simple

GAN, the generator takes a random noise vector as an input and outputs an image [21]. Recent works have proposed encoder-decoder-like generators where the input of the generator is an image [30]. This approach has been used for style transfer in natural images.

More recently, GAN-based segmentation methods have been proposed in the literature. In [26], the authors replace the traditional discriminator with a fully convolutional multiclass classifier. The classifier assigns to each input image pixel one label that corresponds to a semantic class or to fake/real mark. In this way, they use unlabeled images during the training process.

In [2], the discriminator is adapted to distinguish between manually segmented cell microscopy images and generated images from CNNs. The generated (estimated) segmentation images are similar to manually annotated images and therefore are more accurate than those obtained from a simple CNN segmentation model. In addition, such methods have been used as domain adaptation techniques [30] to transform magnetic resonance images (MRI) into computed tomography (CT) images [14] or Differential Interference Contrast to Phase Contrast microscopy images [12]. These transformations could allow for the use of manual segmentations in one modality to segment images acquired in another modality.

The main contribution is a comparison of augmentation, transfer learning, and representation learning for building segmentation CNN models with minimal data. An initial representation is learned either using transfer learning or unsupervised GAN before being transferred to the CNN network and refined on the small number of manually segmented images. The segmentation accuracy of each CNN model is evaluated on images of iRPE cells with the contour/region based metrics for a varying number of annotated images. The novelty lies in quantifying the accuracy contributions of three approaches to the accuracy of a CNN model over a varying number of manually annotated images. In addition, we modified the GAN network to match the traditional encoder-decoder structure of the U-Net model to enable unsupervised representation learning and pre-optimization of the U-Net CNN weights.

3. Materials and Methods

This section is organized as follows. Subsection 3.1 presents the three segmentation approaches. Subsection 3.2 describes the dataset and metrics used for performance evaluations.

3.1. Three Segmentation Approaches

3.1.1 GAN Transfer Learning (TL-GAN)

GAN transfer learning (TL-GAN): We use GANs to extract an abstract unsupervised representation from all unannotated images. This representation is then transferred to the CNN encoder-decoder-based segmentation model before being further fined-tuned using the small number of available manually annotated images. We assume that only a small number of manually segmented images is available (annotations only).

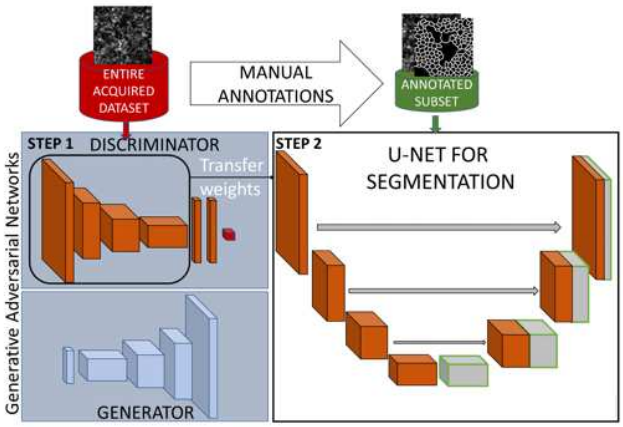


Figure 1. GAN-based transfer learning for a U-Net segmentation. Step-1: All the available data is passed through the GAN. Once the GAN optimization is finished, the discriminator weights are transferred to the encoder part of the U-Net. Step-2: The U-Net is trained on the manually annotated images. All weights in U-Net are optimized.

Figure 1 describes the use of GAN and U-Net CNN-based segmentation to improve cell boundary detection accuracy. All images of iRPE cells are passed through the GAN, so that the discriminator learns an unsupervised abstract representation of the data. The discriminator weights are then transferred to the U-Net encoder. Only the discriminator weights were transferred from the trained GAN to the segmentation U-Net. This is motivated by the GAN design. The generator convolutional weights convert the GAN latent space noise vector into a fake image while the U-Net decoder weights convert the compact segmentation representation at the bottom of the U-Net into a full segmentation map. The U-Net is further trained with manually annotated images by optimizing all weights. The U-Net model is identical to the original paper except we added batch normalization [22]. U-Net was trained to converge to minimum cross entropy loss using the Adam optimizer with default hyperparameters aside from a learning rate of 3×10^{-4} and with early stopping criterion. The GAN model and training procedure is inspired by DCGAN outlined in [21] except we did not modify the Adam beta1 parameter. Instead of following the DCGAN network architecture, we setup our GAN architecture using the encoder-decoder structural elements from U-Net. Specifically, the U-Net encoder was used as the GAN discriminator and the U-Net decoder was

used as the GAN generator. To prevent the discriminator from going to zero we used one-sided label smoothing (noisy labels) [25, 28]. This means fake images have a random label between 0.9 and 1.0 and real images having a random label between 0.0 and 0.1.

3.1.2 COCO Transfer Learning (TL-COCO)

We implemented a second transfer learning approach where all the weights of the network are initialized with pretrained weights from a U-Net model trained to convergence on the COCO dataset. Similar to the GAN-based approach, all encoder-decoder weights are optimized to segment the RPE cells absorbance images.

3.1.3 Data Augmentation (Aug)

Our augmentation approach is based on the following image models: rotation, reflection, translation, and scale. These models were implemented according to [18]. As documented in the past work [18] and based on our a priori knowledge of iRPE cell microscopy imaging, rotation, translation and reflection are the most accuracy-improving transformation for cell microscopy imaging applications. The choice of these augmentation models is motivated by the fact that a microscope objective and a specimen placement introduce geometrical variability that must be ignored during the cell segmentation (i.e., segmentation invariance to geometrical transformations). The choice of augmentation model parameters is empirical.

We compared baseline U-Net with training U-Net using these augmentation models parameterized by $\pm 10\%$, e.g. the scale augmentation is capped at a 10% modification of size per dimension. Table 1 summarizes the augmentation configurations.

Augmentation Model	Parameterization
Rotation	Uniform
Reflection	Bernoulii
Translation	Uniform $\pm 10\%$ Image Size
Scale	Uniform $\pm 10\%$ Image Size

Table 1. Augmentation models

3.1.4 Six CNN Segmentation Configurations

The transfer learning approaches are compared to the baseline model which is a U-Net model trained on only annotated images. Since the augmentation approach can be combined with the transfer learning approaches and the baseline model, we design six CNN segmentation configurations: {TL-COCO, TL-GAN, and baseline} $\times$ {with Aug and without Aug}. Each configuration is trained on different numbers of annotated images (6 randomly selected subsets) to yield a trained model. Each model is trained 10 times to take into account stochastic nature of model training.

3.2. Dataset

The dataset for this study consisted of absorbance microscopy images of human iRPE cells. The red, green, and blue wavelength filtered images from transmitted white light bright-field microscope were converted to absorbance images $absorbance = \log_{10}(\frac{1}{transmittance})$. The absorbance images map to a concentration of pigment according to Beer-Lambert law. The images were split into 1000 tiles of 256×256 pixels with 16 bits-per-pixel (bpp) and the corresponding ground-truth segmentation tiles. Each tile contains an RPE monolayer measuring approximately $0.5\,\mathrm{mm} \times 0.5\,\mathrm{mm}$. Overall, approximately $185\,000$ RPE individual cells were imaged. Since the objective of this study was to evaluate accuracy performance of multiple CNN models trained with a small number of annotated images, we chose 500 tiles for training the segmentation models and 500 tiles for evaluation (testing). To assess sensitivity of CNN model accuracy to the number of annotated images, we varied the number of training examples from 50 tiles to 500 tiles for training the six models described in the previous section. All $80\,403$ available absorbance image tiles of 256×256 pixels were used for the GAN unsupervised training. The absorbance images with their manually created segmentation masks, as well as all un-annotated absorbance images, are available for browsing and downloading from this URL [2].

3.3. Evaluation Metrics

The iRPE cells form a tightly packed mono-layer that becomes the implant. CNN models were trained to segment cell contours or boundaries. A pixel belongs either to a cell interior or its boundary. For evaluation we used the following metrics:

1. Average Contour Dice [8]: The Dice similarity index (or F-1 score) is computed according to the equation below where pixels in A belong to all segmented boundaries while pixels in B belong to all reference boundaries. The average of all test tiles is reported.

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

2. Average Adjusted Rand Index [13, 27]: Individual cell regions are obtained from segmented cell boundaries by applying skeletonization, inversion, and connected

[2] https://isg.nist.gov/deepzoomweb/data/RPEimplants

components operations. The labeled cell regions are compared using the ARI metric to determine labeling agreement. ARI is a more stringent metric because any disconnect in the cell boundary results in the whole cell being mislabeled.

4. Experimental Results

The experimental results are divided into four graphs based on the models {without and with augmentation} and evaluation metrics {Contour Dice and Region Adjusted Rand Index}.

4.1. Models without Augmentation

Figure 2 shows the evaluation results of the three segmentation approaches and 3×6 trained models at different numbers of annotated of images.

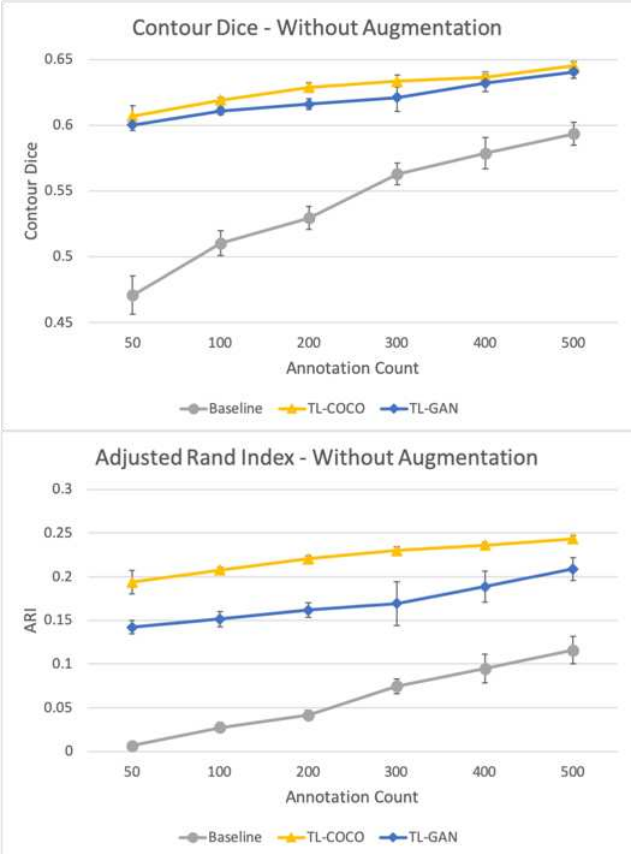


Figure 2. Comparison of segmentation accuracies without augmentation for the baseline U-Net models, U-Net model transfer learned from the COCO dataset, and U-Net model with the GAN discriminator.

Each graph in Figure 2 shows the values of one of the evaluation metrics presented in the previous section as a function of the number of training images for the multiple U-Net models. Error bars show the standard deviation across 10 repetitions of training the model to account for the stochastic nature of the process. No data augmentation was used in Figure 2 to show the pure effects of transfer learning from COCO and GAN. The Contour Dice shows significant improvement regardless of the source of the transfer learning information.

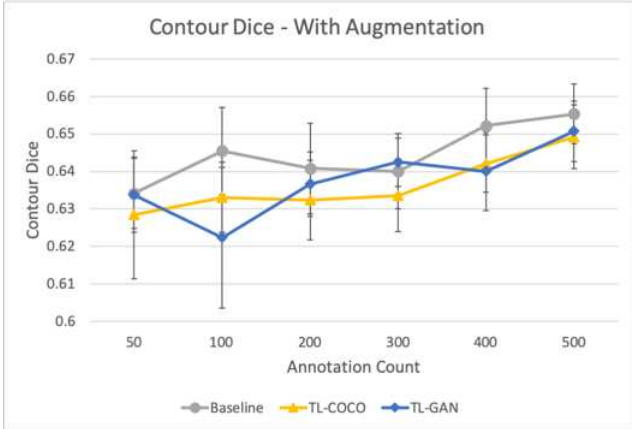
Transfer learning from COCO and GAN provide similar improvements measured by Contour Dice throughout the range of annotation counts. However, TL-COCO consistently achieves the highest accuracy when measured by ARI. One possible explanation is that the COCO pretrain enables the final segmentation model to better detect closed cell boundaries that do not have small boundary gaps. Small gaps cause the labeling routine to consider two cells as one which significantly impacts the ARI metric-based accuracy. The COCO dataset has been created for a semantic segmentation task and contains closed regions which causes the network to be sensitive to object edges. GAN pretraining learns a compact representation of the U-Net encoder, but the GAN only requires the encoder/discriminator to decide whether an image is real or fake. TL-GAN does not provide the pretrained network any indication that closed object regions are important. Additionally, the GAN pretrain only enables transferring the discriminator weights from the GAN to the encoder portion of the segmentation model. Thus, one half of the U-Net model is initialized with random weights and hence the GAN pretrain is transferring less useful information into the final segmentation model than the COCO pretrain.

4.2. Models with Augmentation

Figure 3 shows the evaluation results for baseline, TL-COCO, and TL-GAN configurations with the addition of all aforementioned augmentation models in Table 1.

Figure 3 presents much more complicated dependencies than the ones in Figure 2. The Contour Dice and ARI y-axes are rescaled for visual clarity but comparable between Figure 3 and Figure 2. With the randomness added by data augmentation the metric error bars across the 10 repetitions show significantly more variability. The data augmentation models based on our a-priori knowledge alone provide the most accuracy improvement although that advantage is within the error bars of TL-COCO with augmentation and TL-GAN with augmentation. For Contour Dice based evaluations, injecting domain knowledge via known data invariances using augmentation appears to be the most effective method for improving accuracy, edging out both transfer learning sources. However, for ARI based evaluations, the data augmentation slightly poisons the model accuracy for TL-COCO, reducing the average ARI across all repetitions from 0.24 to 0.19.

Figure 4 shows example segmentation results from each U-Net model. This figure highlights the importance of cell



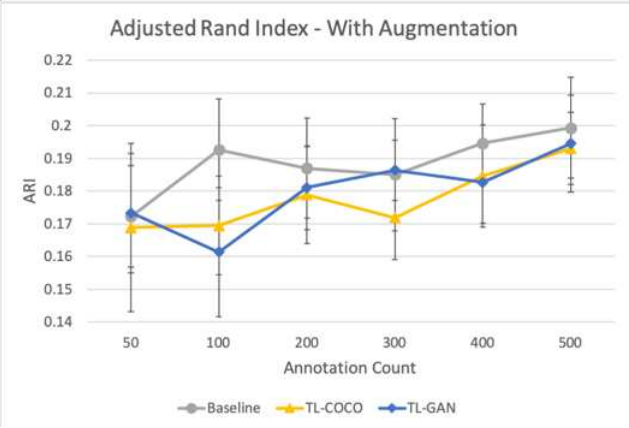


Figure 3. Comparison of segmentation accuracies with augmentation for the baseline U-Net models, U-Net model transfer learned from the COCO dataset, and U-Net model with the GAN discriminator.

edges and closed regions, especially for the ARI metric.

Table 2 provides a more detailed quantitative evaluation of the data presented in Figure 2 and Figure 3. Table 2 summarizes the Contour Dice and ARI measurements for 100 training images, $mean \pm standard\_deviation$ of each segmented test image across all 10 repetitions. The numerical results document the variability of a single models segmentation accuracy per image within the test data. Table 3 shows the same information for 200 annotations. The variance between individual image segmentation accuracy per repetition remains similar in magnitude across all annotation counts. Table 2 and Table 3 also highlight that the within repetition (per image) variation in segmentation accuracy is much higher than the variation in average metric accuracy between repetitions. This could be explained by a higher cell heterogeneity across the iRPE cell images than the heterogeneity of stochastic optimization paths across 10 repetitions of model training.

Configuration	Contour Dice	ARI
Baseline	0.510 ± 0.076	0.027 ± 0.068
Baseline with Aug	0.645 ± 0.086	0.193 ± 0.101
TL-COCO	0.619 ± 0.072	0.208 ± 0.069
TL-COCO with Aug	0.633 ± 0.091	0.169 ± 0.110
TL-GAN	0.611 ± 0.075	0.151 ± 0.105
TL-GAN with Aug	0.622 ± 0.105	0.161 ± 0.113

Table 2. Average per-image Dice and ARI accuracy metrics given 100 annotations

Configuration	Contour Dice	ARI
Baseline	0.529 ± 0.081	0.042 ± 0.087
Baseline with Aug	0.641 ± 0.103	0.187 ± 0.111
TL-COCO	0.629 ± 0.073	0.221 ± 0.076
TL-COCO with Aug	0.632 ± 0.110	0.179 ± 0.115
TL-GAN	0.616 ± 0.078	0.162 ± 0.113
TL-GAN with Aug	0.637 ± 0.095	0.181 ± 0.112

Table 3. Average per-image Dice and ARI accuracy metrics given 200 annotations

5. Discussion

The work presented in this paper focuses on improving cell boundary segmentation by using augmentation and transfer learning from two sources when a small number of manual segmentations is available. The main objective is to compare transfer learning from a very large collection of out-of-domain annotated images (e.g., COCO based transfer learning) or from a sufficiently large collection of in-the-domain un-annotated images and a small number of in-the-domain annotated images (GAN).

The results presented in the previous section indicate that both the representation learned via transfer learning and GAN improved the overall segmentation accuracy drastically when data augmentation is not used. In the TL-GAN configuration, the U-Net network encoder had already seen all of the available training and test images prior to the segmentation task while serving as a discriminator in a GAN setup. As a result, the encoder for the U-Net segmentation model reuses the representation learned while it was part of the GAN generator-discriminator tandem. However, because only the discriminator can be transferred into the U-Net model encoder, the U-Net decoder starts with the same random initialization of weights as a vanilla U-Net. This might explain the gap in accuracy between TL-COCO and TL-GAN without augmentation, since TL-COCO starts the entire U-Net network with previously learned weights instead of just one half of the weights.

The difference in ARI metric between the TL-COCO and TL-GAN results without augmentation might be attributed to the first few layers of the U-Net model with a COCO

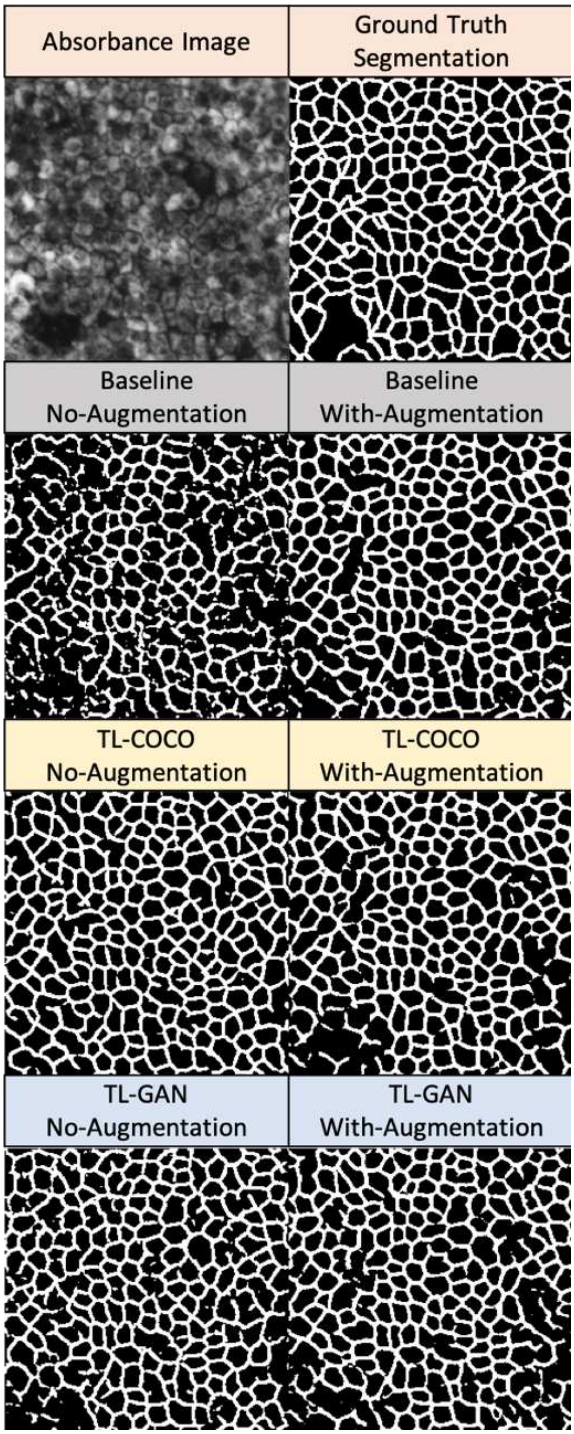


Figure 4. Example segmentation results from each configuration (baseline, TL-COCO, TL-GAN) with and without data augmentation. The images come from the replicate number one of all three model configurations trained on 100 annotations.

pre-train already having been optimized to find closed regions and edges as low-level features in the natural images. Even if the RPE images are different from natural images, the results indicate that this initialization of the convolution kernels improves model accuracy. In practice, the relative accuracy improvements due to each transfer learning approach depends on (1) the generalization of extracted image features from annotated COCO collection to the features characterizing RPE cell boundaries, (2) the number of un-annotated RPE cell images for GAN to learn the underlying representation, and (3) the number of annotated RPE cell images assuming that the CNN model for transfer learning has been fixed.

To overcome a lack of within domain annotated training data, our study focused on quantifying the accuracy of two transfer learning information sources, supervised out of domain data and unsupervised representation learning via GAN. In addition, we evaluated these two sources with and without data augmentation models selected based on a-priori microscopy imaging knowledge of the invariances that should be present in the segmentation model.

The trends of the curves in Figure 2 and Figure 3 indicate that the accuracy gain due to either of the transfer learning source decreases with an increase of the number of training examples and the accuracy gains become insignificant with many annotated images.

We used Contour Dice (or F-1) and ARI metrics to evaluate the segmentation outputs of the different models. The Contour Dice metric provided a measure of the general pixel-level segmentation quality while the ARI metric evaluated the quality and connectedness of the segmented cell borders. The ARI metric harshly punishes the network when it fails to completely connect any cell border since what would have been two cells is now considered one. The area of one of those whole cells is considered incorrect by the ARI metric.

GAN-based methods proposed in the literature have focused on using adversarial losses to modify the output of the U-Net segmentations so that they look similar to manual segmentations [2]. In contrast, we transfer the abstract unsupervised representations learned by GAN models to a segmentation task. Our approach is suitable to the common scenario of training a multi-million parameter CNN model with a few annotated example segmentations.

There is, however, a trade-off between the improvement in accuracy and an increase in the computational cost of training because our approach requires two independent training steps: the first consists of training an transfer learning source model (GAN or COCO) and the second consists of refining a CNN segmentation model using those pre-trained weights. Table 4 shows the compute time required to build both pre-trained models and to refine U-Net starting from the pre-trained weights. These times were generated on a single IBM Witherspoon node containing two 20-core IBM Power9 CPUs and four Nvidia V100 GPUs

Training Configuration	GPU Time
TL-COCO (pretrain + refine)	$4036 + 78$ min
TL-GAN (pretrain + refine)	$3120 + 78$ min
Baseline (refine)	78 min

Table 4. GPU Wall Time

with NVLink2 interconnection fabric. The table numbers account for parallel training across the four GPUs. For example, the COCO pretrain wall time was 1009 minutes using all four GPUs. Including augmentation into pretraining did not affect wall time.

Much more computing time is required during the pretraining phase for the creation of both the GAN and COCO models. The COCO pretrain time could have been avoided if published model weights were available as exist in many model zoos. However, now that we have the pretrained U-Net weights on COCO, they can be reused in any additional segmentation tasks. The U-Net COCO pretrain is a one-time cost whereas the GAN needs to be recomputed for each new dataset.

6. Conclusion

This paper presented a comparison of three approaches, six configurations, and 36 models retrained 10 times in order to understand segmentation accuracy trends. One of the approaches is based on a new restructuring of an encoder-decoder segmentation network (U-Net) into an unsupervised GAN model to enable representation learning with the same network elements which will later be used for segmentation. Promising results were shown when this approach was applied to segmenting individual cell contours from absorbance images of human iRPE monolayer implants. We hypothesize that TL-GAN shows inferior results to TL-COCO because only one half of the final segmentation network could be transferred from the GAN representation. This stands in contrast where the all the weights can be transferred from the COCO segmentation pretrain. All approaches described in this paper could be used to improve any other segmentation task.

Which of the approaches yields higher accuracy improvement depends on generalization of pre-trained models and our a-priori knowledge. One could improve the GAN representation learning pre-optimization by collecting more un-annotated RPE cell images. In comparison, the size and content of the source transfer learning dataset (COCO) is fixed unless more manual effort is invested into expanding the COCO dataset.

7. Disclaimer

Commercial products are identified in this document in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products identified are necessarily the best available for the purpose. Analysis performed [in part] on the NIST Enki HPC cluster.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[2] Assaf Arbelle and Tammy Riklin Raviv. Microscopy cell segmentation via adversarial neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 645–648. IEEE, 2018.

[3] Shagun Arora, Alia Rashid, Micah Chrenek, Qing Zhang, Soojung Park, Hans Grossniklaus, and John Nickerson. Analysis of human retinal pigment epithelium (rpe) morphometry in the macula of the normal aging eye. *Investigative Ophthalmology & Visual Science*, 54(15):2014–2014, 2013.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. pages 1–14, 2015.

[5] K. Bharti and B. Jha. Regenerating retinal pigment epithelial cells to cure blindness: a road towards personalized artificial tissue. *Current stem cell reports*, 1(2):79–91, 2015.

[6] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[9] Dieleman. Classify plankton with deep neural networks. `http://benanne.github.io/2015/03/17/plankton.html`.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Liang Han and Zhaozheng Yin. Transferring microscopy image modalities with conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 99–107, 2017.

[13] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[14] Yuankai Huo, Zhoubing Xu, Shunxing Bao, Albert Assad, Richard G Abramson, and Bennett A Landman. Adversarial synthesis learning enables segmentation without target modality ground truth. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1217–1220. IEEE, 2018.

[15] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[17] M. Brady M. Simon, J. Chalfoun and P. Bajcsy. Do we trust image measurements? variability, accuracy and traceability of image features. 2016.

[18] Michael Majurski, Petre Manescu, Joe Chalfoun, Peter Bajcsy, and Mary Brady. Impact of sampling and augmentation on generalization accuracy of microscopy image segmentation methods. In *3rd IEEE International Workshop on Computer Vision for Microscopy Image Analysis (CVMI)*, 2018.

[19] D. Malacara-Hernndez M. C. Wilson D. Williams P. Rangel-Fonseca, A. Gomez-Vieyra and E. Rossi. Automated segmentation of retinal pigment epithelium. *Journal of the Optical Society of America*, 30(20), 2013.

[20] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, may 2015.

[23] Ethan A Rossi, Piero Rangel-Fonseca, Keith Parkins, William Fischer, Lisa R Latchney, Margaret A Folwell, David R Williams, Alfredo Dubra, and Mina M Chung. In vivo imaging of retinal pigment epithelium cells in age related macular degeneration. *Biomedical optics express*, 4(11):2527–2539, 2013.

[24] M. Chrenek Q. Zhang-B. Bruce M. Klein J. Nickerson S. Bhatia, A. Rashid. Analysis of rpe morphometry in human eyes. *Molecular vision*, 22, 2016.

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[26] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017.

[27] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[28] David Warde-Farley and Ian Goodfellow. 11 adversarial perturbations of deep neural networks. *Perturbations, Optimization, and Statistics*, 311, 2016.

[29] Kun Xu, Hang Su, Jun Zhu, Ji-Song Guan, and Bo Zhang. Neuron segmentation based on cnn with semi-supervised regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2016.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.