RethNet: Object-by-Object Learning for Detecting Facial Skin Problems

Shohrukh Bekmirzaev
lululab Inc
Gangnam-gu, Seoul, Korea
shoh.bek@lulu-lab.com

Seoyoung Oh
lululab Inc
Gangnam-gu, Seoul, Korea
seoyoung.oh@lulu-lab.com

Sangwook Yoo
lululab Inc
Gangnam-gu, Seoul, Korea
sangwook.yoo@lulu-lab.com

Abstract

Semantic segmentation is a hot topic in computer vision where the most challenging tasks of object detection and recognition have been handling by the success of semantic segmentation approaches. We propose a concept of object-by-object learning technique to detect 11 types of facial skin lesions using semantic segmentation methods. Detecting individual skin lesion in a dense group is a challenging task, because of ambiguities in the appearance of the visual data. We observe that there exist co-occurrent visual relations between object classes (e.g., wrinkle and age spot, or papule and whitehead, etc.). In fact, rich contextual information significantly helps to handle the issue. Therefore, we propose REthinker blocks that are composed of the locally constructed convLSTM/Conv3D layers and SE module as a one-shot attention mechanism whose responsibility is to increase network's sensitivity in the local and global contextual representation that supports to capture ambiguously appeared objects and co-occurrence interactions between object classes. Experiments show that our proposed model reached MIoU of 79.46% on the test of a prepared dataset, representing a 15.34% improvement over Deeplab v3+ (MIoU of 64.12%).

1. Introduction

Semantic segmentation has been one of the fundamental and active topic in computer vision for a long time. This topic is of wide interest for real-world applications of autonomous driving, robotics and a range of medical imaging applications. Recent improvements and advances in semantic segmentation enable to emerge various new application areas in skin analysis. For example, facial skin lesion analysis has been attracting a lot of attention as having beautiful skin without troubles is getting popular and influenced on the society nowadays. E -cosmetics which involve the beautification, facial image simulation, digital makeup and accurate facial lesion analysis are fast-growing sector in the marketing.

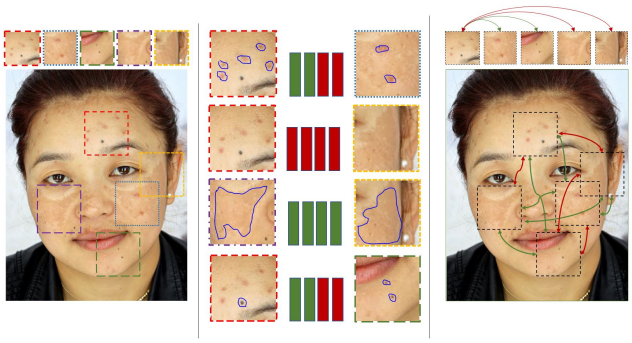

Figure 1. **The concept of object-by-object learning**: In fact, skin lesion objects have visual relations between each other where it helps easily human to judge about what type of skin lesions they are. In the figure, the green and red boxes represent the level of the image patch's similarity. The green and red line describe that there exist positive and negative relations between objects or a group of objects of each patch.

Accurately and early detecting facial skin problems is an important clinical task and automated dermatology can be used to save time and reduce costs [25]. E-cosmetics and dermatological computer-aided systems are developing rapidly behind computer vision progresses [49], [26], [25] , [45], [12], [17], [29]. However, previous methods have shown only limited improvements and visual understanding of individual skin lesion in a dense group is still a challenging task. This is because, it is hard to distinguish some types of facial skin lesions between each other as the fine-grained object categorization problem. Furthermore, facial skin lesions appear ambiguously with different (typically small) sizes, which lead a network to assign wrong classes easily. The use of rich contextual relation information helps to reduce the issue significantly [42], [27]. For example, there are the object-object interactions [27] between some skin lesions where the detection of an object class helps to detect another by their co-occurrence interactions. The detection decisions about individual skin lesions can be switched dynamically through contextual relations among objects. We denote this

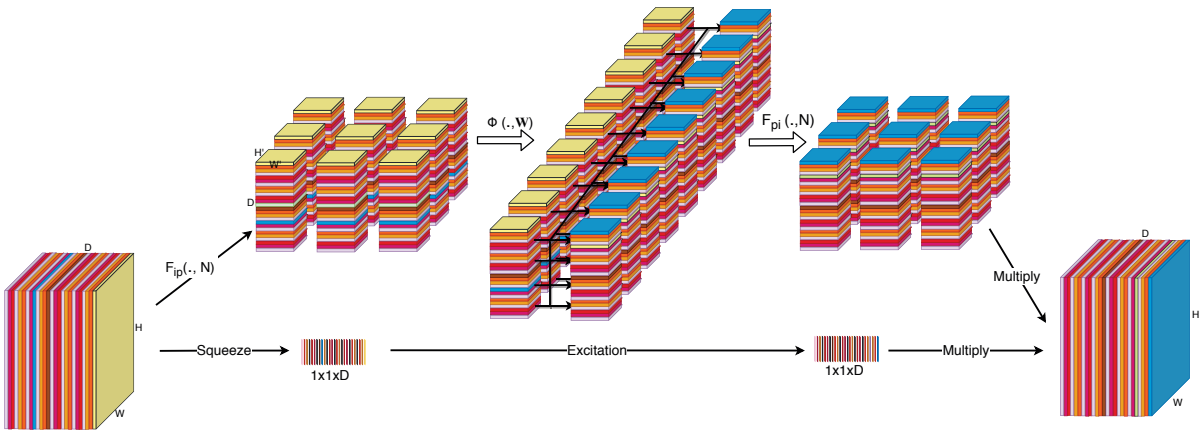

Figure 2. Proposed REthinker block that consists of the locally constructed convLSTM/Conv3D layers and SE module [23] as a one-shot attention mechanism whose responsibility is to improve network's sensitivity in local and global contextual representation that helps to capture ambiguously appeared objects and co-occurrence interactions between object classes.

cognitive process as *object-by-object* decision-making.

We present a *REthinker* module based on the SENet module [23] [22] and locally constructed convLSTM/conv3D unit [41] to increase network's sensitivity in local and global contextual representations. The proposed modules are easy to use and applicable in any standard convolutional neural networks (CNNs). The use of the *REthinker* modules forces networks to capture the contextual relationships between object classes regardless of similar texture and ambiguous appearance they have.

We experiment our proposed modules by modifying current state-of-the-art networks [19], [43], [8] for feature extraction. We originally use the decoder of DeepLabv3+ [7]. Experimental results show that our proposed models outperform the state-of-the-art segmentation networks [34],[50], [7] ,[4], [51], [39] by the high difference of 15.34% MIoU [5], [21], [52] in detecting of facial skin problems in a prepared dataset. Moreover, our models have shown promising results on ISIC 2018 segmentation tasks. The overall contributions of our paper can be summarized as follows:

- We introduce a new concept named "object-by-object" learning, where an object can be identified by looking at other objects.
- We propose a novel residual block called REthinker modules that support "object-by-object" technique by capturing contextual relationships between object classes.
- We develop a novel RethNet architecture that detects skin lesions with higher accuracy than the recent state-of-art segmentation approaches.

2. Related Work

Semantic Scene Understanding: Semantic understanding of visual scenes has become ubiquitous in computer vision. Impressive semantic segmentation approaches are mostly based on the Fully Convolutional Network (FCNs) [33], [11], [30], [32], [53] ,[3] , [37]. One key reason for the success of FCNs is that they use multi-scale (MS) image representations, which are subsequently upsampled to recover the lost resolution. Moreover, atrous convolution has proven to be an effective technique by providing a larger receptive field size without increasing the number of kernel parameters [3]. Spatial pyramid pooling module [18] has been successfully applied with atrous convolutions by state-of-the-art networks [5], [50],[21], [52], [34], [4], [51],[7] , [39] on segmentation and object detection benchmarks [14], [35], [13], [31], [16], [10]. Recently, depthwise separable convolution [8] has known as an efficient technique to reduce the computation complexity in convolutional operations and allow networks to go deeper [5], [39], [21], [52], [34]. Originally, the depthwise separable convolution consists of the *depthwise* and *pointwise* convolutions where depthwise convolution keeps the channels separate and uses the standard convolution operation in each input channel [8].

Contextual relations: H.S. Hock *et al.* [20] introduced early the contextual relations between object scenes by experiments based on the criteria of familiarity, physical plausibility, and belongingness. The scene context information plays a crucial role on the semantic scene understanding. However, the contextual relation information between object classes are often ignored [15], [42], [27]. The relative spatial configurations of particular objects yield the higher-level contextual information while the lower-level contextual information demonstrates the semantic and visual

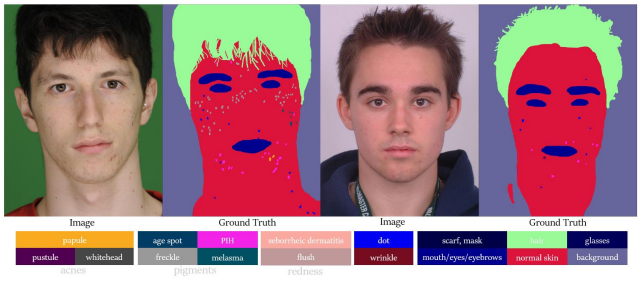

Figure 3. Samples of the SiblingsDB [46] dataset. Few samples with skin problems have been annotated in order to show only the labelling process of facial skin lesions. Note that we do not disclose our MSLD dataset samples.

relationships among objects or group of objects. S. Kumar *et al.* [27] categorize types of the contextual relationships of the scene labeling to *region-region, object-region* and *object-object* interactions and provide a hierarchical framework using Conditional Random Fields (CRFs) for semantic segmentation. CRFs models are widely used to capture the local contextual interactions of image regions [44] , [42], [36], [5], [40]. RNNs are investigated broadly to aggregate global context in the semantic segmentation [48], [2], [38], [15] [28], [54]. SE modules [23] [22] are successfully applied to capture global contextual information by simply exploiting CNN layers.

3. Datasets

We prepare a dataset called "Multi-type Skin Lesion Labelled Database" (MSLD) with pixel-wise labelling of frontal face images. We report that the designing of MSLD is unique in ML community where it is not available such dataset with the labelling of multi-type skin lesions of facial images. We further test the proposed models in the International Skin Imaging Collaboration (ISIC) dataset which holds dermoscopic images with 5 types of skin problems. In this section, we introduce the process of image accumulation and annotation of MSLD dataset and announce about ISIC dataset.

3.1. Image accumulation

We collected a total of 27,790 frontal face images using kiosks in cosmetics stores during the period from April to August in 2018. The kiosks have a standard camera whose image sensor is a 1/3.2 Inch CMOS IMX179. The user's consent is obtained before capturing images. The total number of pixels of the image sensor is 3,288 x 2,512 (8.26Mpixel) with 24-bit depth. The images are captured at autofocus and auto-white balance in the distance of 20 mm.

3.2. Image Annotation

The collected images are studied carefully. Then 412 images have been annotated with the labelling of 11 common types of facial skin lesions and 6 additional classes as in Figure 3 using the PixelAnnotation tool [1]. The skin lesions are *whitehead, papule, pustule, freckle, age spots, PIH* [1], *flush, seborrheic, dermatitis, wrinkle* and *black dot*. The additional classes are normal skin, hair, eyes/mouth/eyebrow, glasses, mask/scarf and background. We report that we do not disclose our (MSLD) dataset as the user's privacy is taken under the responsibility. We use the collected and annotated images as research purposes where they are used only in the training.

3.3. ISIC Dataset

We test the proposed approaches in the TASK 2 [2] of the ISIC, 2018 challenge called "Lesion Attribute Detection" [9] in order to provide further experiments. The goal of the task is to predict 5 skin lesion attributes from dermoscopic images. These lesion attributes are *pigment networks, negative network, streaks, milia-like cysts, globules, and dots.* The lesion classes that have visual similarity can be seen as a fine-grained classification problem. There are 2594 images for training, 100 images for validation, and 1000 images for the test.

4. Methods

4.1. Squeeze and Excitation module

Squeeze and Excitation network (SENet) [23] has been introduced as a winner of the ILSVRC 2017 classification task in the top-5 error of 2.251%. SE blocks Figure 4 [a-b] have proven to be an effective channel-wise attention mechanism [6] , which enables the network to perform dynamic channel-wise feature recalibration. SE module is computationally cheaper and helps the network to learn contextual higher-level features by the aggregated transformations of the global pooling. The SE block consists of squeeze and excitation operations where the *squeeze* operation uses the global pooling to transform global spatial information to channel-wise statistics as a channel descriptor. The *excitation* operation performs a self-gating mechanism based on 2 fully connected (FC) layers to capture channel-wise dependencies from the channel descriptor. Finally, the channel-wise dependencies are used to exploit the previous input transformation by the multiplication operation. The role of the SE module in our proposed module is to pass high-level contextual information outside of this region for context-dependent decision making.

[1] Post inflammatory hyperpigmentation (PIH) caused usually by inflammation or acnes

[2] https://challenge2018.isic-archive.com/task2/

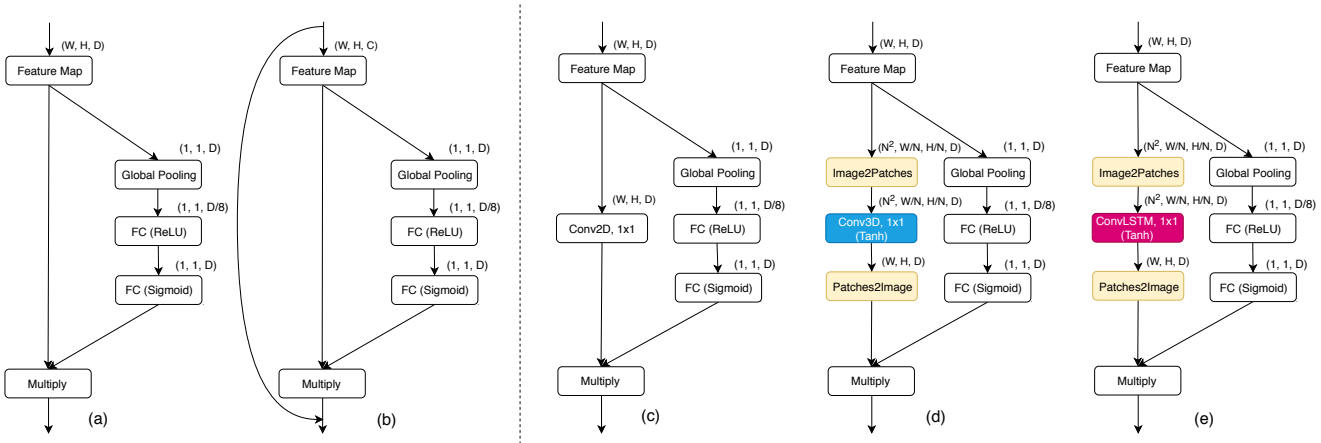

Figure 4. The figure represents comparison of SE blocks (a-b), baseline block (c) and REthinker blocks (d-e): The proposed REthinker blocks are designed under locally constructed Conv3D (d) and convLSTM (e) layers.

4.2. REthinker module

REthinker modules consist of SE module and the locally constructed convLSTM/conv3D layers as an one-shot attention mechanism as in Figure 4 [d-e], which are both responsible for extracting contextual relations from features. Precisely, as the global pooling of SE module aggregates the global spatial information, the SE module passes more embedded higher-level contextual information across large neighborhoods of each feature map. Whereas, the locally constructed convLSTM/conv3D layers encode lower-level contextual information across local neighborhoods elements of fragmented feature map (patches) while further take spatial correlation into consideration distributively over patches. The output of locally constructed convLSTM/conv3D recieves 3D , $U_d \in R^{H\times W\times D}$ feature maps from the residual blocks and passes a transformed 3D , $U'_d \in R^{H\times W\times D}$ feature map. The locally constructed convLSTM/conv3D is identified as follows:

$$U'_d = F_{pi}([\Phi(F_{ip}(U_d, N)|v_t)|h_t], N) \tag{1}$$

Where, F_{ip} is the *image2patches* operator function, $F_{ip} : R^{H\times W\times C} \rightarrow R^{N^2\times H'\times W'\times D}$ to provide local spatiotemporal 4D data, $v_t = F_{ip}(U_d, N)$. Practically, the feature maps are transformed to patches over channel as spatiotemporal data. Here, $H'\times W'$ is a patch size $(H' = H/N, W' = W/N)$ to be assumed as an object or a group of objects. The given N is the dimensional slicing coefficient over the spatial dimensions (W, H) of the feature map. Thus, Φ is conv3D or convLSTM operator function, $\Phi : R^{N^2\times H'\times W'\times D} \rightarrow R^{N^2\times H'\times W'\times D}$ to be applied with keeping the depth of the feature map. The convLSTM/conv3D serves to encode spatiotemporal correlations between features by viewing sequentially objects or a group of objects of patches and passes the output as spatiotemporal data $h_t = \Phi(v_t)$ to the *patches2image* operator whose identification function here is $F_{pi} : R^{N^2\times H'\times W'\times D} \rightarrow R^{H\times W\times D}$ and $U'_d = F_{pi}(h_t, N)$. Note that $h_t = \Phi(v_t, h_{t-1})$ performs as the hidden states in the convLSTM. The gates i_t, h_t, o_t of the convLSTM are identified as follows:

$$\begin{aligned}
i_t &= \sigma(w_{vi} \odot v_t + w_{hi} \odot h_{t-1} + w_{ci} \circ c_{t-1} + b_i) \\
f_t &= \sigma(w_{vf} \odot v_t + w_{hf} \odot h_{t-1} + w_{cf} \circ c_{t-1} + b_f) \\
c_t &= f_t \circ c_t - 1 + i_t \circ \tanh(w_{vc} * v_t + w_c \odot h_{t-1} + b_c) \\
o_t &= \sigma(w_{vo} \odot v_t + w_{ho} \odot h_{t-1} + w_{co} \circ c_t + b_o) \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned} \tag{2}$$

Where, '$\odot$ ' represents the convolution operator and '$\circ$' denotes the Hadamard product. The output of SE module is used to exploit the output of locally constructed convLSTM/conv3D by the channel-wise multiplication operation as the output of REthinker module whose feature map represents long-range local and global contextual information to enable the context-dependent decision making.

4.3. RethNeT

An Encoder Search: In practice, the REthinker blocks are applicable in any standard CNNs. We consider to employ current state of the art networks [19], [43], [8]. We integrate the modern architectures with our proposed REthinker blocks to improve the network's sensitivity in local and global contextual representations enabling object-by-object learning technique. We experiment ResNet [19], ResNeXt [47] and Xception [8], [7].

A Decoder Search: We believe that the rich contextual information is a key to capture ambiguously appeared ob-

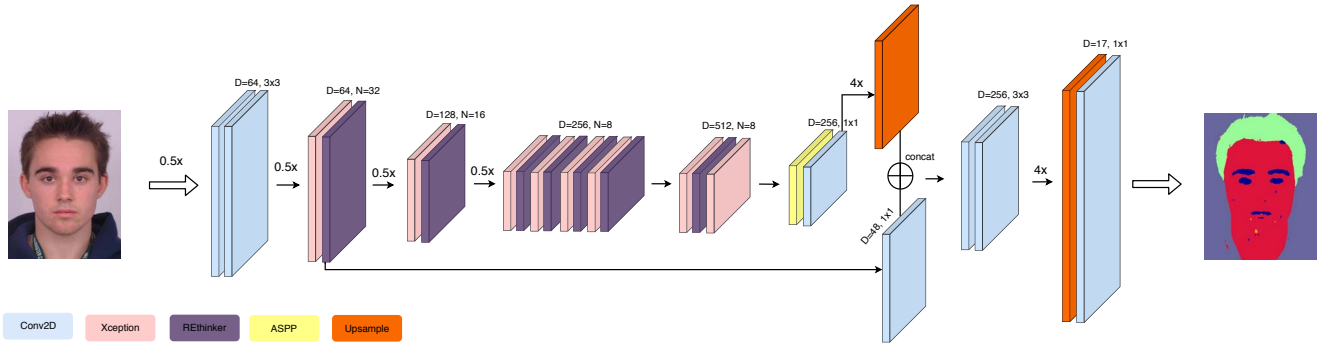

Figure 5. Our proposed RethNet based RE-Xception and the decoder of the DeepLabv3+ [7]. Where *D* and *N* is the depth of future maps and the dimensional slicing coefficient over the spatial dimensions respectively.

.

Models	Mean IoU(%)	Pixel Acc(%)
DenseASPP [50]	58.31	89.31
PSPNet+ResNet-101 [18]	63.24	91.92
DeepLabv3Plus + ResNet-101 [7]	63.74	92.51
DeepLabv3Plus +ResNeXt-101 [33]	64.64	93.14
DeepLabv3Plus + Xception [7]	64.12	94.08
DeepLabv3Plus + Xception+SE	65.49	94.12
DeepLabv3Plus + Xception+baseline-c	65.52	94.21
RethNet + baseline-c	62.11	92.44
RethNet + Rethinker-d	**76.56**	**96.45**
RethNet + Rethinker-e	**79.46**	**96.11**

Table 1. Experimental results on test samples of our MSLD dataset.

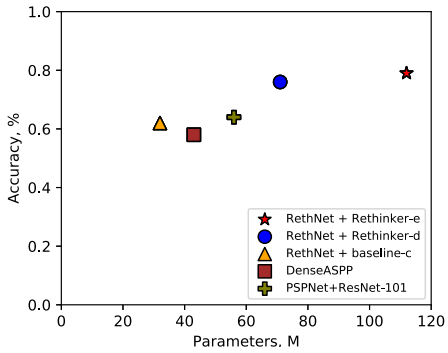

Figure 6. The number of parameters in the reference models and proposed modes with accuracy comparison on the test of MSLD dataset.

jects and co-occurrence interactions between object classes where that is usually obtained in encoders. Therefore, we do not consider the decoder path. We select the decoder of DeepLabv3+ [7] to recover object segmentation details of individual skin lesions.

RethNet: We simply investigate RethNet with the combining of the Xception module and REthinker modules as in Figure 5 . We modify Xception as follows: **(1)** We add REthinker module after each Xception blocks without spatial loss of feature maps. **(2)** We remove the final block of the *entry flow* of Xception. **(3)** We keep the patch size as 4x4 in each REthinker module in order to "see" future maps wider in ConvLSTM/conv3D with simply increasing time steps. **(4)** The number of parameters is minimized in the *middle flow* and *exit flow*. As suggested in [7], **(5)** the max-pooling operation is replaced by depthwise separable convolutions with striding and the batch normalization and ReLU is applied after each 3 x 3 depthwise convolution of the Xception module.

5. Experiments

5.1. Implementation

For experimental comparisons, we use the standard networks that are DeepLabv3+, PSPNet, DenseASPP as the reference networks. All models of networks were implemented using the TensorFlow framework and trained on a single NVIDIA GeForce GTX 1080 Ti GPUs, Intel(R) Core(TM) i7-8700K CPU @ 3.20GHz. In the experiments, 374 images are used for training and 38 images are for testing. We use the standard data augmentation techniques whose participants are the random rotation, random zooming, and random horizontal flipping during the training. The input and ground-truth images of the dataset is resized to 730 x 960 and random cropping applied by 512 x 512 as the input of the network during the training. We follow the same training protocols as suggested in [7], [18]. In all experiments, the softmax cross-entropy is applied for loss and the momentum optimizer is used, whose base learning rate is set to 0.001

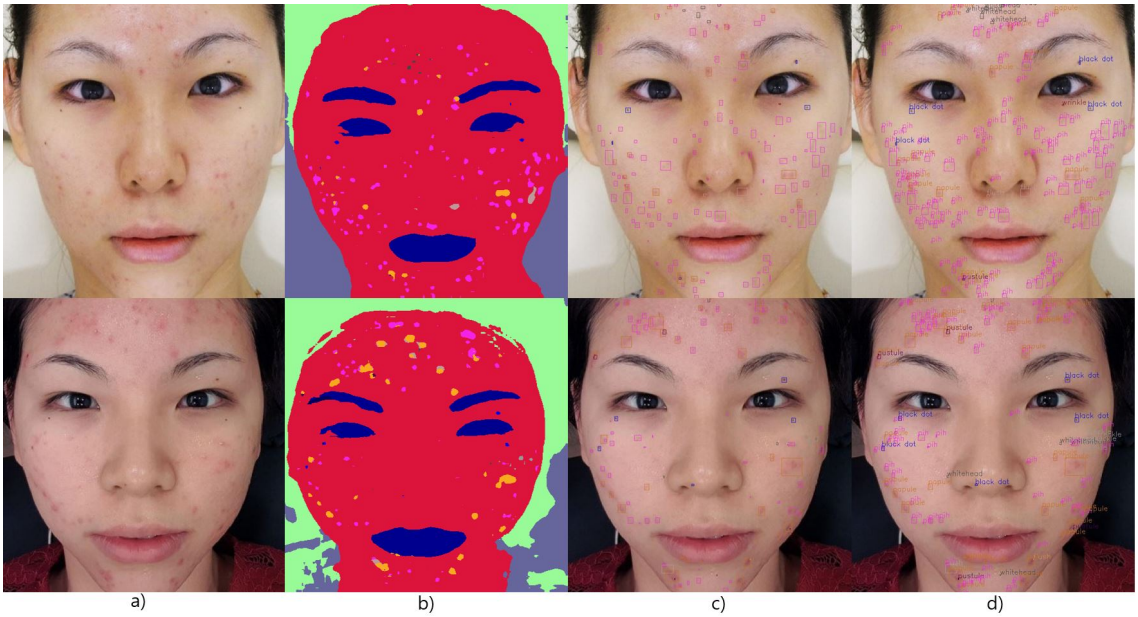

Figure 7. The visualization results of **RehtNet + Rethinker-e:** The real test images (a) are obtained from the fionaseah.com with an agreement of the author. The candidate of the test images is suffered from basically, PIH (caused by acne) and papules where the results images (b-d) show that the most of facial skin lesions are correctly predicted.

Models	Rank	Jaccard (%)	Dice (%)
Ensemble [24]	1	**0.483**	**0.651**
Unet + ASPP +DenseNet169 [24]	2	0.464	0.629
Unet + ASPP + Resnetv2 [24]	3	0.455	0.616
Unet + ASPP +ResNet151 [24]	5	0.436	0.598
DeepLabv3Plus + Xception	-	0.451	0.614
DeepLabv3Plus + Xception+SE	-	0.469	0.627
DeepLabv3Plus + Xception+baseline-c	-	0.456	0.616
RethNet + baseline-c	-	0.441	0.592
RethNet + Rethinker-d	-	0.473	0.639
RethNet + Rethinker-e	-	**0.475**	**0.644**

Table 2. Experimental results on test samples of ISIC 2018 challenge in the task 2. The evaluation metrics in this task are average Jaccard Index (mIoU) and Dice coefficient following by the proposed metrics of the challenge. The rank represents positions on the test leaderboard of the challenge [24]. Note that the Jaccard index metric, also referred to as the Intersection over Union (IoU).

decreased by a factor of 10 every 50 epoch of total 200 epoch with decay 0.9.

5.2. Results

Thanks to REthinker blocks, It shows significant improvements in the facial skin lesion detection task of MSLD dataset. As Table 1 summarized results, we made a further improvement by DeepLabv3Plus + Xception+ SE and DeepLabv3Plus + Xception+ baseline-c, which showed a better result of 65.49 and 65.52 MIoU than all other reference models except the proposed models. Our proposed network RethNet+ REthinker-e blocks achieved MIoU of 79.46% on the test, where it is the top in all experiments.

Computational cost: We compare proposed models with the reference models in terms of the number of parameters and accuracy (Figure 6). Even though our RethNet + REthinker-e block is the top on the list with 112M parameters, it reached high performance in the accuracy with a big difference (e.g 15 % of mIoU greater than the best reference model, 14% of mIoU higher than the baseline model). We further report inference time of the RethNet + REthinker-e block that whose running time for per image inference by the 512x512 of resolution is an average 2.7 sec on a single GPU and 11 sec on CPU of those mentioned hardware sources.

5.3. ISIC challenge

As Table 2 represents, our RethNet + REthinker-e block showed 47.5% of Mean-Jaccard on the test of ISIC dataset, where it outperforms all competitive single models except an ensemble model (”48.3% of Mean-Jaccard”). Note that the top-ranked models on the test leaderboards of the challenge use broadly preprocessing techniques such as image enhancing, polluting dermoscopic images with random hairs, and data augmentation. However, we apply only common techniques of data augmentation during training and test images are evaluated without any preprocessing techniques.

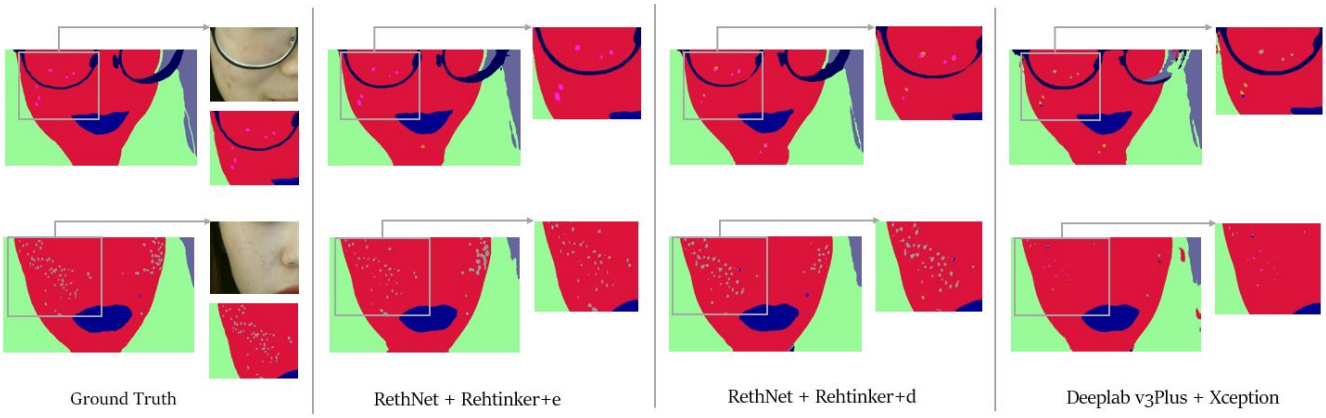

Figure 8. The ground truth and inference results of proposed models and the best reference model in MSLD dataset . We keep the face entity of the MSLD dataset.

5.4. Contribution Discussion

In fact, the fine-grained classification has been becoming an open issue in the computer vision community so far. It is a real challenge to differentiate classes that have similar visual context. Especially the problem is more common in medical imaging applications. We consider solving the problem with a novel straightforward technique by our application ("Detecting multi-type facial skin lesions") where there is not yet accurate and solid application or method to detect correctly and differentiate multi-type facial skin lesions. The proposed blocks are easy to use and possible to apply to any standard CNNs with considering time complexity by limiting the number of the blocks.

6. Conclusion

We propose successfully an efficient network architecture to address detecting multi-type facial skin lesions by a novel *object-by-object* learning technique. Experimental results show that our proposed model outperformed state-of-the-art segmentation networks by a high gap in the MSLD dataset. Furthermore, our model takes promising results on the ISIC 2018 segmentation task. In the future, we consider the time complexity of Rethinker blocks and try to design more light-weight models.

References

[1] A. Br. Pixel annotation tool. 2017. 3

[2] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3555, June 2015. 3

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. 2

[4] L. 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. *CoRR*, abs/1606.00915, 2016. 2

[5] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 2, 3

[6] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T. Chua. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. *CoRR*, abs/1611.05594, 2016. 3

[7] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. 2, 4, 5

[8] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. 2, 4

[9] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. 3

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 2

[11] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *CoRR*, abs/1503.01640, 2015. 2

[12] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115 EP –, Jan 2017. 1

[13] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object

classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 2

[14] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010. 2

[15] H. Fan, X. Mei, D. Prokhorov, and H. Ling. Multi-level contextual rnns with attention model for scene labeling. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3475–3485, Nov 2018. 2, 3

[16] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, Sept. 2013. 2

[17] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529 – 1538, 2018. 1

[18] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014. 2, 5

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2, 4

[20] H. S. Hock, G. P. Gordon, and R. Whitehurst. Contextual relations: The influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, 16(1):4–8, Jan 1974. 2

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 2

[22] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. *CoRR*, abs/1810.12348, 2018. 2, 3

[23] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. 2, 3

[24] N. A. Koohbanani, M. Jahanifar, N. Z. Tajeddin, A. Gooya, and N. M. Rajpoot. Leveraging transfer learning for segmenting lesions and their attributes in dermoscopy images. *CoRR*, abs/1809.10243, 2018. 6

[25] K. Korotkov. automatic change detection in multiple pigmented skin lesions, 2014. 1

[26] K. Korotkov and R. Garcia. Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, 56(2):69 – 90, 2012. 1

[27] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1284–1291 Vol. 2, Oct 2005. 1, 2, 3

[28] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. *CoRR*, abs/1511.04510, 2015. 3

[29] H. Liao. A deep learning approach to universal skin disease classification. 2015. 1

[30] G. Lin, C. Shen, I. D. Reid, and A. van den Hengel. Efficient piecewise training of deep structured models for semantic segmentation. *CoRR*, abs/1504.01013, 2015. 2

[31] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[32] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. *CoRR*, abs/1509.02634, 2015. 2

[33] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. 2, 5

[34] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[35] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[36] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 891–898, Washington, DC, USA, 2014. IEEE Computer Society. 3

[37] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1520–1528, Washington, DC, USA, 2015. IEEE Computer Society. 2

[38] A. Salvador, M. Bellver, M. Baradad, F. Marqués, J. Torres, and X. Giró i Nieto. Recurrent neural networks for semantic instance segmentation. *CoRR*, abs/1712.00617, 2017. 3

[39] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 2

[40] F. Shen, R. Gan, S. Yan, and G. Zeng. Semantic segmentation via structured patch prediction, context crf and guidance crf. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5178–5186, July 2017. 3

[41] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015. 2

[42] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, Jan 2009. 1, 2, 3

[43] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 2, 4

[44] A. Torralba, K. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. *NIPS*, 10 2004. 3

[45] N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima, S. Akazaki, K. Hori, and Y. Miyake. Image-based skin color and texture analysis/synthesis by extracting

hemoglobin and melanin information in the skin. *ACM Trans. Graph.*, 22(3):770–779, July 2003. 1

[46] T. F. Vieira, A. Bottino, A. Laurentini, and M. De Simone. Detecting siblings in image pairs. *The Visual Computer*, 30(12):1333–1345, Dec 2014. 3

[47] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. 4

[48] Z. Yan, H. Zhang, Y. Jia, T. Breuel, and Y. Yu. Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation. *CoRR*, abs/1603.04871, 2016. 3

[49] L. Yang, S.-H. Lee, S.-G. Kwon, H.-J. Song, and K.-R. Kwon. Skin pigment recognition using projective hemoglobin-melanin coordinate measurements. *Journal of Electrical Engineering and Technology*, 11:1825–1838, 11 2016. 1

[50] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5

[51] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. *CoRR*, abs/1803.08904, 2018. 2

[52] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *CoRR*, abs/1707.01083, 2017. 2

[53] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1529–1537, Washington, DC, USA, 2015. IEEE Computer Society. 2

[54] Y. Zhuang, F. Yang, L. Tao, C. Ma, Z. Zhang, Y. Li, H. Jia, X. Xie, and W. Gao. Dense relation network: Learning consistent and context-aware representation for semantic image segmentation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3698–3702, Oct 2018. 3