RWIDE: A Real-World Image Dehazing Dataset

Xiao Lv[1] **Ying Yang**[2] **Chuan Ma**[1] **Tao Xiang**[1]
[1]College of Computer Science, Chongqing University, China
[2]Agency for Science, Technology and Research (A*STAR), Singapore
{xiaolv, chuan.ma, txiang}@cqu.edu.cn, senereone@gmail.com

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.

Haze degrades the quality of captured images, significantly impacting the performance of various image processing algorithms and vision-driven applications such as image segmentation, object detection, and video surveillance. Current research typically employs synthetic- or artificial-haze image datasets, constraining their utility in real-world contexts. Existing natural-haze image datasets are often deficient in paired samples and fail to offer a broad spectrum of scene diversity. To address these limitations, we introduce the Real-World Image DEhazing dataset (RWIDE), which is the first real-world hazy image dataset with annotations for multiple scenes. RWIDE is primarily collected in "Fog City" Chongqing for nearly two years. As a city nestled in the mountains at the confluence of the Yangtze and Jialing rivers, Chongqing offers optimal high-humidity conditions ideal for capturing haze.

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 Image Processing and Security Laboratory at Chongqing University, on behalf of Chongqing University.

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 is supported in part by the National Natural Science Foundation of China under Grants U20A20176, 62072062, and 62302072.

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 contains pairs of haze-free and hazy images for the same scene. Each pair also has annotations describing the ground and sky scene types. The ground scenes in RWIDE are categorized into six types: Mountains and Hills (MH), Lakes and Rivers (LR), Forests and Jungles (FJ), Buildings and Cityscape (BC), Roadscape (RS), and Snowscape (SS). The sky scenes in RWIDE are classified into five types: No Sky Visible (NSV), Overcast with a Clear Sky (OCS), Sunny with a Clear Blue Sky (SCBS), Overcast with a Cloudy Sky (OCCS), and Sunny with a Blue Sky and some Clouds (SBSC).

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

RWIDE consists of 2,455 pairs of haze-free and hazy images, divided into two parts: RWIDE-α with 305 pairs and RWIDE-β with 2,150 pairs. Additionally, we attach annotations for RWIDE.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Annotation distribution and the number of pairs.

Label	RWIDE-α	RWIDE-β_1	RWIDE-β_2	RWIDE-β_3
Mountains and Hills (**MH**)	17	645	54	24
Lakes and Rivers (**LR**)	30	270	45	9
Forests and Jungles (**FJ**)	112	528	171	82
Buildings and Cityscape (**BC**)	198	273	78	18
Roadscape (**RS**)	175	261	55	250
Snowscape (**SS**)	0	249	13	23
No Sky Visible (**NSV**)	4	249	73	138
Overcast with a Clear Sky (**OCS**)	161	267	110	41
Overcast with a Cloudy Sky (**OCCS**)	98	279	119	26
Sunny with a Clear Blue Sky (**SCBS**)	18	294	56	20
Sunny with a Blue Sky and some Clouds (**SBSC**)	24	411	42	25

These annotations categorize various scene types based on two criteria. (1) Ground scene classification is based on the area where the haze occurs. They are classified into six types: MH, LR, FJ, BC, RS, and SS. (2) Sky scene classification relies on haze-free image sky statuses. They are classified into five types: NSV, OCS, SCBS, OCCS, and SBSC. Table 1 shows the number of images for each annotation type.

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 RWIDE dataset comprises RWIDE-α and RWIDE-β, captured respectively by digital cameras and webcams. Our dataset construction process is bias-free and non-discriminatory, devoid of any personally identifiable information. All data is sourced from publicly available channels. However, it is worth noting that the dataset may not fully represent all types of hazy images found in natural settings, such as those from different angles (e.g., indoor or remote sensing scenarios). Instead, RWIDE-α is primarily suited for human-shot hazy images, while RWIDE-β is more suitable for webcam or surveillance perspectives.

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 (image pair) includes a hazy image and a haze-free image of the same scene in JPG format (.jpg file), along with annotations detailing the ground scene type and sky scene type, as outlined in Table 1 (.csv file).

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

Yes. Annotations are available for each image pair. Table 1 offers a comprehensive breakdown of instance counts and distribution across various annotation labels. Ground scene classifications (MH, LR, FJ, BC, RS, and SS) are interdependent due to overlap. The inclusion of specific scene elements warrants a "1" label; otherwise, it is labeled "0". Conversely, sky scene classification (NSV, OCS, OCCS, SCBS, and SBSC) remains independent. Detailed information can be found at:
`https://github.com/Emily-29/RWIDE`.

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. Each instance in the dataset consists of a pair of hazy and haze-free 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 do not offer predefined partitions for training, validation, and testing sets.

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

Stringent quality checks were performed to eliminate any inconsistencies, alignment discrepancies, or annotation errors in the dataset. We maintain a dedicated plan for ongoing maintenance, promptly addressing and rectifying any reported errors post-release, thereby ensuring continuous 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 developed independently without relying on external resources. It does not necessitate guarantees or constraints regarding the existence, stability, or licensing of external resources.

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. Our dataset construction process is bias-free and non-discriminatory, devoid of any personally identifiable information. All data is sourced from publicly available channels.

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 does not contain any 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.

Our dataset comprises RWIDE-α and RWIDE-β, captured respectively by digital cameras and webcams. Each instance is directly observable and collected 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?

Our dataset comprises RWIDE-α and RWIDE-β. Image pairs in RWIDE-α are manually captured using the ProRAW Max format of an Apple iPhone 15 Pro. Image pairs in RWIDE-β are handpicked from static webcams worldwide. These webcams are sourced from freely available and legal websites [1–4]. We design a two-step progressive method for the image alignment problem. Initially, a lens mask is used for primary alignment, followed by advanced alignment utilizing SuperPoint [5] and LightGlue [6]. The alignment accuracies of the instances are verified based on the Percentage of Correct Keypoints (PCK) metric, following the methodology [7, 8] 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 of the paper and members of the Image Processing and Security Laboratory at Chongqing University were involved in the data collection process. They were supported in part by the National Natural Science Foundation of China under Grant U20A20176, Grant 62072062, and Grant 62302072. Their average monthly stipend is about 400 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.

Images in RWIDE-α were collected from October 2023 to May 2024, while those in RWIDE-β were collected from January 2022 to December 2023. The data collection period aligns with 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.

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

No.

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.

We design a two-step progressive method for the image alignment problem. Initially, a lens mask is used for primary alignment, followed by advanced alignment utilizing SuperPoint [5] and LightGlue [6]. The alignment accuracies of the instances are verified based on the Percentage of Correct Keypoints (PCK) metric as following the methodology [7, 8] described in the main body of the paper. 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.

Yes, the unaligned "raw" images are accessible at:
`https://github.com/Emily-29/RWIDE`.

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 release 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?

This dataset has the potential to be utilized for tasks related to hazy image restoration, such as image dehazing and content 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.

While the dataset does not explicitly impose restrictions on its usage for specific tasks, it is primarily designed to support research focused on hazy image restoration, including image dehazing and content restoration. It may not be ideal for tasks unrelated to image degradation due to haze, such as tasks requiring clear, non-hazy images or those focused on other types of image distortions.

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 accessible to the public through our GitHub repository at:
`https://github.com/Emily-29/RWIDE`.

How will the dataset will 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/Emily-29/RWIDE. Dataset DOI is at: https://doi.org/10.5281/zenodo.11401411.

When will the dataset be distributed?

The dataset will be 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) license. This permits users to freely use, share, and modify the dataset, provided that they properly attribute the original author, distribute any derived works under the same license, and use it exclusively for non-commercial purposes. For detailed information about this license, please refer to the official Creative Commons website at: https://creativecommons.org/licenses/by-nc-sa/4.0.

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 independently created without utilizing any external datasets or sources. Therefore, there are no associated licensing terms, access points, or fees.

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 of the paper as well as the members of the Image Processing and Security Laboratory at Chongqing University, China.

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

Users can reach out to the authors via the email addresses provided in the paper or through the GitHub repository. They can also contact the Image Processing and Security Laboratory at Chongqing University, accessible through the homepage at: https://sites.google.com/site/xiangtaooo/home.

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)?

In case of any reported labeling errors, the dataset will be promptly revised, with updates 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.

As the dataset does not relate to individuals, there are no specific restrictions or requirements concerning 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 are reported and subsequently updated, older versions of the dataset containing those errors will no longer 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 have the opportunity to expand the dataset by contacting us to validate and incorporate their additions. Approved additions will be integrated into the dataset. Similarly, individuals interested in contributing additional annotations can write annotations for each instance and reach out to us for validation. Approved annotations will be included in the official GitHub repository for distribution.

Any other comments?

N/A.

References

[1] Windy. Accessed: 2023. [Online]. Available: https://www.windy.com/

[2] Camera FTP. Accessed: 2023. [Online]. Available: https://www.cameraftp.com/cameraftp/publish/publishedcameras.aspx

[3] Skyline webcams. Accessed: 2023. [Online]. Available: https://www.skylinewebcams.com/

[4] Snowbasin. Accessed: 2023. [Online]. Available: https://www.snowbasin.com/the-mountain/web-cams/

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[6] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[7] X. Zeng, G. Howe, and M. Xu, "End-to-end robust joint unsupervised image alignment and clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3854–3866.

[8] R. Feng, C. Li, H. Chen, S. Li, J. Gu, and C. C. 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*, 2023, pp. 5013–5022.