MOTIVATION.

[question1] 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.
[answer1] This dataset was created to be a drop-in replacement of Omniglot, which is more challenging. Omniglot can hardly push further the state-of-the-art since recent methods achieved almost perfect performances. Furthermore, Omniglot was not intended to be a realistic dataset: the characters were drawn online and do not look natural. The associated task would be the classical N-way-K-shot few-shot classification task [6, 27, 12].

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] Haozhe Sun created the dataset, under the supervision of Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR. ChaLearn also supported the development of the software.

[question3] 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.
[answer3] ANR (Agence Nationale de la Recherche, National Agency for Research, https://anr.fr/), grant number 20HR0134 and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] The instances are 32_32 RGB images of synthetic printed characters.

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] OmniPrint-meta[X] is a collection of five datasets. These 5 datasets, called OmniPrint-meta[1-5], share the same set of characters and data split and only differ in transformations and styles. For each OmniPrint-meta[X] dataset, there are 1409 classes (characters) in total. Each class has 20 image instances. In consequence, each OmniPrint-meta[X] dataset has 1409 _ 20 = 28180 images. There are 28180 _ 5 = 140900 images in total.

[question7] 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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer7] These datasets are synthesized from the data synthesizer OmniPrint, thus they can be viewed as a

sample of instances from all the possible images given the nuisance parameters (fonts, styles, noises, etc.). OmniPrint-meta[X] are representative of such images because the synthesis parameters of each instance were uniformly sampled, no further selection was performed. The involved scripts are Arabic, Armenian, Balinese, Latin, Bengali, Devanagari, Ethiopic, Georgian, Greek, Gujarati, Hebrew, Hiragana, Katakana, Khmer, Lao, Mongolian, Myanmar, N'Ko, Oriya, Russian, Sinhala, Tamil, Telugu, Thai and Tibetan.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Each instance is a 32_32 RGB image. Each image contains one single character from a certain script, rendered in a particular way (background, foreground, distortions, noises).

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Yes, there is a label (character) associated with each instance. Furthermore, the metadata is provided
for each instance, which can also serve as labels for specific tasks. The metadata includes e.g., the font, background, stroke width (if applicable), blur radius, margins, rotation angle, shear, text color, etc., and the alphabet of the character.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] No. All of the metadata is provided for each instance.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] All relationships are contained in the labels and metadata, all provided.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] Yes, there is a recommended data split in the context of N-way-K-shot learning, between meta-train,
meta-validation and meta-test. For each of the 5 OmniPrint-meta[X] datasets, there are 1409 classes (characters), each class contains 20 image instances. The first 900 classes belong to meta-train, then 149 classes belong to meta-validation, the last 360 classes belong to meta-test. This data split is chosen in order to imitate the proportion of meta-train/meta-validation/meta-test of the popular Vinyals split [33] of Omniglot [16]. The recommended data split is provided via a data loader which forms the episodes of few-shot learning.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] We intentionally introduced various transformations and noises to each image instance. The transformation
parameter space is large so there is little chance that two instances are identical.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that

is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
[answer14] The 5 datasets OmniPrint-meta[X] are self-contained. They will exist, and remain constant, over time
once we release them after the NeurIPS 2021 meta-learning challenge.
[question15] Does the dataset contain data that might be considered confidential (for example, 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.
[answer15] The OmniPrint-meta[X] datasets were considered confidential before the NeurIPS 2021 meta-learning
challenge, they have been publicly released.
[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16] No.
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] YES
[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]
[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]
[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.
[answer19]
[question20] Any other comments?
[answer20]

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] Each instance is synthesized by OmniPrint. Each instance is an image and is directly observable.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] The data are synthesized using the data synthesizer OmniPrint. The involved Unicode characters were manually selected from the Unicode standard, which constitutes a set of characters from several languages around the world. The involved fonts were downloaded from a manually-defined list of URLs, the downloaded fonts were then filtered by a python program in order to filter corrupted fonts. Several distortions and noises were involved, including affine and perspective transformations, random elastic transformations, natural background, foreground text filling, etc.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] The data is synthesized by a data synthesizer OmniPrint. The sampling is uniformly random in the given transformation parameter space.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] The data is synthesized by a computer software. However the design and implementation of the software, the choice of characters and fonts involve the authors of this paper.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] The five datasets were synthesized on May 22, 2021.

[question26] Were any ethical review processes conducted (for example, 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.

[answer26] N/A

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

[SkipB] YES

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27]

[question28] 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.
[answer28]
[question29] 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.
[answer29]
[question30] 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).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.
[answer31]
[question32] Any other comments?
[answer32]

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.
[answer33] No preprocessing/cleaning/labeling was performed. The datasets are made available as they were synthesized. No feature extraction or removal of instances was done.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] N/A
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35] N/A
[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37] No, however a variant of these datasets will be used by the NeurIPS 2021 meta-learning challenge.
[question38] 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.
[answer38] Yes, the link is https://github.com/SunHaozhe/OmniPrint-datasets. This repository is also used to announce any necessary information related to the OmniPrint datasets e.g., potential changes of the dataset hosting address.
[question39] What (other) tasks could the dataset be used for?
[answer39] Besides few-shot learning classification tasks, the five OmniPrint-meta[X] datasets can be used for classification tasks of a large number of characters, and for transfer learning (each dataset being used either as a source domain or a target domain). Furthermore, as the metadata can serve as labels, other kinds of classification or regression problems can also be considered e.g., classification of fonts, classification of languages, regression of rotation angle, regression of horizontal shear, etc. Finally, the datasets can be used to study disentangling the label (class character) from the nuisance variables (font, style, distortions).
[question40] 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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
[answer40] The datasets can be used without further considerations.
[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.
[answer41] Not that we know of.
[question42] Any other comments?
[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
[answer43] The datasets are made available to everyone via the Internet.
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44] The OmniPrint-meta[X] datasets are publicly released via Kaggle Datasets. The digital object identifier (DOI) is 10.34740/kaggle/dsv/2763401. The access information and any necessary updates

are announced via https://github.com/SunHaozhe/OmniPrint-datasets.

[question45] When will the dataset be distributed?

[answer45] The datasets have been released after the NeurIPS 2021 meta-learning challenge.

[question46] 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.

[answer46] The datasets OmniPrint-meta[1-5] are distributed via Kaggle datasets. They are licensed under a Creative Commons license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/. This comes with the following guarantee disclaimer: Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

[question47] 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.

[answer47] No.

[question48] 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.

[answer48] No.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The authors of this paper are responsible for supporting the datasets.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The preferred way to contact the maintainers is to raise issues on https://github.com/SunHaozhe/OmniPrint-datasets. In case of emergency, the authors of this paper can be contacted via email: omniprint@chalearn.org.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] Any necessary information or updates will be accessible via https://github.com/SunHaozhe/OmniPrintdatasets.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] No. New needs will be met by synthesizing new datasets.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.

[answer54]

[question55] 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 dataset consumers.

[answer55] N/A

[question56] 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 dataset consumers? If so, please provide a description.

[answer56] Any necessary information or updates will be accessible via https://github.com/SunHaozhe/OmniPrintdatasets.

[question57] Any other comments?

[answer57] No.