FinSeg: Finger Parts Semantic Segmentation using Multi-scale Feature Maps Aggregation of FCN

Adel Saleh[1], Hatem A. Rashwan[1], Mohamed Abdel-Nasser[1,3], Vivek K. Singh[1], Saddam Abdulwahab[1], Md. Mostafa Kamal Sarker[1], Miguel Angel Garcia[2] and Domenec Puig[1]

[1]*Department of Computer Engineering and Mathematics, Rovira i Virgili University, Tarragona, Spain*

[2]*Department of Electronic and Communications Technology, Autonomous University of Madrid, Madrid, Spain*

[3]*Electrical Engineering Department, Aswan University, 81542 Aswan, Egypt*

adelsalehali.alraimi, hatem.rashwan, mohamed.abelnasser, vivekkumar.singh, mdmostafakamal.sarker,

Keywords: Semantic Segmentation, Fully Convolutional Network, Pixel-wise Classification, Finger Parts.

Abstract: Image semantic segmentation is in the center of interest for computer vision researchers. Indeed, huge number of applications requires efficient segmentation performance, such as activity recognition, navigation, and human body parsing, etc. One of the important applications is gesture recognition that is the ability to understanding human hand gestures by detecting and counting finger parts in a video stream or in still images. Thus, accurate finger parts segmentation yields more accurate gesture recognition. Consequently, in this paper, we highlight two contributions as follows: First, we propose data-driven deep learning pooling policy based on multi-scale feature maps extraction at different scales (called FinSeg). A novel aggregation layer is introduced in this model, in which the features maps generated at each scale is weighted using a fully connected layer. Second, with the lack of realistic labeled finger parts datasets, we propose a labeled dataset for finger parts segmentation (FingerParts dataset). To the best of our knowledge, the proposed dataset is the first attempt to build a realistic dataset for finger parts semantic segmentation. The experimental results show that the proposed model yields an improvement of 5% compared to the standard FCN network.

1 INTRODUCTION

Semantic segmentation is an important task in image recognition and understanding. It is considered as a dense classification problem. The main task in Semantic segmentation is to assign a unique class to every pixel in an image. Deep learning approaches have been used in several applications, such as human activity recognition, object recognition, image classification (Saleh et al., 2018b), time-series forecasting (Abdel-Nasser and Mahmoud, 2017) as well as semantic segmentation. Recently, convolutional neural networks (CNNs) have obtained significant results in image understanding tasks. However, these approaches still exhibit obvious shortcomings when they come to dense prediction tasks, e.g., semantic segmentation. The main reason for the shortcomings is that these models include repeated steps of pooling and convolution can cause losing much of finer image information.

One way of handling this shortcoming is to learn an up-sampling operation (deconvolution) to generate the feature maps of higher-resolution. Indeed, those deconvolution operations can not recover the lost low-level visual after the down-sampling operations. For this reason, they are unable to precisely generate a high resolution output. Indeed, the low-level visual structure is essential for a proper prediction on the boundaries and details alike. Recently, the work proposed in (Chen et al., 2018) applied dilated convolution filters to deal with larger receptive fields without down-sampling the image. The aforementioned approach is successful, but it has two limitations. First, the dilated convolution uses a coarse sub-sampling of features, which likely causes a loss of important details. Second, it performs convolutions on a large number of detailed feature maps that have high dimensional features, which yields additional algorithmic complexity.

Several applications necessitate accurate segmentation methods, such as activity recognition, navigation, and human body parsing (Saleh et al., 2018a; Liang et al., 2018). One of the important applications is gesture recognition that is the ability to under-

Saleh, A., Rashwan, H., Abdel-Nasser, M., Singh, V., Abdulwahab, S., Sarker, M., Garcia, M. and Puig, D.
FinSeg: Finger Parts Semantic Segmentation using Multi-scale Feature Maps Aggregation of FCN.
DOI: 10.5220/0007382100770084
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 77-84
ISBN: 978-989-758-354-4

standing human hand gestures by detecting and counting finger parts in a video stream or in still images. In this paper, we attempt to deal with such small objects (i.e., finger parts). Consequently, it is essential to extract extra information from different image scales (e.g., fine to coarse features). Thus, we propose to enforce the low level layers to learn these *fine-to-coarse* features. This is achieved by feeding different resolutions of input images to the network. This will be advantageous information for solving finger parts semantic segmentation task, and it can help the model to overcome scale variations, which is considered as high-level knowledge. However, the question here is which scale will be more beneficial for extracting high-level information for an accurate finger parts segmentation. Thus, after feeding images of different scales, our proposed model can learn to weight the generated feature maps at different scales. These feature maps are up-sampled to a unified-scale and then pooled to feed them to next layers, as shown in Figure 1. The main contributions of this paper can be summarized as follows:

- We propose a novel deep aggregation layer based on a multi-scale segmentation network which combines coarse semantic features with fine-grained low-level features in a parallel style to generate high-resolution semantic feature maps. The proposed model is called *FinSeg*.
- With the lack of realistic labeled finger parts datasets, we release a dataset for finger parts semantic segmentation (called *FingerParts* dataset). As far as we know, this is the first available dataset for finger parts segmentation using a high resolution real images.

2 RELATED WORKS

Recently, the most successful methods for the semantic segmentation task are related to deep learning models, specifically CNNs. In (Girshick et al., 2014), a region-proposal-based method has been used to estimate segmentation results. In turn, the authors of (Long et al., 2015; Chen et al., 2018) have shown the effective feature generation of CNNs and presented semantic segmentation based on the fully convolutional networks (FCNs). It worth to note that FCN becomes a standard deep network for different applications, such as image restoration (Eigen et al., 2013), image super-resolution (Dong et al., 2014) and depth estimation (Eigen and Fergus, 2015; Eigen et al., 2014). However, the main limitation of networks based on the FCN architecture is the low-resolution prediction. Thus, many works proposed different techniques to tackle this limitation in order to generate high-resolution predictions. For instance, conditional random field (CRF) has been used as a post layer for coping with this problem. This is done by generating a middle resolution score feature map and then refining boundaries using a dense CRF. In addition, an atrous convolution layer has been proposed in (Chen et al., 2014). The atrous layers are convolution filters with different rates to extract the key features of input images in different scales. In (Zheng et al., 2015), a robust end-to-end fashion parsing method is proposed by adding recurrent layers in order to improve the performance of the FCN network.

Furthermore, many deconvolution based methods have been proposed in (Badrinarayanan et al., 2015; Noh et al., 2015) to learn how to up-sample low resolution prediction by taking into account the advantage of middle layer features in the FCN network. For example, the work proposed in (Chen et al., 2014) added prediction layers to middle layers to generate prediction scores at multiple resolutions. Then the multi-resolution predictions are averaged to generate the final prediction. But, this model was trained in multi-stage style rather than end-to-end manner. In turn, other methods, such as SegNet (Badrinarayanan et al., 2015), (Sarker et al., 2018; Singh et al., 2018) and U-Net (Ronneberger et al., 2015) have used skip-connections in the decoder architecture to add information from feature maps extracted of the middle layers to the deconvolution layers.

Unlike the aforementioned methods, the proposed FinSeg model exploits the multi-scale features in the low-level layers in order to predict coarse-to-fine semantic features extracted from different resolution of an input image. In addition, unlike the standard FCN network, FinSeg uses the residual network, namely ResNet101, instead of the VGG network. In addition, we use the skip-connections of all encoder layers to add feature maps to all decoder layers as shown in Figure 1.

3 PROPOSED MODEL

We propose a deep semantic segmentation model (FinSeg) based on a new aggregation layer. FinSeg accepts an input image at different resolutions, extracts feature maps of every scale, weights each extracted feature maps, pools them and then feeds the final feature maps through long range connections to achieve a high-resolution semantic segmentation of finger parts. Below, we describe the steps of our model.

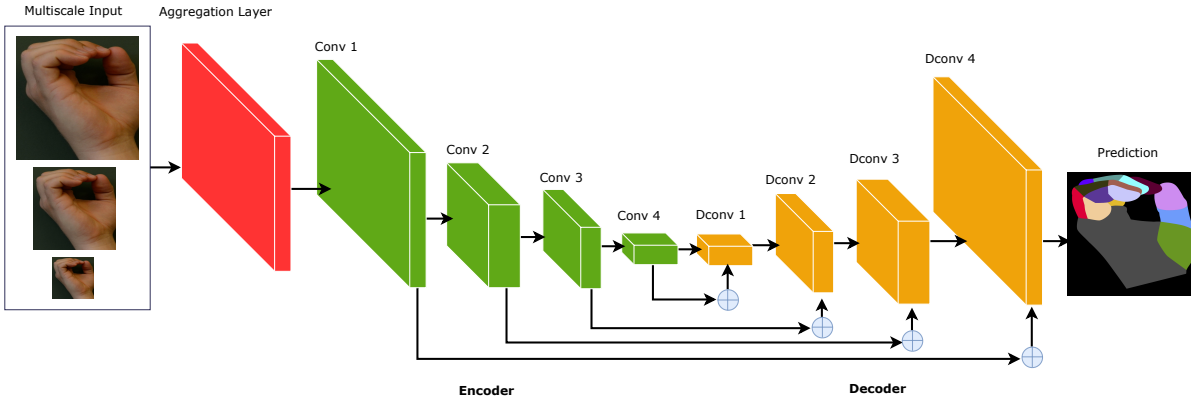

Figure 1: The main structure of the proposed model (FinSeg). Red block refers to the generation of the feature maps from the proposed aggregation block (shown in details in Figure 2).

3.1 FinSeg Architecture

As shown in Figure 1, the proposed model has an encoder-decoder architecture. In general, the encoder reduces the spatial dimension through pooling layers along with summarizing the input images. In turn, the decoder recovers the object mask and spatial dimension. Following (Ronneberger et al., 2015), we use skip-connections from the encoder to the decoder in order to recover the object details in the decoder stage by transferring low level feature from lower layers to the higher ones.

3.2 Aggregation Layer

We show the architecture of the aggregation layer in Figure 2. As shown, an image I is fed into the model with s scales. The input images $I_1, I_2 ... I_s$ are fed to a parallel sequence of convolution layers. Shared convolution filters are applied on the images of different scales. After feeding images of different scales through first parallel layers of the model, the resulted feature maps have different sizes. Since, it is not possible to aggregate feature maps with different sizes, the multi-scale feature maps are up-sampled to the largest dimension and aggregated in one feature map. After aggregation, the resulted feature maps are then fed into the next aggregation layer and this procedure is repeated k times.

Fully connected layer (FC) of s inputs and $s \times nl$ outputs is used to learn the weights of the aggregation alyer, where s is the number of scales and nl is the number of internal sequent layers of the aggregation layer. We propose a fully automated procedure that can learn how to give a high weight for the more important scaled feature maps and suppress others. In this study, $s = 3$ and $nl = 3$ are the optimum values that yield the best results. The FC layer learns to weight the resulted feature maps of each scale (see Figure 2). A softmax function is used as an activation for each resulted s weights. In this work FC is initialized with an input vector $w = [1/3; 1/3; 1/3]$. It is obvious that we start with giving an equal weight for all scales.

Suppose that the final aggregated feature maps extracted at a layer l can be expressed as follows:

$$F_{l,i} = \sum_{i=1}^{s} w_{l,i} F_{l,i-1}$$

under the constraint of $\sum_{i=1}^{s} w_{l,i} = 1$, where $F_{l,i-1}$ is the feature maps of the previous scale $i-1$, and $i \in 1....s$ with $l \geq 2$. The resulted $F_{l,i}$ is then fed into the convolution layer of the next internal layer.

3.3 Encoder and Decoder of FinSeg

Encoder: After calculating the multi-scale aggregated feature maps, they are fed into the encoder network. The encoder consists of four convolution layers followed by a max-pooling layers (down-sampling layers) to encode input into feature representations at different levels as shown in Figure 1. The encoder layers are adapted from the pre-trained ResNet101 network (the first four layers only).

Decoder: It consists of up-sampling and summing followed by regular convolution operations. To recover original image dimensions by up-sampling, we use the bi-linear interpolation. Thus, we expand the feature maps dimensions to meet the same size with the corresponding blocks of the encoder and then apply skip connections by summing the feature maps of the decoder layer with the ones generated from the corresponding encoder layers.

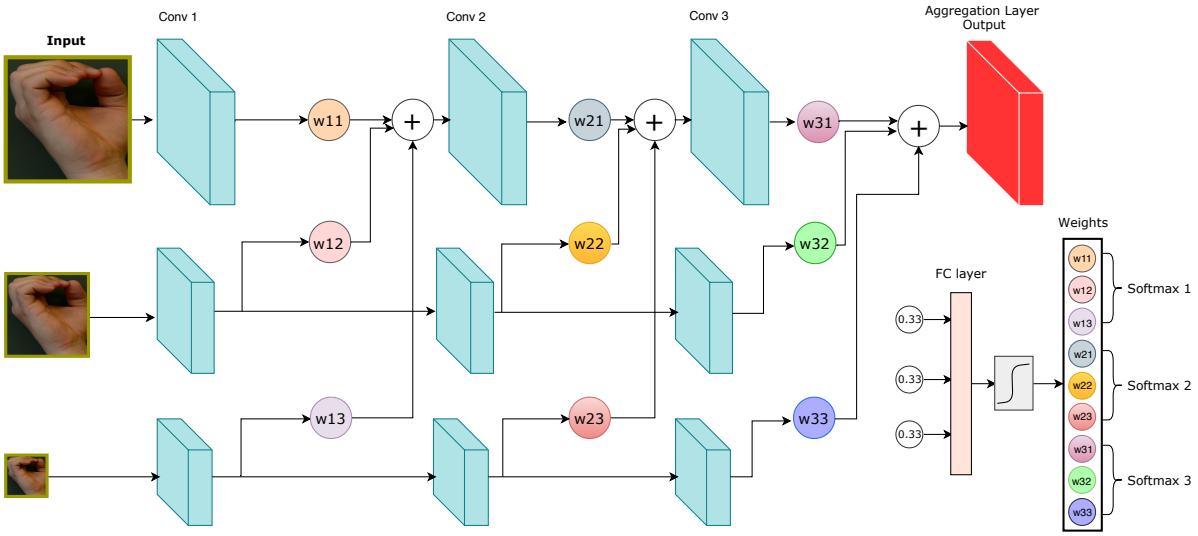

Figure 2: The architecture of the aggregation layer. Feature maps are aggregated at the largest scale in each internal layer.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 FingerParts Dataset

In this paper, we introduce to a new dataset based on real hand images (called FingerParts) that can be used for the human palm and finger parts segmentation task. The FingerParts dataset contains 1100 real images and their corresponding annotations. We have ordered human made annotations, which is in general perfect. Number of hands per image is ranging from one hand to three hands in most cases. These images can contain backside or frontal views of different hands as shown in Figure 3.

Furthermore, 1000 images were taken from a public dataset for hand gesture recognition (Kawulok et al., 2014; Nalepa and Kawulok, 2014; Grzejszczak et al., 2016). In addition, 100 images were collected by scrapping images from *Google Image*. The results of scrapping were manually checked in order to avoid repeated and non-relevant images. The number of classes in the dataset is 17: a class for the background, 3 classes per finger $(3 \times 5 = 15)$ and one for each palm. Information about key-points is also available. There is 16 key points information per hand (i.e., 15 for the fingers parts and one for the palm). In Table 1, we show a comparison between the FingerParts dataset with prior state-of-the-art datasets. It is obvious that our dataset is based on realistic images and it can be used for semantic segmentation and gesture recognition tasks.

Data Augmentation

In this study, we applied data augmentation by scaling the input images by a random value varying between 0.5 and 2.0. In addition, we applied illumination changes via a gamma correction operator with values varying from 0.5 to 3.0 with a step of 0.5. Random horizontal flipping was also applied. Furthermore, we added extra synthetic backgrounds to the input images to expose the model to more difficult tasks. In total, we have 58,380 images for training and 4935 for testing.

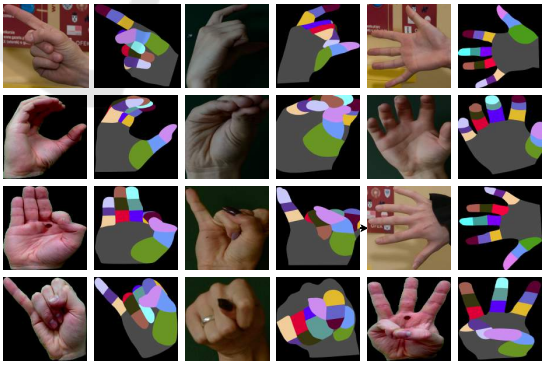

Figure 3: Samples of the proposed FingerParts dataset.

4.2 Training Procedure

In each iteration, FinSeg reads a batch of 8 images, resizes them to 512x512 and normalizes them. The normalization step consists of 3 steps: 1) the input image is divided by 255. This step makes values of each RGB image varies between 0 and 1.0,

Table 1: Quantitative comparison of our proposed dataset, FingerParts, with public datasets of hand segmentation task.

Dataset	Number of Images	Segmentation Task	Real/Synthetic	Key Point
(Zimmermann and Brox, 2017)	41258	Yes	Synthetic	Yes
(Kawulok et al., 2014)	899	No	Real	Yes
MU HandImages (Barczak et al., 2011)	2425	No	Real	Yes
FingParts(our)	1100	Yes	Real	Yes

2) centralization of image values through subtracting $[0.485, 0.456, 0.406]$ from RGB channels respectively is applied, and 3) the RGB channels are divided by $[0.229, 0.224, 0.225]$. Those values were used on ImageNet dataset for classification task and fixed (empirically) from computer vision community. An initial learning rate of 0.01 with weight decay of 10^{-8} were used in the training procedure. SGD was chosen as an optimizer and with a value of 0.99 for the momentum parameter. In this work, the cross-entropy is used as a loss function. It is defined as:

$$CE = -\sum_{i} y_i' log(y_i)$$

where y_i is the probability for predicted class i and y_i' is the true probability for that class.

Although, the proposed aggregation layer add some algorithmic complexity to the proposed model by multi-scale layers, it converges in the same number of iterations of the standard FCN model (See Figure 4). However, the training process is more expensive in terms of time consumption.

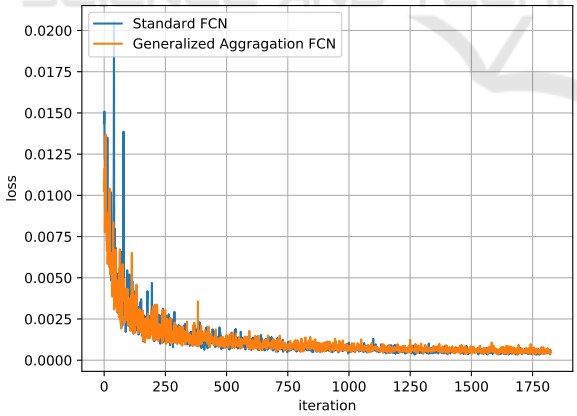

Figure 4: The convergence of the proposed model and the FCN model.

4.3 Evaluation Metrics

In this work, we use two metrics to assess the performance of the proposed model: the *intersection over Union* (IoU) and *pixel accuracy*. In literature, IoU is referred to as the Jaccard index, which is basically a metric to calculate the percent overlap between the target mask and the prediction output.

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

We also use the *pixel accuracy* metric. This metric reports the percent of pixels in the image which were correctly classified. The pixel accuracy is calculated for each class separately as well as globally over all classes. It can be defined as follows:

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

4.4 Experimental Results and Discussion

We evaluate our approach on the proposed dataset (FingerParts). To present the usefulness of automatically selecting of feature maps scales, we choose the FCN model of (Chen et al., 2014) as baseline. In this section, we compare the results of three variations of the aggregation layer of the proposed model (AvrageAggr, AggrFCNSoftmax and AggrFCNRelu) with the ones of FCN model. The first variation of the proposed aggregation layer (AvrageAggrFCN), we apply aggregation is by averaging of feature maps of different scales with the same internal layer. The second variation (AggrFCNSoftmax), we use a softmax function as an activation function applied on the weights resulting of the FC layer. In the third variation (AggrFCNRelu), we add a Relu after every internal convolution layer of the aggregation block.

Table 2 hows the experimental results of the proposed model with the proposed dataset. The baseline model, FCN, yielded an IoU of 0.58 and an accuracy of of 87%. AvrageAggr gave an improvement of 4% in IoU values (only after average the feature maps extracted at different scales). However, for the accuracy, there was a small improvement ($< 0.5\%$).

Learning a weight for the resulted feature maps at a scale is a generalized form of aggregation, and it has more potential to find optimized weights. According to results shown in Table 2, predicting the weights of each feature map using an FCN layer yields better results than the baseline model. An improvement of 5% with AggrFCNSoftmax was achieved. Another experiment were conducted to check Relu function as an activation function with AggrFCNRelu yielded an

IoU improvement of about 3%. Thus, the best results was achieved when we use the softmax function for estimating the weight values of each scale.

Qualitative results of some of these experiments are shown in Figure 5. As shown, and supporting our quantitative results, the proposed model with AggrFCNSoftmax (using aggregation of FC and sofmax layers) present visual improvements of finger parts segmentation with our dataset, compared to the FCN model and the two other variations of the proposed model (AggrFCNRelu and AvrageAggr).

Table 2: The performance of the three variants of the proposed model (AggrFCNSoftmax, AggrFCNRelu and AvrageAggr) and the FCN model.

Method	IoU	Accuracy
FCN (Chen et al., 2014)	0.5833	87.32
AvrageAggr	0.6231	87.64
AggrFCNSoftmax	**0.6307**	**88.13**
AggrFCNRelu	0.6151	87.91

A Case Study

To assess the performance of the proposed model on a concrete case, we select an image randomly (see Figure 6) from the dataset.Then, we analyze the performance of the proposed model under different conditions: illumination changing, background changing, and image flipping. With no effects on the input image, our model achieved an *IoU* of 0.5515. Applying illumination effect based on nonlinear Gamma correction with different values ($\gamma \in \{0.5, 1.0, 1.5, 2.5\}$) causes a degradation in the performance of our model (*IoU* drops to 0.5515). This degradation can be explained by the disappearance of small parts in Figure 6-(col 1-2). Another issue was investigated by changing the background and image flipping. Our experiments show that the changing in the background *IoU* reduces to 0.5501 (see Figure 6-(cols. 3-4)), while image flipping reduces the *IoU* to 0.5493 (see Figure 6-(cols. 5-6)). As shown, the change of the *IoU* value around 0.55 under different conditions, such illumination changes, adding background and image flipping. Consequently, we can say that the change on the global context of the input images has insignificant impact on the final decision of the proposed model. It is important to note that different finger parts are discriminated using their relative location to the palm more than their appearance. Thus, we can conclude that the model learns how to extract global shape information from the input images.

5 CONCLUSIONS

In this paper, we have proposed a novel deep learning based model for finger parts semantic segmentation. The proposed model is based on generating features maps with different resolution of an input image. These features maps are then aggregated together using automated weights estimated from fully connected layer. The estimated weights are used to assign a high weight for the more important scaled feature maps and suppress others. The generated feature maps are fed into an encoder-decoder network with skip-connections to predict the final segmentation mask. In addition, we have introduced a new dataset that can help to solve finger parts semantic segmentation problem. To the best of our knowledge, FingerParts is first dataset for finger parts semantic segmentation with real high resolution images. The proposed model outperformed the standard FCN network with an improvement of 5% in terms of the *IoU* metric. Future work will include the use of the segmented fingers parts to improve the accuracy of gesture recognition methods.

REFERENCES

Abdel-Nasser, M. and Mahmoud, K. (2017). Accurate photovoltaic power forecasting models using deep lstm-rnn. *Neural Computing and Applications*, pages 1–14.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*.

Barczak, A., Reyes, N., Abastillas, M., Piccio, A., and Susnjak, T. (2011). A new 2d static hand gesture colour image dataset for asl gestures.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer.

Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658.

Eigen, D., Krishnan, D., and Fergus, R. (2013). Restoring an image taken through a window covered with dirt or

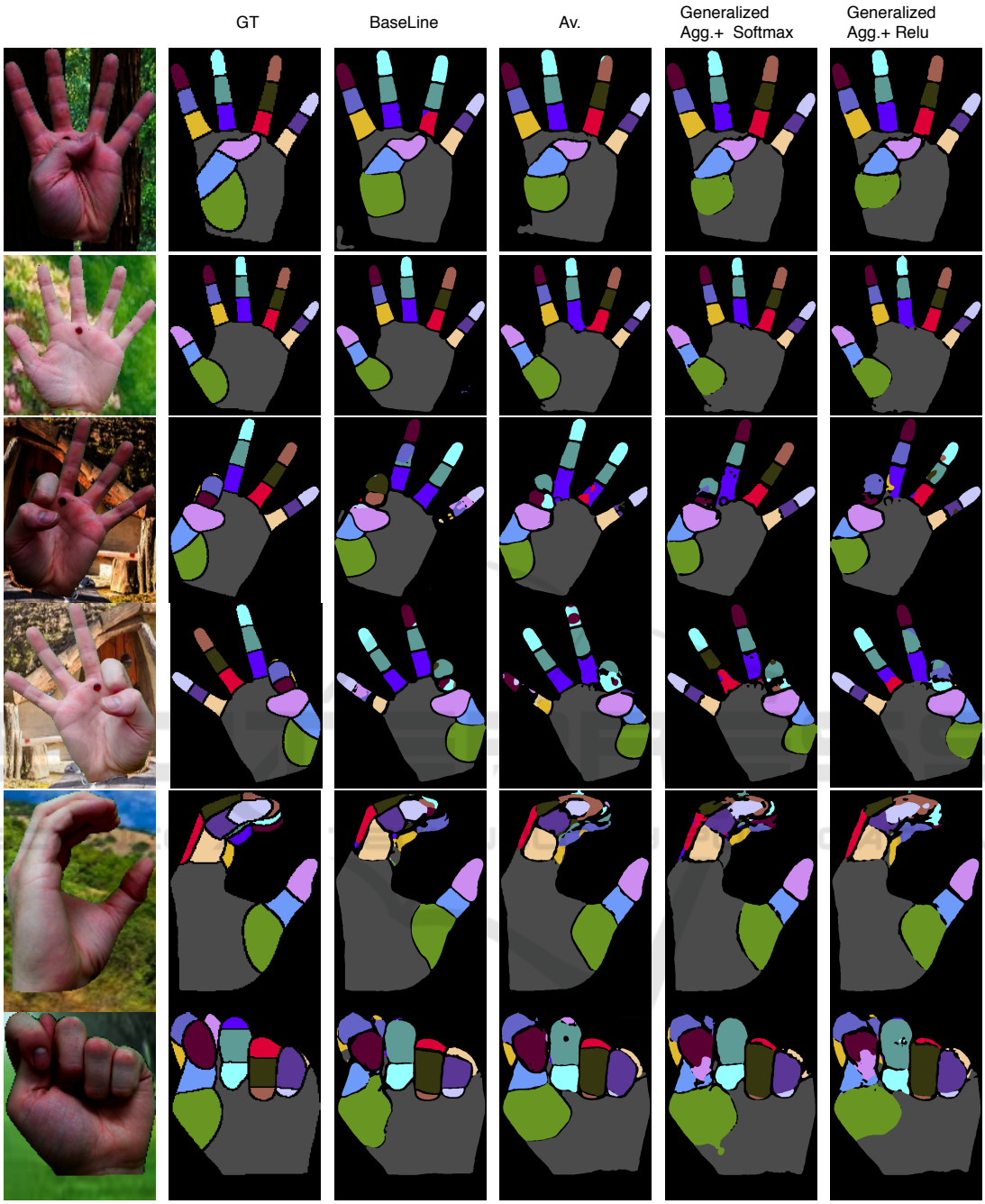

Figure 5: A visual comparison between the different versions of proposed model (FinSeg) and the FCN model with the FingerParts dataset. Input images (col. 1), ground-truth (col. 2), results of the FCN model (col. 3), results of AvrageAggr (col. 4), results of AggrFCNSoftmax (col. 5), and results of AggrFCNSoftmax (col. 6).

rain. In *Proceedings of the IEEE international conference on computer vision*, pages 633–640.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Grzejszczak, T., Kawulok, M., and Galuszka, A. (2016). Hand landmarks detection and localization in color images. *Multimedia Tools and Applications*, 75(23):16363–16387.

Kawulok, M., Kawulok, J., Nalepa, J., and Smolka, B.

Figure 6: Analyzing the performance of the proposed model under different conditions: illumination changing (cols. 1-2), background changing (cols. 3-4), and image flipping (cols. 5-6).

(2014). Self-adaptive algorithm for segmenting skin regions. *EURASIP Journal on Advances in Signal Processing*, 2014(1):170.

Liang, X., Gong, K., Shen, X., and Lin, L. (2018). Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Nalepa, J. and Kawulok, M. (2014). Fast and accurate hand shape classification. In *International Conference: Beyond Databases, Architectures and Structures*, pages 364–373. Springer.

Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Saleh, A., Abdel-Nasser, M., Garcia, M. A., and Puig, D. (2018a). Aggregating the temporal coherent descriptors in videos using multiple learning kernel for action recognition. *Pattern Recognition Letters*, 105:4–12.

Saleh, A., Abdel-Nasser, M., Sarker, M. M. K., Singh, V. K., Abdulwahab, S., Saffari, N., Garcia, M. A., and Puig, D. (2018b). Deep visual embedding for image classification. In *Innovative Trends in Computer Engineering (ITCE), 2018 International Conference on*, pages 31–35. IEEE.

Sarker, M., Kamal, M., Rashwan, H. A., Banu, S. F., Saleh, A., Singh, V. K., Chowdhury, F. U., Abdulwahab, S., Romani, S., Radeva, P., et al. (2018). Slsdeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. *arXiv preprint arXiv:1805.10241*.

Singh, V. K., Romani, S., Rashwan, H. A., Akram, F., Pandey, N., Sarker, M., Kamal, M., Barrena, J. T., Saleh, A., Arenas, M., et al. (2018). Conditional generative adversarial and convolutional networks for x-ray breast mass segmentation and shape classification. *arXiv preprint arXiv:1805.10207*.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537.

Zimmermann, C. and Brox, T. (2017). Learning to estimate 3d hand pose from single rgb images. Technical report, arXiv:1705.01389. https://arxiv.org/abs/1705.01389.