Supplementary Material

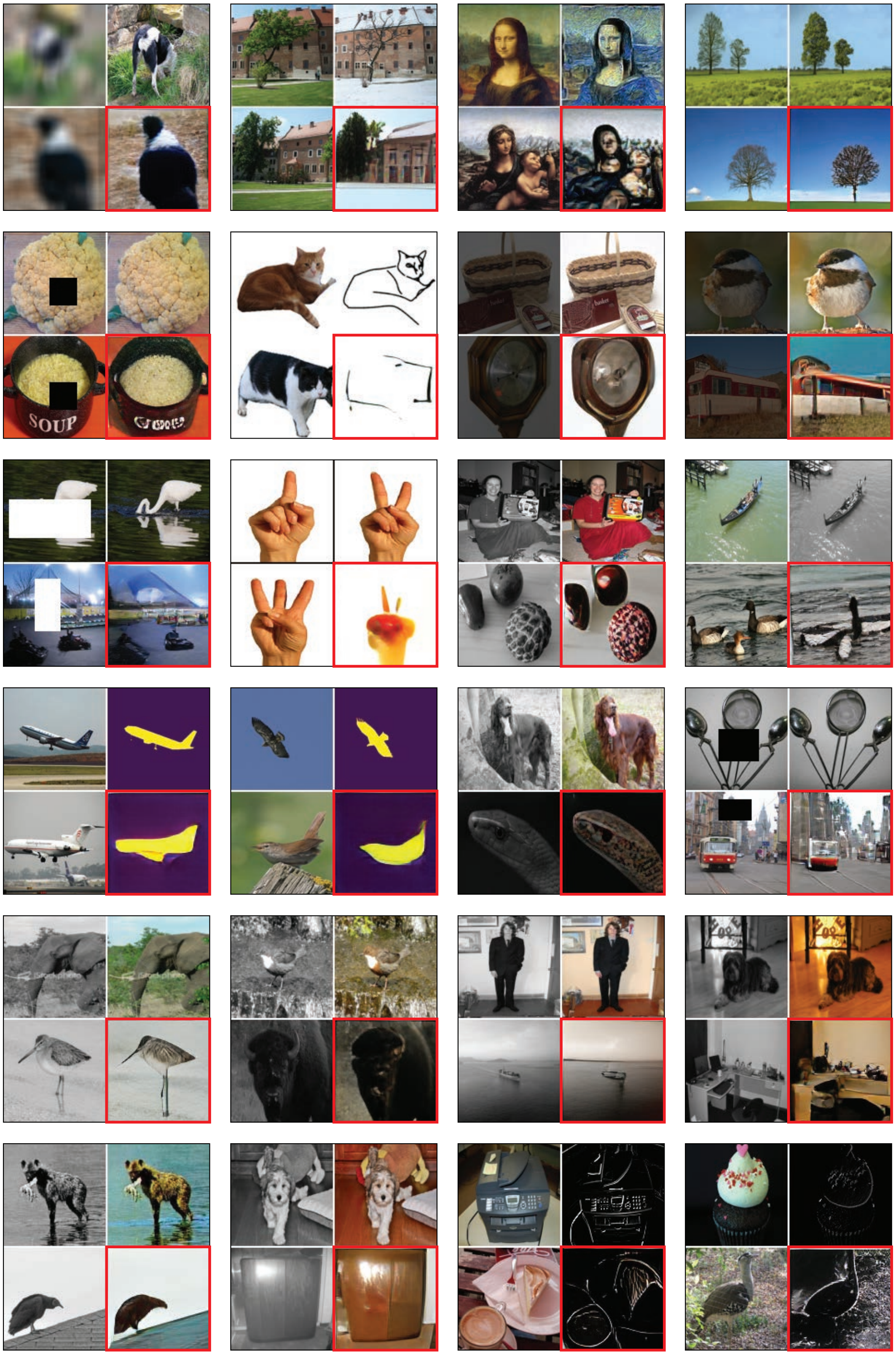

Figure 12: **Additional visual prompting results**. For each visual prompt, the result is marked in red.

We include more information about the experimental study, as well as the Computer Vision Figures Dataset datasheet. Additional visual prompting results are included in Figure 12.

6 Experiments

6.1 Downstream Computer Vision Tasks

Qualitative model comparison. We include a qualitative comparison of visual prompting results when applied to Foreground Segmentation using different inpainting models (see Figure 13). Compared to other models, MAE-VQGAN outputs more smooth and accurate segmentation results.

Single Object Detection. We include qualitative results of MAE-VQGAN when applied to Single Object Detection (see Figure 14). As mentioned in Section 4.2, the raw output is rounded and postprocessed using morphological operations to convert the foreground segmentation to a box.

The effect of the input-output example. We fix the input query and change the input-output examples and visual prompt MAE-VQGAN (see Figure 15). Different examples could lead to slightly different synthesis results. However, as long as the example is meaningful (e.g, not fully background or foreground) the model results are plausible regardless of the choice of example.

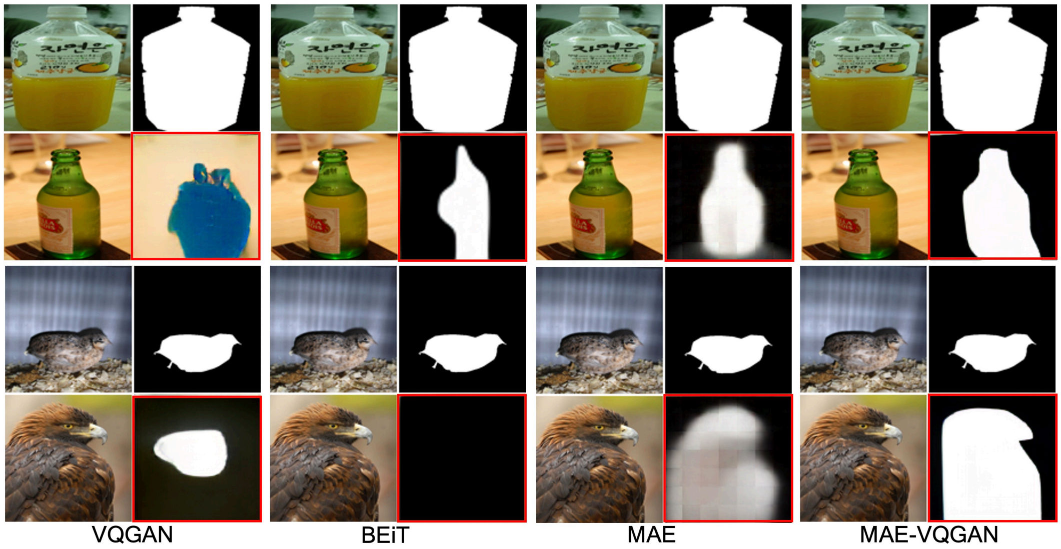

Figure 13: **Visual Prompting for Foreground Segmentation using different models**. Compared to other models, MAE-VQGAN produces more smooth and accurate results.

Visual Prompting results on computer vision tasks. We include the mean and standard deviation for each Visual Prompting model on Pascal 5i (see Table 6).

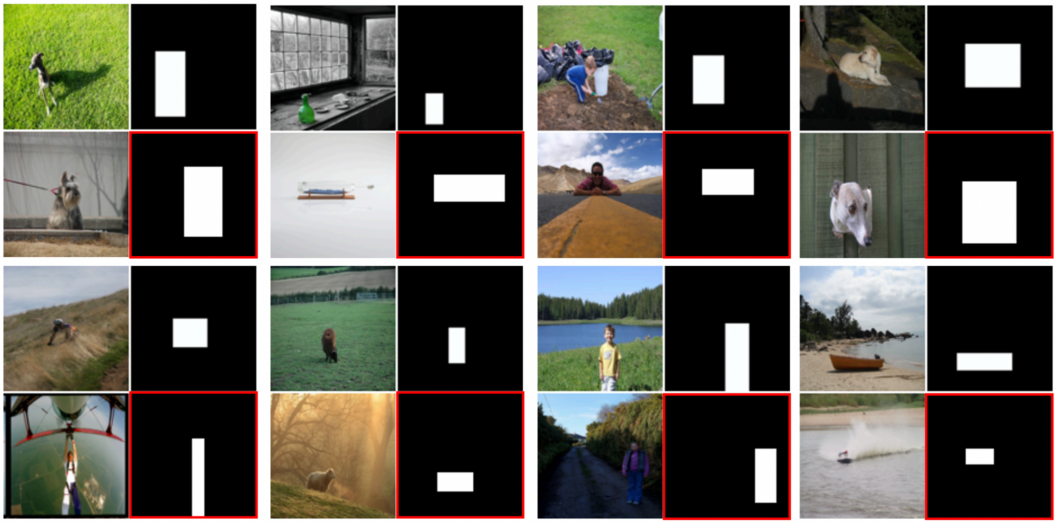

Figure 14: **Visual prompting applied for Single Object Detection.** The raw MAE-VQGAN results are rounded and post processed using morphological operations.

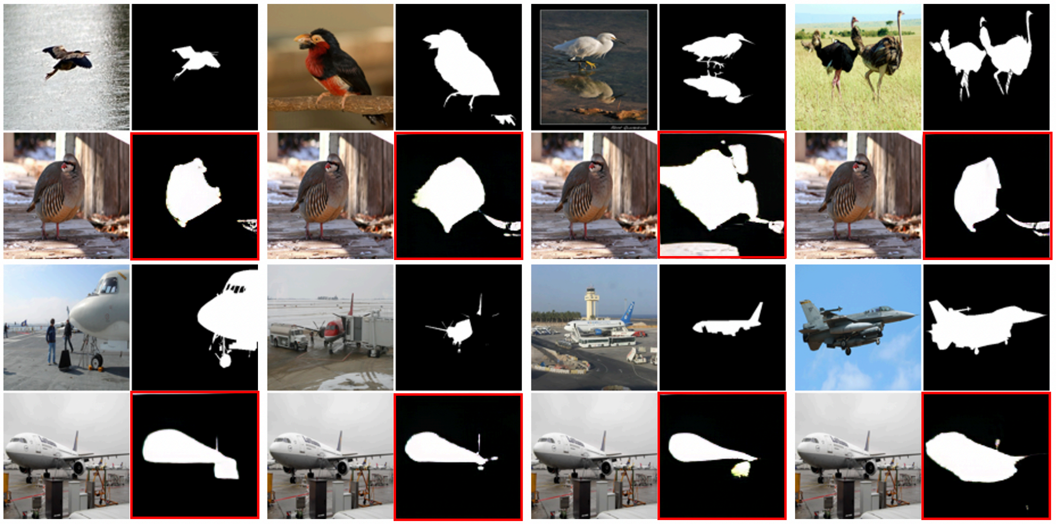

Figure 15: **The effect of input-output examples on prompting results.** We fix the input query and change the input-output example. Different input-output examples could lead to slightly different results.

Table 6: **Visual prompting results on computer vision tasks.** We report mean and standard deviation of mIOU scores for Foreground Segmentation and Single Object Detection.

Model	Foreground Segmentation		Single Object Detection	
	Mean	Std	Mean	Std
Copy	14.91	1.93	12.76	0.54
BEiT (IN-21k)	0.79	0.24	0.21	0.08
VQGAN (IN-1k)	9.13	1.33	5.09	0.07
MAE (IN-1k)	4.28	1.73	1.65	0.22
MAE-VQGAN (IN-1k)	5.26	1.84	3.04	0.24
BEiT (Figures)	3.95	0.87	0.12	0.06
VQGAN (Figures)	14.85	1.78	2.28	0.18
MAE (Figures)	19.57	3.62	5.39	0.32
MAE-VQGAN (Figures)	**27.17**	2.27	**25.00**	0.47

7 The Computer Vision Figures Dataset Datasheet

In Section 3.3 we presented the Computer Vision Figures dataset. The full dataset will be made available upon acceptance. Next, we include the dataset datasheet.

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 Computer Vision Figures (Figures) dataset was proposed enable our approach for Visual Prompting. In our setup, a Visual Prompt is a single image that have a grid-like figure structure that stitches together images coming from different distributions, like natural images and segmentation masks. Therefore, a model trained on a standard dataset (e.g., ImageNet [42]) might struggle to process these grid-like images. To mitigate the domain gap, we collected a new dataset with images that more closely resemble the proposed Visual Prompt structure.

The dataset was collected from Arxiv, the open-access web archive for scholarly articles from a variety of academic fields. Arxiv sources are publicly available to download and the Computer-Vision partition contains images that more closely resemble a grid structure, as shown in Figure 3.

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

N/A

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.

N/A

Any other comments?

No.

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.

Each instance in the dataset is an image, which is a single arXiv paper figure.

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

The data is comprised of $88,645$ images, partitioned to 90% train and 10% validation. We include descriptive statistics of the data in Table 7. We find that around 40% of the images in the data does not include any embedded annotation. The majority of the data (84%) is comprised of grid, figure-like images. To obtain these statistics, 100 random Figures images were manually labeled by a human tagger as a single image type ("single") or a grid-like image type ("grid"). In addition, if the image contained an annotation, the tagger picked the annotation category from a predefined list (e.g, mask, box, pose, heatmap) or "Other" if the annotation doesn't match any of these categories.

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 based on Arxiv papers figures from 2010 to 2022 in the the Computer-Vision partition "cs.CV". It only consists of figures with at least one natural image. To remove unrelated source

Table 7: **Computer Vision Figures Dataset Annotations Statistics.** The results were obtained by manually labeling 100 images.

Annotation	Type	%
No Annotation	Grid	37
Other	Grid	15
Box	Grid	11
Mask	Grid	9
Heatmap	Grid	8
Box	Single	6
No Annotation	Single	5
Pose	Grid	4
Other	Single	3
Pose	Single	1
Heatmap	Single	1

images like graphs or charts, we manually tagged 2000 images and trained a binary image classifier to assign a high score to source images in a figure-like structure with at least one natural image. We then used the classifier over the entire data to keep only the most informative source images, coming from $23,302$ different papers. In a manual review of 100 random figures, we found that the dataset is 96% clean of graphs or charts.

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 is in PNG image format with resolution up to 1024×1024. Each image has an accompanied Arxiv paper id and an indicator to an associated train/val partition.

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

There are no labels associated with each instance.

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.

All instances are complete.

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.

Images can have a similar or different arXiv source paper and we release this information.

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.

The dataset was randomly split into $90\%/10\%$ train and validation. In this paper, we only made use of the train partition for training on unlabeled figures. In future works, the validation partition may be used in different ways, like hyper-parameters tuning or for the evaluation of generative models.

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

Noise. In a manual review of 100 random figures, we found that the dataset is 96% clean of unintended graphs or charts. Naturally, different figures could have potential overlap and we acknowledge this potential redundancies.

Overlap with Computer Vision Datasets. The Computer Vision Figures dataset is a collection of figures from different computer vision papers. Therefore, some of the images may overlap with computer vision datasets test, with or without annotations. To evaluate to what extent this is the case, we randomly sampled 100 Pascal 5i validation images and for each image computed its 10 nearest neighbors in Figures. Given the pascal image and a Figures image, we used the OpenCV template matching function ("cv.matchTemplate") that operates in a sliding window to compute similarity in pixel space using mean squared error. We also attempted to use an ImageNet pretrained ResNet50 to compute similarity and found that they work worse, likely because the Figures dataset contains out-of-distribution images. A human tagger examined the results and did not find duplicates. In 72% of the cases the nearest neighbors retrieved were irrelevant, and in 28% there was at least one Figures image that is somewhat similar (e.g, the pascal image contains a person and an image of a different person was retrieved). We conducted an additional independent check to evaluate the overlap. For each image of 100 random Pascal 5i test images we computed the 5 nearest neighbors in the Figures datasets using CLIP embeddings. Out of the 100 images, we found only a single one contained in a Figure image, and this image did not have any associated ground-truth annotation. Therefore, we conclude that even if there are potential overlaps they are likely very small and insignificant.

Imprecise Embedded Annotations. Instances in the dataset are images that may include an embedded annotation. Naturally, this annotation might be noisy, incomplete or just imprecise.

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.

We will share and release links to the relevant Arxiv sources and figures, as well as an automated code for downloading and preprocessing the sources which is consistent with the Arxiv license (Non-exclusive license to distribute). Arxiv is a reliable source that is backed up by the entire research community. As commonly done, we may allow an alternative download option of the dataset on a "fair use" basis.

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.

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.

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?

No.

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 was directly observable, e.g, automatically extracted from Arxiv sources.

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 was downloaded in accordance with the offical Arxiv guidlines for data access: `https://arxiv.org/help/bulk_data`.

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

The dataset is based on Arxiv paper figures from 2010 to 2022 in the the Computer-Vision partition "cs.CV". It only consists of figures with at least one natural image. To remove unrelated source images like graphs or charts, we manually tagged 2000 images and trained a binary image classifier to assign a high score to source images in a figure-like structure with at least one natural image. We then used the classifier over the entire data to keep only the most informative source images, coming from $23,302$ different papers. In a manual review of 100 random figures, we found that the dataset is 96% clean of graphs or charts.

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 team of students who worked together on this project. None of the team members were compensated for this work, beyond their regular compensation for the position they held.

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 papers and their accompanied sources were collected by Arxiv as part of their normal work protocols between 2010 to 2022. The dataset was created in 2022 based on these sources.

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.

N/A

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?

No.

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.

To remove unrelated source images like graphs or charts, we manually tagged 2000 images and trained a binary image classifier to assign a high score to source images in a figure-like structure with at least one natural image. We then used the classifier over the entire data to keep only the most informative source images, coming from $23,302$ different papers. In a manual review of 100 random figures, we found that the dataset is 96% clean of graphs or charts. Each kept figure image was then resized to a resolution of up to 1024×1024 and saved in a PNG format.

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.

We will release a list of links to the raw figures source files.

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

The software to download, extract, and preprocess the images will be made publicly available.

Any other comments?

No.

Uses

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

The dataset was used for unsupervised learning algorithms, e.g, for pretraining inpainting models.

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.

A partial list of papers that use the dataset will be made available after the review stage.

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

Potential other use cases include generative modeling, image retrieval and explainabilty.

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?

We do not anticipate any negative biases in the data or potential harms. Additionally, since Arxiv is moderated we do not anticipate the presence of offensive content.

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

N/A

Any other comments?

No.

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 made publicly available after the review period.

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 will be distributed as a comma-separated (.csv) file describing the Arxiv links, partition, and paper id, containing the original figure sources. Additionaly, we will provide a direct download of the dataset in the form of a tarball on a "fair-use" basis. The dataset DOI will be the same as the one of this work.

When will the dataset be distributed?

The dataset will be made available after the review period.

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 comma-separated (.csv) file and accompanying code will be distributed under the MIT license.

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.

Arxiv holds a (non-exclusively) license to distribute all submitted papers and sources (`https://arxiv.org/licenses/nonexclusive-distrib/1.0/license.html`). Publishing the download links for Arxiv sources is in compliance with this license. As commonly done, we may allow an alternative download option of the dataset on a "fair use" basis.

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?

No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the paper will be maintaining the dataset, which will be hosted on GitHub.

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

We will post the contact information after the review period.

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

No.

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)?

There are no plans to update the dataset at this time.

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.

N/A

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.

No.

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.

Contributions will be made possible using standard open source tools, submitted as pull request to the relevant GitHub repository.

Any other comments?

No.