A. Skin Model Creation

Each skin component (epidermis, dermis, hypodrmis, hair, and blood network) was modeled in Houdini. We defined a set of mutable parameters corresponding to the properties of the skin components and used a random set of the parameters to generate each model. The process was automated using a Python script within Houdini, where 100 versatile and different models of each component were generated. The output files were stored in the format of an OBJ file for subsequent steps. The list of all mutable parameters and their predefined ranges are provided in Table 1, while the quantities of parameter choices are listed in Table 2.

Variation	Skin Generation Parameters	min	max	unit
Original Models	Hypodermis surface noise amplitude	0	1	
	Hypodermis surface noise roughness	0	0.7	
	Hypodermis surface noise turbulence	0	6	
	Deep blood network start number of points	20	60	
	Deep blood network start scatter seed	0	10	
	Deep blood network end number of points	200	800	
	Deep blood network end scatter seed	0	10	
	Upper blood network number of points	20	60	
	Upper blood network scatter seed	0	10	
	Dermis thickness	1	4	mm
	Epidermis thickness	0.02	0.15	mm
	Dermoepidermal Junction (DEJ) papillary count	2000	4000	
	Dermoepidermal Junction (DEJ) papillary seed	0	10	
	Epidermis surface noise offset	-10	10	mm
	Hair noise offset	0	30	
	Hair density value	0	30	
	Hair scatter seed	0	10	
	Hair relax iterations	0	5	
	Hair length	0	15	
	Hair bend curve angle	30	90	degrees
	Hair bend curve random angle	0	30	degrees
	Hair rotation amount	0.25	0.55	
	Hair uniform thickness	0.55	0.65	
	Hair mid thickness	0.15	0.22	
Hair Artifact	Hair density value	50	90	
	Hair scatter seed	0	20	
	Hair length	4	15	
	Hair bend curve angle	0	60	degrees
	Hair bend curve random angle	30	90	degrees
Blood Vessels Artifact	Deep blood network start number of points	40	60	
	Deep blood network start scatter seed	0	20	
	Deep blood network end number of points	500	900	
	Deep blood network end scatter seed	0	20	
	Upper blood network number of points	40	60	
	Upper blood network scatter seed	0	20	
	Dermis thickness	0.3	2.3	mm

Table 1. Ranges for all mutable parameters in the procedural generation of skin models in Houdini. Unless otherwise noted, the parameters represent relative units within the software's internal system rather than absolute measurements.

B. Artifact Generation

Each artifact was created as follows.

(a) **Calibration chart:** a sphere representing the calibration chart was added to the scene using Mitsuba with a random size and location. However, both variables were kept within pre-defined limits (radius: 2-5 mm, y-location of center: 1-5 mm from the edges, x-location of center: 2-4 mm from the edges) to prevent overlapping with the skin lesion. In addition, this artifact was added only to the scene when the lesion covered approximately less than 50% of the skin surface, ensuring that it did not obscure the lesion. The calibration charts were randomly generated in two different shades of blue or yellow.

(b) **Ruler:** we added 15 vertically oriented rectangles and one horizontally oriented rectangle using Mitsuba and placed them on the top of the epidermis to represent the ruler artifact in the dermoscopic images. The ruler's position was randomized in both the x and y directions, with the y-location and x-location restricted to 2-6 mm and 2-5 mm from the edges, respectively. Similarly to the calibration chart, the ruler artifact was only added in scenes with skin lesions smaller than a specified size.

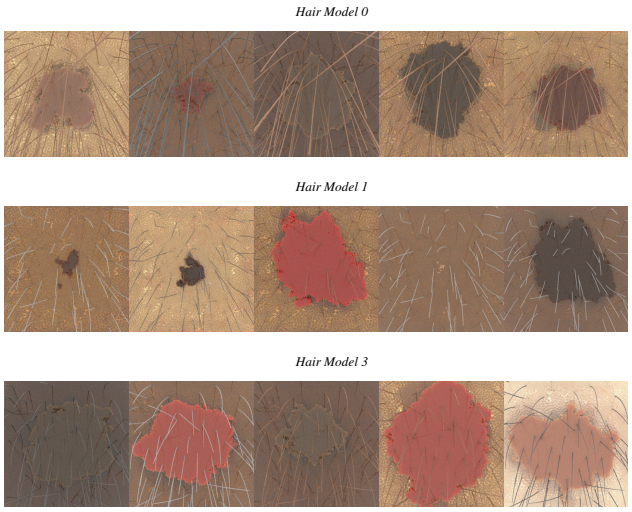
(c) **Dark frame:** the dark frame artifact was generated by placing a black torus beneath the sensor in Mitsuba. The radius of the torous was changed as a function of the distance between the sensor and the skin model to simulate various degrees of dark frame coverage. Similarly to the calibration chart and ruler, the dark frame was added only to the scenes with small lesions.

(d) **Hair:** images with hair artifact were generated by adjusting the mutable parameters related to the hair properties in Houdini. This process involved increasing the pre-defined ranges for hair scatter seed and hair density by a factor of two and generating the hair models at greater curve angles compared to the baseline models. The other skin components remained the same.

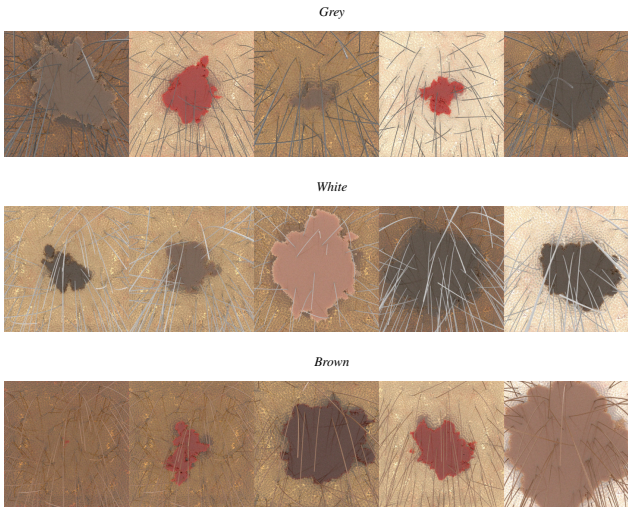
(e) **Blood vessels:** the blood vessel artifact was added by modifying the blood network and dermis layers in Houdini. We increased the number of random seeds, start and ending points for both the deep and upper blood networks. To enhance the visibility of the blood vessels on the skin, we reduced the thickness of the dermis layer compared to the baseline models and limited the melanosome fraction to small values (less than 0.05 corresponding to only 5 variations instead of 50 variations used for the other images) representing lighter skin tones. The other skin components and optical properties remained the same.

Parameter	Variations
Each skin model component (epidermis, dermis, hypodermis, hair, blood network)	100
Lesion model (each stored at 10 different timepoints)	20
Melanosome fraction (i.e. epidermis material)	50
Blood fraction (i.e. dermis material)	4
Hair material	3
Hypodermis material	1
Blood network material	1
Lesion material	20

Table 2. Summary of the parameter variations for the simulated models and their optical properties.



(a) Appearance variation of the same hair geometry with different hair color albedo and background.



(b) Albedo variations (hair colors) with different geometries.

Figure 1. Sample OASIS property variations.

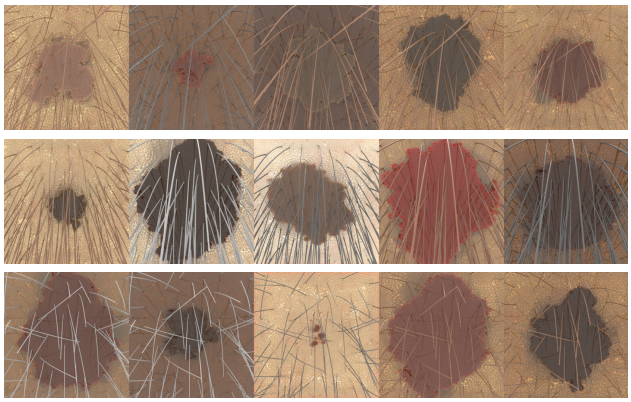
A summary of artifact parameter ranges is available in Table 3.

C. Parameter Variation

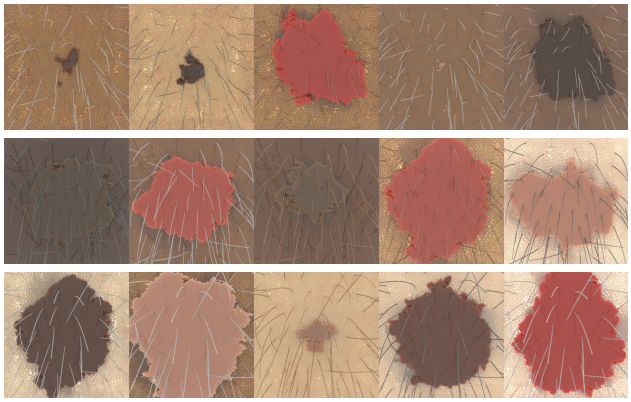
In Fig. 1 and 2 we show how parameters of the models can be connected to interpretable concepts about hair albedo (color) and geometry. As a result, every OASIS images comes with a rich set of annotations describing the properties (e.g.,"thin grey hair artifact").

D. Data Splits

For the artifacts annotations in ISIC (hair, cal Chart, frame, ruler), first, we obtained labels in the train, validation and test datasets (see Tab. 4), where annotations came from Bissoto [1]. We then controlled the proportion of artifacts present in the training and validation data (from 0% (no images with an artifact of a specific type present) to 100% (all images with an artifact of a specific type present). For example, hair (60:0) refers to an experiment where 60% hair-artifact present images were retained in the train and validation sets, respectively. We then supplemented the re-



(a) Geometry variation: thick hair examples.



(b) Geometry variation: thin hair examples.

Figure 2. Sample OASIS property variations (hair thickness).

Artifact	Parameter Range
Calibration chart	
Radius	2-5 mm
Center (y-location)	1-5 mm from the edges
Center (x-location)	2-4 mm from the edges
Color	blue or orange
Ruler	
Radius	2-5 mm
Center (y-location)	2-6 mm from the edges
Center (x-location)	2-5 mm from the edges
Dark frame	
Radius	11-13 mm

Table 3. Summary of the parameter ranges of the artifacts that were added as an object in Mitsuba.

		With	Without	Total			With	Without	Total
Hair	Train	1056	759	1815	Frame	Train	681	1134	1815
	Val	149	110	259		Val	85	174	259
	Test	313	207	520		Test	194	326	520
		With	Without	Total			With	Without	Total
CalChart	Train	135	1680	1815	Ruler	Train	927	888	1815
	Val	17	242	259		Val	115	144	259
	Test	194	326	520		Test	242	278	520

Table 4. Breakdown of ISIC artifact-present and artifact-absent images by train, validation, and test split.

maining images with OASIS examples. For instance, hair (60:40) refers to the experiment where 40% of the amount of hair artifact present examples were replaced with OASIS images. The hold-out test sets consisted only patient images across all the experiments. This approach allowed us to systematically control the amount of patient data and synthetic data present, and evaluate the effect of synthetic data augmentation on segmentation performance. We also provide a breakdown by artifact-present and artifact-absent images in the train, validation, and test sets in ISIC in Tab. 4.

E. Additional Segmentation Results

In Fig. 3 we include additional results of the replacement experiment reported in the manuscript, except where now evaluation is run on the full ISIC test set (including both artifact-present and artifact-free images). We observe similar performance patterns, i.e., that synthetic data improves performance for darker skin examples when only partial patient data is available and does not have a marked effect in lighter skin tomes. In all scenarios except for calibration chart (row 2), addition of synthetic data does not reduce performance.

F. Multi-Reader Multi Case Study

A multi-reader multi-case study is a type of study design that allows capturing variability across readers and examples, commonly used in imaging studies. We relied on the iMRMC software [2] to which we provided a matrix of Dice values for each of the five readers to calculate the mean and standard deviation.

References

[1] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *CVPR Workshops*, 2020. 2

[2] RST Catalog. iMRMC: Software for the Statistical Analysis of multi-reader multi-case studies, 2022. 3

[3] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv*, 2019. 4

Patient Dataset	Patient-to-Synth Ratio	Training Size		Validation Size		Test Size	
		Patient	Synth	Patient	Synth	Patient	Synth
Hair	100:00	1815	0	259	0	520	0
Hair	0:00	759	0	110	0	520	0
Hair	20:00	970	0	139	0	520	0
Hair	40:00	1181	0	169	0	520	0
Hair	60:00	1392	0	199	0	520	0
Hair	80:00	1603	0	229	0	520	0
Hair	0:100	759	1056	110	149	520	0
Hair	20:80	970	845	139	120	520	0
Hair	40:60	1181	634	169	90	520	0
Hair	60:80	1392	423	199	60	520	0
Hair	80:20	1603	212	229	30	520	0
calChart	100:00	1815	0	259	0	520	0
calChart	0:00	1680	0	242	0	520	0
calChart	20:00	1707	0	245	0	520	0
calChart	40:00	1734	0	248	0	520	0
calChart	60:00	1761	0	252	0	520	0
calChart	80:00	1788	0	255	0	520	0
calChart	0:100	1680	135	242	17	520	0
calChart	20:80	1707	108	245	14	520	0
calChart	40:60	1734	81	248	11	520	0
calChart	60:80	1761	54	252	7	520	0
calChart	80:20	1788	27	255	4	520	0
Frame	100:00	1815	0	259	0	520	0
Frame	0:00	1134	0	174	0	520	0
Frame	20:00	1270	0	191	0	520	0
Frame	40:00	1406	0	208	0	520	0
Frame	60:00	1542	0	225	0	520	0
Frame	80:00	1678	0	242	0	520	0
Frame	0:100	1134	681	174	85	520	0
Frame	20:80	1270	545	191	68	520	0
Frame	40:60	1406	409	208	51	520	0
Frame	60:80	1542	273	225	34	520	0
Frame	80:20	1678	137	242	17	520	0
Ruler	100:00	1815	0	259	0	520	0
Ruler	0:00	880	0	144	0	520	0
Ruler	20:00	1073	0	167	0	520	0
Ruler	40:00	1258	0	190	0	520	0
Ruler	60:00	1444	0	213	0	520	0
Ruler	80:00	1629	0	236	0	520	0
Ruler	0:100	888	927	144	115	520	0
Ruler	20:80	1073	742	168	92	520	0
Ruler	40:60	1258	557	190	69	520	0
Ruler	60:80	1444	371	213	46	520	0
Ruler	80:20	1629	186	236	23	520	0

Table 5. Number of synthetic (OASIS) and patient (ISIC) images within the data subsets for each experiment.

Figure 3. Multi Reader Multi Case (MRMC) evaluation of Dice segmentation performance (↑) when artifact-present training examples in the ISIC [3] are replaced with OASIS images. Results are reported with mean and standard deviation across the full ISIC test set (images with and without artifacts). We observe that at limited data regimes, the model trained with synthetic examples containing artifacts performs better for darker skin tones (tan2, int1). In light skin tones, all methods perform similarly. Row 1: hair, Row 2: Cal. Chart, Row 3: frame, Row 4: ruler. Blue: partial patient data (with only a fraction of artifacts present), orange: patient data supplemented with synthetic OASIS examples, and green: full patient data baseline.