Datasheet for Hyper-Skin Dataset

Pai Chet Ng, Zhixiang Chi, Yannick Verdie,
Juwei Lu, Konstantinos N. Plataniotis

June 2023

1 Datasheet for Hyper-Skin

The documentation is prepared in accordance with the datasheets for datasets guidelines [1]. Following the guidelines, we aim to provide transparency and facilitate a comprehensive understanding of the Hyper-Skin dataset for researchers and users.

2 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The Hyper-Skin dataset was created with the purpose of advancing hyperspectral skin analysis, particularly on consumer devices. The specific task in mind was to enable the reconstruction of hyperspectral cubes from RGB or multispectral images, allowing for a deeper understanding of various physiological characteristics of human facial skin. The dataset aimed to fill a gap in consumer-level skin analysis by providing comprehensive spectral coverage, including both the visible spectrum (400nm - 700nm) and near-infrared spectrum (700nm - 1000nm). This allowed researchers and developers to explore beyond the visual appearance of selfies and gain valuable insights into important skin properties such as melanin concentration, hemoglobin levels, moisture content, collagen content, and subcutaneous blood vessels. By addressing the limitations of consumer cameras that primarily capture RGB data, the Hyper-Skin dataset aimed to facilitate new possibilities for skin analysis applications and contribute to advancements in understanding skin health and well-being.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The Hyper-Skin dataset was created by the research team at the Multimedia Laboratory, University of Toronto, under the supervision of Professor Plataniotis. The dataset collection process strictly adhered to an approved protocol,

following the rigorous research ethics guidelines set by the University. These guidelines ensure the protection of participant rights, confidentiality, and privacy throughout the data collection process. The research team obtained necessary approvals from the relevant authorities, including the University's Human Research Ethics Program (HREP), to ensure compliance with ethical standards.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the Hyper-Skin dataset was funded by Huawei Technology Canada. The dataset creation was part of a collaborative research project between Huawei and the University of Toronto. The funding for this dataset was provided directly by Huawei as part of their research collaboration with the University.

3 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the Hyper-Skin dataset represent hyperspectral data at different spectral bands, which contain information about the reflectance properties of the skin. The dataset consists of two types of data pairs: (RGB, VIS) and (MSI, NIR), each serving different purposes in skin analysis.

- The (RGB, VIS) data pair includes 31-band hyperspectral data at the visible (VIS) spectrum ranging from 400nm to 700nm. It is accompanied by RGB images that serve as inputs for hyperspectral reconstruction research at the visible spectrum. This data pair allows for the analysis of surface-level skin characteristics such as melanin concentration, blood oxygenation, pigmentation, and vascularization.
- The (MSI, NIR) data pair consists of 31-band hyperspectral data at the near-infrared (NIR) spectrum ranging from 700nm to 1000nm. It includes multispectral data containing RGB images and an infrared image at 960nm, facilitating hyperspectral reconstruction research from MSI to NIR. This data pair enables the study of deeper tissue properties such as water content, collagen content, subcutaneous blood vessels, and tissue oxygenation.

Each instance in the dataset corresponds to a specific facial image sample obtained from the data collection process, representing individual facial skin samples for hyperspectral skin analysis. Additionally, the dataset may include associated metadata such as demographic information (e.g., age, gender, and skin type) to provide further context and characteristics related to the instances.

The Hyper-Skin dataset focuses on capturing and analyzing HSI data of skin samples, offering different ranges of hyperspectral data for comprehensive investigations of various skin features. It is designed to support research and analysis in the field of skin imaging and related applications.

How many instances are there in total (of each type, if appropriate)?

The Hyper-Skin dataset consists of a total of 306 instances of facial data collected from 51 participants. Each instance represents a hyperspectral face data sample with a spatial resolution of 1024x1024 and 448 spectral bands ranging from 400nm to 1000nm. The dataset was created by capturing 6 faces from each participant, including 2 facial expressions (neutral and smile) and 3 different facial positions (left, front, right). Since the spatial resolution of the hyperspectral face instance is 1024x1024, each instance contains millions of spectra per image, providing detailed spectral information about the facial skin.

For each type of data pair, namely the (RGB, VIS) and (MSI, NIR) data types, the number of instances remains the same. Therefore, the Hyper-Skin dataset consists of 306 instances for both data types. The resulting hyperspectral data at the visible band spans from 400nm to 700nm and has dimensions of 1024x1024x31, with a spectral spacing of 10nm between each band. Similarly, the resulting hyperspectral data at the NIR band spans from 700nm to 1000nm and also has dimensions of 1024x1024x31, with a 10nm spectral spacing.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The Hyper-Skin dataset represents a sample of instances rather than containing all possible instances. The larger set from which the sample is drawn is the population of potential facial skin data that could be collected. The sample of instances in the dataset was carefully selected to represent a diverse range of facial characteristics and conditions. However, it is important to note that the dataset may not be fully representative of the entire population. The selection of participants and the specific facial expressions and positions captured may introduce some biases or limitations in terms of representativeness.

To address the need for a more diverse dataset, the research team is actively working with the research ethics board in our university to schedule another data collection campaign. The objective is to increase the diversity of the dataset by including a broader range of participants, considering factors such as age, gender, ethnicity, and skin type. This additional data collection effort aims to enhance the representativeness of the dataset and provide a more comprehensive resource for skin analysis research. The decision to conduct further data collection reflects our commitment in improving the dataset and expand its

coverage. By incorporating a wider range of instances, the dataset will become more robust and valuable for various applications in the field of hyperspectral skin analysis.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in the Hyper-Skin dataset consists of resampled 31-band hyperspectral data, derived from the raw hyperspectral data. The raw hyperspectral data has a spatial resolution of 1024x1024 pixels, providing detailed information about the skin at a fine level of granularity. It is captured in 448 spectral bands, covering the wavelength range from 400nm to 1000nm, representing the unprocessed, high-dimensional measurements of skin reflectance across the entire spectrum.

To make the dataset more manageable and accessible, the 448-band data is resampled into two types of 31-band data. The first type covers the visible (VIS) spectrum from 400nm to 700nm, while the second type covers the near-infrared (NIR) spectrum from 700nm to 1000nm. Corresponding to these two types of 31-band data, we have created two types of data pairs: (RGB, VIS) and (MSI, NIR). The (RGB, VIS) data pair includes the Red, Green, and Blue channels of the visible spectrum, while the (MSI, NIR) data pair consists of multispectral data containing the RGB channels along with an infrared image at 960nm.

The primary dataset shared with interested researchers will include the two types of data pairs, comprising a total of 306 pairs for each type. The raw dataset, which encompasses the 448 spectral bands, can be provided upon request due to its large size of over 2TB. This approach ensures that the researchers have access to the essential data pairs while managing the storage and distribution of the raw data efficiently.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, there is a target associated with each instance in the Hyper-Skin dataset. For each data pair, the target is the 31-band hyperspectral data either in the visible (VIS) spectrum or the near-infrared (NIR) spectrum. The input data for each instance is either the corresponding RGB data or the multispectral (MSI) data, depending on the data pair.

The presence of the target labels allows researchers to train models and algorithms to learn the mapping between the input data (RGB or MSI) and the corresponding 31-band hyperspectral data. This facilitates the development and evaluation of methods for hyperspectral data reconstruction and analysis, leading to potential advancements in consumer device-based skin analysis applications.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was

unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, there is no missing information from individual instances in the Hyper-Skin dataset. Each instance is complete and does not have any intentionally removed or redacted information.

During the data collection process, all necessary information was obtained and included in the dataset. The dataset was carefully curated to ensure that each instance contains the required hyperspectral data, as well as the corresponding RGB or MSI data. The participants provided the necessary facial images with different facial expressions and positions, allowing for comprehensive analysis of the skin.

The dataset aims to provide a complete and representative collection of hyperspectral data for facial skin analysis. As such, there are no missing values or intentionally withheld information in the individual instances. However, it's important to note that the dataset may not include additional personal or demographic information about the participants, such as age, gender, or skin type, to protect privacy and comply with ethical guidelines.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No, relationships between individual instances are not made explicit in the Hyper-Skin dataset. The instances in the dataset are independent facial samples obtained from different participants. Each instance represents a standalone hyperspectral data of a facial image, and there are no explicit connections or relationships between these instances within the dataset.

The Hyper-Skin dataset is designed to provide researchers with a comprehensive collection of hyperspectral data for studying skin properties and conducting related analyses. While the dataset does not incorporate explicit relationships between instances, it offers valuable insights into the characteristics and properties of individual facial skin samples, enabling researchers to explore various aspects of skin analysis and related applications.

Researchers utilizing the dataset can focus on analyzing individual instances independently or incorporate their own methodologies to establish relationships between instances, if desired, by considering external factors or additional datasets that provide such connections.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes, the Hyper-Skin dataset includes explicit data splits for training, validation, and testing purposes. Specifically, 4 participants were randomly chosen from the participant pool to form the testing set. This testing set includes a total of 24 instances, covering 2 types of facial expressions (neutral and smile)

and 3 different face poses (left, front, right). For the validation set, 3 participants were selected, contributing 18 instances. The training set, consisting of the remaining participants' data, allows researchers to train their models on a substantial amount of diverse samples. The validation set, comprising images from specific participants, aids in fine-tuning hyperparameters and monitoring model performance during the development process.

The rationale behind this data split is to ensure a diverse representation of participants, facial expressions, and face poses in each subset. By randomly selecting participants for the testing set and a separate group of participants for the validation set, we mitigate potential biases that may arise if data from the same participant were present in multiple subsets. This approach promotes fairness and generalizability when evaluating models' performance on unseen participants.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The Hyper-Skin dataset has undergone rigorous preprocessing steps to minimize errors, noise, and redundancies. To obtain spectral reflectance data from the RAW hyperspectral data, radiometric calibration was applied. This involved capturing a white reference image and a dark reference image. The white reference image represents a spectrally neutral surface, providing a consistent reflectance value across all spectral bands. The dark reference image was obtained by triggering the camera with the lens closed.

In the data preprocessing phase, several steps were taken to ensure data quality. First, the dark reference values were subtracted from the RAW data to remove any noise or offset present in the images. This subtraction helps improve the accuracy of the spectral reflectance values. Second, the data was divided by the white reference values to normalize and convert it to reflectance values. This normalization step ensures that the reflectance values are consistent across the dataset and facilitates meaningful comparisons between instances.

Regarding redundancies, the dataset does not contain any repeated participants. Each participant contributes a unique set of images, encompassing different facial expressions and face poses. This lack of redundancies ensures that each instance in the dataset provides distinct information and adds to the overall diversity of the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The Hyper-Skin dataset is self-contained and does not rely on external resources such as websites, tweets, or other datasets. The dataset itself is the primary resource provided for research and analysis purposes. As the dataset is self-contained, there are no external resources linked to it that may pose guarantees, require archiving, or impose fee restrictions. Users can rely solely on the dataset and the associated GitHub repository for their research and analysis without any additional dependencies on external resources.

To access the dataset, interested users are required to digitally sign an End User License Agreement (EULA) online. The EULA outlines the terms and conditions for using the dataset, including provisions for using only the authorized images in future publications. Users can find detailed instructions for requesting the dataset in the GitHub repository provided at https://hyper-skin-2023.github.io/dataset/. The GitHub repository contains the code for data loading, evaluation, and benchmarking. It serves as a supplementary resource for users to access and download the code associated with the dataset. The repository is publicly available and can be accessed without any restrictions.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the Hyper-Skin dataset does not contain data that might be considered confidential. It does not include any information protected by legal privilege or doctor-patient confidentiality. The dataset solely consists of hyperspectral data derived from facial images and does not include any content from individuals' non-public communications. The dataset is focused on capturing and analyzing the reflectance properties of the skin for research and analysis purposes in the field of skin imaging.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the Hyper-Skin dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or otherwise cause anxiety. The dataset solely consists of hyperspectral data derived from facial images and does not include any content that could be considered offensive or harmful. The dataset is intended for research and analysis purposes in the field of skin imaging and does not contain any explicit or objectionable content.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people. It consists of facial skin data collected from 51 participants, and each instance in the dataset represents a specific facial skin sample obtained from the data collection process.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, the dataset identifies subpopulations based on age and gender. The demographic information of the participants, including age and gender, is recorded and associated with each instance in the dataset. The subpopulations are identified by considering the age and gender information of the participants.

For age subpopulation, the dataset includes participants from various age groups. The age distribution within the dataset covers a range from 20s, 30s, 40s, 50s, and so on. The distribution of age groups is shown in Figure 1. Regarding gender subpopulation, the dataset encompasses participants of different genders, including male and female. The distribution of gender within the dataset aims to maintain a balance and representativeness. The distribution of gender and the number of participants belonging to each gender is shown in Figure 1. By including information about age and gender subpopulations, the dataset allows researchers to analyze and study potential variations or correlations in skin characteristics based on these demographic factors.

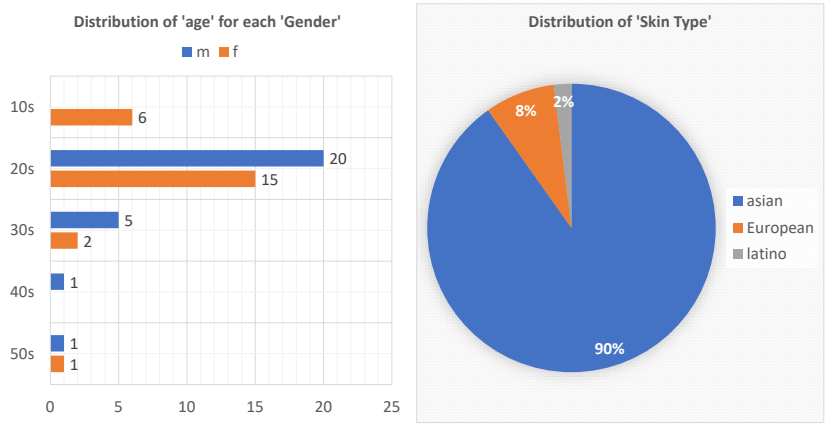

Figure 1: The bar chart and pie chart represent the participant demographics in terms of age, gender, and skin type. The bar chart shows the distribution of age for each gender, with the total count of participants being 51. The pie chart shows the distribution of skin types among the participants, with the majority being Asian (46 out of 51).

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, it is not possible to directly identify individuals from the dataset. Personal information, including names and addresses, is not collected or stored in the database. Instead, a unique ID number is assigned to each participant to ensure anonymity. This dataset follows strict protocols to protect participants' privacy and confidentiality. The database is securely stored and can only be accessed by authorized personnel, specifically the Principal Investigator (PI) and the research team. Access to the dataset is granted by the PI to interested researchers who have signed the End User License Agreement (EULA). The EULA ensures that researchers understand and adhere to the privacy and confidentiality measures in place. Any future publications or presentations resulting from this study will not disclose any personal information that could identify participants. The dataset is carefully anonymized to safeguard the privacy of individuals. The focus is on protecting participant confidentiality and ensuring that the study adheres to ethical guidelines and regulations.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, the dataset does not contain any sensitive data that could reveal racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, locations, financial or health data, biometric or genetic data, forms of government identification, or criminal history. The dataset focuses solely on the hyperspectral data of facial images and does not include any personally identifiable information or sensitive attributes of the participants.

The collection and handling of the data strictly adhere to privacy and ethical guidelines to ensure the protection of participants' sensitive information. Confidentiality and privacy measures are in place to safeguard the data and maintain the anonymity of the individuals involved in the study.

4 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance in the dataset was acquired through direct measurement with a hyperspectral camera to capture the skin's reflectance across a wide range of spectral bands, ranging from 400nm to 1000nm. The

acquisition of the data does not involve reported responses from subjects or indirect inference/derivation from other data sources. Instead, it relies on the direct capture of the skin's reflectance using calibrated hyperspectral imaging techniques.

To ensure the accuracy and validity of the acquired data, calibration procedures were implemented. This involved capturing a white reference image, which provides a consistent reflectance value across all spectral bands, and a dark reference image to remove noise or offset. Radiometric calibration techniques were applied to subtract the dark reference values and divide the data by the white reference values, resulting in normalized spectral reflectance data. These calibration procedures help validate and verify the accuracy and reliability of the hyperspectral data. By employing rigorous calibration techniques, any sources of errors or noise in the measurements were minimized, ensuring the quality and validity of the acquired data.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data acquisition for the hyperspectral dataset was performed using the Specim FX10 hyperspectral camera, which is designed for industrial machine vision applications. We rented the FX10 camera from Channel Systems, a local distributor in Manitoba, Canada [2]. The rental package included the FX10 camera, halogen light, white reference, and tripods necessary for the data collection process. The rental agreement was reviewed by the Risk and Management Department, and insurance coverage was provided for the equipment during the rental period. The FX10 camera offers various advantages, such as free wavelength selection from 220 bands, small size, fast optics, and high speeds. It has a spectral range of 400-1000nm, 448 spectral bands with a full width at half maximum (FWHM) of 5.5nm, and a spatial sampling of 1024 pixels. The camera supports different camera interface options, including CameraLink and GigE Vision.

A customized scanner was requested from the vendor to move the FX10 camera instead of mounting the camera on the LabScanner, which is used to move the captured object while keeping the camera fixed. This arrangement was necessary to capture data effectively. The FX10 camera and the scanner were controlled by a motor and the LUMO recorder software installed on a computer, as shown in Figure 2. The FX10 camera was connected to the computer through the GigE port, while the scanner was connected through the USB port. To enable the USB driver for the scanner, a Pleora Ebus GigE Vision license was purchased. This setup allowed the data to be acquired efficiently while minimizing data loss due to limited bandwidth and interference.

The data acquisition procedures and equipment were carefully validated and tested to ensure reliable and accurate data capture. Table 1 shows the parameter settings, such as the frame rate, exposure time, spectral binning, and spatial binning, were configured based on the specific requirements of the study. The

working distance between the FX10 camera and the human face was set to 40cm, allowing for capturing a full image with a spatial resolution of 1024 × 1024. The chosen parameter settings and working distance were selected to optimize data quality and capture the required spectral and spatial information.

Table 1: Data acquisition parameter settings

Equipment	**Parameter**	**Value**
LUMO recorder	Frame rate	45Hz
	Exposure time	22ms
	Spectral binning	1
	Spatial binning	1
Motor controller	Length	300mm
	Speed	12.3 mm/s
	Working distance	40cm

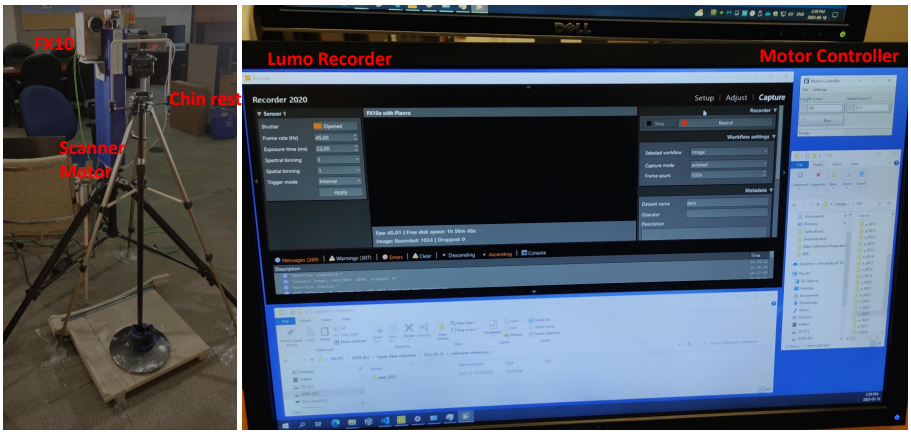

Figure 2: The left image shows the customized scanner with the mounted FX10, connected to the computer via a GigE port for the camera and a USB port for the scanner. The right image displays the LUMO recorder and motor controller software, which control the FX10 and the scanner.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The Hyper-Skin dataset applied a deliberate sampling strategy to capture face images at different orientations, representing diverse participants. Our sampling strategy has the following rules to incorpoarte diverse participants and different face orientation:

- Selection of Participants: The dataset consists of images collected from 51 participants. To ensure diversity, a sampling strategy could have involved randomly selecting participants from a larger pool of potential participants. By randomly selecting participants, the dataset aims to minimize potential bias and ensure a representative sample.

- Pool Selection for Different Orientations: The dataset includes face images captured at different face poses, such as left, front, and right. To achieve this, a sampling strategy might have involved dividing the participants into different pools based on their willingness and ability to pose their faces in different orientations. For example, participants who were comfortable and capable of posing in a left orientation formed one pool, while those posing in a front or right orientation formed separate pools.

- Random Selection within Pools: Within each pool, a random selection process could have been applied to choose a specific number of participants. This random selection ensures that the dataset includes face images from a diverse range of participants who are willing to pose in the respective orientations.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process for the Hyper-Skin dataset involved the participation of individuals from various roles. The following individuals were involved:

- Students: Students from the University of Toronto actively participated in the data collection process. They were engaged in tasks such as operating the hyperspectral camera, assisting participants during the data collection sessions, and ensuring proper data handling procedures. Students involved in the data collection process were compensated through various mechanisms, such as research assistantships, internships, or work-study programs. The compensation for students was in accordance with the university's guidelines and standards.

- Research Team: The research team, consisting of faculty members, researchers, and technical experts, played a crucial role in overseeing the data collection process. They provided guidance, supervision, and quality control to ensure accurate and reliable data collection. The research team members were typically salaried employees or researchers affiliated with the University of Toronto. They received compensation based on their employment contracts or research agreements.

- Participants: Individuals from diverse backgrounds voluntarily participated in the data collection campaign. Participants were recruited through online forums, email communication, physical posters, and social media

advertisements. They contributed their facial data by undergoing the data collection procedure. To express gratitude for their participation, each participant received a $20 gift card as a thank-you gesture. The gift card number was tracked using a tracking form, with only the last three digits recorded for identification purposes. In the event that a participant chose to withdraw from the study, they were eligible for compensation on a pro-rated basis. The compensation rate was set at $20 per hour for the time they had spent in the study up until the point of withdrawal. This ensured fairness and recognized the value of their contribution, regardless of their duration of participation.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data for the Hyper-Skin dataset was collected over a specific timeframe consisting of two phases.

- The first phase, which involved the verification of the data collection setup and the collection of sample face images from the lab group, took place approximately one month before the actual campus-wide data collection. The purpose of this phase was to ensure the effectiveness of the data collection setup and identify any areas for improvement before launching the full-scale data collection.
- The second phase encompassed the actual data collection with recruited participants over a period of three weeks. Participants were contacted individually to schedule their participation times. The data collection phase lasted for two full weeks, including weekends. During this time, a total of 51 participants contributed their facial data to the dataset.

The timeframe for data collection aligns with the creation timeframe of the data associated with the instances. The data was collected in a timely manner within the specified timeframe, ensuring that the instances are representative of the participants who participated in the study during that period.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, the research ethics review was conducted by the Research Ethics Board (REB) at the University of Toronto. The REB is responsible for ensuring that research involving human participants adheres to ethical guidelines and regulations. The study protocol, informed consent procedures, data handling practices, participant confidentiality measures, and potential risks or ethical considerations associated with the research were all evaluated during the review

process. The goal was to safeguard the rights, welfare, and privacy of the participants involved. Following the review, the REB provided a favorable approval for the study, indicating that it met the necessary ethical standards and requirements. This approval signifies that the research protocol adequately addresses ethical concerns and provides appropriate protections for the participants. Supporting documentation, including the approved research protocol and informed consent forms are included in the supplementary materials.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relate to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected directly from the individuals involved in the study. Participants willingly contributed their facial data by participating in the data collection process. The data collection procedures were carried out in person, following the established protocols and consent procedures. No third parties or external sources were involved in obtaining the data for this particular dataset.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Yes, individuals involved in the data collection process were duly notified. Prior to their participation, participants received a comprehensive informed consent form that outlined the study's purpose, data collection procedures, and their rights as participants. The informed consent form explicitly covered the use of their facial data, the measures implemented to safeguard confidentiality and privacy, and the voluntary nature of their involvement. To inform individuals about the data collection process, we utilized email notifications, a screenshot of the email is shown below. The emails were sent to the participants through the department's administrative staff, centrally managing the distribution to prevent overwhelming them with excessive emails.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, the individuals in question provided their consent for the collection and use of their data. Prior to participating in the data collection process, participants were required to review and sign an informed consent form, which explicitly outlined the purpose of the study, the data collection procedures, and

FW: Invitation to Participate in our data collection campaign and earn free gift card

ECE Graduate Office <ecegradoffice@utoronto.ca>
Thu 3/2/2023 2:46 PM

1 attachments (297 KB)
[28.02.2023] - poster-section6-2.pdf;

Dear ECE Graduate Students,

The following opportunity may be of interest to you. Please see the attached for details.

Sincerely,
The ECE Graduate Office

Dear Students,

We are excited to invite you to participate in our upcoming data collection campaign, which will be held from March 20 to March 29, in the Multimedia Lab, BA4157, Bahen Building. We will be compensating each participant with $20 (in the form of a gift card) for your participation, plus an additional $10 bonus for those with good performance.

The data collection campaign will involve acquiring hyperspectral images, standard RGB images, and skin images from a hyperspectral camera, a smartphone camera, and a cosmetology camera. The purpose of this data collection campaign is to reconstruct hyperspectral images from RGB images captured by consumer-grade smartphones or any end-user devices and apply advanced hyperspectral imaging (HIS) analysis techniques to predict users' skin information, such as hemoglobin, oxygen saturation, melanin, skin hydration, and skin sebum.

If you are interested in participating in our study, please register your participation and book your slot here.

Figure 3: Screenshot of the email notification to the students.

their rights as participants. The consent form obtained their explicit agreement for the collection, storage, and use of their facial data for research purposes. One sample consent form is included in the supplementary materials.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Yes, the consenting individuals were provided with a mechanism to revoke their consent in the future or for certain uses. As part of the informed consent process, participants were informed about their right to withdraw their consent at any time and were provided with information on how to do so. Typically, the mechanism for revoking consent involves contacting the research team or the designated point of contact indicated in the informed consent form. Participants may be provided with contact details such as an email address or a phone number to communicate their decision to revoke consent.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

At present, a formal analysis of the potential impact of the Hyper-Skin dataset and its use on data subjects, such as a data protection impact analysis, has not been conducted. However, we recognize the importance of such an analysis in ensuring the responsible and ethical handling of data. Therefore, we intend to conduct a data protection impact analysis as a future action. This analysis will involve a systematic assessment of the dataset's potential impact on data subjects, taking into account factors such as data sensitivity, privacy risks, and security measures. By conducting this analysis, we aim to identify and address any potential risks or concerns associated with the dataset and its use, thereby ensuring the protection of data subjects' privacy and rights.

5 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, several preprocessing steps were conducted on the Hyper-Skin dataset to ensure data quality and compatibility for further analysis. As part of the preprocessing steps, radiometric calibration was conducted on the Hyper-Skin dataset to obtain spectral reflectance data from the RAW hyperspectral data.

During data preprocessing, the dark reference values were subtracted from the RAW data. This subtraction step helped eliminate any noise or offset that may have been introduced during the imaging process. Lastly, the RAW data was divided by the white reference values. This division step served to normalize the data and convert it to reflectance values. By dividing the data by the reflectance of the white reference image, we were able to obtain the spectral reflectance data.

These radiometric calibration steps were crucial in ensuring that the Hyper-Skin dataset contained accurate and reliable spectral reflectance information. By subtracting the dark reference values and normalizing the data using the white reference values, we minimized potential noise, offset, and variations introduced during data acquisition. The resulting spectral reflectance data provided a more consistent and standardized representation of the facial skin characteristics across the hyperspectral bands.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, the raw data of the Hyper-Skin dataset was saved in addition to the preprocessed, cleaned, and labeled data. Recognizing the potential for unanticipated future uses and the need for flexibility in data analysis, we have retained the raw data. However, due to its large size of approximately 2TB, it is not included in the dataset directly. Researchers interested in accessing the raw data can make a request, and we will provide instructions and facilitate the sharing process to ensure they have access to the complete dataset, including the raw data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the software used to preprocess, clean, and label the instances in the Hyper-Skin dataset is available. To access the software used for preprocessing, cleaning, and labeling, you can find the Python scripts and relevant libraries on the project's GitHub repository at https://github.com/hyperspectral-skin/Hyper-Skin-2023. The repository contains the necessary code and instructions to reproduce the preprocessing steps and apply them to similar datasets.

6 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The Hyper-Skin dataset has been primarily used for internal exploration and the study of hyperspectral skin analysis. It has not been utilized in any published work or specific tasks outside of the initial research scope. The dataset has

served as a valuable resource for conducting experiments, developing algorithms, and exploring potential applications in the field of hyperspectral imaging. However, the findings and results derived from these internal investigations have not been formally published or shared externally at this time.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

As of now, there is no repository or specific link available that directly links to papers or systems that use the Hyper-Skin dataset. The dataset has primarily been used for internal exploration and analysis, and there have not been any published papers or systems developed using this dataset that are publicly accessible. However, in the future, if any publications or systems arise that utilize the Hyper-Skin dataset, we plan to compile a central repository to summarize all the works that use our dataset for their research study. This repository will serve as a valuable resource for researchers and interested individuals to access and explore the studies conducted using the Hyper-Skin dataset.

What (other) tasks could the dataset be used for?

The Hyper-Skin dataset can be potentially used for various tasks and research areas beyond its current applications. Some possible tasks that the dataset could be used for include:

- Skin-related research: The dataset can be utilized for studying various aspects of human skin, such as skin conditions, aging effects, skin tone analysis, or the impact of external factors on skin appearance.
- Hyperspectral data analysis: The dataset offers an opportunity for researchers to delve into hyperspectral data analysis techniques, including dimensionality reduction, feature extraction, and spectral-spatial analysis.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The composition and collection of the Hyper-Skin dataset, as well as its preprocessing/cleaning/labeleling, do not inherently introduce any biases or issues that might impact future uses. However, it is essential for future users to be aware of potential considerations to avoid any unintended consequences or undesirable harms.

One aspect to consider is the need for responsible and ethical use of the dataset. It is crucial to ensure that any analysis, modeling, or applications built upon the dataset adhere to principles of fairness, non-discrimination, and respect

for privacy. This includes avoiding any uses that may perpetuate stereotypes, lead to unfair treatment of individuals or groups, or result in unintended biases.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Yes, there are tasks for which the Hyper-Skin dataset should not be used. Specifically, this dataset is not intended for face recognition or biometric applications. While the dataset contains facial images, its primary purpose is for exploring hyperspectral analysis of skin rather than identifying individuals or performing biometric authentication. The Hyper-Skin dataset lacks the necessary annotations, metadata, and specific data collection procedures required for robust face recognition or biometric applications. It does not include identity labels, or other attributes typically necessary for training face recognition algorithms. Therefore, attempting to use this dataset for such tasks could lead to inaccurate or misleading results. It is crucial to respect the intended use and limitations of the dataset to avoid any misapplication or misuse. Instead, the Hyper-Skin dataset should be used for research and analysis related to hyperspectral skin analysis, skin condition assessment, or similar domains that leverage the spectral properties of skin data.

7 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No, the Hyper-Skin dataset will not be distributed to third parties outside of the university and the company which funded the collection of this data. The distribution of the dataset is currently limited to internal use within the research team responsible for its creation. This ensures that the dataset is utilized in a controlled and responsible manner, adhering to the terms and conditions set forth by the university. Any future distribution of the dataset to external parties would require appropriate considerations, such as obtaining consent, ensuring data privacy and security, and complying with relevant legal and ethical guidelines.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The Hyper-Skin dataset will be distributed to interested researchers in a format that is convenient and accessible. To ensure data security and control access, the dataset will be hosted on the university's managed server. Access to the dataset will be restricted to authorized individuals, and interested researchers can request access by signing the End User License Agreement (EULA) digitally at the following link: https://hyper-skin-2023.github.io/dataset/.

At present, the dataset does not have a digital object identifier (DOI). However, researchers will still be able to access the dataset through the password-protected channels centrally managed by the university. This approach ensures that proper data governance measures are in place and that the dataset is securely distributed within the intended research community.

When will the dataset be distributed?

The Hyper-Skin dataset will not be distributed openly or made publicly available. Instead, researchers can request access to the dataset by signing an End User License Agreement (EULA) digitally. Once the request is made and approved, the dataset will be made available to the approved researchers. Therefore, the dataset is ready to be shared whenever a request is made and approved by the PI.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Yes, the Hyper-Skin dataset will be distributed under applicable terms of use (ToU). The exact license and ToU will be defined in the End User License Agreement (EULA) that researchers are required to sign in order to access the dataset. The EULA will outline the terms and conditions for the use of the dataset, including any copyright restrictions, intellectual property rights, and permitted uses. The EULA is available at this link: https://hyper-skin-2023.github.io/dataset/.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No, there are no third-party IP-based or other restrictions imposed on the data associated with the instances of the Hyper-Skin dataset. The dataset is solely owned and managed by the research team at the Multimedia Laboratory, University of Toronto, who is responsible for its creation, and no external parties have imposed any additional restrictions or limitations on its use.

As a result, there are no specific licensing terms or fees associated with third-party restrictions. Researchers who gain access to the dataset will be subject to the terms and conditions outlined in the End User License Agreement (EULA) provided by the entity responsible for the dataset. The EULA will govern the permitted uses and any applicable restrictions, ensuring that the dataset is used in compliance with the authorized terms of use.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and pro-

vide a link or other access point to, or otherwise reproduce, any supporting documentation.

There are no export controls or other regulatory restrictions apply specifically to the Hyper-Skin dataset or its individual instances. Researchers and users of the dataset are advised to familiarize themselves with the applicable laws, regulations, and export control requirements of their respective jurisdictions. It is recommended to consult legal counsel or appropriate authorities to ensure compliance with any relevant export control or regulatory restrictions that may apply.

8 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The Hyper-Skin dataset is supported, hosted, and maintained by the Multimedia Laboratory at the University of Toronto. The laboratory operates under the direction of Professor Plataniotis, a full professor at the Department of Electrical and Computer Engineering (ECE). The ECE department provides the necessary infrastructure and resources to ensure the continued hosting and maintenance of the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The owner/curator/manager of the Hyper-Skin dataset can be contacted through the following email addresses:

- Prof. Kostas Plataniotis (PI): kostas@ece.utoronto.ca
- Main Investigator: pcng@utoronto.ca

Please feel free to reach out to them via email for any inquiries or requests regarding the dataset.

Is there an erratum? If so, please provide a link or other access point.

As of now, there is no known erratum associated with the Hyper-Skin dataset. If any erratum or corrections are identified in the future, they will be documented and made available through an appropriate channel, such as a link or access point provided by the dataset owner/curator.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will be periodically reviewed and updated as needed. This includes adding new instances, or removing erroneous instances. The frequency

of updates will depend on various factors such as the identification of errors or the availability of new data. Updates to the dataset will be managed by the dataset owner/curator, who will oversee the review process and implement necessary changes. The specific communication method for updates will be determined by the dataset owner/curator and may include channels such as a mailing list, or a repository like GitHub. Users will be notified about updates through the designated communication channel to ensure they are aware of any changes or improvements made to the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

The Hyper-Skin dataset contains sensitive information related to individuals. Participants were informed about the retention of their data during the informed consent process. The data retention period, which is 10 years, was clearly communicated to the participants, specifying the duration for which their data would be stored before being deleted.

To enforce the limits on data retention, strict data management practices will be followed. The dataset owner/curator will ensure that the data associated with the instances is securely stored and protected during the specified retention period. After the designated time has elapsed, the data will be permanently and securely deleted to comply with the stated limits on data retention.

These data retention limits will be enforced through robust technical and administrative measures, such as secure storage systems, access controls, and regular audits to monitor compliance. The dataset owner/curator will take all necessary precautions to safeguard the privacy and confidentiality of the individuals involved and adhere to applicable data protection regulations and guidelines.

Will older versions of the dataset continue to be supported/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Since the Hyper-Skin dataset is currently in its first version, there are no older versions to support, host, or maintain. As the dataset evolves and new versions are released in the future, the focus will be on supporting and maintaining the latest version. Any updates or enhancements to the dataset will be made available in subsequent versions, ensuring that users have access to the most recent and comprehensive data.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not,

why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

At present, there is no specified mechanism in place for external users to directly extend, augment, build on, or contribute to the Hyper-Skin dataset. However, interested researchers or individuals who wish to collaborate or contribute to the dataset can reach out to the dataset owner or the project team, as mentioned previously, through the provided contact information.

If contributions are considered and deemed appropriate, a collaborative process can be established to evaluate and integrate the proposed additions or modifications to the dataset. This process may involve verification and validation steps, such as assessing the quality, relevance, and integrity of the contributed data before incorporating it into the dataset.

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. I. au2, and K. Crawford, "Datasheets for datasets," 2021.

[2] "Channel systems canada," 2023. [Online]. Available: https://channelsystems.ca/