Deep Image Segmentation by Quality Inference

Kelwin Fernandes
INESC TEC
Faculty of Engineering
University of Porto
Portugal
Email: kafc@inesctec.pt

Ricardo Cruz
INESC TEC
Faculty of Engineering
University of Porto
Portugal
Email: rpcruz@inesctec.pt

Jaime S. Cardoso
INESC TEC
Faculty of Engineering
University of Porto
Portugal
Email: jaime.cardoso@inesctec.pt

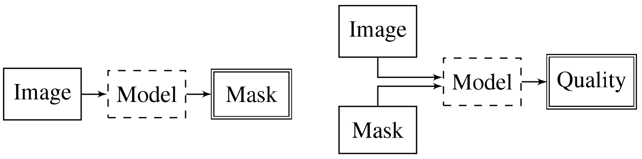
Abstract—**Traditionally, convolutional neural networks are trained for semantic segmentation by having an image given as input and the segmented mask as output. In this work, we propose a neural network trained by being given an image and mask pair, with the output being the quality of that pairing. The segmentation is then created afterwards through backpropagation on the mask. This allows enriching training with semi-supervised synthetic variations on the ground-truth. The proposed iterative segmentation technique allows improving an existing segmentation or creating one from scratch. We compare the performance of the proposed methodology with state-of-the-art deep architectures for image segmentation and achieve competitive results, being able to improve their segmentations.**

I. Introduction

Segmentation of images into its constituent parts is a decades-old problem. Traditional methods range from the usage of color threshold to clustering, and iterative methods such as region growing and active contours. However, all these methods require strong human supervision and tuning to find the right parameters.

The advent of machine learning, in particular convolutional neural networks like SegNet [1], has allowed semantic segmentation – where the parameters of the model are optimized automatically in a supervised manner on the object of interest. These new methods lack the iterative nature of previous techniques. The downside of such methods is the great amount of data required for training. Furthermore, applying these models to slightly different contexts, without re-training or fine-tuning, proves problematic.

Opposite to how machine learning algorithms are trained, as humans, we do not have a single ground-truth solution on our daily tasks, but a spectrum of alternative choices that are able to fulfill our goal to a certain degree. From an economics and social choice perspective, this decision process usually involves a utility function that reflects our satisfaction degree about a solution [2]–[4]. Based on such utility function, we are able to apply local (and non-local) updates in order to fulfill the requirements. In this work, we propose a novel segmentation paradigm: the convolutional neural network is trained to learn the quality of an image-segmentation pair. For a given dataset of images, multiple possible synthetically-created segmentations of varying qualities are used in the training process. The model not only has more information, but the problem complexity is reduced. The output, instead of having size $2^{\text{width}\times\text{height}}$ which is the area of the entire segmentation probability mask, the output is a single number corresponding to the quality of the given segmentation. This is represented in a diagram in Fig. 1.



(a) Traditional approach (b) Proposal

Fig. 1: Diagram representing segmentation flows.

Once training is performed, the segmentation process is no longer done in a single forward-pass like in the traditional approach. In our proposal, the segmentation process also makes use of the network as an oracle of the current segmentation quality to refine the mask in an iterative fashion. In order to do this, we rely on the backpropagation algorithm. This iterative process is inspired by previous segmentation techniques such as region growing, and the human visual system; human design evolves steps of "anticipated emergence" – sketching, in particular, involves seeing-moving-seeing steps [5]. The proposed model is an iterative process that can, not only produce segmentations from scratch, but also improve on those provided by an existing model.

II. State-of-the-art

Many traditional computer vision techniques have involved iterative processes. This is the case, for example, of region growing and active contours (also known as snakes).

In **region growing** [6], the segmentation is initialized from a seed point $R(0)$ at time 0, and then grows to include its neighbor pixels $\mathbf{N}$, $R(t+1) = \bigcup_j R(t) \cup N_j$ according to a user-provided logical predicate $P(R \cup N_j)$. In **active contours** [7], a curve composed of discrete points $\mathbf{v}(s) = \{(x(s), y(s))\}$, indexed by $s \in [0, 1]$, is found by minimizing an energy function $E_{\text{snake}} = \int_0^1 [E_{\text{internal}}(\mathbf{x}(s)) + E_{\text{external}}(\mathbf{x}(s))]\, ds$. The E_{internal} acts as a regularizer punishing many oscillations in the curve, while E_{external} is a function of the intensity or gradients within the image and can be both negative (repellent) or positive (attractor).

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

These traditional techniques have recently been surpassed by convolutional neural networks, which are capable of "semantic segmentation"; i.e. the ground-truth is used to guide the learning process.

The most widely used architectures are based on an **encoder-decoder** two-phased neural network; the image is first compressed into a smaller semantic representation, usually using convolutions and pooling (the encoding phase), and then decompressed into the final segmentation, usually using convolutions transposes (the decoding phase). The first example of this was SegNet [1].

A big problem in the encoding-decoding strategy is in avoiding the so-called checkerboard problem. Some detail is lost during the encoding step, which prevents the decoding step of doing as good a job as it could in refining the segmentation. This can result in a segmentation with a checkerboard effect. Since the encoding-decoding phases of the neural network are symmetric, U-Net [8] created so-called "skip-layers" where each decoding layer ℓ does not only receive as input the activation output of the previous layer $\ell - 1$, but also of the symmetric layers $L - \ell$ from the encoding phase, where L is the number of layers and $\ell > \frac{L}{2}$. In summary, each encoding layer computes the usual function $a^{(\ell)} = f(a^{(\ell-1)})$ and each decoding layer computes the function $a^{(\ell)} = f(a^{(\ell-1)}, a^{(L-\ell)})$ which is also using information from the encoding phase. It should be pointed out that U-Net was the best performing model in the ISBI 2016 Skin Lesion Analysis Towards Melanoma Detection Challenge [9], which supports its choice as one of the baseline models in this work.

Another important landmark in avoiding the checkerboard effect are **dilated kernels** (originally known as atrous convolutions). DeepLab [10], which makes use of such kernels, ranks first place in many benchmarks, including PASCAL VOC. There are no distinct encoding and decoding phases which produce activation maps of varying size. In this model, the activation maps remain the same size across the network. Filters are interconnected to the layers in a way, that each weight is shared across the same activation, so that the activation produced can have the same dimension along the network. Such a model is also used as a baseline in this work.

Also, worth mentioning is that iterative segmentation already exists in the form of recurrent neural networks adapted for segmentation [11]. The current work is innovative in that it is far simpler than any previous approach, since it most resembles traditional architectures used for segmentation.

III. Deep Segmentation by Quality Inference

The main idea of this paper can be summarized as follows:

1) learn to evaluate the quality of a certain segmentation mask for a given image;
2) use the model learned in 1) to find a local optimal segmentation by walking in the space of segmentation masks.

The proposed idea is illustrated in Fig. 2, where the image-mask is iteratively fed to the quality oracle in order to estimate the correspondence between them. Then, a search procedure is used to discover potential improvements to the input mask. This process is repeated until a convergence criterion (e.g. desired quality, improvement tolerance, number of iterations) is met.

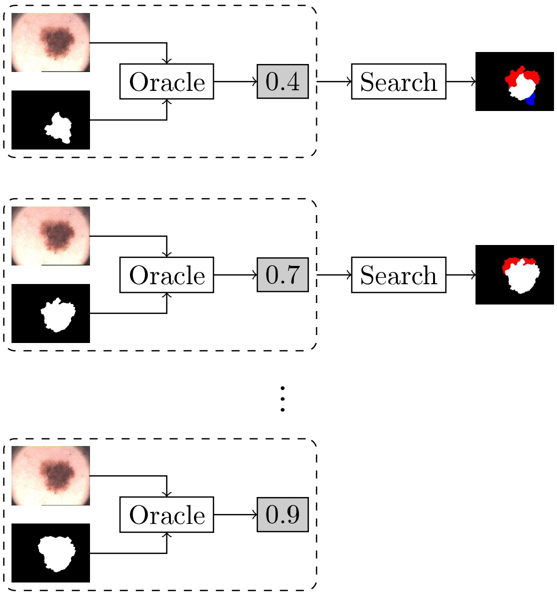


Fig. 2: Illustration of the iterative process of quality estimation and improvement. The search procedure indicates that red/blue regions should be added/removed from the input mask to improve the quality estimated by the oracle.

We propose to achieve 1) by using Deep Convolutional Neural Networks and 2) by using gradient ascent (backpropagation) over the input mask. We argue that it is more robust to learn to evaluate the quality of a given image/segmentation pair than to learn how to segment the image. Also, the quality concept has the potential to be more generic and easily transferred between tasks.

A. *Quality Inference as Deep Similarity Learning*

In this section, we address several aspects of the construction of a model capable of predicting the quality of a semantic image segmentation mask. How to express a utility of an entity (e.g. commodity, good, segmentation mask) is an open problem in economics and social choice [4]. The main choices for modeling utilities are pairwise preferences [12] and cardinal/ordinal functions [13]. These paradigms can be mapped to similarity learning as regression and ranking similarity learning respectively. In order to simplify the learning process (i.e. decision models and optimization), we model the utility (quality) of a segmentation mask as a cardinal function. Thereby, we are interested in regression models where pairs of objects – image i and mask m – are given to the model, being the quality $\hat{q}$ the outcome of the model. In our case, the utility is a measure of quality or correspondence between the mask and the image. This can be learnt by minimizing the

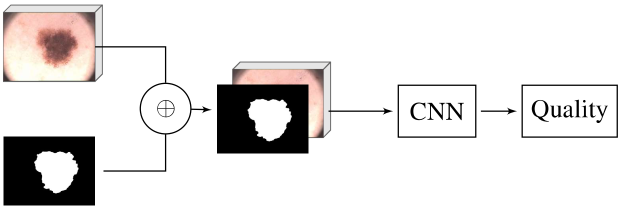


Fig. 3: Diagram representing a potential single-mixed stream approach to the problem.

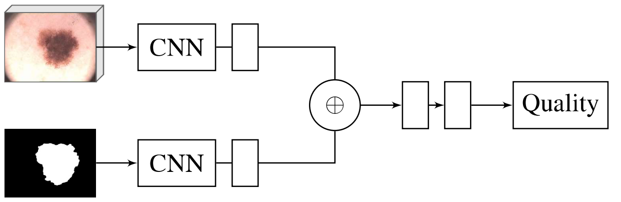


Fig. 4: Diagram representing a potential dual stream approach to the problem.

regularized loss function

$$\min_{\theta} \sum_{k} \mathcal{L}_{\theta}\left(f(i_k, m_k), q_k\right) + \lambda \mathcal{R}(\theta), \tag{1}$$

where $\mathcal{L}$ can be instantiated to the squared error of the estimated quality and the corresponding ground-truth quality and $\mathcal{R}$ is a regularizer of the model complexity (e.g. L_2). In this approach, we assume that, during training, a quality function for each image-mask pair is available.

The most straightforward strategy to solve this problem would be to use a traditional Convolutional Neural Network (CNN) where the mask is appended to the image as an additional channel (see Fig. 3). These two data sources (i.e. image and mask) belong to different categories (real-valued and binary) but are handled by the same operation (i.e. convolution) which may difficult the learning process.

An alternative approach would be to have separate streams for the input image and masks (see Fig. 4), being merged in the final dense blocks by concatenating their latent representation. The main drawback of this model would be that as we move deeper through the network, the intrinsic loss of resolution would limit the analysis of low-level patterns. Moreover, since each stream works with a different type of data, it is not clear how similar would be their latent representation.

In this work, we propose a deep architecture to tackle this problem, allowing an early integration of the information from image and masks. The main intuition behind this architecture is (i) having two streams that attempt to model the regions defined as foreground and background respectively by the input mask; (ii) streams communicate – "gossips" – to each other in order to increase/decrease their confidence on the recognition of their corresponding regions. In the rest of this section, we formalize the proposed architecture and its training procedure.

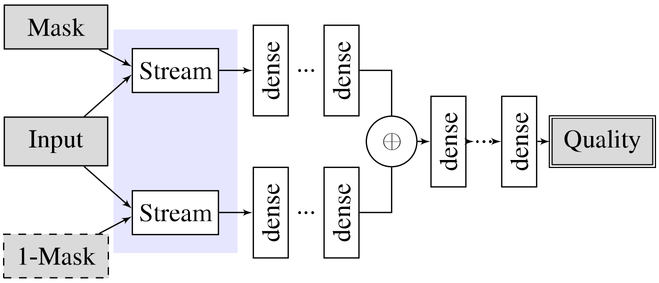


Fig. 5: Diagram of the general Gossip network. 1 - Mask denotes the complement of the mask.

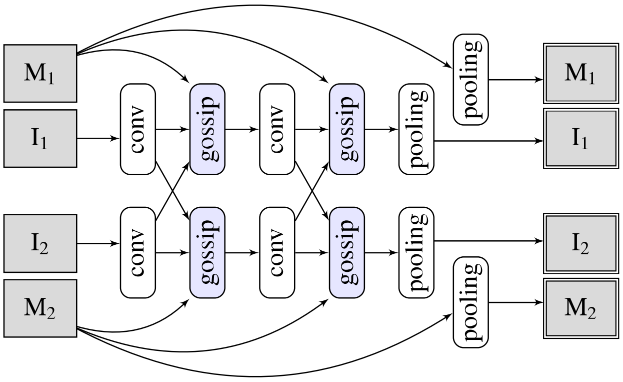


Fig. 6: Diagram of the two streams containing the gossip blocks.

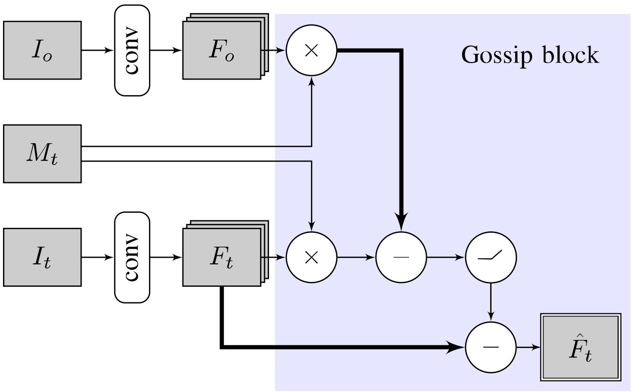


Fig. 7: Diagram of the gossip block. Thick arrows define the first argument of the operations that are not commutative.

1) Gossip Networks: Gossip networks are structured in such a way that the foreground and background representation is modeled by a pair of streams. This architecture is best described in three degrees of scale:

1) The **general architecture** is composed of the initial split into background and foreground streams, followed by dense layers, which produce the final quality score (see Fig. 5).
2) Within each **stream**, gossip blocks are consecutively intertwined with traditional convolutions; pooling is applied at the end of the stream to feature maps and masks (see Fig. 6).
3) At the lower-level, the **gossip block** combines the information from the two streams (see Fig. 7). The gossip block receives the feature maps obtained by 2D convolutional layers for the corresponding stream S and reciprocal layer $\hat{S}$. Then, for the region of interest of each channel, we penalize the activations where the reciprocal channel has stronger activations than the current one. We set the non-linearity term of these layers to the penalty term in order to favor the propagation of gradients to the original source pixels at inference. Namely, we avoid the problem of dead units where gradients are zero [14].

This type of double-helix connection between streams seen in the diagram was used to ensure an early interaction between both streams in order to reinforce/penalize their assumptions on each resolution-level.

The propagation of gradients through max-pooling blocks is sparse, leading to an unstable refinement of the segmentation masks (see section III-C). Thereby, we decided to use average pooling that ensures gradients are propagated through all the pixels in the block. Also, we restrain the activations of the convolutions to the valid regions in order to avoid unbalanced magnitude of the gradients in the edges of the image.

B. Training

Here, we describe how to efficiently train the proposed architecture in order to cover the space of potential segmentation masks in an efficient manner.

1) Similarity Metric: The similarity metric used in this work was the Sørensen-Dice coefficient D, often referred to as simply the Dice Coefficient. This index is given by the intersection over the union of the true and predict masks,

$$D(Y,\hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}. \qquad (2)$$

The index may be seen as a kind of F_1 score, hereby ensuring both positives and negatives (mis)classifications are captured equally in the metric.

2) Transformations: Two levels of transformations were applied: at the level of the ground-truth image and mask pairs, and then further transformations were applied to synthetically generate different segmentations of varying degrees of quality similarities – the ability to perform this latter data augmentation process is one of the key features that makes the Gossip architecture stand out from the current state-of-the-art.

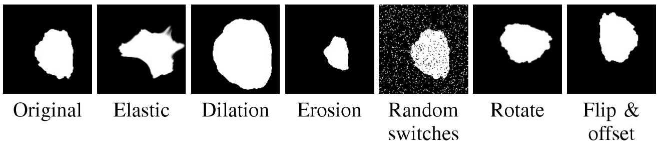


Fig. 8: Examples of synthetically created segmentations.

Traditional data augmentation was applied to the **ground-truth** image and mask pairs. These transformations encompass random horizontal and vertical flips, horizontal and vertical shifts, random zoom scales, as well as shear and contrast stretching deformations. These transformations were customized for each trained dataset.

Furthermore, the Gossip architecture has the ability to be trained for new segmentation masks created **synthetically** – with their corresponding similarity metric. The following transformations have been used. Notice all of them have one or multiple parameters, which are listed in brackets.

- Elastic deformation (α, σ, α'), which is a type of local, affine distortion [15]
- Morphological erosion and dilation (size)
- Random pixel switching (#pixels)
- Rotations (angle)
- Flip transformations (horizontal and/or vertical) and horizontal and vertical mask shifts (xoffset, yoffset).

These transformations are illustrated in Fig. 8. The parameters of these transformations were grid-searched in order to provide a balanced range of qualities of Dice. There is an inverse relation between the magnitude of each one of these parameters and the similarity index, but quantifying this relation is not straightforward. For this reason, the following procedure was applied.

First, the impact each combination of transformations and respective parameters had on the similarity index is computed empirically. For each transformation, the parameters are drawn by grid-search, and a similarity index $D(Y, Y')$ is computed between the ground-truth mask Y and the synthetically created Y'. Dice was discretized into B bins (in our case, $B = 8$). A frequency distribution p_{ib} was then found, representing the number of times p that the parameter combination i resulted in Dice b.

A couple of these transformations are stochastic (elastic and random pixel switch), therefore these two transformations were repeated 10 times for each ground-truth mask.

A second distribution is then computed from which we can sample parameters in order to ensure the similarity index is being drawn equitably across all bins (so that dice is evenly represented). This was performed by finding β which minimized the system composed of B equations, representing each bin b as

$$\begin{cases} \boldsymbol{\beta} \cdot \mathbf{p}_1 = \frac{1}{B} \\ \vdots \\ \boldsymbol{\beta} \cdot \mathbf{p}_B = \frac{1}{B}. \end{cases} \tag{3}$$

This was solved as a non-negative linear square problem, in order to ensure that each β represented a probability, which was then used to draw the parameters.

Notice there is a different $\boldsymbol{\beta}$ for each dataset and for each transformation. The reason why a different $\boldsymbol{\beta}$ was computed for each transformation was to ensure that each transformation was used uniformly. First, a transformation is chosen randomly, then the parameters are chosen using the respective p distribution. Otherwise, some transformations might have been used more than others. Notice that horizontal/vertical flips and mask shift offsets transformations were combined into a single transformation (see the previous bullet list) because the limited number of parameters made it impossible to create an equitable distribution for flip transformations alone.

C. Improving Segmentation by Backpropagation

The Gossip network f is trained to predict the quality of an image-mask pair, $\hat{q} = f(i, m)$ for a given image i and mask m pair.

During training, the Gossip network uses traditional gradient descent by computing the gradients of the loss function $\mathcal{L}$ relative to each weight w_i within the network, $\frac{\partial \mathcal{L}(q,\hat{q})}{\partial w_i}$, where q is the expected quality for the given image and mask pair. Each weight is then updated in the opposite direction of the gradient, $w_i \leftarrow w_i - \alpha \frac{\partial \mathcal{L}}{\partial w_i}$, using a learning rate α. This step is known as backpropagation.

On segmentation inference, inspired by the literature on generating adversarial examples [16], we propose improving a given segmentation by performing backpropagation on the predicted segmentation mask $\hat{m}$ by maximizing the predicted quality $\hat{q}$. An initial mask $\hat{m}_{ij}$ is then updated iteratively by gradient ascent,

$$\hat{m}_{ij} \leftarrow \hat{m}_{ij} + \alpha \frac{\partial \hat{q}}{\partial \hat{m}_{ij}}. \tag{4}$$

We have found, as reported below, that this technique works well both for improving existing segmentations, as well as creating segmentations from scratch, by starting with a black mask.

Some architectural design choices were based on allowing using gradients to update the mask. In order to avoid coarse gradients, average pooling has been used, rather than the more traditional maximum pooling approach. The derivative of average pooling is the averaging constant, while the derivative of maximum pooling is 1 for one pixel (the activation pixel) and 0 for all others. Empirically, maximum pooling would result in a very coarse segmentation.

For better convergence, gradients are normalized so that $\frac{\partial \hat{q}}{\partial \hat{m}_{ij}} \in [-1, 1]$ per mask and a sigmoid smoothness, $S(x) = \frac{1}{1+\exp -kx}$, is also applied to the gradients. We used a fixed number of backpropagation iterations on the segmentation mask. The sigmoid smoothness k, the number of backpropagation iterations, and the fixed update rate α were found empirically for each dataset using the validation set.

TABLE I: Summary of the datasets used in this paper. FS denotes the average foreground area.

Dataset	Ref.	# Imgs.	% FS
SmartSkins	[17]	80	37.5
PH2	[18]	200	49.1
ISBI 2017	[9]	2750	9.3
Teeth-UCV	[19]	100	23.7
Breast-Aesthetics	[20]	120	19.1
Cervix-HUC	[21]	287	5.8
Cervix-MobileODT	[22]	1613	17.1
Mobbio	[23]	2164	5.1

An alternative to backpropagation would have been using an exhaustive or heuristic exploration of the space of segmentation masks using the network as a fitness oracle. While these techniques would be able to discover non-local improvements, backpropagation stands as an efficient exploration strategy when the decision function is known and C^1 (differentiable).

IV. Experiments

In this section, we provide a detailed analysis of the experiments and results. First, we describe the datasets and baselines used in the validation of the proposed strategy. Then, we validate the performance of the proposed strategy on these datasets and on cross-database applications, where models are trained and validated on different datasets. We made the source code available for reproducibility purposes[1].

A. Data

We validated the performance of the proposed architectures on six real-life biomedical datasets. The datasets cover applications on the segmentation of melanoma, teeth, breast, cervix, and iris. Further details and sample images are shown in Table I and Fig. 9 respectively. These datasets were chosen to provide a range of segmentation of diverse complexity used in real clinical applications.

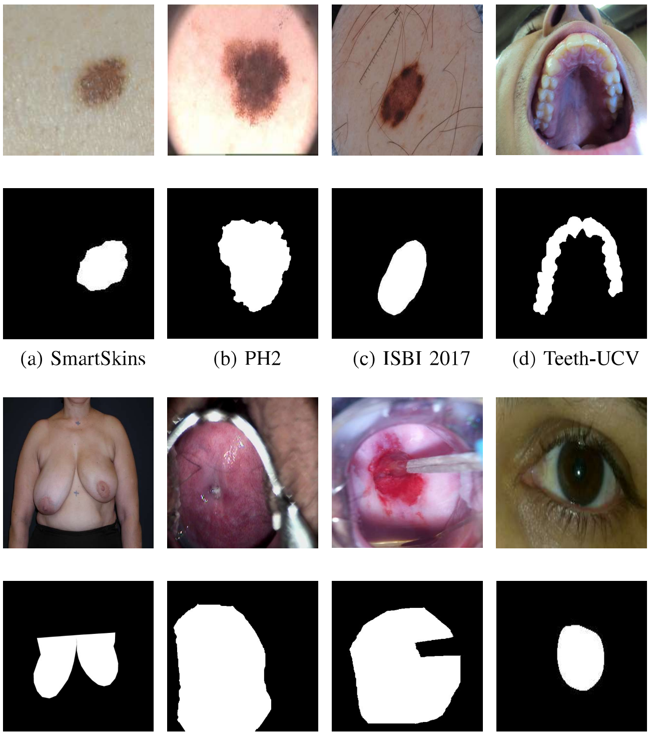
The goal of the first three datasets (i.e. SmartSkins, PH2, and ISBI 2017) is to segment skin lesions in dermoscopic images. The task on the Teeth-UCV, Breast-Aesthetics and Mobbio databases is to segment teeth, breasts, and iris from the background on natural RGB images. Finally, the object of interest in Cervix-HUC and Cervix-MobileODT datasets is the cervix, being the images acquired using digital colposcopy with several modalities (i.e. Hinselmann, Schiller and Green filter [24]).

We divided all the datasets into training, validation and test sets following the standard 60-20-20 partitioning. All images were first resized to 128 × 128 for easy comparison.

B. Models

In order to validate the performance of the proposed technique, we compared our results with the state-of-the-art

[1] https://github.com/kelwinfc/segmentation-by-quality



(a) SmartSkins (b) PH2 (c) ISBI 2017 (d) Teeth-UCV

(e) Breast Aesthetics (f) Cervix-HUC (g) Cervix-MobileODT (h) Mobbio

Fig. 9: Sample images and masks from the several datasets used for training.

U-Net [8] and U-Net with Dilated Convolutions (DilatedNet) [10]. For each model, we choose the best number of blocks (i.e. two convolutional layers and one pooling layer) on the validation set within the interval 2, 3 and 4. We use 32 filters on the first convolutional layer and double the value on each level as typically done in the literature [8]. The loss function for the Gossip Network was the mean squared error, and the networks were trained with 2, 3 and 4 gossip-gossip-pooling blocks, with one and two dense layers per each stream and in the final common section. ReLU activations were used on the intermediate dense layers and a sigmoid activation on the final layer to predict a quality value between 0 and 1.

We trained the models using Adam [25] for a maximum number of 500 iterations. In order to avoid overfitting, early-stopping was used after 50 iterations without improvement, and the best validation model was used.

C. Results

First, we explore the performance of the model in the most extreme scenario, where the initial segmentation is completely empty (i.e. black mask). Fig. 10 shows the performance of the network after N iterations of refinement. As can be seen in the Fig. 10, the network converges to a good solution on about 20 iterations. The remaining steps of the optimization focus on minor details with little impact on the overall performance. Some degenerated cases were observed where the network performance decayed after some iterations. This effect is the result of miss-estimations of the quality function, where the network was not able to learn the right direction for improvement. Fig. 11 illustrates the network progress at several stages of the refinement. As can be seen in the figure, the proposed approach emulates the traditional region-growing strategy [6], where the mask is progressively extended.

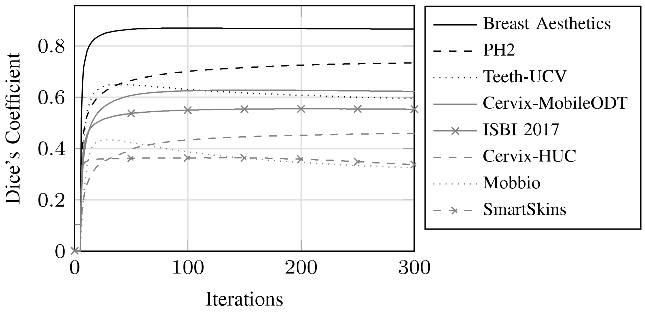


Fig. 10: Average Gossip Network performance after N iterations of refinement starting from empty masks.

The second scenario we explored was the iterative refinement of base segmentation done by the UNet and DilatedNet architectures. In this case, we choose the best number of refinement steps on the validation set, being 100 the maximum number of iterations. As can be seen in Table II, the proposed strategy improved the performance of the UNet and DilatedNet architectures in all databases.

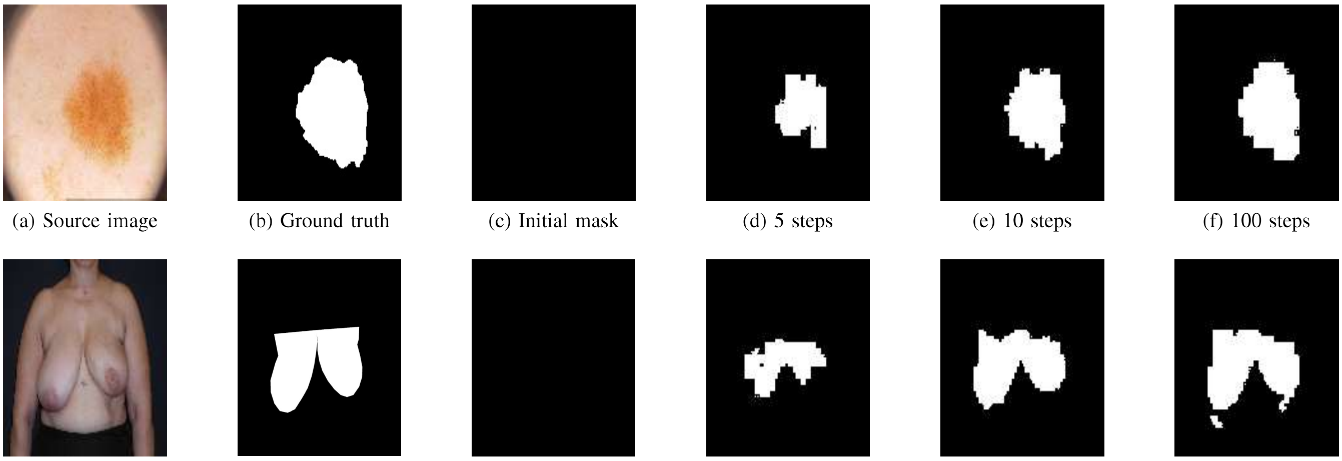
The main drawback of the proposed strategy is that, being an iterative procedure, it requires more time to segment an image than a single-shot model. However, as can be seen in Figure 10, the proposed methodology was able to achieve good results after a few iterations, even when starting from a blank

TABLE II: Contrasting models with and without Gossip enhancement, in terms of Dice's coefficient. Best results per database are presented in bold.

Dataset	U-Net		Dilated-Net	
	Original	W/Gossip	Original	W/Gossip
SmartSkins	76.62	79.45	76.35	**83.36**
PH2	83.70	84.09	85.52	**86.41**
ISBI 2017	71.35	**76.52**	72.06	76.11
Teeth-UCV	85.85	85.91	86.03	**86.14**
Breast Aesthetics	93.08	93.31	94.03	**94.15**
Cervix-HUC	77.25	**77.26**	75.37	75.37
Cervix-MobileODT	88.24	**88.25**	86.38	**88.25**
Mobbio	67.91	68.23	69.90	**70.11**

TABLE III: Cross-database model performance in terms of Dice's coefficient

Source	Target	U-Net		Dilated-Net	
		Original	W/Gossip	Original	W/Gossip
PH2	**SmartSkins**	76.87	81.21	75.71	**81.60**
PH2	**ISBI 2017**	64.44	67.02	66.13	**72.10**
Cervix-MobileODT	**Cervix-HUC**	57.94	**57.99**	32.18	36.00
PH2	**Cervix-HUC**	44.44	50.28	60.42	**60.62**



(a) Source image (b) Ground truth (c) Initial mask (d) 5 steps (e) 10 steps (f) 100 steps

(g) Source image (h) Ground truth (i) Initial mask (j) 5 steps (k) 10 steps (l) 100 steps

Fig. 11: Iterative refinement of images from PH2 and Breast Aesthetics datasets, respectively, using Gossip Networks. Initial masks are completely void.

mask.

Finally, we validate the performance of the proposed model on cross-database scenarios, where the model was trained for a given task and validated on a different dataset. This is common in applications where training data is synthetic due to field restrictions (e.g. aerospace) and cross-sensor applications. Results of this experiment are presented in Table III. In the first two cases, we use datasets for melanoma segmentation. We can observe large gains achieved by the Gossip network being initialized by the U-Net and Dilated-Net masks. For the validation of cervix segmentation (i.e. Cervix-MobileODT to Cervix-HUC), we observe a drop in the model performance when compared to training on the Cervix-HUC dataset directly. However, the Gossip Network achieves better performance than its counterparts. The last case covers cross-domain transitions, from melanoma to cervix segmentation. We observe a gain of about 6% when comparing the U-Net and Gossip Networks. The intuition behind these gains is that the notion of segmentation quality is more robust to changes in the data distribution. Namely, some concepts associated to predicting the quality of a segmentation such as the alignment between edges, the difference of contrast between foreground and background can be easily transferred among tasks.

V. Conclusion

This paper addresses the problem of semantic image segmentation with deep neural networks. We propose a new paradigm, based on similarity learning techniques, that tries to learn a quality function that maps an image-mask pair to the corresponding segmentation quality. Using the proposed architecture and, in combination with backpropagation, the proposed strategy is able to improve segmentation masks by maximizing the expected quality. By framing the problem as a regression task, we reduce the output complexity. Moreover, we are able to exploit the dataset size by learning from a large number of synthetically-generated candidate segmentation masks with their corresponding quality values.

We validated the proposed strategy in several biomedical applications and achieved good results when compared with the state-of-the-art U-Net and Dilated-Net architectures, with negligible computational expense. Also, we validated the proposed approach on cross-database scenarios and achieved promising results.

As future work, we intend to explore pairwise approaches based on the triplet loss [26], where the learning process is driven by comparing the outcome of pairs of masks for a single image. The proposed network could also be used for dynamic ensembles of models from which to produce the final segmentation. Namely, the importance of each model on the decision can rely on the quality predicted by the proposed network. Smaller details that could be improved are the fixed learning rate and the fixed number of iterations used on the segmentation by backpropagation part of the work.

Acknowledgment

This work was funded by the Project “NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant numbers SFRH/BD/93012/2013 and SFRH/BD/122248/2016.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *arXiv preprint arXiv:1511.00561*, 2015.

[2] P. C. Fishburn, “Utility theory for decision making,” Research analysis corp McLean VA, Tech. Rep., 1970.

[3] J. S. Dryzek and C. List, "Social choice theory and deliberative democracy: a reconciliation," *British journal of political science*, vol. 33, no. 1, pp. 1–28, 2003.

[4] J. Lang, L. Van Der Torre, and E. Weydert, "Utilitarian desires," *Autonomous agents and Multi-agent systems*, vol. 5, no. 3, pp. 329–363, 2002.

[5] R. Oxman, "The thinking eye: visual re-cognition in design emergence," *Design Studies*, vol. 23, no. 2, pp. 135–164, 2002.

[6] J. Fan, D. K. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," *IEEE transactions on image processing*, vol. 10, no. 10, pp. 1454–1466, 2001.

[7] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[9] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[11] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[12] K. Fernandes, J. S. Cardoso, and H. Palacios, "Learning and ensembling lexicographic preference trees with multiple kernels," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 2140–2147.

[13] D. Ellsberg, "Classic and current notions of" measurable utility"," *The Economic Journal*, vol. 64, no. 255, pp. 528–556, 1954.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[15] P. Y. Simard, D. Steinkraus, J. C. Platt *et al.*, "Best practices for convolutional neural networks applied to visual document analysis." in *ICDAR*, vol. 3, 2003, pp. 958–962.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[17] M. J. M. Vasconcelos, L. Rosado, and M. Ferreira, "Principal axes-based asymmetry assessment methodology for skin lesion image analysis," in *International symposium on visual computing*. Springer, 2014, pp. 21–31.

[18] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5437–5440.

[19] K. Fernandez and C. Chang, "Teeth/palate and interdental segmentation using artificial neural networks," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2012, pp. 175–185.

[20] J. S. Cardoso and M. J. Cardoso, "Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment," *Artificial intelligence in medicine*, vol. 40, no. 2, pp. 115–126, 2007.

[21] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*. Springer, 2017, pp. 243–250.

[22] https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening.

[23] A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira, "Mobbio: a multimodal database captured with a portable handheld device," in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, vol. 3. IEEE, 2014, pp. 133–139.

[24] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Temporal segmentation of digital colposcopies," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2015, pp. 262–271.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.