2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Fast Skin Lesion Segmentation via Fully Convolutional Network with Residual Architecture and CRF

Wenfeng Luo
Department of Computer Science and Engineering
South China University of Technology
Email: irlyue@outlook.com

Meng Yang*
School of Data and Computer Science
Sun Yat-Sen University
Email: yangm6@mail.sysu.edu.cn
*Corresponding author

***Abstract*—Melanoma is known to be the most fatal form of skin cancers. In order to achieve automated diagnosis of such disease, a system is needed to accurately locate suspicious skin lesions using images captured by standard digital cameras. Recently, there exists a trend for the use of Fully Convolutional Network(FCN) to perform image segmentation task. In this paper, we propose a FCN-based processing pipeline that incorporates a deep neural net and a graphical model, to attain a segmentation mask of lesion region from normal skin. Our method extends the residual network by adding a transposed convolution layer to yield a FCN architecture. We demonstrate that the noisy outcome from FCN can be refined by a fully connected Conditional Random Field(CRF). Our model enjoys three major advantages over existing algorithms: simpler process pipeline, state-of-art accuracy in terms of segmentation sensitivity(95.6%) and fast inference time.**

***Index Terms*—Melanoma, Fully Convolutional Network, Transposed Convolution, Image Segmentation, Conditional Random Field**

I. Introduction

Melanoma is known to be the most dangerous form of skin cancer, due to the unlimited growth of pigment-producing melanocytes. Around the world, over 230,000 people were newly diagnosed as melanoma in 2012 [1]. And the active number had climbed up to 3.1 million over the next 3 years, causing nearly 60,000 deaths [1]. Despite high mortality, it's almost always curable if recognized in an early stage and treated carefully. Over the years, many methods have been come up with to ease early detection, including the ABCD signs(asymmetry, border irregularity, color pattern and diameter) [2] and ugly duckling [3].

In order to ease the pain of screening melanoma in the massive population, an automated system should be developed to locate the suspicious lesion region in pixel accuracy. Some early screening systems make use of the images captured by digital dermatoscope [4], which is a special device designed for dermatologists to exam skin lesion. A review of existing methods using dermoscopy images can be found in [5].

Recently, researchers have shifted their focus to systems that can take advantage of photos acquired by a standard digital camera. It's commonly treated as a semantic segmentation task in the field of computer vision. It's much more demanding for detection system considering the illumination effect and random noise in these images. Multi-stage Illumination Modeling [6] is commonly used as a preprocessing step to get rid of such illumination variation as shadows and bright spots. Traditionally, many handcrafted features were proposed to segment the clinical images. Majority of these features are based on color information, either in single channel [7] or multiple channels [8]. TDLS [9], referred to as the TD lesion segmentation algorithm, utilizes texture distinctiveness(TD) to locate skin lesion.

Nowadays, deep learning methods, especially deep convolutional networks, have achieved state-of-art results in image classification [10], object detection [11], image segmentation [12]. Many of the techniques have been applied in medical imaging field. To segment fundus image with low contrast, [13] trains a deep neural network on blood vessel dataset and conducts experiments on several variants to produce segmentation mask, including structured prediction, where multiple pixels are classified simultaneously. Work of [14] explores a cascade architecture of deep CNN to better exploit local textures and global context on images of brain tumor. In order to combine local texture and global structure information, a patch-wise model [15] extracts convolutional features from both a small local patch and a zoomed-out bigger patch centered at the target pixel. And patch-wise training is performed at different pixel locations. Following the success of fully convolutional network [12], a FCN [16] has been used to segment dermoscopic images and achieves pretty satisfying results.

Although many of the proposed models have achieved decent results in skin datasets, they either require complex processing procedures or suffer from pool inference runtime. In this paper, we come up with a novel processing pipeline with three major modules: data augmentation and preprocessing, a fully convolutional network with skip connections, post processing using Conditional Random Filed(CRF). Only the last two steps are necessary at test time. Our model manages to recover the precise border from the coarse outcome of a deep FCN. With a segmentation sensitivity of 95.6%, our

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

method outperforms all existing models on non-dermoscopic dataset. In addition, our processing pipeline is really time-efficient and runs to completion within 0.5 second during evaluation. Section II delves into each of the three processing modules. We show our experimental result in section III. Finally, conclusions are drawn in Section IV.

II. Process Pipeline

Our processing pipeline begins with image augmentation and preprocessing as to the insufficiency problem of training data. And a FCN extended from residual network is trained to generate segmentation mask. The outcome from FCN is not accurate enough due to the inherent invariance property of deep convolution network, so a fully connected Conditional Random Field(CRF) is applied as a post processing procedure.

A. Data Augmentation and Preprocessing

Deep neural networks are able to outperform traditional models based on handcrafted features due to its enormous capacity. But deep networks can only perform at their best given enough training data. Otherwise, they'll suffer from severe overfitting problem. Since there are only 126 images from DermQuest database [17], extra efforts are needed so as to make the best of available training images. We adopt the augmentation procedure as in Fig. 2.

For each training image, we would like to have the lesion appear in different locations during training. So we first calculate the tight bounding-box inside which the lesion resides. And then we randomly crop an image region that contains the whole bounding-box. This is quite different from just randomly cropping an image as in some image classification task since it fails to maintain the completeness of a lesion region. Next, we resize the cropped image region to a fixed size of 224*224. The network structure described below is by no means restricted to this specific input dimension. The resizing procedure is to support training with more than one image.

In order to let the network be aware of the illumination variation, we randomly adjust the brightness and contrast of the input image, leading the network to learn the best way of recovering the image without manual elimination at the beginning. The introduction of these two noises actually increases the richness of training data.

Lastly, it's a good practice to normalize the image to have zero mean and unit variance before feeding to a CNN. We loose the guided filter as in [15] since it's in itself a convolution operation. Through data augmentation, a single skin lesion may actually appear in different location, have different scale and brightness condition during training. We apply data augmentation on the fly so no extra storage is needed. Our experimental result shows the augmentation techniques are essential.

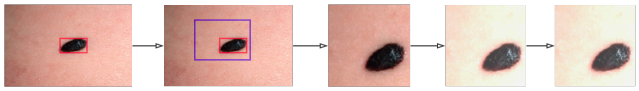

Fig. 2. Data augmentation(from left to right): calculating bounding-box(red), random crop(blue), resizing, random noise and normalization

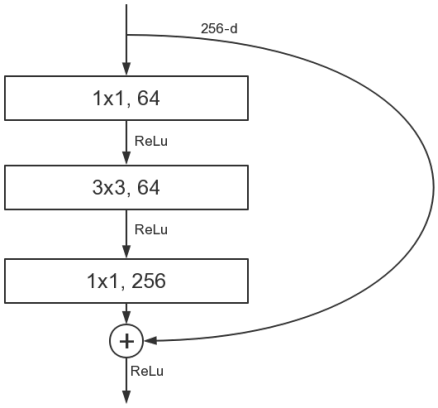

Fig. 3. An example of residual block unit, adopted from [10].

B. Fully Convolutional Network

The fully convolutional network(FCN) consists of a backbone network and a transposed convolution layer. The backbone network is made up of consecutive stack of convolution layers and downsamples the input image to attain an abstract feature representation. Then the transposed convolution is used to transform the downsampled feature map back to its initial dimension, giving the predictions at every pixel location in a single forward. This FCN setting significantly saves lots of repeated convolution computation compared to the patch-wise model [15] and thus yields a great boost of model runtime. Together with pixel-level ground truth, we can compute the cross entropy loss and train the network end-to-end.

1) Residual Network: We favor the Residual Network [10] as our backbone network over other deep CNNs. ResNet yields better deep feature representation due to its state-of-art result on image classification. Besides, ResNet looses most of the pooling layers to alleviate the inherent invariance properties of CNN. Moreover, the residual connections ease off the effect of gradient vanish and thus help optimization in some extent. Considering the size of our segmentation problem, we use only the first two block of resnet-50 [10]. Inside residual network, there exists connections skipping 3 layers, which in total form a residual block(see Fig.3).

Table. I summaries the detailed architecture. Every residual block is made up of different number of block units, each of which contains 3 convolution layers with 1x1, 3x3 and 1x1 kernel size. The skip connection happens inside each block unit. And downsampling of stride 2 is implemented in the last unit of each block. To sum up, the backbone network consists of 23 convolution layers in total, which indeed has pretty large field of view.

2) Transposed Convolution: A common approach to scale up an image is interpolation. For example, bilinear interpolation takes into consideration the four nearest pixels and applies

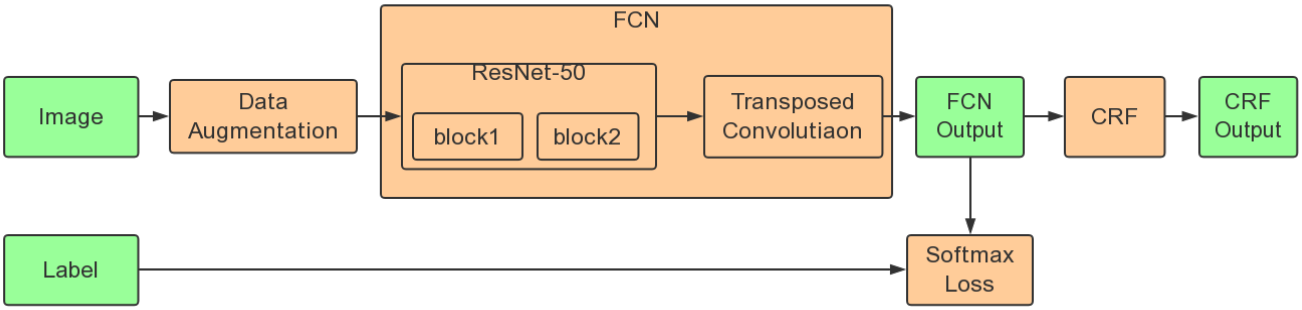

Fig. 1. ResNet-based Fully Convolutional Network Processing Pipeline

TABLE I
ARCHITECTURE OF FULLY CONVOLUTIONAL NETWORK

Layer Name	Output Size	Kernels	Stride
conv1	112×112	7×7,64	2
block1	56×56	3×3, max pool	2
	28×28	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	2
block2	14×14	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	2
transposed conv	224×224	32×32,2	16

a linear transformation to generate an upscaled version of the input.

In fact, bilinear interpolation is a special case of transposed convolution. Historically, some sources may call it fractionally strided convolution or de-convolution(which is now widely considered inappropriate), but we'll stick with transposed convolution. Given the downsampled feature output from backbone network, transposed convolution is capable of reconstructing the spatial resolution of the initial image size. And the whole model ends up perfectly in a encoder-decoder fashion instead of two separate process. The transposed convolution kernel can also be initialized randomly and trained together with backbone network so as to learn to best form the segmentation mask.

Since the feature map from backbone network has been downscaled by a factor of 16, we need a transposed convolution layer with kernel size of 32*32 and stride of 16. We need only two filters in the transposed convolution layer for this binary classification problem.

C. Post processing - Fully Connected CRF

The outcome from out deep network is pretty noisy, and it has difficulties in recovering some local structure, especially around the edge of lesion. It's actually a common problem with fully convolution network since the result is in itself an upscaled version. Besides, some small disconnected areas formed by false positives should be filtered out in the final predictions.

To get a better result and smooth out some of the false positives, we use a graphical model of fully connected Conditional Random Field [18], to refine the outcome of FCN. Here, the fully connected CRF is defined on the whole set of pixels, each of which is linked to the other pixels in the image. Typically, there're some energy terms associated with reasonable assumptions. For instance, pixels with similar color structure or closed to each other are more likely to be assigned with the same label. Following [19] and [18], the model defines the energy as

$$E(\mathbf{y}) = \sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j) \tag{1}$$

where $\mathbf{y}$ is the label assignment of all pixels.

We adopt the same potential as in [18]. The first term, so-called unary potential $\psi_u(y_i) = -\log P(y_i)$, comes from FCN output probability $P(y_i)$ for pixel i. The pairwise potential is fully defined as

$$\psi_p(y_i, y_j) = \mu(y_i, y_j) \sum_{m=1}^{K} w^{(m)} \kappa^{(m)}(f_i, f_j) \tag{2}$$

where μ is the label compatibility function, f_i are the feature input at pixel i for kernel $\kappa^{(m)}$, and $w^{(m)}$ are the kernel weights. We include two Gaussian kernels concerning bilateral position(denoted as p) and color pattern(denoted as I)

$$\kappa^{(1)}(p_i, p_j, I_i, I_j) = \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) \tag{3}$$

$$\kappa^{(2)}(p_i, p_j) = \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right) \tag{4}$$

where the appearance kernel $\kappa^{(1)}$ focuses both on pixel nearness and color similarity, while the smoothness kernel $\kappa^{(2)}$ focuses only on pixel positions. For label compatibility function μ, we simply use the Potts model, $\mu(y_i, y_j) = [y_i \neq y_j]$. The iteration dynamics of an example image is shown in Fig. 4. We believe the fully connected CRF could serve as an alternative for the hole filling method.

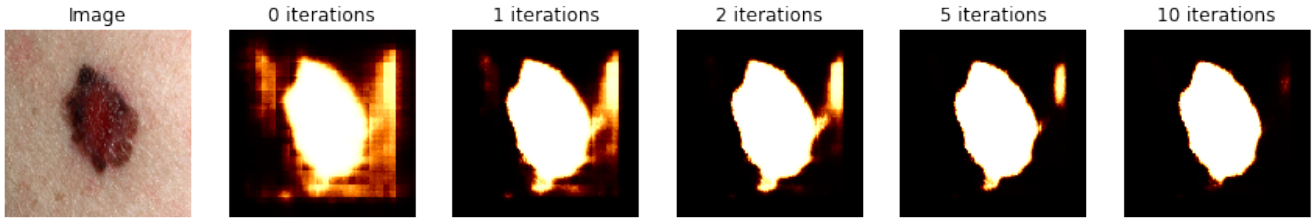

Fig. 4. Visualization of probability heat map during fully connected CRF iterations. The brighter the pixel value is, the more likely it's lesion skin.

III. Experiments

In this section, we report our model's performance on a publicly available skin lesion dataset, DermQuest [17], which comes with pixel-level label mask. The dataset includes 126 images(66 melanoma, 60 non-melanoma). We conduct k-fold cross-validation experiments on 126 images. The dataset is randomly split into 4 folds of equal size. And the result is reported on one fold while using the remaining three folds as training data. This is repeated on all 4 folds to get a complete evaluation of the dataset. Our model is implemented in Python and Tensorflow, on a device with 32 GB of RAM and NVIDIA GeForce GTX Titan XP GPU card. We use a slightly different version of ResNet-50 from slim package [20] and a public implementation [21] of fully connected CRF. The fully connected CRF runs at CPU and requires no more than 0.5 second.

In addition, since it's a skewed classification task considering the number of pixels from normal skin and lesion region, we adopt a weighted sum while calculating the final loss and train the whole network with Adam optimizer [22] from scratch. We use a learning rate of 0.001 at the beginning and decrease it by a factor of 0.5 to optimize the network parameter at a batch size of 4.

Following [15], we evaluate our model at three commonly used metrics: sensitivity, specificity and pixel accuracy. Mathematically, these metrics are defined as:

$$sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$specificity = \frac{TN}{TN + FP} \tag{6}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP, TN, FP and FN each stand for number of true positive, true negative, false positive and false negative predictions.

A. Necessity of Data Augmentation

To prove that the data augmentation dose make a difference, we train two separate FCNs, one using augmentation techniques and another not. In this experiment, we only report results from FCN output. From Table. II, we can see that network trained without augmentation embraces nearly perfect segmentation result in training set, but it behave poorly at test images in terms of segmentation sensitivity. We argue that the network is actually overfitted to the training examples because of the limited size of available data. On the other hand, the augmentation techniques ease off the overfitting effect and the network performs much better, with an increase of over 10 percentage points on segmentation sensitivity.

TABLE II
Results from models trained with or without Data Augmentation

Datasets	Augmentation	Sensitivity (%)	Specificity (%)	Accuracy (%)
Training set	×	96.6	100.0	99.7
Test set	×	**83.0**	99.6	98.5
Training set	√	97.5	97.9	97.9
Test set	√	**95.6**	97.4	97.3

TABLE III
Segmentation Results for All Lesion Images

Segmentation Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)
L-SRM [23]	89.4	92.7	92.3
Otsu-R [7]	87.3	85.4	84.9
Otsu-RGB [24]	93.6	80.3	80.2
Otsu-PCA [25]	79.6	**99.6**	98.1
TDLS [9]	91.2	99.0	98.3
Patch-wise Model [15]	95.0	98.9	**98.5**
Our method(w/o CRF)	**95.6**	97.4	97.3
Our method(w CRF)	**95.1**	98.0	97.8

B. Quantitative Evaluation

We include results from six other segmentation methods during comparison. As can be seen in Table. III, our model achieves state-of-art segmentation sensitivity of 95.6% without any post-processing and comparable results on two other metrics. In spite of slight degradation on sensitivity after CRF post-processing, we believe it's worthwhile to trade for higher accuracy results on the other two metrics in return.

C. Model Runtime Evaluation

To estimate model inference time accurately, we run through all models in the same device with same configuration. We first warm up the GPU(or CPU) by running each model 10 times and then run for another 10 loops to estimate the mean

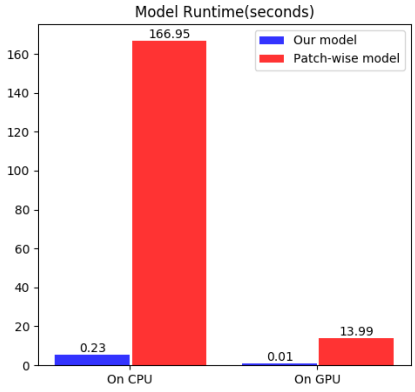

Fig. 5. Model runtime compared with patch-wise model [15]. The input image size is 400*600 and the patch-wise model [15] needs to loop around 1800 times with a batch size of 128. Here, only the inference time of neural network is considered.

runtime. To get prediction mask from a 400*600 input image, we only need to forward the image through our network one single time. But the patch-wise model [15] has to loop around 1800 times even with a batch size of 128. As shown in Fig. 5, our model runs much faster and operates at 5 fps even at a CPU configuration.

D. Qualitative Result

Apart from quantitative result and runtime efficiency, our method also produces qualitative segmentation border compared to [15]. The result is shown in Fig. 6. The ground truth, result from model [15] and result from our method are displayed in the first, second and third column.

IV. Conclusion

In this paper, we propose a new processing pipeline to automate segmentation of skin lesion images taken by normal digital camera. Before feeding the input to our fully convolutional network, we introduce several augmentation techniques to enrich our limited training data. And the segmentation job can be done in a single forward of the FCN. To better maintain the local structure and filter out small area of disconnected false positives, a fully connected conditional random field is used to refine the predictions from FCN. Our method achieves state-of-art segmentation sensitivity of 95.6% and comparable results on the other two commonly used metrics. Moreover, our processing pipeline is pretty time-efficient and outperforms existing methods by orders of magnitude.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China(Grant no. 61772568), Guangzhou Science and Technology Program(Grant no. 201804010288), and Shenzhen Scientific Research and Development Funding Program(Grant no. JCYJ20170302153827712).

References

[1] [Online]. Available: https://en.wikipedia.org/wiki/Melanoma

[2] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, "The abcd rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions," *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551 – 559, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0190962294700613

[3] Mascaro, J. JM, and M. JM, "The dermatologist's position concerning nevi: A vision ranging from "the ugly duckling" to "little red riding hood"," *Archives of Dermatology*, vol. 134, no. 11, pp. 1484–1485, 1998. [Online]. Available: + http://dx.doi.org/

[4] H. C. Engasser and E. M. Warshaw, "Dermatoscopy use by us dermatologists: A cross-sectional survey," *Journal of the American Academy of Dermatology*, vol. 63, no. 3, pp. 412 – 419.e2, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0190962209013516

[5] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *CoRR*, vol. abs/1011.0640, 2010. [Online]. Available: http://arxiv.org/abs/1011.0640

[6] J. Glaister, R. Amelard, A. Wong, and D. A. Clausi, "Msim: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1873–1883, July 2013.

[7] P. G. Cavalcanti, Y. Yari, and J. Scharcanski, "Pigmented skin lesion segmentation on macroscopic images," in *2010 25th International Conference of Image and Vision Computing New Zealand*, Nov 2010, pp. 1–7.

[8] P. G. Cavalcanti and J. Scharcanski, "Automated prescreening of pigmented skin lesions using standard cameras." *Comp. Med. Imag. and Graph.*, vol. 35, no. 6, pp. 481–491, 2011.

[9] J. Glaister, A. Wong, and D. A. Clausi, "Segmentation of skin lesions from digital images using joint statistical texture distinctiveness," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1220–1230, April 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[13] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, Nov 2016.

[14] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. C. Courville, Y. Bengio, C. Pal, P. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *CoRR*, vol. abs/1505.03540, 2015. [Online]. Available: http://arxiv.org/abs/1505.03540

[15] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, S. M. R. Soroushmehr, K. Ward, and K. Najarian, "Skin lesion segmentation in clinical images using deep learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 337–342.

[16] Y. Yuan, M. Chao, and Y. C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1876–1886, Sept 2017.

[17] [Online]. Available: http://www.dermquest.com

[18] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014. [Online]. Available: http://arxiv.org/abs/1412.7062

[19] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *CoRR*, vol. abs/1210.5644, 2012. [Online]. Available: http://arxiv.org/abs/1210.5644

[20] [Online]. Available: https://www.tensorflow.org/

[21] [Online]. Available: https://github.com/lucasb-eyer/pydensecrf

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

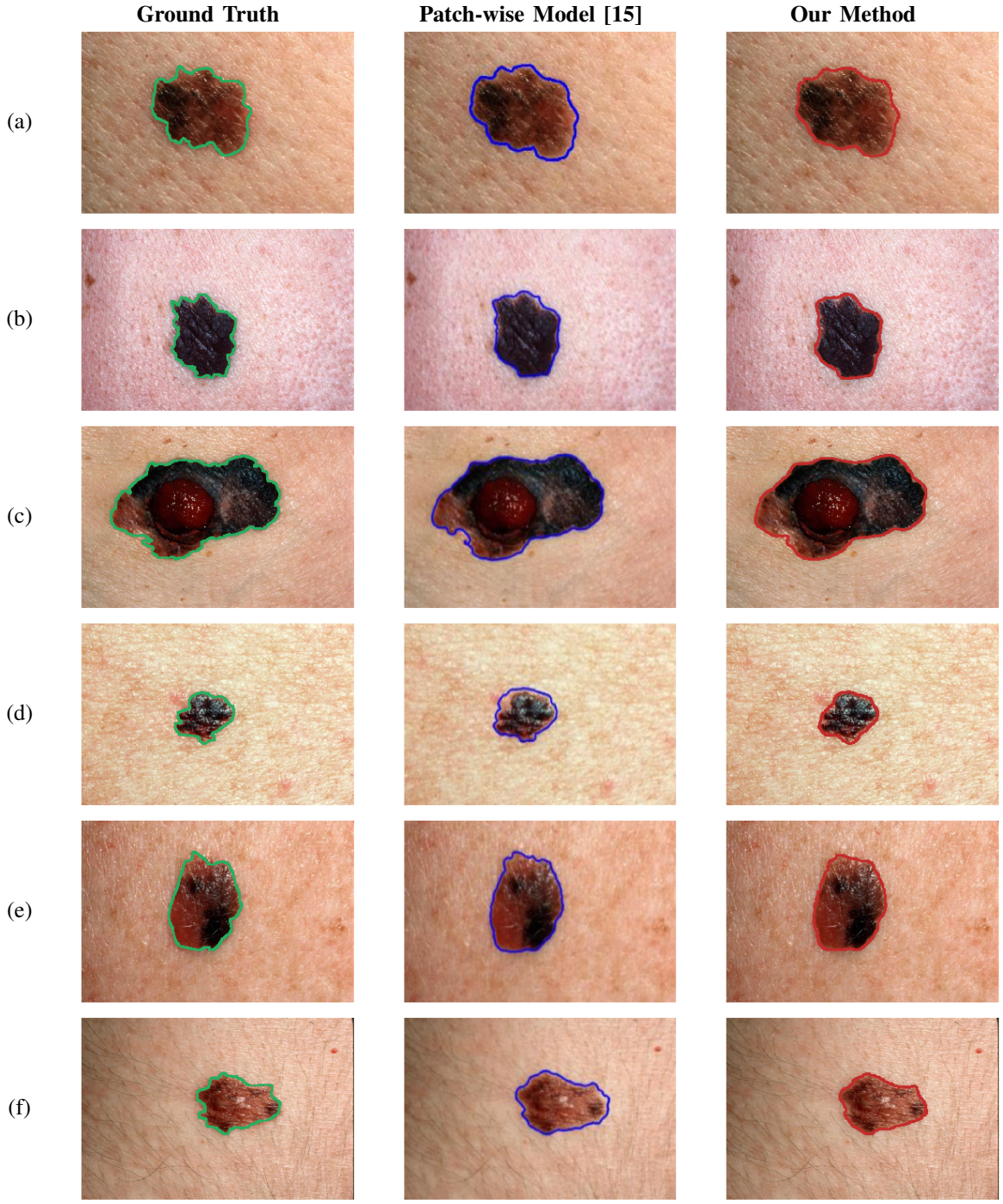

Fig. 6. Demonstration of sample images. Left column: border of ground truth(green). Middle column: result from patch-wise model [15](blue). Right column: result from our method(red).

[23] M. Emre Celebi, H. A. Kingravi, H. Iyatomi, Y. Alp Aslandogan, W. V. Stoecker, R. H. Moss, J. M. Malters, J. M. Grichnik, A. A. Marghoob, H. S. Rabinovitz, and S. W. Menzies, "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology*, vol. 14, no. 3, pp. 347–353, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1600-0846.2008.00301.x

[24] P. G. Cavalcanti, J. Scharcanski, and C. B. O. Lopes, "Shading attenuation in human skin color images," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammoud, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, and L. Avila, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 190–198.

[25] P. G. Cavalcanti and J. Scharcanski, "Automated prescreening of pigmented skin lesions using standard cameras," *Computerized Medical Imaging and Graphics*, vol. 35, no. 6, pp. 481 – 491, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611111000395