Domain Adaptation for Holistic Skin Detection

Aloisio Dourado, Frederico Guth, Teofilo de Campos and Li Weigang
Universidade de Brasilia (UnB), Departamento de Ciência da Computação - CIC
Email: t.decampos@st-annes.oxon.org

***Abstract*—Human skin detection in images is a widely studied topic of Computer Vision for which it is commonly accepted that analysis of pixel color or local patches may suffice. However, we found that the lack of contextual information may hinder the performance of local approaches. In this paper, we present a comprehensive evaluation of holistic and local Convolutional Neural Network (CNN) approaches on in-domain and cross-domain experiments and compare them with state-of-the-art pixel-based approaches. We also propose combining inductive transfer learning and unsupervised domain adaptation methods evaluated on different domains under several amounts of labelled data availability. We show a clear superiority of CNN over pixel-based approaches even without labeled training samples on the target domain and provide experimental support for the superiority of holistic over local approaches for human skin detection.**

I. Introduction

Human skin detection is the task of identifying which pixels of an image correspond to skin. The segmentation of skin regions in images has several applications: video surveillance, people tracking, human-computer interaction, face detection and recognition and gesture detection, among many others [1], [2].

Before the boom of Convolutional Neural Networks (CNNs), most approaches were based on skin-color separation or texture features, as in [3] and [4]. Despite all the advances that deep fully convolutional neural networks have brought for image segmentation, some common criticism is still made to argue that pixel-based approaches are more suitable for skin detection. Namely: the need for large training datasets [5]; the specificity or lack of generalization of neural nets; and prediction time [6].

In this paper, to address the first criticism (on the need for large training datasets), we propose a new Domain Adaptation strategy that combines Transfer Learning and Pseudo-Labeling [7] in a cross-domain scenario that works under several levels of target domain label availability. We evaluate the proposed strategy under several cross-domain situations on four well-known skin datasets. We also address the other criticisms with a series of comprehensive in-domain and cross-domain experiments. Our experiments show the effectiveness of the proposed strategy and confirm the superiority of FCN approaches over local approaches for skin segmentation. We can improve the F_1 score on skin segmentation using little or no labelled data from the target domain with the proposed strategy.

Our main contributions are:

1) the proposal of a new Domain Adaptation strategy that combines Pseudo-Labeling and Transfer Learning for cross-domain training;
2) a comparison of holistic versus local approaches on in-domain and cross-domain experiments applied to skin segmentation with an extensive set of experiments;
3) a comparison of CNN-based approaches with state-of-the-art pixel-based ones; and
4) experimental assessment of the generalization power of different human skin datasets (domains).

II. Background

A. Fully Convolutional Networks (FCN)

In opposition to patch-based classification [8], where each pixel is classified using a patch of the original image that surrounds it, the FCN-based approach for image segmentation introduced by [9] considers the context of the whole image. The FCNs are Convolutional Neural Networks (CNN) in which all trainable layers are convolutional. These networks can perform segmentation taking the whole image as an input signal and generate full image segmentation results in one forward step of the network, without requiring to break the image into patches. Because of that, FCNs are faster than the patch-based approaches, and overcame the state-of-the-art on PASCAL VOC, NYUDv2, and SIFT Flow datasets, by using Inductive Transfer Learning from ImageNet.

Following the success of FCNs, [10] proposed the U-Net architecture, that consists of an encoder-decoder structure initially used in biomedical 2D image segmentation. In U-Net, the encoder path is a typical CNN, where each down-sampling step doubles the number of feature channels. What makes this architecture unique is the decoder path, where each up-sampling step concatenates the output of the previous step with the output of the down-sampling with the same image dimensions. With this strategy, the U-Net is able to model contextual information, increasing robustness level of detail.

B. Transfer Learning and Domain Adaptation

Following the notation of [11], a domain $\mathcal{D}$ is composed of a d-dimensional feature space $\mathcal{X} \subset R^d$ with a marginal probability distribution $P(\mathrm{X})$. A task $\mathcal{T}$ is defined by a label space $\mathcal{Y}$ with conditional probability distribution $P(\mathrm{Y}|\mathrm{X})$. In a conventional supervised machine learning problem, given a sample set $\mathrm{X} = \{\mathrm{x}_1, \cdots, \mathrm{x}_\mathrm{n}\} \in \mathcal{X}$ and the corresponding labels $\mathrm{Y} = \{\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{n}\} \in \mathcal{Y}$, $P(\mathrm{Y}|\mathrm{X})$ can be learned from feature-label pairs in the domain. Suppose we have a source domain $\mathcal{D}^s = \{\mathcal{X}^s, P(X^s)\}$ with a task $\mathcal{T}^s = \{\mathcal{Y}^s, P(Y^s|X^s)\}$ and a target domain $\mathcal{D}^t = \{\mathcal{X}^t, P(X^t)\}$ with a task $\mathcal{T}^t = \{\mathcal{Y}^t, P(\mathrm{Y}^t|\mathrm{X}^t)\}$. If the two domains correspond ($\mathcal{D}^s = \mathcal{D}^t$) and the two tasks are the same

($\mathcal{T}^s = \mathcal{T}^t$), we can use conventional supervised Machine Learning techniques. Otherwise, adaptation and/or transfer methods are required.

If the source and target domains are represented in the same feature space ($\mathcal{X}^s = \mathcal{X}^t$), but with different probability distributions ($P(\mathrm{X}^\mathrm{s}) \neq P(\mathrm{X}^t)$) due to domain shift or selection bias, the transfer learning problem is called homogeneous. If ($\mathcal{X}^s \neq \mathcal{X}^t$), the problem is heterogeneous TL [11], [12].

Domain Adaptation. Domain Adaptation is the problem where tasks are the same, but data representations are different or their marginal distributions are different (homogeneous). Mathematically, $\mathcal{T}^s = \mathcal{T}^t$ and $\mathcal{Y}^s = \mathcal{Y}^t$, but $P(\mathrm{X}^s) \neq P(\mathrm{X}^t)$. Most of the literature on domain adaptation for visual applications is dedicated to image classification [12]. To extend the domain adaptation concepts to the image segmentation problem, we treat the Skin Segmentation as a pixel-wise classification problem.

Inductive Transfer Learning. When source and target domains are different ($\mathcal{D}^s \neq \mathcal{D}^t$), models trained on $\mathcal{D}^s$ may not perform well while predicting on $\mathcal{D}^t$ and if tasks are different ($\mathcal{T}^s \neq \mathcal{T}^t$), models trained on $\mathcal{D}^s$ may not be directly applicable on $\mathcal{D}^t$. Nevertheless, when $\mathcal{D}^s$ maintains some kind of relation to $\mathcal{D}^t$ it is possible to use some information from $\{\mathcal{D}^s, \mathcal{T}^s\}$ to train a model and learn $P(\mathrm{Y}^t|\mathrm{X}^t)$ through a processes that is called Transfer Learning (TL) [11].

According [12], the Transfer Learning approach is called inductive if the target task is not exactly the same as the source task, but the tasks are in some aspects related to each other. For instance, consider an image classification task on ImageNet [13] as source task and a Cats vs Dogs classification problem as a target task. If a model is trained on a dataset that is as broad as ImageNet, one can assume that most classification tasks performed on photographies downloaded from the web are subdomains of ImageNet which includes the Cats vs Dogs problem (i.e. $\mathcal{D}^{\mathtt{cats}\times\mathtt{dogs}} \subset \mathcal{D}^{\mathtt{ImageNet}}$), even though the tasks are different ($\mathcal{Y}^{\mathtt{ImageNet}} = \mathbb{R}^{1000}$ and $\mathcal{Y}^{\mathtt{cats}\times\mathtt{dogs}} = \mathbb{R}^2$). This is the case of a technique to speed up convergence in Deep CNNs that became popularised as *Fine Tuning* for vision applications.

In deep artificial neural networks, fine tuning is done by taking a pre-trained model, modifying its final layer so that its output dimensionality matches $\mathcal{Y}^t$ and further training this model with labelled samples in $\mathcal{D}^t$. In this work, we compare our proposed domain adaptation approach to inductive transfer learning applied to the skin segmentation problem.

Unsupervised Domain Adaptation. Domain adaptation methods are called unsupervised (also known as transductive TL) when labeled data is available only on source domain samples.

Several approaches have been proposed for unsupervised DA, most of them were designed for shallow learning methods [14]. The methods that exploit labeled samples from the source domain follow a similar assumption to that of Semi-Supervised Learning methods, with the difference that test samples come from a new domain. This is the case of [15] and [16]. Both methods start with a standard supervised learning method trained on the source domain in order to classify samples from the target domain. The classification results are taken as pseudo-(soft)labels and used to iteratively improve the learning method in a way that it works better on the target domain.

When labeled samples are not available at all, it is possible to perform unsupervised transfer learning using methods that perform feature space transformation. Their goal is to align source and target domain samples to minimise the discrepancy between their probability density functions [17]. Style transfer techniques such as that of [18] achieve a similar effect, but their training process is much more complex.

Semi-supervised learning. Semi-supervised learning methods deal with the problem in which not all training samples have labels [19], [20]. Most of these methods use a density model in order to propagate labels from the labeled samples to unlabeled training samples. This step is usually combined with a standard supervised learning step in order to strengthen the classifiers, c.f. [21], [22]. There are several semi-supervised learning approaches for deep neural networks. Methods include training networks using a combined loss of an auto-encoder and a classifier [23], discriminative restricted Boltzmann machines [24] and semi-supervised embeddings [25]. [7] proposed a simple yet effective approach, known as Pseudo-Labelling, where the network is trained in a semi-supervised way, with labeled and unlabeled data in conjunction. During the training phase, for the unlabeled data, the class with the highest probability (pseudo-label) is taken as it was a true label. To account for the unbalance between true and pseudo labels, the loss function uses a balancing coefficient to adjust the weight of the unlabeled data on each mini-batch. As a result, pseudo-label works as an entropy regularization strategy. These methods assume that training and test samples belong to the same domain, or at least that they are very similar ($\mathcal{D}^s \approx \mathcal{D}^t$). The original Pseudo-Labelling method only considers in-domain setups. As we are dealing with homogeneous DA, in this work we extend Pseudo-Labeling to cross-domain and evaluate this extension under several target label availability constrains, from semi-supervised learning to fully unsupervised .

C. Related Works on Skin Detection

[6] achieved state-of-the-art results in skin segmentation using correlation rules between the YCb and YCr subspaces to identify skin pixels on images. A variation of that method was proposed by [26], who claimed to have achieved a new state-of-the-art plateau on rule-based skin segmentation based on neighborhood operations. [27] compared different color-based and CNN based skin detection approaches on several public datasets and proposed an ensemble method.

In contrast to Domain Adaptation for image classification, it is difficult to find literature focused on domain adaptation methods for image segmentation [12], especially for the skin detection problem. [28] use agreement of two detectors based on skin color thresholding, applied to selected images from several manually labeled public datasets for human activity recognition, but do not explore their use in cross-domain

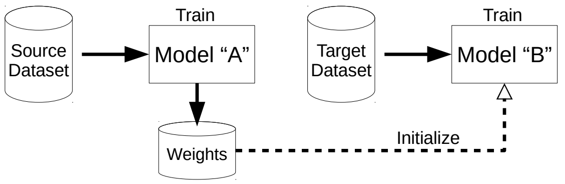

Fig. 1. Inductive Transfer Learning by “fine tuning” parameters of a model to a new domain. Model “A” parameters are trained on the source dataset. Model “B” parameters are initialized from Model “A” parameters. Model “B” is then “fine tuned” to the new domain.

setups. [29] also use two independent detectors, with their parameters selected by maximising agreement on correct detections and false positives to dynamically change a classifier on new data automatically without any user annotation.

In this work we compare two CNN approaches (one patch-based and one fully convolutional) with above mentioned state-of-the-art pixel-based methods for in-domain skin detection. We also compare the two CNN approaches to each other in cross-domain setups, even in the absence of target-domain labeled data. Unfortunately, previous state-of-the-art pixel-based skin segmentation papers do not present results on cross-domain setups.

III. Methods

In order to exploit domain adaptation techniques to address training data avaiability problem for skin segmentation, we evaluate conventional transductive transfer learning using fine tuning, our cross-domain extension applied to the Pseudo-Labeling approach of [7] and our proposed combined approach that uses both inductive transfer learning and unsupervised or semi-supervised DA. All source code developed to perform the training and the evaluations, along side the resulting models weights are available from http://cic.unb.br/~teodecampos/.

For inductive transfer learning with deep networks, we use the learnt parameters from the source domain as starting point for optimisation of the parameters of the network on the target domain. The optimisation first focuses on the modified output layer, which is intimately linked with the classification task. Other layers are initially frozen, working as a feature extraction method. Next, all parameters are unfrozen and optimisation carries on until convergence. Figure 1 illustrates this process.

In this work, we propose a method that relates the pseudo-label approach of [7], but instead of using the same model and domain for final prediction and pseudo-label generation, we use a model trained in a different domain to generate pseudo labels for the target domain. These pseudo-labels are then used to fine-tune the original model or to train another model from scratch in a semi-supervised manner. We call this technique **cross-domain pseudo-labelling** as illustrated in Figure 2. This approach allows us to train the final model with very few labeled data of the target domain. In the worst case scenario, the model can be trained with no true label at all, in a fully unsupervised fashion. This still takes advantage of Entropy Regularization of the pseudo-label technique.

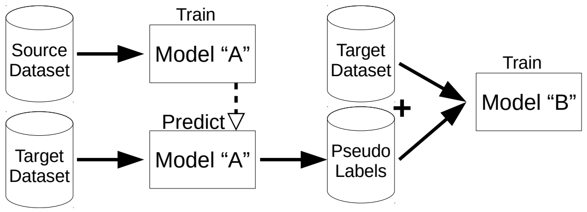

Fig. 2. Semi-supervised and unsupervised Domain Adaptation by cross-domain pseudo-labelling. Model “A” is trained on the source dataset and it is used to predict labels on the target dataset. Then, the target dataset and previously predicted labels are used to train Model “B”. When no labels are available on the target dataset, the process is fully unsupervised.

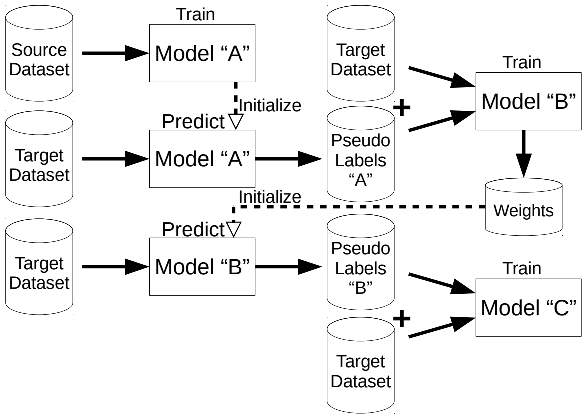

Fig. 3. Combined transfer learning and domain adaptation approach. Model “A” is trained on the source dataset and it is used to predict labels on the target dataset. Then, target dataset and previously predicted labels are used to train Model “B” which is fine tuned on the target dataset before be used to generate a new set of more accurate pseudo-labels.

Our last approach consists in combining fine tune and pseudo labeling approaches in order to improve final model performance. Figure 3 illustrates this procedure. We use weights obtained from a cross-domain pseudo-label model (Model “B”) to fine tune a model that will be used to generate a more accurate set of pseudo-labels. These new pseudo-labels are then used in one in-domain pseudo-label training round to get the final model (“Model C”). The intuition behind this approach is that using a more accurate set of labels jointly with weights of a better model should lead to better results. Because of the fine tuning step, which requires at east some labels from the target dataset, this approach is semi-supervised.

We evaluated two approaches for skin segmentation, a local (patch-based) convolutional classification method and a holistic (FCN) segmentation method. The patch-based approach uses the raw values of a small region of the image to classify each pixel position based on its neighbourhood. Inspired by the architecture described by [8], we use a 3 convolutional layer network with max pooling between convolutions, but we add ReLU activation function in the inner layers, to reduce the vanishing gradient effect. As input, we use a patch of 35×35 pixels and 3 channels, to allow the network to capture the surroundings of the pixel. This patch size is similar to that used by [8] (32×32), but we chose an odd number to focus

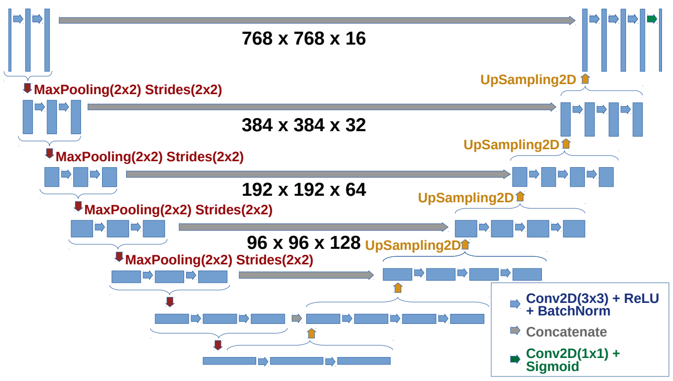
768 x 768 x 16
MaxPooling(2x2) Strides(2x2)
UpSampling2D
384 x 384 x 32
MaxPooling(2x2) Strides(2x2)
UpSampling2D
192 x 192 x 64
MaxPooling(2x2) Strides(2x2)
UpSampling2D
96 x 96 x 128
UpSampling2D
MaxPooling(2x2) Strides(2x2)
Conv2D(3x3) + ReLU + BatchNorm
Concatenate
Conv2D(1x1) + Sigmoid

Fig. 4. Our variation of the U-Net architecture for holistic image segmentation.

the prediction in the center of the patch. The output of the network consists of two fully connected layers and a sigmoid final activation for binary classification. For this approach, the images are not resized. To reduce the cost of training while maintaining data diversity, data subsampling is used so that only 512 patches are randomly selected from each image. For prediction, all patches are extracted in a sliding window fashion, making one prediction per pixel.

Due to its simplicity and performance [27], we choose to use the U-Net as the holistic segmentation method to be evaluated in this paper. Our model follows the general design proposed by [10], but we use a 7-level structure with addition of batch normalization between the convolutional layers, as shown in figure 4. We also use an input frame of 768 × 768 pixels and 3 channels to fit most images, and same size output. Smaller images are framed in the center of the input and larger ones are resized in a way that its larger dimension fits the input frame. For evaluation purposes, predictions are done over the images restored to their original sizes. In both local and holistic models, the image pixels are normalized to 0 to 1 and the sigmoid activation function applied to the output. In both models we also used data augmentation, randomly varying pixels values in the HSV colour space. For the U-Net model we also used random shift and flip. To favor sharpness of predicted bonderies, for the loss function, we used a modified (and differentiable) Sørensen–Dice coefficient [30]. For each given class label, let $\vec{p} \in [0,1]^{\mathcal{I}}$ be the vector of predicted probabilities for each pixel (where $\mathcal{I}$ is the number pixels in each image), $\vec{q} \in \{0,1\}^{\mathcal{I}}$ be the binary vector that indicates, for each pixel, if that class has been detected, based on $\vec{p}$, and $\vec{g}$ be the ground truth binary vector that indicates the presence of that label on each pixel. The derived loss function from the Sørensen–Dice coefficient is given by equation (1), where s is the smoothness parameter that was set to $s = 10^{-5}$.

$$DiceLoss(\vec{p}, \vec{g}) = 1 - \frac{s + 2\vec{p} \cdot \vec{g}}{s + |\vec{p}| + |\vec{g}|} \tag{1}$$

IV. Experiments and Results

The main goal of our experiments is to evaluate the performance of homogeneous transductive fine-tuning, cross-domain pseudo-labelling, and a combined approach in several domains and under different availability of labeled data on the target domain. In our experiments, we used four well-known datasets dedicated to skin segmentation: Compaq [31] – a very traditional skin dataset with 4,670 images of several levels of quality; SFA [32] – a set of 1,118 face images obtained from two distinct datasets, some of them with white background; Pratheepan [33] – 78 family and face photos, randomly downloaded using Google; and VPU [28] – 290 images extracted from video surveillance cameras. As most of the datasets do not provide a standard image-based train/test split, we adopted the same test split reported by the authors of SFA [32], which uses 15% of the images for testing and the remaining for training on all these datasets.

TABLE I
Same domain results on the SFA dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
Faria (2018) [26]	-	-	92.88	39.58	55.51
SegNet (2020) [27]	-	-	-	-	88.90
U-Net (2020) [27]	-	-	-	-	84.80
DeepLab (2020) [27]	-	-	-	-	93.90
Our patch-based	91.14	82.17	89.71	91.00	90.35
Our U-Net	**97.94**	**92.80**	**96.65**	**95.89**	**96.27**

TABLE II
Same domain results on the Compaq dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
Branc.(2017) [6]	-	-	43.54	**80.46**	56.50
SegNet (2020) [27]	-	-	-	-	73.70
U-Net (2020) [27]	-	-	-	-	68.60
DeepLab (2020) [27]	-	-	-	-	**81.70**
Our patch-based	90.18	46.00	58.92	73.59	65.45
Our U-Net	**92.62**	**54.47**	**68.49**	71.64	70.03

A. In-domain evaluations

Here we compare our CNN-based local (patch-based) and holistic (U-Net) models in same-domain training situations to previous color-based local models [6], [26], [28] and CNN-based holistic solutions [27]. We evaluate using: Accuracy (Acc), Intersection Over Union (IoU), Precision, Recall and F_1 Score, in order to compare to previous works. Results are shown on tables I, II, III and IV, given the availability of result data in original publications. Our patch-based CNN suparssed all previous color-based models and our holist U-Net achieved state-of-the-art scores compared to others recent CNN models for the datasets in study. This confirms the superiority of the deep learning models over color-based ones and the superiority of holistic over local approaches. The results also show that the datasets have different levels of difficulty, being VPU the most challenging and SFA the least challenging, considering the most used F1 criteria.

B. Cross-domain baseline results

The cross-domain capabilities of our models and generalization power of domains are shown on table V, which presents source only mean F_1 scores results without any transfer or

TABLE III
SAME DOMAIN RESULTS ON THE PRATHEEPAN DATASET (IN %).

Model	Acc	IoU	Prec	Recall	F_1
Branc.(2017) [6]	-	-	55.13	81.99	65.92
Faria (2018) [26]	-	-	66.81	66.83	66.82
SegNet (2020) [27]	-	-	-	-	80.20
U-Net (2020) [27]	-	-	-	-	71.3
DeepLab (2020) [27]	-	-	-	-	**87.50**
Our patch-based	87.12	55.57	59.83	**82.49**	69.36
Our U-Net	**91.75**	**60.43**	**72.91**	74.51	73.70

TABLE IV
SAME DOMAIN RESULTS ON THE VPU DATASET (IN %).

Model	Acc	IoU	Prec	Recall	F_1
SMig.(2013)	-	-	45.60	**73.90**	56.40
SegNet (2020) [27]	-	-	-	-	32.80
U-Net (2020) [27]	-	-	-	-	33.2
DeepLab (2020) [27]	-	-	-	-	62.80
Our patch-based	93.48	14.14	46.34	42.82	44.51
Our U-Net	**99.04**	**45.29**	**57.86**	71.33	**63.90**

TABLE V
CROSS-DOMAIN MEAN F_1 SCORES (%) OBTAINED WITHOUT TRANSFER NOR ADAPTATION.

Model	Source Domain	Target Domain			
		SFA	Compaq	Prathee.	VPU
U-net	SFA	-	18.92	44.98	11.52
	Compaq	**86.14**	-	**75.30**	23.67
	Prathee.	80.66	**63.49**	-	**36.68**
	VPU	14.83	44.71	48.02	-
Patch	SFA	-	54.80	62.92	21.60
	Compaq	71.28	-	72.59	19.94
	Prathee.	80.04	62.68	-	13.74
	VPU	82.63	51.48	58.34	-

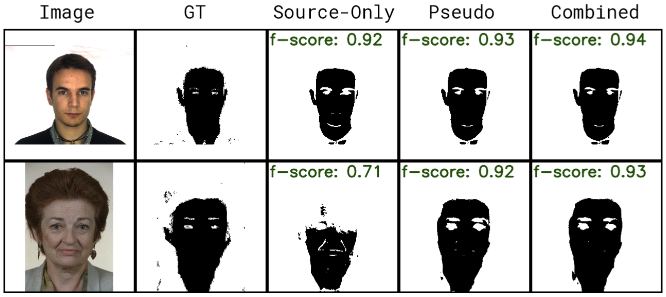

Fig. 5. Domain adaptation from Compaq to SFA using no real labels from target. From left to right: target test image, ground truth and results with source only, domain adaptation based on cross-domain pseudo-labels and the combined domain adaptation + transfer learning approach.

adaptation to target dataset. As we can see, source dataset Compaq in conjunction with the U-Net Model presented the best generalization power on targets SFA and Pratheepan. Source dataset Pratheepan also in conjunction with the U-Net Model did better on targets Compaq and VPU. These source-only setups surpassed the respective color-based approaches shown on previous tables, except for the VPU dataset.

Note that the patch-based model surpassed U-Net when using source domains with low generalization power like SFA and VPU. For example, using VPU as source domain and SFA as target, patch-based reached mean F_1 score of 82.63%, while U-Net only got 14.83%. Using SFA as source and Compaq as target, patch-based also surpassed U-Net (54.80% vs. 18.92%). These results are expected, since SFA and VPU are datasets of very specific domains with little variation in the type of scenes between their images (SFA images are close-ups on faces and VPU images are typical viwes from conference rooms or surveillance cameras). On the other hand, Compaq and Pratheepan include images with a wide range of layouts. Therefore, SFA and VPU only offer relevant information at a patch level for skin detection, their contexts are very specific, which hinders their generalisation ability. If the goal is to design a robust skin detector and avoid negative transfer, our results show that it is better to use Compaq or Prateepan as source samples.

C. Domain Adaptation Results

Following the recommendation in the previous section, we performed domain adaptation experiments using Compaq and Pratheepan as source datasets. Given the superiority of CNN-based holist approaches, for now on, we focus on our U-Net model. Table VI presents the F_1 scores obtained by the methods and settings we evaluated. For each source→target pair, we indicate in bold face which result was better than the target-only method. We evaluated the effect of the amount labeled target samples given and present results ranging from no labels (0%), i.e. an unsupervised domain adaptation setting, to all labels (100%) given in the target training set, i.e., an inductive transfer set up. Target only results are provided for comparison purposes, i.e, within domain experiments with the number of training labels ranging from 5 to 100%. The target only results are expected to be an upper bound in performance when 100% of the training labels are used because there is no domain change, but they may suffer from the reduced training set size in comparison to the domain adaptation settings.

Compaq has confirmed our expectations of being the most generalizable source dataset, not only for being the most numerous in terms of sample images but also due to their diversity in appearance. The use of Compaq as source lead to very good results on SFA and Pratheepan as targets. These results are illustrated in figures 5 and 6, respectively, which show the effects of using different domain adaptation methods

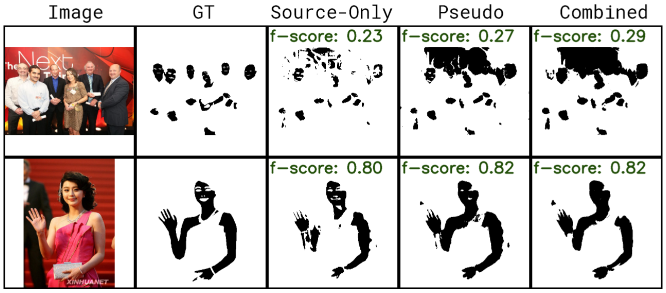

Fig. 6. Domain adaptation from Compaq to Pratheepan using no real labels from target (same setting as Figure 5).

TABLE VI

U-Net mean F_1 scores under different scenarios and domain adaptation approaches.

Source	Target	Approach	Target Training Label Usage				
			0%	5%	10%	50%	100%
Target only	SFA	Target only	-	93.49	94.50	95.72	**96.27**
	Compaq		-	**66.84**	**67.78**	**69.37**	70.03
	Pratheepan		-	46.36	59.86	69.04	73.70
	VPU		-	41.27	53.44	63.18	63.90
Compaq	SFA	Source only	86.14	-	-	-	-
		Fine-tuning only	-	92.89	94.04	**95.86**	95.98
		Cross-domain pseudo-label only	88.80	88.90	89.69	93.22	-
		Combined approach	**89.24**	90.05	90.36	94.57	-
	Pratheepan	Source only	75.30	-	-	-	-
		Fine-tuning only	-	72.52	74.69	76.47	**77.16**
		Cross-domain pseudo-label only	75.58	75.52	77.18	**80.08**	-
		Combined approach	**76.80**	**75.67**	**77.84**	79.87	-
	VPU	Source only	**23.67**	-	-	-	-
		Fine-tuning only	-	**51.51**	46.50	67.47	**69.62**
		Cross-domain pseudo-label only	02.67	02.86	02.68	02.77	-
		Combined approach	02.66	02.68	02.67	02.66	-
Pratheepan	SFA	Source only	80.66	-	-	-	-
		Fine-tuning only	-	**93.68**	**94.70**	**95.69**	95.99
		Cross-domain pseudo-label only	82.50	83.36	83.63	90.60	-
		Combined approach	**82.96**	84.12	84.47	92.93	-
	Compaq	Source only	**63.49**	-	-	-	-
		Fine-tuning only	-	64.88	66.10	68.97	**70.52**
		Cross-domain pseudo-label only	39.50	41.26	44.69	62.39	-
		Combined approach	34.72	36.22	39.05	57.06	-
	VPU	Source only	**36.68**	-	-	-	-
		Fine-tuning only	-	**51.61**	**60.19**	**68.15**	**69.44**
		Cross-domain pseudo-label only	02.66	02.66	02.67	02.77	-
		Combined approach	02.65	02.66	02.67	02.74	-

with no labels from target dataset. Note that when using Compaq as source and Pratheepan as target, the gain of the domain adaptation approaches is very expressive when compared to target only training. Domain adaptation methods got better results using any amount of labels on the target training set, being the combined approach the best option in most cases. Using 50% of training data our cross-domain pseudo-label approach was better than regular supervised training with 100% of training data. Besides that, all the results of domain adaptation methods with no labels were better than the state-of-the-art results of color-based approaches presented in Section IV-A.

When VPU is the target dataset, Pratheepan outperformed Compaq as source dataset. However, the pseudo-labels caused negative transfer, leading to very bad results when domain adaptation was used. The results with fine-tuning were better than regular supervised training with all evaluated amounts of training labels. In this scenario, the reference color-based approach by [28] was beaten starting from 10% of training label usage. Results with 5, 10 and 50% are shown for two sample images in Figure 7.

Still with Pratheepan as source dataset, but with Compaq as target, the "source only" result was reasonable and surpassed the color-based approach. However, we observed that domain adaptation methods did not remarkably improve the results from regular supervised training. Figure 8 shows the results of fine-tuning from Pratheepan to Compaq.

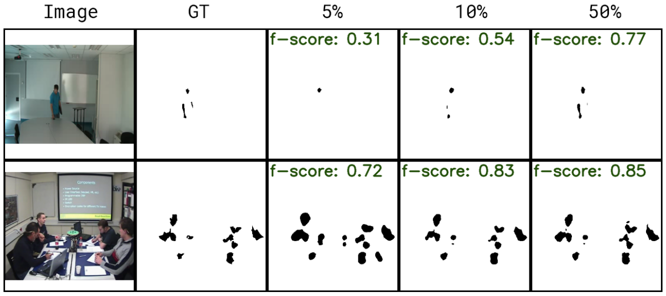

Fig. 7. Adaptation from Pratheepan to VPU with fine-tuning TL. From left to right: target test image, ground truth and resutls with 5, 10 and 50% of labels on the target training set.

D. Discussion

Although most approaches for skin detection in the past have assumed that skin regions are nearly textureless [1], [3], [6], [34], [35], our results give the unintuitive conclusion that texture and context play an important role. A holistic segmentation approach like fully convolutional networks, taking the whole image as input, in conjunction with adequate domain adaptation methods, has more generalization power than local approaches like color and patch-based. The improvement level and best domain adaptation approach varies depending on how close target and source domains are and on the diversity of the samples in the source dataset. The closer the domains and the higher the source variety, the higher the improvement. For example, a very positive transfer from Compaq→SFA

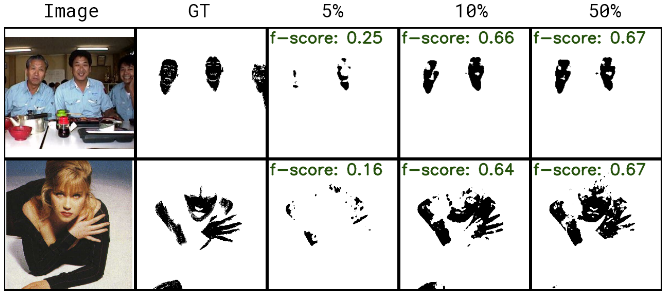

Fig. 8. Adaptation from Pratheepan to Compaq with fine-tuning using different amounts of labels on the target training set (following the same setting as Figure 7).

was observed because Compaq is more diverse and includes samples whose appearance is somewhat similar to those of SFA. This is intuitive, as these approaches depend on the quality of the pseudo-labels. When the transition between domains goes from specific to diverse datasets, the pseudo-labels are expected to be of low quality, thus, not contributing to the target model training. On these situations, fine-tuning has showed to be more effective, although with the drawback of requiring at least some few labeled images for training.

Domain Adaptation methods have also showed improvements when compared to regular supervised training in cases where the target has few images, like Pratheepan and VPU. The level of improvement depends on the amount of labeled target training data and on the similarity of source and target domains. The higher the amount, the lower the improvement, and the higher the similarity, the higher the improvement. Figure 9 shows a comparison of regular supervised training versus the combined approach in the Compaq→Pratheepan scenario with 5, 10 and 50% of the target training samples with labels. This scenario is good for the pseudo-label approach, since Compaq has more diversity than Pratheepan. Note the superiority of combined approach in all levels of target labels availability.

Figure 10, on the other hand, shows the comparison of regular supervised training versus the fine-tune approach in the Pratheepan → VPU scenario. As Pratheepan does not cover scenes that occur on VPU, the fine-tune approach perform better than cross-domain pseudo-labels in this scenario.

Another important aspect to be addressed is the criticism for the applicability of CNN approaches to real-time applications. The criticism is probably valid for patch-based CNN approaches, but it does not hold for our FCN holistic approach. The average prediction time of our patch-based CNN, using a simple Nvidia GTX-1080Ti, with a frame size of 768 × 768 pixels, is 7 seconds per image which is indeed not suitable for real-time applications. However, our U-Net prediction time is 80 ms per frame for the same setup, i.e., 12.5 images are processed per second (without parallel processing). [6] has reported prediction time of about 10ms per frame with frame size of 300 × 400 pixels (8× faster on images that are 5× smaller), which is indeed a bit faster, at a penalty of producing worse results.

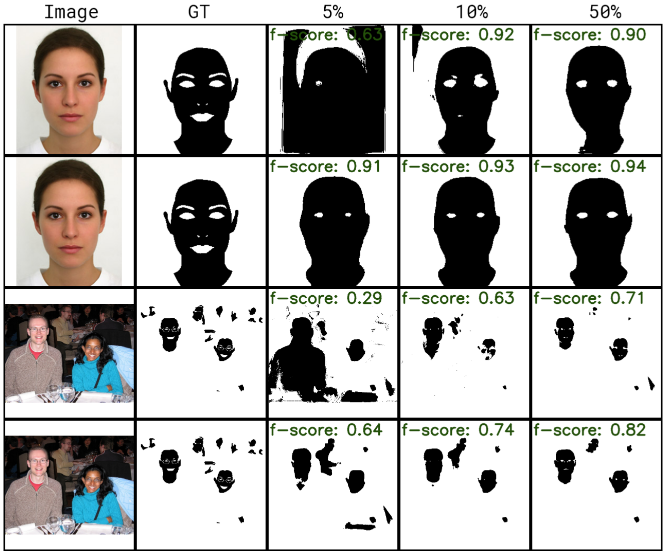

Fig. 9. Comparison of source only vs. domain adaptation combined approach in the Compaq→Pratheepan scenario with different proportions of labeled target training samples. For each target test image, the first row is regular supervised training and the second is the combined domain adaptation approach.

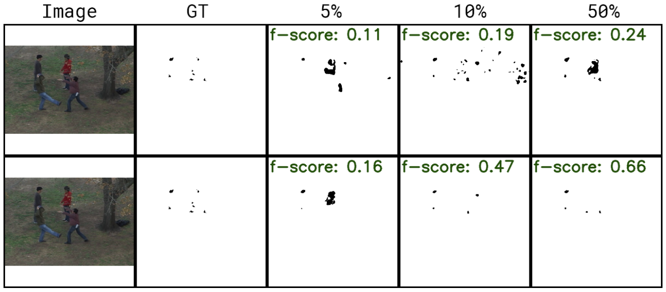

Fig. 10. Comparison of source only vs. fine-tune in the Pratheepan → VPU scenario with different proportions of labeled target training samples. For each target test image, the first row is regular supervised training and the second is the fine-tuning approach.

V. Conclusions

In this work we refuted some common criticisms regarding the use of Deep Convolutional Networks for skin segmentation. We compared two CNN approaches (patch-based and holistic) to the state-of-the-art pixel-based solutions for skin detection in in-domain situations. As our main contribution, we proposed novel approaches for semi-supervised and unsupervised domain adaptation applied to skin segmentation using CNNs and evaluated it with a extensive set of experiments.

Our evaluation of in-domain skin detection approaches on different domains/datasets showed the expected and incontestable superiority of CNN based approaches over color based ones. Our U-Net model obtained F_1 scores which were on average 30% better than the state-of-the-art recent published color based results. In more homogeneous and clean datasets, like SFA, our F_1 score was 73% better. Even in more difficult and heterogeneous datasets, like Prathepaan and VPU, our U-Net CNN was more than 10% better. We experimentally came to the conclusion that a holistic approach like U-net, besides being much faster, gives better results than a patch-based local approach. We also concluded that the common critique of lack of generalization of CNNs does not hold true against our experimental data. With no labeled data on the target domain, our domain adaptation method's F_1 score is an improvement of 60% over color-based results for homogeneous target datasets like SFA and 13% in heterogeneous datasets like Pratheepan.

Acknowledgments

Dr. de Campos acknowledges the support of CNPq fellowship PQ 314154/2018-3.

References

[1] K. B. Shaik, P. Ganesan, V. Kalist, B. Sathish, and J. M. M. Jenitha, "Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space," *Procedia Computer Science*, vol. 57, pp. 41–48, 2015.

[2] M. R. Mahmoodi and S. M. Sayedi, "A comprehensive survey on human skin detection," *International Journal of Image, Graphics & Signal Processing*, vol. 8, no. 5, pp. 1–35, 2016.

[3] Q. Huynh-Thu, M. Meguro, and M. Kaneko, "Skin-Color-Based Image Segmentation and Its Application in Face Detection," in *MVA*, 2002, pp. 48–51.

[4] V. K. Shrivastava, N. D. Londhe, R. S. Sonawane, and J. S. Suri, "Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features," *Comput. Methods Prog. Biomed.*, vol. 126, no. C, pp. 98–109, Apr. 2016.

[5] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106 – 1122, 2007.

[6] N. Brancati, G. D. Pietro, M. Frucci, and L. Gallo, "Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering," *Computer Vision and Image Understanding*, vol. 155, pp. 33 – 42, 2017.

[7] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning (WREPL)*, July 2013, pp. 1–6.

[8] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017, first appeared as a preprint in 2014 at arXiv:1411.4038.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[11] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[12] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," in *Domain Adaptation in Computer Vision Applications*, G. Csurka, Ed. Cham: Springer International Publishing, 2017, pp. 1–35, preprint available at arXiv:1702.05374.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[14] G. Csurka, *Domain adaptation in computer vision applications*. Springer, 2017.

[15] M. Long, J. Wang, G. Ding, and P. Yu, "Transfer learning with joint distribution adaptation," in *Proc 14th Int Conf on Computer Vision, Sydney, Australia*, 2013, pp. 2200–2207.

[16] N. FarajiDavar, T. de Campos, and J. Kittler, "Adaptive transductive transfer machines: A pipeline for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition. Springer International, 2017, pp. 115–132.

[17] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, June 2016, pp. 2414–2423.

[19] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[20] K. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, Massachusetts: MIT press, 2012.

[21] C. Leistner, A. Saffari, J. Santner, and H. Bischof, "Semi-supervised random forests," in *Proc 12th Int Conf on Computer Vision, Kyoto, Japan, Sept 27 - Oct 4*. IEEE, 2009, pp. 506–513.

[22] A. Criminisi and J. Shotton, *Semi-supervised classification forests*. Springer, 2013, ch. 8, pp. 95–107.

[23] M. A. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in *Proceedings of the 25th international conference on Machine learning - ICML*. Helsinki, Finland: ACM Press, 2008, pp. 792–799.

[24] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proceedings of the 25th international conference on Machine learning - ICML*. Helsinki, Finland: ACM Press, 2008, pp. 536–543.

[25] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML. New York, NY, USA: ACM, 2008, pp. 1168–1175.

[26] R. A. D. Faria and R. Hirata Jr., "Combined correlation rules to detect skin based on dynamic color clustering," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, vol. 5, INSTICC. SciTePress, 2018, pp. 309–316.

[27] A. Lumini and L. Nanni, "Fair comparison of skin detection approaches on publicly available datasets," *Expert Systems with Applications*, vol. 160, December 2020, DOI: 10.1016/j.eswa.2020.113677.Preprint available at arXiv:1802.02531.

[28] J. C. San Miguel and S. Suja, "Skin detection by dual maximization of detectors agreement for video monitoring," *Pattern Recognition Letters*, vol. 34, no. 16, pp. 2102 – 2109, 2013.

[29] C. O. Conaire, N. E. O'Connor, and A. F. Smeaton, "Detector adaptation by maximising agreement between independent data sources," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–6.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[31] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *Proc IEEE Conf on Computer Vision and Pattern Recognition, Fort Collins CO, June*, vol. 1, 1999, pp. 274–280 Vol. 1.

[32] J. P. B. Casati, D. R. Moraes, and E. L. L. Rodrigues, "SFA: A Human Skin Image Database based on FERET and AR Facial Images," in *IX Workshop de Visão Computacional*, 2013, p. 5.

[33] P. Yogarajah, J. Condell, K. Curran, A. Cheddad, and P. McKevitt, "A dynamic threshold approach for skin segmentation in color images," in *Proc IEEE Int Conf on Image Processing ICIP, Hong Kong, September 26-29*, Sept 2010, pp. 2225–2228.

[34] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551–564, June 1999.

[35] Son Lam Phung, A. Bouzerdoum, and D. Chai, "A novel skin color model in ycbcr color space and its application to human face detection," in *Proceedings. International Conference on Image Processing*, vol. 1, Sep. 2002, pp. I–I.