S2MGen: A Synthetic Skin Mask Generator for Improving Segmentation

Subhadra Gopalakrishnan[1], Trisha Mittal[1], Jaclyn Pytlarz[1], Yuheng Zhao[2,*]
[*1]Dolby Laboratories Inc. [2]University of Southern California
{subhadra.gopalakrishnan, trisha.mittal, jaclyn.pytlarz}@dolby.com
{yuhengz}@usc.edu

Abstract

Skin segmentation is an important and challenging task which finds use in direct applications such as image editing and indirect downstream tasks such as face detection or hand gesture recognition. However, the availability of diverse and high-quality training data is a major challenge. Annotation of dense segmentation masks is an expensive and time consuming process. Existing skin segmentation datasets are often limited in scope: they include downstream task-specific datasets captured under controlled conditions, with limited variability in lighting, scale, ethnicity, and age. This lack of diversity in the training data can lead to poor generalization and limited performance when applied to real-world images. To address this issue, we propose a tunable generation pipeline, Synthetic Skin Mask Generator (S2MGen), which allows for the creation of a diverse range of body positions, camera angles, and lighting conditions. We explore the impact of these tunable parameters on skin segmentation performance. We also show that improvements can be made to the performance and generalizability of models trained on real world datasets, by the inclusion of synthetic data in the training pipeline.

1. Introduction

Skin Segmentation is the process of separating skin pixels or regions in an image from non-skin pixels (clothes, hair, background). Detection of skin is an important precursor step in other human-centric downstream tasks such as medical image analysis [36], preserving skin pixels in image editing applications, gesture recognition [20], face recognition [32, 61], content moderation [33], etc.

However, segmentation of skin is a challenging problem due to the diversity of input images. Some variability factors include ethnicity, gender, age, clothing, and cosmetics. Skin is affected by factors such as illumination, background colours, camera characteristics, image composition, shadows, and highlights. To reasonably guarantee a consistently acceptable performance over these variations, it is important that the training dataset is extensive and representative of this diversity.

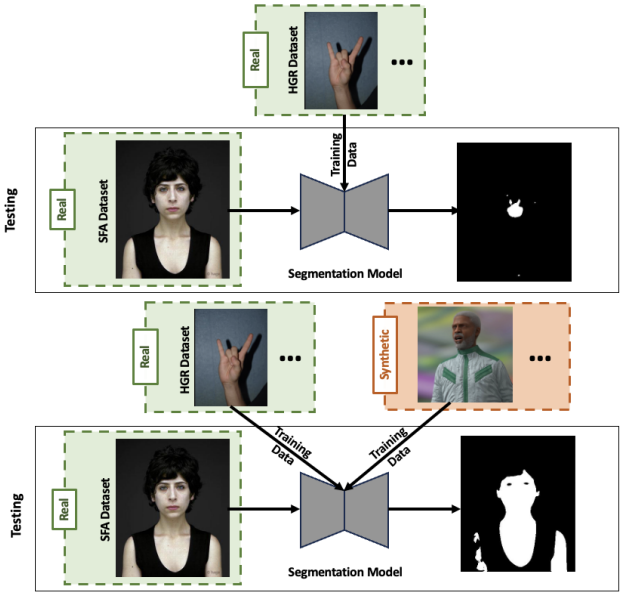


Figure 1. **Using synthetic data for skin segmentation:** We propose S2MGen, a tunable synthetic skin mask generation pipeline and explore the use of the generated data on segmentation performance. On top, we show the performance of the segmentation model trained on a real dataset, HGR [28] and tested on a image from another real dataset, SFA [7]. On bottom, we show the improved performance on the same test image when synthetic data is injected during training.

A tedious step in developing segmentation deep learning models is the collection and annotation of training datasets. This is a time-consuming process, mostly involving manual labelling. State of the art segmentation algorithms have been developed on large scale datasets such as COCO-stuff (164k images) [6], ADE20k (20k images) [81], LIP (50k images) [19] etc. In contrast, most skin segmentation datasets have been constrained in terms of dataset size [65], label quality [26] and data diversity [7, 28]. In our work,

[*]Work done as an intern at Dolby Laboratories Inc.

we look into synthetic data as a source of easy-to-obtain high-quality annotations and analyse its usage in improving cross-dataset performance.

Synthetic data generation for segmentation has been explored as a potential alternative to tedious hand labelling of dense pixel annotations. Datasets generated using 3D modeling software [59, 71] have shown significant progress in improving performance of real training data or replacing real data completely. Additionally, this can also reduce/eliminate privacy issues involved in collecting human-centric data. More recently, foundational models have also been used to generate synthetic data [79], [16] and corresponding segmentation labels [44], however initial experiments for segmenting skin from unlabelled data using foundational segmentation models have not yielded satisfactory results (Appendix, Figure 12). An often overlooked aspect of synthetic training data generation is the effect of different generation parameters on model performance. To the best of our knowledge, we are the first work to explore synthetic data generation along with tunable parameters specifically for skin segmentation.

One of the more challenging aspects of realistic human rendering is producing high quality skin. Sub-surface scattering to represent skin translucency [74], modeling skin reflectance [62, 69], representing skin textures such as pores, wrinkles, scars etc. are some of the common challenges in realistic skin rendering. Recently, high-quality character generation pipelines such as MetaHuman Creator [70] by Unreal Engine, Daz3d [12] and Human Generator [4] in Blender [5] have been proposed. This opens up the possibility of fast, large-scale, procedural rendering of high-quality data for human-centric deep learning applications.

Main Contributions: The following are the novel contributions of our work.

1. We propose the S2MGen pipeline, a high-quality synthetic data generator for skin segmentation with various tunable parameters (such as clothing, camera angles, backgrounds).
2. We present analysis on the effects of the tunable parameters on performance.
3. We demonstrate the effectiveness of pre-training with synthetic data to enhance performance in situations with limited real-world data and across different domains.

2. Related work

In this section, we present some background in related domains. In Section 2.1 and Section 2.2, we introduce prior approaches in skin segmentation. We then motivate the importance and use of augmenting synthetic data with real data in training deep learning models (Section 2.3) and also prior work in closing the gap between real and synthetic data using various domain adaptation techniques (Section 2.4).

2.1. Skin Segmentation Approaches

Traditionally, segmentation of skin areas in an image has been tackled using color dependent approaches such as explicit thresholding in various color spaces [8, 11, 63], representing skin color distributions using gaussian mixture models [26, 55, 77] etc. Phung *et al*. [47] introduced neural networks to skin segmentation by classifying skin vs. non skin pixels using a Multi-Layer Perceptron architecture on the chrominance channels.

Recent advancements in skin segmentation have been achieved using Fully Convolutional Networks (FCNs) [38] which allow for better modelling of the complexity of real world scenes. Tarasiewicz *et al*. [66] propose a lightweight U-Net [58] modified to learn a larger receptive field for skin segmentation. Yi *et al*. [23] utilize body segmentation labels in a dual-task semi-supervised learning setting to generate robust skin masks. As mentioned in the previous section, current available skin datasets are mostly small and heavily biased. Dourado *et al.* [14] study the effect of domain adaptation between these different datasets. More information related to skin segmentation algorithms can be found in [27, 39, 43].

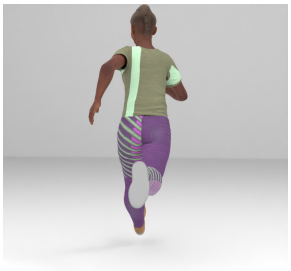
2.2. Skin Segmentation Datasets

Multiple datasets have been been proposed for the task of skin segmentation in prior literature. However, as covered in Section 1, many of these datasets have been developed for very particular downstream tasks such as face detection, hand gesture recognition etc., resulting in very constrained datasets. For example, the SFA dataset [7] consists of 1118 images of close-up shots of faces. Similarly, the HGR dataset [28] consists of 899 images of hand gesture images. The Abdomen dataset [67] is a collection of 1200 exposed torso images curated with skin segmentation labels to aid in robotic abdominal surgeries. Another issue prevalent in larger, more general skin datasets is imprecise labelling quality. For example, the COMPAQ [26] (4,670 images) and the VisUUAL (46,775 images) Dataset [22] provide labels generated by automatic tools or other algorithms resulting in noisy ground truth. Another pain point is dataset size, with datasets such as Pratheepan [65] containing only 78 images, mostly used in algorithm evaluation. We maintain the ECU dataset [46] (4000 images) as our primary baseline because of the relatively larger size and better quality annotations. For our work, we focus on publicly available datasets with precise annotations (ECU, HGR, Abdomen, SFA and Pratheepan).

2.3. Synthetic data for deep learning

A previously explored solution to augment limited annotated real-world data, is to generate and use synthetic data (datasets like the SYNTHIA dataset [59]); Synthetic data has shown major successes in applications like semantic segmentation [10], crowd counting [73] and depth esti-

(a) Face Render (b) Portrait Render



(c) Full body Render

Figure 2. **Qualitative Samples of Diverse Aspects of the S2MGen Pipeline:** In 2a, we show the shortest focal length lens, focused on faces. 2b has a portrait focal length lens, showcasing colored hair, random facial expression, and complex background. Figure 2c is a full body focal length framing with a plain background and demonstrating the random poses and clothing patterns.

mation [35] to mention a few. More specifically, there has been a lot of focus on the generation of human synthetic data to help in applications like 2D and 3D human pose estimation [51, 57, 71, 82], action recognition [53, 54, 56], pedestrian detection [40,48,49], and, face detection [30,31]. Some large-scale human synthetic datasets include SURREAL [71], face analysis dataset [75] and SynFace [50].

2.4. Closing the Domain Gap b/w Datasets

While applications in various domains have long enjoyed the benefits of synthesizing training data with graphics, the domain gap between real and synthetic data has remained a problem, especially for human faces. To address this issue, prior work has tried to bridge this gap with domain adaptation [3, 83], and knowledge transfer [13, 25] techniques. Among supervised techniques, the simplest is transfer learning, where models are pretrained on larger synthetic sets and fine-tuned on the limited real datasets available. The pre-training reduces the amount of real data required. Another approach, Balanced Gradient Contribution (BGC) [60], explored by the authors of the SYNTHA Dataset [59] as a method of domain adaptation, uses gradient updates with controlled perturbations from a noisy domain. A major chunk of the unsupervised domain adaptation methods are based on feature alignment, for example, by adversarial training [24, 68, 72, 80], minimizing MMD (Maximum Mean Discrepancy) [3] etc. Some approaches use self-training; one such work by Zou *et al.* [83] uses pseudo-labeling to label the target set before re-training the model on these labels.

3. S2MGen Pipeline

In this section, we describe the dataset generation process to create synthetic humans and corresponding masks for body and skin. For the dataset generation pipeline, we heavily utilize the 3D modeling software Blender, specifically the Human Generator add-on, to generate the synthetic humans.

3.1. Human Generator

The Human Generator Blender add-on (version 3) is used to create the basic human structure. It includes 52 poses, random facial expressions, random gender, and 18 variations on starting humans. On top of the baseline structure, we override the random generator with new ranges for features such as wrinkles, body types, facial hair, freckles etc. Hair and eyebrow color is randomly set with a distribution biased towards natural hair colors. Some examples of renders are shown in Figure 2.

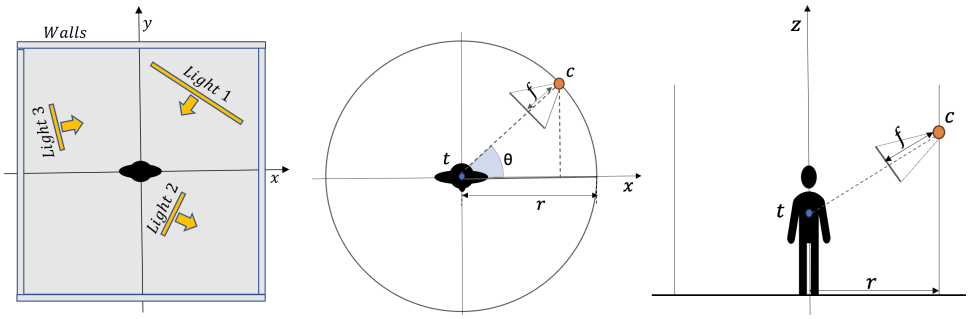
For dataset generation speed, the same human model is maintained for multiple camera angles - changing poses, colors, and clothing in between. This reduces the processing time to around 2 frames per second. The presence of clothing is represented by a boolean random variable (RV) $P_{clothed}$. The probability of adding clothing is a tuneable parameter and is given by $p_{clothed}$: $P_{clothed} = X < p_{clothed}$, where $X \sim \mathcal{U}(0, 1)$

For renders that include clothing, outfits in the Human Generator baseline set are used. Patterns and colors on the clothing are randomly adjusted to increase variability and diversity.

3.2. Virtual Environment

The virtual environment consists of a bounding room, background images, and adjustable lighting. Four walls, a ceiling, and floor surround the human model. The size of the room is parameterized. Since the ambient lighting in the room depends heavily on reflected light off the walls, changing the distance between the subject and the walls results in more diverse subject lighting.

To further improve realism, photographs are added to the walls in the synthetic environment as shown in Figure 2b. Around 1.5k images were scraped from Pexels [45] to use as backgrounds for the rendered images. These backgrounds were manually cross-checked to make sure they did not contain any people or skin patches. The content of the background images is extremely varied in colors and textures and specifically contains skin-like content such as wood and sand, which serves as an important source of hard negative samples for model training. The presence of



(a) Environment and lighting setup (b) Camera position top-down view (c) Camera position front view

Figure 3. **Camera, Subject, and, Lighting Locations for the Synthetic Data Generation Pipeline:** In 3a, we depict the virtual environment and the lighting setup, in 3b and 3c we show the camera position from top-down and front view.

background images in the walls is represented by a boolean RV $P_{background}$. The probability of choosing a background from this dataset versus leaving the wall blank is determined by $p_{background} : P_{background} = x < p_{packground}$, where $x \sim \mathcal{U}(0, 1)$

Although parameterizable, this particular pipeline consists of three lighting locations. All three area lights are placed inside the room: one large light for fill, and two small lights for depth and dimension (Figure 3a). The locations of the lights are randomly set along a uniform distribution in x, y, and z. Rotation of the smaller lights is randomized uniformly. However, the largest light is tracked to maintain angle towards the subject. This ensures a baseline ambient lighting for a properly exposed capture. The energy of each light is uniformly sampled between 50 and 1000.

3.3. Camera Position and Direction

The position and the orientation of the camera in the Blender 3D space can be represented as the Euclidean coordinates $c = (x_{cam}, y_{cam}, z_{cam})$ and the Euler rotation angles respectively. The orientation is fixed by allowing the camera to always track a point t on the synthetic human.

For a fixed z_{cam}, we allow x_{cam} and y_{cam} to lie on a circle, represented in polar coordinates, parameterized by the radius r and angle θ, as $x_{cam} = r \cdot \cos(\theta)$ and $y_{cam} = r \cdot \sin(\theta)$.

The generated synthetic human is centered at the origin. For this dataset, the radius r remains constant. However, the focal length of the camera, f, is allowed to change. The height z_{cam}, is uniformly sampled as: $z_{cam} \sim \mathcal{U}(z_{min}, z_{max})$. Hence, the camera position c, is limited to a cylindrical surface around the generated synthetic human as shown in Figure 3b and Figure 3c.

For camera angle, θ could be sampled from a uniform distribution Θ over a range of values ($\Theta \sim \mathcal{U}(0, 2\pi)$). However, to more accurately model typical photography composition, the camera angles are constrained based on the distribution of real world content. Most photography of humans includes a semi-front facing subject. Therefore, instead of evenly sampling the camera location along this cylindrical surface, the camera location is biased to face the front of the generated synthetic human by sampling θ from a transformed RV Θ, as described below. At $\theta = \frac{3\pi}{2}$ radians, the camera is directly in front of the generated human.

$$\Theta = \frac{3\pi}{2} + \pi D X^{p_{front}} \tag{1}$$

where, $X \sim \mathcal{U}(0, 1)$ and $D \sim \mathcal{U}(\{-1, 1\})$.

p_{front} is a tuneable parameter that controls the angle variance away from front-facing images. The RV D controls the direction from this center, sampling either clockwise or anti-clockwise from $\frac{3\pi}{2}$. As p_{front} increases, the likelihood of front-facing images increases.

The camera orientation is constrained to focus on the synthetic human at a particular focal point t. This point could be parameterized to move vertically in a randomized fashion, to focus on different parts of the generated human. However, for the purposes of this dataset, where mimicking typical photography composition is desired, a fixed focal point on the upper body is maintained.

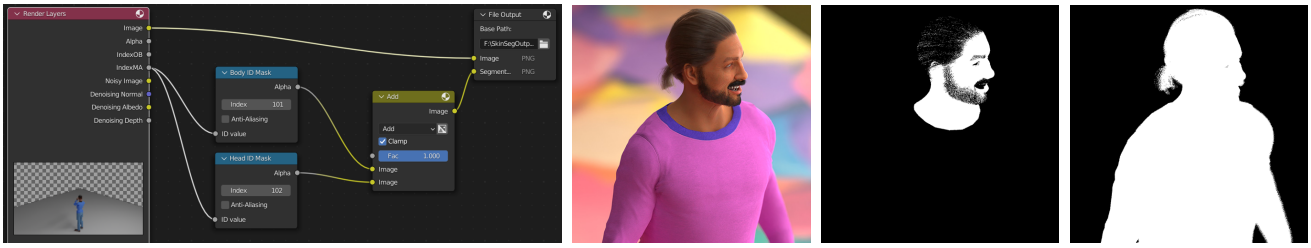
The focal length f is sampled within a minimum (f_{min}) and maximum (f_{max}) focal length. To bias the sampled focal length for a higher probability of close ups vs. farther away shots, we use:

$$F = (f_{max} - f_{min}) X^{1/p_{focal}} + f_{min} \tag{2}$$

At $p_{focal} = 1$, we regress to the uniform distribution $\mathcal{U}(f_{min}, f_{max})$. As p_{focal} increases, the likelihood of zoomed in images increases. Plots showing the effect of p_{focal} and p_{front} are included in the appendix.

3.4. Image and Mask Rendering

The dataset images are rendered using Blender's still image rendering pipeline with the Cycles render engine [1]. For the core rendering algorithms, the python Blender module, *bpy* [2], is used. Material properties are set as tags to render the segmentation masks. Each of the human textures



(a) Blender node structure for skin mask (b) Image render (c) Skin mask (d) Person mask

Figure 4. **Mask Rendering Pipeline of S2MGen:** In 4a we show the Blender node structure for creating the skin mask, and in 4b, 4c, and 4d we show an example of a synthetic rendered image along with the corresponding skin and person segmentation masks.

is labeled with a pass-index ID. Then, using the Blender node structure and a multi-rendering pathway, each index is converted into an alpha mask and added together. An example of the skin mask node structure is shown in Figure 4a and the image render in Figure 4b. The binary renders generated from this pipeline are the masks used for skin segmentation training, as seen in Figure 4c and 4d. The same Blender node workflow shown for skin mask generation is used for person segmentation masks, with the exception of increasing layers of *addition* nodes to compensate for the limited two-input architecture of Blender's nodes.

4. Experiments

In this section, we experiment with the parameterizable features of the S2MGen pipeline. We explore its performance on the skin segmentation task and use the ECU [7], HGR [28], Abdomen [67], SFA [7] and Pratheepan [65] datasets for the following experiments. For the ECU and the Abdomen dataset, we use the existing train-test splits. For the HGR, Abdomen, SFA and Pratheepan dataset we split them into 80% training and 20% evaluation.

Model Architecture Details: The baseline skin segmentation model is a U-Net [58] architecture with skip-connections, slightly modified to have two segmentation heads, one for person segmentation and one for skin segmentation. As previously explored in prior works [21, 23], utilizing additional semantic guidance, such as person segmentation masks, can help in boosting performance of skin detection algorithms. Hence, for training on the synthetic dataset, we use a multi-task learning setup with person segmentation as the auxiliary task. While training on real datasets, we do not backpropogate losses through the person segmentation pathway.

The model is trained for 20 epochs using an image patch size of 256x256. We use the Adam optimizer with a constant learning rate of 10^{-5} on an Nvidia A100 GPU with 40GB RAM. For evaluation of the test performance we use the mean Intersection over Union (mIoU). We use the Cross Entropy loss with inverse frequency weighting for class balancing as described in Minhas et al. [42].

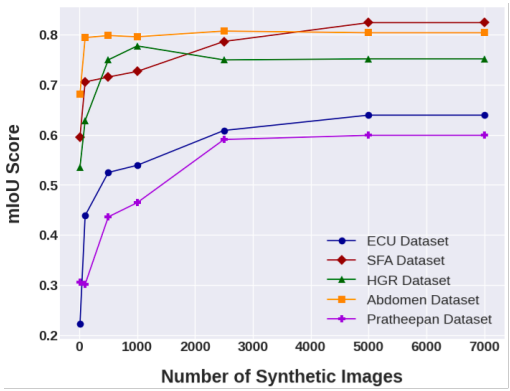
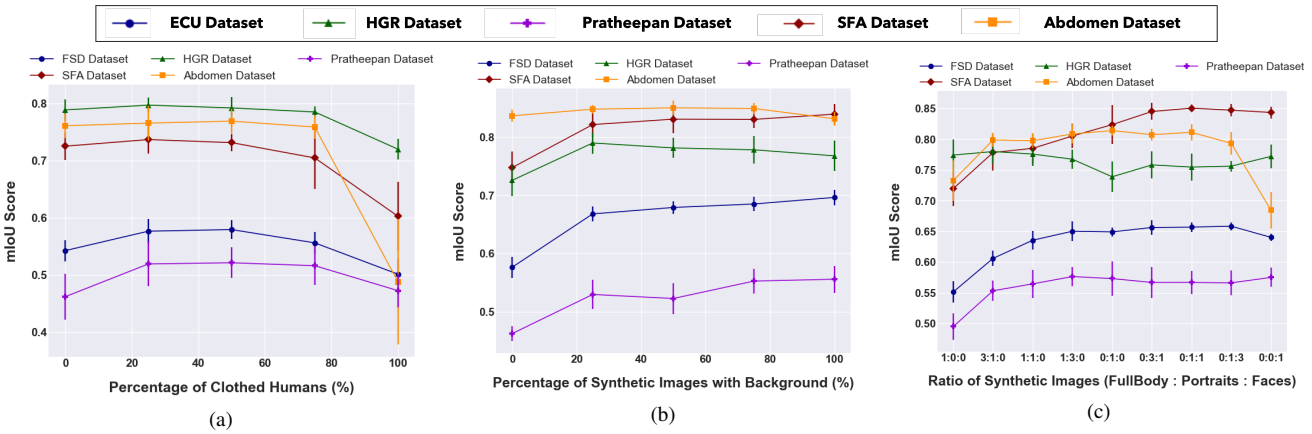


Figure 5. **Effect of Synthetic Dataset Size on Skin Segmentation:** We analyze the impact of adding more synthetic images for training the skin segmentation model and testing the performance on the real-world skin segmentation datasets.

4.1. Effect of the Synthetic Dataset Size

We vary the number of training samples generated from S2MGen to study the effects of dataset size on performance on real world dataset. We vary the synthetic training samples from 25 images to 7500 images. We present our findings in Figure 5.

We observe a general trend of performance increase, quickly converging to a gradual saturation as the number of synthetic training samples increases. This shows a promising correlation between the information learned from synthetic data and the performance on real datasets. Saturation on all datasets occurs around 5000 samples. The Abdomen and the HGR datasets, however, plateau almost immediately (2000 samples). This is likely because these two datasets are constrained in terms of body parts (torso and hands respectively), allowing the model to converge quickly. In contrast, although the SFA dataset is also constrained to a single body part (faces), it benefits from more training samples. This could be attributed to a larger diversity of features on the face (eyes, nose, facial hair, lips etc.).



(a) (b) (c)

Figure 6. **Evaluating Differently Tuned Parameters of S2MGen:** We generate multiple versions of synthetic datasets from S2MGen and to train and evaluate model performance.

4.2. Evaluating S2MGen's Tunable Parameters

To understand the impact of various tunable parameters in the proposed synthetic data generation pipeline, we generate multiple versions of our synthetic dataset and evaluate the corresponding model performance. We explore clothing, backgrounds, and focal length/framing. Our findings are shown in Figure 6. We believe that such analysis can be used to tailor a synthetic dataset based on the requirements of the downstream task. To account for SGD instability due to random seed initialization [64], we train each condition with 10 different random seeds and plot the mean and variance. The exact parameters used in the data generation pipeline and experiment details are in the Appendix.

In this first tuning experiment, we vary the percentage of unclothed and clothed synthetic humans in our training samples ($p_{clothed}$). We observe that ECU and Pratheepan datasets have optimal performance around 50% clothing ratio. Although these datasets do not possess any unclothed humans, the lack of clothing diversity in our synthetic datasets forces the model to learn stronger representations from the unclothed synthetic humans. We see optimal performance at around 25% for the HGR and SFA datasets. This could be attributed to a lower amount of clothing to skin ratio in these datasets. We see a huge drop in performance and model stability for the Abdomen dataset when the synthetic dataset contains fully clothed humans. This could be explained by the fact all the images in the Abdomen dataset contain visible unclothes torso, that is basically unrepresented at this clothing ratio in the synthetic humans.

In the second experiment, we vary the percentage of images with a plain (Figure 2c) vs. a complex (Figure 2b) background. As can be seen in Figure 6b, an increase in background probability generally increases performance, especially in the ECU, Pratheepan and SFA datasets. However, we don't observe such trends for HGR and Abdomen dataset. This could be attributed to pre-dominantly plain background in these datasets.

In the final tuning experiment, we study the effects of varying focal length and framing between facial, portraits, and full body framed shots on the performance of skin segmentation (Figure 2). The images in the full body dataset are generated with a long focal length lens and have a high probability of seeing a full body than the rest. The portraits dataset is generated with a medium focal length lens. These images are roughly focused on the upper torso of the generated human. The faces dataset is generated close up, with short focal length settings and the camera tracking the face area. The exact details of f_{min}, f_{max} and p_{focal} used for generation of these datasets are included in the appendix. The results are shown in Figure 6c where the ratio of images sourced from the three datasets are varied. As expected, we see performance increase on the SFA dataset with more samples of portraits/faces than full body. As observed in 5, performance on the Abdomen dataset converges very quickly with a low number of training samples, which might explain why performance only drops at close to 100% face probability. Interestingly, the performance of the HGR dataset is not affected even at a very high probability of sampling from just the faces dataset. This phenomenon is also noticed in real world datasets as shown in Table 1, where a network trained just on the SFA dataset performs well on the HGR dataset. The trends of the ECU and Pratheepan dataset are alike because the test images of these datasets are quite similar in nature.

4.3. Cross-Dataset performance

An important aspect of model performance is generalizability to a diverse set of images. To that effect, we perform cross-dataset analysis, where we train on one real world dataset, and perform inference on other real world datasets. Furthermore, to analyze the impact of synthetic

Training Dataset	ECU	SFA	HGR	Abdomen	Pratheepan
ECU	0.7759	0.8393	0.8838	0.7924	0.6014
ECU + S2MGen	**0.7928**	**0.8679**	**0.8826**	**0.8509**	**0.6890**
SFA	0.6100	0.8762	**0.7679**	0.7626	0.4534
SFA + S2MGen	**0.6673**	**0.8955**	0.7505	**0.8339**	**0.6054**
HGR	0.5826	0.1682	0.8734	0.6767	0.4510
HGR+S2MGen	**0.7347**	**0.8207**	**0.9103**	**0.7347**	**0.6692**
Pratheepan	0.5755	0.7300	0.7461	0.7544	0.4352
Pratheepan+S2MGen	**0.6789**	**0.8249**	**0.7700**	**0.7583**	**0.6301**
Abdomen	0.5709	0.7505	0.6480	0.8906	0.4462
Abdomen+S2MGen	**0.7012**	**0.8488**	**0.7015**	**0.9061**	**0.6246**
S2MGen only	0.6720	0.8463	0.7672	0.8138	0.5493

Table 1. **Cross Dataset Skin Segmentation Analysis:** We perform cross-dataset analysis, where we train on one real world dataset, and perform inference on other real world datasets. We also experiment with pretraining on a larger synthetic dataset and using the smaller real world dataset to finetune the model.

data on cross-dataset performance, we pre-train on a larger number of synthetic samples generated from S2MGen and use the smaller real world dataset to finetune the model.

We summarize our results in Table 1. To begin, even without using the synthetic datasets, we notice that for the task of skin segmentation, models trained on one dataset can generalize across other datasets reasonably well. This can be attributed to models heavily relying on color information for predicting skin masks. While skin luminance varies substantially, skin hue and saturation is limited to a much smaller spectrum of the color space. However, there is a still a drop in cross-domain performance likely due to a domain gap among datasets. We observe that by finetuning a model pretrained on synthetic dataset, we are able to close this performance gap significantly.

4.4. Qualitative Analysis

In Figure 11, we show qualitative results for skin segmentation with and without finetuning with the Pratheepan dataset. We see reasonable performance training with only synthetic data. When trained on the Pratheepan dataset alone, the segmentation performs poorly. However, we see considerable improvements when pre-training on synthetic data and finetuning with real data. Qualitative results for the other datasets are covered in the Appendix.

4.5. Real to Synthetic Domain Gap

We experiment with Supervised Domain Adaptation (SDA) and Unsupervised Domain Adaptation (UDA) to bridge the gap between real and synthetic data. We summarize these results in Table 2. We report source (S2MGen) only and target (ECU) only performance values. The considerable difference is likely attributed to domain gap. Hence, we experiment with multiple domain adaptation techniques as explained below.

For Supervised Domain Adaptation, we explore two

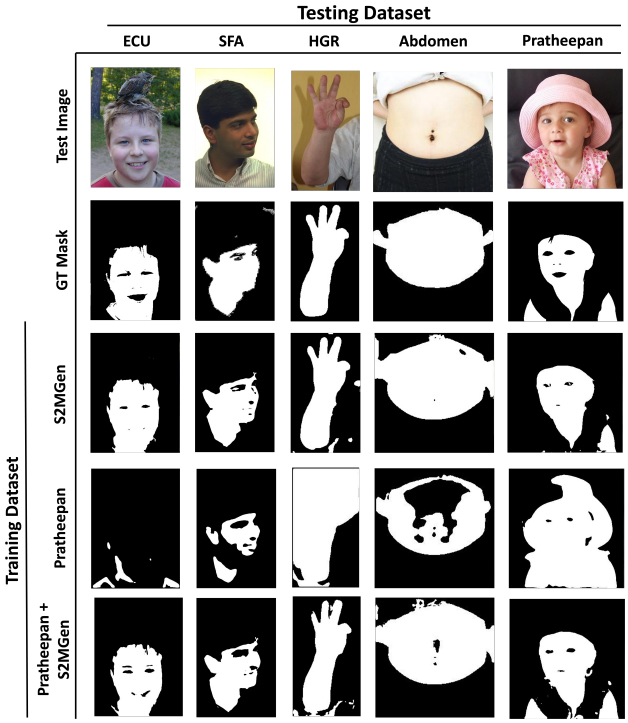


Figure 7. **Qualitative Examples:** We show some examples of performance gain we observe when doing cross-dataset and in-dataset inference on the Pratheepan dataset with and w/o pretrained model on synthetic data generated from S2MGen.

	Approach	IoU	Acc	F1-score
Source only (S2MGen Dataset)	Only Skin Mask	0.6341	0.9123	0.7582
	Skin + Person Mask	0.6720	0.9182	0.7886
UDA	DANN [17]	0.5279	0.8473	0.6700
	PixMatch [41]	0.7048	0.9297	0.8139
	FDA [78]	0.5984	0.8959	0.7299
SDA	Finetuning	**0.7955**	**0.9534**	**0.8773**
	BGC [60]	0.7856	0.9485	0.8689
Target only (ECU Dataset)		0.7759	0.9461	0.8616

Table 2. **Closing the Domain Gap b/w Synthetic and Real Dataset:** We experiment with Supervised Domain Adaptation (SDA) and Unsupervised Domain Adaptation (UDA) to bridge the gap between real and synthetic dataset.

methods: finetuning on the target domain and Balanced Gradient Contribution (BGC) [60]. We also explore Unsupervised Domain Adaptation, specifically with three methods, FDA [78], DANN [17] and PixMatch [41].

We notice comparable results between finetuning vs. BGC, with finetuning performing slightly better. From the UDA results, we see that self-training/pseudolabel based domain adaptation (pixMatch) performs better on our dataset than alignment in high-level (DANN) and low-level feature spaces (FDA).

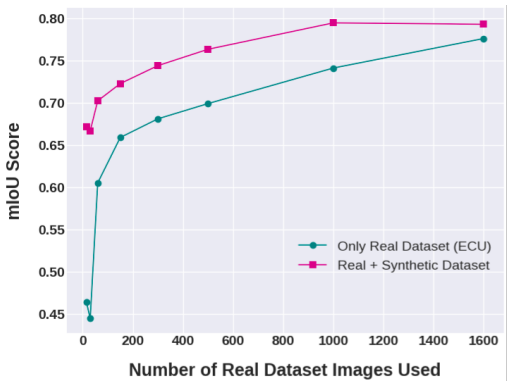


Figure 8. **Effect of Constrained Real Dataset:** We analyze the impact of limiting the real dataset used for training with (blue) and without (pink) the synthetic data pretraining.

Training Dataset	Skintone					
	1	**2**	**3**	**4**	**5**	**6**
Only Real	0.7712	0.7788	0.7735	0.7845	0.8050	0.6522
Real + S2MGen (finetuning)	0.7861	0.7912	0.7936	0.8038	0.8419	0.7681
Real + S2MGen (BGC)	0.7694	0.7845	0.7810	0.7927	0.8130	0.6933

Table 3. **Effect of synthetic dataset on skintones:** We demonstrate the effects of mitigating bias in skin tones by integrating our synthetic data, leading to a more balanced representation in a previously skewed dataset.

We observed that the domain adaptation methods that perform well in reducing the domain gap in datasets like SYNTHIA [59] don't necessarily translate well to our dataset. This could be attributed to the higher complexity of real world scenes for the skin segmentation task.

4.6. Constrained Dataset Size

In this section, we analyse how effectively the synthetic data pretraining can improve performance of very small amounts of real data. We vary the number of real images from the ECU dataset used for finetuning from 15 to 1600. The pretrained model is the same from Section 4.5. These results are shown in Figure 8. We notice the largest performance gains at the smallest real dataset size with the gain slowly decreasing as real dataset size increases. For example, we obtain similar performance at training on 1000 real world images from scratch vs. finetuning on only 300 real images - reducing the required dataset size by atleast 3x.

4.7. Limited Skin Tone Diversity

Prior work [76] has established the presence of bias in the ECU [46] dataset that manifests itself as subpar performance of images containing people with skintone type 6 in the Fitzpatrick [15] scale. In this section, we observe the effects of synthetic data injection on this observed bias. We experiment with a finetuning setup similar to the above sections and Balanced Gradient Contribution (BGC). The results are shown in the Table 3. We observe that adding synthetic data either by finetuning or by BGC, reduces the performance gap between skintone 6 with the rest of the skintones, as a result, mitigating bias on that skintone.

5. Conclusion, Limitations, and Future Scope

In this work, we presented a tunable pipeline, S2MGen, for the procedural generation of synthetic humans and skin segmentation masks. We investigated the impact of synthetic data on skin segmentation performance.

We found the optimal number of synthetic images that allowed for performance convergence across multiple datasets. We then analyzed the effect that different tunable parameters (background, clothing and camera angles) have on the performance on real-world datasets. Each real-world dataset has notably different trends due to their unique tasks (focusing solely on torso, gesture recognition, etc). For maximum generalizability, we recommend tunable parameter values that perform well across all datasets. We conducted experiments with multiple supervised and unsupervised domain adaptation methods, and found finetuning and pseudolabeling to perform better than other methods.

Through experimentation we also conclude that incorporating these large amounts of synthetic datasets in the training pipeline for skin segmentation models can help with better performance and also create generalizable models (increased cross-dataset performance).

However, our synthetic data generation pipeline, S2MGen, still has limitations that we continue to explore. The generated humans are not yet diverse with regards to age, clothing variety, tattoos and accessories (like jewelry), and skin related blemishes (acne, freckles and pigmentation). A lot can be done for achieving more realism in the synthetically generated images; multiple humans, more enriched facial expressions, and incorporating MoCap data for improving pose variety.

While in this work, we focused on skin segmentation, we can also use the S2MGen pipeline to create high-quality labels for tasks like person + part segmentation, face detection, pose estimation, and action recognition with minimal effort. This will enhance the variety of existing datasets for these tasks, and we can also improve performance through multi-task learning. For eg., training the model for both person and skin masks, resulting in a 6% improvement in skin segmentation. This flexibility would allow us to study task correlations and their mutual benefits. We also see potential in the use of S2MGen for balancing bias in real-world skin datasets.

Ethical Considerations: In accordance with ethical considerations and in compliance with data usage policies, this research exclusively employs experimentation with publicly available datasets.

References

[1] Cycles render engine. https://www.cycles-renderer.org/. Accessed: 2023-11-13. 4

[2] Python blender module. https://pypi.org/project/bpy/. Accessed: 2023-11-13. 4

[3] Mahsa Baktashmotlagh, Mehrtash Har, i, and Mathieu Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17(108):1–30, 2016. 3

[4] Blender. Humgen3d. https://www.humgen3d.com. Accessed: 2023-11-13. 2

[5] Blender Foundation. Blender. https://www.blender.org. Accessed: 2023-11-13. 2

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context, 2018. 1

[7] João Paulo Brognoni Casati, Diego Rafael Moraes, and Evandro Luís Linhari Rodrigues. Sfa: A human skin image database based on feret and ar facial images, Jan 1970. 1, 2, 5

[8] D. Chai and K.N. Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564, 1999. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 14

[10] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019. 2

[11] Ying Dai and Yasuaki Nakano. Face-texture model based on sgld and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017, 1996. 2

[12] Daz 3D. Daz3d. https://www.daz3d.com. Accessed: 2023-11-13. 2

[13] Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–52, 2018. 3

[14] Aloisio Dourado, Frederico Guth, Teofilo Emidio de Campos, and Li Weigang. Domain adaptation for holistic skin detection, 2020. 2

[15] Thomas B. Fitzpatrick. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology*, 124(6):869–871, 06 1988. 8

[16] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. 2

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016. 7, 14

[18] Github. Langsam. https://github.com/luca-medeiros/lang-segment-anything. Accessed: 2024-03-29. 15

[19] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing, 2017. 1

[20] Junwei Han, G.M. Award, A. Sutherland, and Hai Wu. Automatic skin segmentation for gesture recognition combining region and support vector machine active learning. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 237–242, 2006. 1

[21] Kooshan Hashemifard, Pau Climent-Perez, and Francisco Florez-Revuelta. Weakly supervised human skin segmentation using guidance attention mechanisms, 2023. 5

[22] Kooshan Hashemifard and Francisco Florez-Revuelta. From garment to skin: The visuaal skin segmentation dataset. In Pier Luigi Mazzeo, Emanuele Frontoni, Stan Sclaroff, and Cosimo Distante, editors, *Image Analysis and Processing. ICIAP 2022 Workshops*, pages 59–70, Cham, 2022. Springer International Publishing. 2

[23] Yi He, Jiayuan Shi, Chuan Wang, Haibin Huang, Jiaming Liu, Guanbin Li, Risheng Liu, and Jue Wang. Semi-supervised skin detection by network with mutual guidance, 2019. 2, 5

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[25] Taotao Jing, Haifeng Xia, and Zhengming Ding. Adaptively-accumulated knowledge transfer for partial domain adaptation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1606–1614, 2020. 3

[26] M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 274–280 Vol. 1, 1999. 1, 2

[27] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007. 2

[28] Michał Kawulok. 1, 2, 5

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 15

[30] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[31] Adam Kortylewski, Andreas Schneider, Thomas Gerig, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891*, 2018. 3

[32] Prem Kuchi, Prasad Gabbur, P. Bhat, Sumam David, and S. Smieee. Human face detection and tracking using skin color modeling and connected component operators. *IETE Journal of Research*, 48, 03 2003. 1

[33] Yung-Ming Kuo, Jiann-Shu Lee, and Pau-Choo Chung. The nude image identification with adaptive skin chromatic distribution matching scheme. In *2010 2nd International Conference on Computer Engineering and Technology*, volume 7, pages V7–117–V7–120, 2010. 1

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, Mar. 2020. 15

[35] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1738–1764, 2020. 3

[36] James Ren Hou Lee, Maya Pavlova, Mahmoud Famouri, and Alexander Wong. Cancer-net sca: tailored deep neural network designs for detection of skin cancer from dermoscopy images. *BMC Medical Imaging*, 22(1):1–12, 2022. 1

[37] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2022. 15

[38] Chang-Hsian Ma and Huang-chia Shih. Human skin segmentation using fully convolutional neural networks. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 168–170, 2018. 2

[39] Mohammad R. Mahmoodi and Masoud S. Sayed. A comprehensive survey on human skin detection. *International Journal of Image, Graphics and Signal Processing*, 8(5):1, 05 2016. Copyright - Copyright Modern Education and Computer Science Press May 2016; Last updated - 2017-04-15. 2

[40] Javier Marin, David Vázquez, David Gerónimo, and Antonio M López. Learning appearance in virtual scenarios for pedestrian detection. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 137–144. IEEE, 2010. 3

[41] Luke Melas-Kyriazi and Arjun K. Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training, 2021. 7, 14

[42] Komal Minhas, Tariq M. Khan, Muhammad Arsalan, Syed Saud Naqvi, Mansoor Ahmed, Haroon Ahmed Khan, Muhammad Adnan Haider, and Abdul Haseeb. Accurate pixel-wise skin segmentation using shallow fully convolutional neural network. *IEEE Access*, 8:156314–156327, 2020. 5

[43] Loris Nanni, Andrea Loreggia, Alessandra Lumini, and Alberto Dorizza. A standardized approach for skin detection: Analysis of the literature and case studies. *Journal of Imaging*, 9(2), 2023. 2

[44] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation, 2023. 2

[45] Pexel. Pexel. https://www.pexels.com. Accessed: 2023-11-13. 3

[46] S.L. Phung, A. Bouzerdoum, and D. Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005. 2, 8

[47] Son Lam Phung, D. Chai, and A. Bouzerdoum. A universal and robust human skin color model using neural networks. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, volume 4, pages 2844–2849 vol.4, 2001. 2

[48] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. 3

[49] Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Learning people detection models from few training samples. In *CVPR 2011*, pages 1473–1480. IEEE, 2011. 3

[50] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10880–10890, October 2021. 3

[51] Weichao Qiu. *Generating human images and ground truth using computer graphics*. University of California, Los Angeles, 2016. 3

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 15

[53] Hossein Rahmani and Ajmal Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2458–2466, 2015. 3

[54] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2017. 3

[55] Yogesh Raja, Stephen J. McKenna, and Shaogang Gong. Segmentation and tracking using colour mixture models. In Roland Chin and Ting-Chuen Pong, editors, *Computer Vision — ACCV'98*, pages 607–614, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. 2

[56] Arun V Reddy, Ketul Shah, William Paul, Rohita Mocharla, Judy Hoffman, Kapil D Katyal, Dinesh Manocha, Celso M de Melo, and Rama Chellappa. Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances. *arXiv preprint arXiv:2303.10280*, 2023. 3

[57] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *Advances in neural information processing systems*, 29, 2016. 3

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2, 5

[59] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), June 2016. 2, 3, 8, 14

[60] German Ros, Simon Stent, Pablo F. Alcantarilla, and Tomoki Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation, 2016. 3, 7

[61] Ming-Jung Seow, D. Valaparla, and V.K. Asari. Neural network based skin color model for face detection. In *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, pages 141–145, 2003. 1

[62] C. So-Ling and Ling Li. A multi-layered reflection model of natural human skin. In *Proceedings. Computer Graphics International 2001*, pages 249–256, 2001. 2

[63] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 236–241, 1996. 2

[64] Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization, 2021. 6

[65] Wei Ren Tan, Chee Seng Chan, Pratheepan Yogarajah, and Joan Condell. A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics*, 8(1):138–147, 2012. 1, 2, 5

[66] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. Skinny: A lightweight u-net for skin detection and segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2386–2390, 2020. 2

[67] Anirudh Topiwala, Lidia Al-Zogbi, Thorsten Fleiter, and Axel Krieger. Adaptation and Evaluation of Deep Leaning Techniques for Skin Segmentation on Novel Abdominal Dataset. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 752–759. IEEE, 2019. 2, 5

[68] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 3

[69] Valery Tuchin. Tissue optics light scattering methods and instruments for medial diagnosis. *SPIE*, 13, 01 2000. 2

[70] Unreal Engine. Metahuman creator. https://www.unrealengine.com/en-us/metahuman. Accessed: 2023-11-13. 2

[71] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 2, 3

[72] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 3

[73] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[74] Henrik Wann, Jensen Stephen, R. Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. 35, 09 2002. 2

[75] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, October 2021. 3

[76] Han Xu, Abhijit Sarkar, and A. Lynn Abbott. Color invariant skin segmentation, 2022. 8, 15

[77] Ming-Hsuan Yang and Narendra Ahuja. Gaussian mixture model for human skin color and its applications in image and video databases. In Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman, editors, *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 458 – 466. International Society for Optics and Photonics, SPIE, 1998. 2

[78] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation, 2020. 7, 14

[79] Jiacheng Ye, Chengzu Li, Lingpeng Kong, and Tao Yu. Generating data for symbolic language with large language models, 2023. 2

[80] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 778–786, 2019. 3

[81] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1

[82] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 3

[83] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3