UDC-SIT: A Real-World Dataset for Under-Display Cameras

Kyusu Ahn[1,3] **Byeonghyun Ko**[2] **HyunGyu Lee**[2]
Chanwoo Park[2] **Jaejin Lee**[1,2]

[1]Dept. of Data Science, Seoul National University, Seoul, Republic of Korea
[2]Dept. of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[3]Research Center, Samsung Display Co., Ltd., Yongin, Republic of Korea
kyusu.ahn@snu.ac.kr, {byeonghyun, hyungyu}@aces.snu.ac.kr, {99chanwoo, jaejin}@snu.ac.kr

2023

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.

Under Display Camera (UDC) faces challenges related to image degradation, including issues such as low transmittance, blur, noise, and flare. Despite its significance, there has been a lack of real-world datasets in the UDC domain. Only synthetic images are available that do not accurately represent real-world degradation. As far as we know, it is the first real-world UDC dataset to overcome the problems of the existing UDC datasets.

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

The dataset is created by the authors of the paper as well as the members of the Thunder Research Group at Seoul National University, including Woojin Kim, Gwangho Choi, Sangsoo Im, Gyuseong Lee, Dongyoung Lee, Gyeongje Jo, Yeonkyoung So, Jiheon Seok, Jaehwan Lee, Donghun Choi, and Daeyoung Park, on behalf of universities and research institutions.

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.

This work was supported in part by the Institute for Information & communications Technology Promotion (IITP) grant (No. 2018-0-00581, CUDA Programming Environment for FPGA Clusters) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00222663). This work was also supported in part by the Samsung Display Co., Ltd. ICT at Seoul National University provided research facilities for this study.

Any other comments?

N/A.

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 dataset consists of pairs of undistorted and UDC-distorted images for the same scene. Each pair also has annotations that tell the image-capturing conditions, such as light sources,

day/night, indoor/outdoor, and flare components.

How many instances are there in total (of each type, if appropriate)?
There are a total of 2,340 pairs of undistorted and UDC-distorted images.

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 dataset is carefully collected, considering important factors, such as indoor/outdoor environments, daytime/nighttime conditions, flare components (glares/shimmers/streaks) [1], and various light sources. However, images were intentionally captured without personally identifiable information (e.g., vehicle license plates and human faces) to prioritize privacy protection. The dataset does not aim to fully represent the entire spectrum of distortions caused by all UDC products. Instead, it serves as sample images captured specifically by Samsung Galaxy Z-Fold 3 [4] UDC, one of the many UDC products.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Every instance (i.e., image pair) contains the following components:

- An undistorted image and a UDC-distorted image for the same scene in the RAW format (.NPY file).
- Annotations that tell the image-capturing conditions, such as light sources, day/night, indoor/outdoor, and flare components (.TXT file).

Is there a label or target associated with each instance? If so, please provide a description.
We offer annotations for each image pair. Table 1 provides a detailed overview of the total count and distribution of instances of different annotation labels. Note that an instance can have multiple annotation labels. The parenthesized number beside a label is the encoding of the label. The instances are categorized based on image degradation factors, such as indoor/outdoor, day/night, glare/shimmer/streak, and light sources. If images are captured indoors without any window or access to natural sunlight, there would be no distinction between daytime and nighttime. In such cases, we label them as "No distinction." Detailed information can be found at: https://github.com/mcrl/UDC-SIT.

Table 1: Annotation distribution and the number of instances.

Label	# of pairs
Indoor (1)	1,754
Outdoor (3)	586
No distinction (1)	1,340
Day (2)	649
Night (3)	351
Glare (0 or 1)	2,037
Shimmer (0 or 1)	1,899
Streak (0 or 1)	1,067
No flare (0)	273
Natural light (1)	175
Artificial light (2)	1,639
Both (3)	253
Total	**2,340**

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.
There is no missing information.

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.

There are no explicit relationships between individual instances. An instance in the dataset is a pair of undistorted and UDC-distorted images.

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.

We plan to release the dataset with predefined splits into training, validation, and test sets. The dataset is randomly divided into three sets with the following counts: 1,864 for training, 238 for validation, and 238 for testing. This random distribution guarantees that the various annotation types are appropriately represented in each set.

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

Stringent quality checks were conducted to ensure the absence of image pair inconsistencies, alignment discrepancies, and annotation errors in the dataset. We have a dedicated maintenance plan to promptly address and rectify any error the user reports after releasing the dataset publicly, ensuring ongoing data integrity and usability.

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 dataset is self-contained and created without any reliance on external resources. It does not require guarantees or restrictions related to external resources' existence, stability, or licensing.

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. To prioritize privacy protection, the dataset comprises images intentionally captured without any personally identifiable information, such as vehicle license plates or human faces.

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 dataset contains no content that is offensive, insulting, threatening, or likely to cause anxiety.

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

No.

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.

N/A

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.

N/A

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.

N/A.

Any other comments?

N/A.

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 our dataset is directly observable and collected using smartphone cameras without relying on the reported information or indirect inference.

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 in our dataset is collected using a smartphone with a UDC (e.g., Samsung Galaxy Z-Fold 3 [4]) and another smartphone with a non-UDC (e.g., Samsung Galaxy Note 10 [3]). We developed a software program for the alignment of paired images. It uses the discrete Fourier transform. The alignment accuracies of the instances are verified based on the Percentage of Correct Keypoints (PCK) metric following the methodology presented by Feng *et al.* [2], as described in the main body of the paper.

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

No. It is not a sample from a larger set.

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 authors and the members of the Thunder Research Group at Seoul National University created the dataset. They were supported in part by the Institute for Information & communications Technology Promotion (IITP) grant (No. 2018-0-00581, CUDA Programming Environment for FPGA Clusters) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00222663). They were also supported in part by the Samsung Display Co., Ltd. Their average monthly stipend is about 2,000 USD.

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 images are collected from January to May 2023. The data collection period matches the data creation timeframe.

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.

The dataset does not contain any content related to ethical issues. Thus, there is no need to go through the review process of an institutional review board. Instead, a lawyer (the first author's friend) confirmed excluding personally identifiable information from the dataset.

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

No, it does not contain images related 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)?

N/A.

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.

N/A.

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.
N/A.

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).
N/A.

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.
N/A.

Any other comments?
N/A.

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.
The alignment between image pairs is achieved by using the discrete Fourier transform. The instances are labeled with proper annotations, as described in Table 1.

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.
The unaligned "raw" images will be available to individuals upon request.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
We plan to publicly open the software for aligning the images in our GitHub repository. The cleaning and labeling process is not performed using software but through crowdsourcing.

Any other comments?
N/A.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
No, the dataset has not been used for any tasks yet.

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.
No, there are no papers or systems that use the dataset yet.

What (other) tasks could the dataset be used for?
The dataset could potentially be used for tasks related to UDC image restoration.

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?
There is no issue with the dataset leading to unfair treatment of individuals or groups or causing undesirable harm.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset does not impose any restrictions on its usage for specific tasks. Its primary purpose is to facilitate the research on UDC image restoration tasks.

Any other comments?
N/A

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.

The dataset will be publicly available via our GitHub repository.

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

- The dataset is available at: https://github.com/mcrl/UDC-SIT.
- Dataset DOI: https://doi.org/10.5281/zenodo.79769 04.

When will the dataset be distributed?

The dataset will be distributed and made publicly available before the start of the NeurIPS conference.

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.

The dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). This means that users are allowed to freely utilize, share, and modify this work under the condition of properly attributing the original author, distributing any derived works under the same license, and utilizing it exclusively for non-commercial purposes. Detailed information about this license can be found in the official Creative Commons website.

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, no third parties have imposed IP-based or other restrictions on the data associated with the instances. The dataset was created independently, without utilizing any external datasets or sources. Thus, no relevant licensing terms, access points, or fees are associated.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

N/A.

Maintenance

Who will be supporting/hosting/ maintaining the dataset?

The dataset will be supported/hosted/maintained by the authors and the members of the Thunder Research Group at Seoul National University.

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

Users can contact the authors using the email addresses provided in the paper or through the GitHub repository. In addition, they can contact the Thunder Research Group at Seoul National University.

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

There are currently no reported errata for the dataset.

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)?

If users report any labeling errors, the dataset will be promptly updated, and the revisions will be communicated through our GitHub repository.

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.

Since the dataset does not relate to individuals, there are no specific restrictions or requirements regarding data retention.

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

If errors have been reported and updated, older versions of the dataset containing the errors will not continue to be supported/hosted/maintained. However, if new instances are added to the existing dataset, the older versions will continue to be supported/hosted/maintained.

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.

Others can extend the dataset. They can contact us to validate the addition and add it to our dataset. Approved addition will be added to the dataset. If someone wants to contribute additional annotations to the dataset, they can write annotations for each instance and contact us for validation. Approved annotations will be added to the official GitHub repository for distribution.

Any other comments?

N/A.

References

[1] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[2] Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Jinwei Gu, and Chen Change Loy. Generating aligned pseudo-supervision from non-aligned data for image restoration in under-display camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5013–5022, 2023.

[3] Ltd. Samsung Electronics Co. Samsung galaxy note 10, 2019. Available at https://www.samsung.com/my/smartphones/galaxy-note10/specs/.

[4] Ltd. Samsung Electronics Co. Samsung galaxy z fold 3, 2021. Available at https://www.samsung.com/global/galaxy/galaxy-z-fold3-5g/specs/.