MULTIMODAL DATASET UPGRADING: A NEW CHALLENGE FOR DATA ANNOTATION

Haiwen Huang[1,2] **Dan Zhang**[1,4] **Andreas Geiger**[2,3]
[1]Bosch Lab, University of Tübingen
[2]Autonomous Vision Group, University of Tübingen
[3]Tübingen AI Center [4]Bosch Center for Artificial Intelligence
{haiwen.huang, a.geiger}@uni-tuebingen.de, Dan.Zhang2@de.bosch.com

ABSTRACT

In recent years, many large-scale datasets become available, yet their annotations are coarse and noisy. In this paper, we propose a novel task of *multimodal dataset upgrading* to enhance the quality of multimodal annotations. Distinguishing from traditional annotation efforts that focus on creating labels from scratch, multimodal dataset upgrading seeks to refine existing annotations by increasing annotation granularity, reducing errors, and improving multimodal alignment. We propose a framework for tackling multimodal data upgrading, consisting of generating candidates for upgrading and cross-modality matching to select the upgraded data. We further provide a case study on open-vocabulary segmentation datasets where by improving the class name quality, we achieve significant performance enhancements in state-of-the-art open-vocabulary segmentation models. As an initial exploration, we hope this paper showcases the benefits of data upgrading and opens up new avenues for research in data problems for foundation models.

1 INTRODUCTION

In recent years, many large-scale datasets such as Youtube-8M (Abu-El-Haija et al., 2016), Conceptual Captions (Sharma et al., 2018), and LAION (Schuhmann et al., 2021) have become available and widely used. Despite their large scale, most of these datasets are webscraped data with noisy annotations such as image-caption pairs. It therefore remains a question whether we can dig more value out of these datasets by generating more accurate and fine-grained annotations. In this paper, we propose a new task called *data upgrading*, which is a type of data annotation that aims to improve the data annotation quality for increased usefulness of the dataset. The improvement of data quality can be in many folds such as higher granularity, fewer annotation errors, and more aligned multimodal matching. The overall goal is to enhance the original annotations, which in turn contributes to stronger foundation models and downstream models.

In this paper, we specifically address the problem of upgrading multimodal datasets. Multimodal data are central to most foundational models and the various modalities allow humans to interact with the model in various ways (Radford et al., 2021; Rombach et al., 2022). We first provide a general framework to tackle the multimodal data upgrading problem by dividing it into two subtasks: candidate generation and cross-modality matching. We then discuss some typical examples and potential solutions of multimodal data upgrading. Furthermore, we conduct a case study on segmentation datasets such as ADE20K (Zhou et al., 2017) and Cityscapes (Cordts et al., 2016). By examining the problems of current annotations and improving the class names, we show that current state-of-the-art open-vocabulary segmentation models perform significantly better when using our upgraded names. Finally, we conclude with a discussion on the limitations of our approach and the promising research opportunities in this new challenge for data annotation.

2 A FRAMEWORK FOR MULTIMODAL DATA UPGRADING PROBLEM

Based on the common use cases of datasets, the goals of data upgrading can be identified as improving the correctness and the granularity of the dataset, thereby enriching it with more valuable

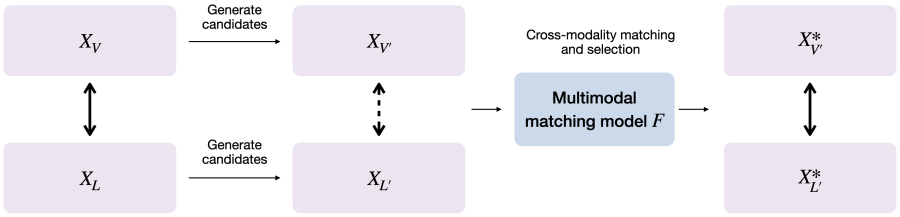

Figure 1: An illustration of the proposed framework for multimodal data upgrading

Goal	Input data (X)	Candidates (X')	Final outputs (X^*)
Improving Correctness	2D masks with names	masks ($X_{V'} = X_V$) with new candidate names ($X_{L'}$)	aligned masks and their names
	3D masks with names	new mask proposals ($X_{V'}$) with names ($X_{L'} = X_L$)	aligned 3D masks with their names
	VQA datasets	new mask proposals ($X_{V'}$) with referring expressions ($X_{L'}$)	aligned masks with corresponding expressions
Improving Granularity	Image-caption datasets	mask proposals ($X_{V'}$) and candidate names ($X_{L'}$)	images with masks and their corresponding names
	Images with class agnostic masks	candidate object names ($X_{L'}$)	images with masks and their corresponding names
	Videos with captions	object trajectories ($X_{V'}$) and verbs for movement ($X_{L'}$)	object trajectories and the words describing their movements

Table 1: Examples of multimodal data upgrading.

information. For multimodal data upgrading, the "correctness" goal includes both the correctness in different modalities as well as their alignment, posing a distinct challenge from unimodal data upgrading. And the "granularity" goal means to inject more information into the dataset by adding more annotations such as hierarchies in names or dense visual annotations.

To this end, we propose a general framework to tackle the multimodal data upgrading problem by decomposing it into two subtasks, as illustrated in Fig. 1. First is the *candidate generation task*, where we create candidate proposals for data in one or multiple modalities. The generated candidates serve as the candidate pool for finer granularity or error correction. Second is the *cross-modality matching task*, where the goal is to match and select the best-matching candidates from different modalities. Formally, given a multimodal dataset $X = \{X_V, X_L\}$ where X_V, X_L are paired (denoted as $X_V \leftrightarrow X_L$) and are in two different modalities, e.g., images and their captions. We first generate candidate pools $X' = \{X_{V'}, X_{L'}\}$ in both modalities, e.g., masks and names. We note that the candidate generation should be information non-decreasing, i.e., X' should preserve all the information in X and may also augment it with additional information such as name synonyms or mask proposals. Next, we do cross-modality matching with a multimodal matching model $F : X_{V'} \times X_{L'} \to [0, 1]$ to match and select pairs in X' based on certain rules. For example, when $X_{V'}$ and $X_{L'}$ are mask proposals and candidate names, given a candidate name $x^*_{L'}$, we can select the best matching mask proposal $x^*_{V'} = \arg\max_{V'} F(x_{V'}, x^*_{L'})$ and keep this pair if $F(x^*_{V'}, x^*_{L'}) > 0.5$. After cross-modality matching and selection, we get our upgraded dataset $X^* = \{X^*_{V'}, X^*_{L'}\}$ where $X^*_{V'} \leftrightarrow X^*_{L'}$. In Table 1, we provide many examples of data upgrading under this framework. We further note that for the choice of the cross-modality matching function F, it is often the case that we will not have a pre-trained model ready for use at the same granularity. For example, CLIP Radford et al. (2021) are great for image-caption matching, but are suboptimal for dense tasks like mask-name matching. In this case, it requires to either fine-tune the pre-trained models or train from scratch. In the next section, we further provide a detailed study of a specific data upgrading problem.

3 Case study: Upgrading class names in segmentation datasets

3.1 Problems of current class names in segmentation datasets

Class names in currently established segmentation datasets have not been designed with multimodal tasks in mind. In fact, most datasets are labeled with class names that serve merely as identifiers to distinguish classes within a dataset, rather than descriptive labels for tasks like open-vocabulary generalization. As shown in Fig.2, existing class names are often inaccurate, too general, or lack enough context for state-of-the-art open-vocabulary segmentation models (Yu et al., 2023) to perform well.

3.2 Data upgrading pipeline

In this case study, we aim to upgrade the class names of current segmentation datasets. Following the framework in the last section, we divide the task into the candidate generation task and the cross-

modality matching and selection task. Specifically, for a segmentation dataset $X = \{X_V, X_L\}$ where X_V and X_L are masks and their corresponding class names, we first generate new name candidates $X_{L'}$ and keep the masks unchanged, i.e., $X_{V'} = X_V$, and then train a cross-modality model to match masks and name candidates as in Case 3 in the last section.

3.2.1 CANDIDATE GENERATION

We use GPT-4 (OpenAI, 2023) for creating a pool of class name candidates. To this end, a naive solution is to prompt GPT-4 with the original class name and ask it to generate synonyms and hierarchical concepts. However, since the original names are often too general, GPT-4 does not have sufficient knowledge to generate high-quality candidates. Therefore, we propose to exploit the visual contents for generating "context names" that assist GPT-4 in comprehending the category's meaning prior to generating candidate names.

As shown in Fig. 3, for each category, we use an image captioning method to process all training images that contain that specific category based on ground-truth annotations. From the generated captions, we further extract nouns by text parsing and filtering as done in CaSED (Conti et al., 2023). We aggregate and sort the extracted nouns based on their frequency and designate the top 10 most recurrent nouns as the *context names* for each category. We observe that these names offer deep insights into the typical traits associated with the category and we use them as additional inputs alongside the original class names to prompt GPT-4.

Figure 2: **Problems of class names in current datasets** (MS COCO, ADE20K, and Cityscapes). Class name upgrading significantly enhances open-vocabulary segmentation performance.

3.2.2 CROSS-MODALITY MATCHING AND NAME SELECTION

Due to a lack of high-quality datasets for the vision-language segmentation task, to perform cross-modality matching and selection, we propose to train a matching model on the datasets to be upgraded. Specifically, we repurpose the meta-architecture of Mask2Former Cheng et al. (2022) for cross-modality matching by using the text embeddings as the input queries to the transformer decoder and the frozen CLIP vision encoder as the visual backbone, as illustrated in Fig. 3. We also incorporate groundtruth masks as attention biases in the transformer decoder to provide guidance of the regions to be matched. Our matching objective is the quality of the predicted masks, i.e., IoU between the predicted masks and the target groundtruth masks. This is based on the assumption that a candidate name that well describes the segment in the image should help the segmentation model recover the ground-truth mask of the segment. Our training objective consists of both the matching loss and the classification loss as regularization to cluster name embeddings from the same class.

3.3 EVALUATING THE UPGRADED CLASS NAMES

To evaluate the upgraded class names, we first conduct a human evaluation test where we ask 20 vision-language researchers about their preferences between the upgraded names and the original names when presented with the corresponding image segments. In 100 image segments we study, our upgraded names are preferred in 76% cases, showing a clear advantage.

Next, we propose to use the performance on the downstream task of open-vocabulary segmentation to evaluate the name quality. Intuitively, if the name quality is higher, i.e., better matching the visual contents, the same open-vocabulary segmentation models should perform better. Specifically, we use three state-of-the-art pretrained open-vocabulary segmentation models and evaluate them on datasets

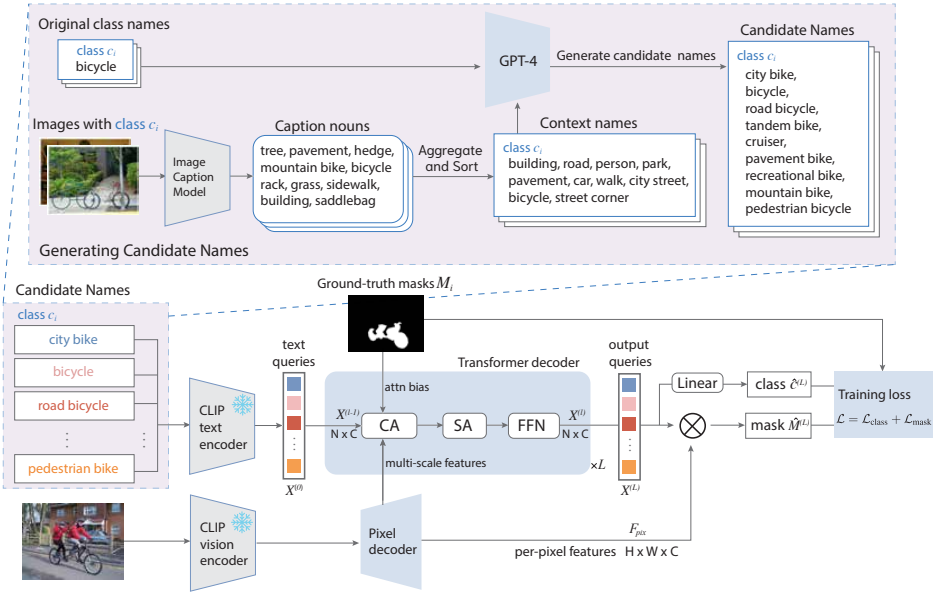

Figure 3: An overview of candidate generation and cross-modality matching for name upgrading.

Model	Test names	ADE20K PQ	ADE20K AP	ADE20K mIoU	Cityscapes PQ	Cityscapes AP	Cityscapes mIoU
ODISE	Original class names	21.88	13.94	29.16	39.72	27.73	49.60
	OpenSeg names	21.63	13.95	28.88	43.26	**28.45**	54.53
	Upgraded names	**23.69**	**14.38**	**31.64**	**43.61**	28.38	**57.42**
MasQCLIP	Original class names	23.46	12.80	30.32	33.78	18.11	45.35
	OpenSeg names	23.70	12.84	31.17	35.05	17.79	46.46
	Upgraded names	**25.00**	**12.93**	**32.30**	**35.63**	**18.19**	**50.07**
FC-CLIP	Original class names	24.30	16.79	32.14	39.42	22.76	53.08
	OpenSeg names	25.35	17.30	33.06	43.68	26.93	56.15
	Upgraded names	**28.10**	**17.88**	**37.35**	**45.90**	**29.79**	**62.55**

Table 2: **Open-vocabulary segmentation evaluation.** We use ODISE Xu et al. (2023a), MasQ-CLIP Xu et al. (2023b), and FC-CLIP Yu et al. (2023) pre-trained checkpoints released by the original papers. Our results demonstrate that open-vocabulary performances of the pre-trained models were underestimated due to the inappropriate class names.

ADE20K Zhou et al. (2017) and Cityscapes Cordts et al. (2016) with different class names. We compare our upgraded names with the original class names and OpenSeg names Ghiasi et al. (2022) which are modified from the class names by manually inspecting the segmentation benchmarks. We use standard metrics, panoptic quality (PQ), Average Precision (AP), and mean Intersection-over-Union (mIoU), for assessing panoptic, instance, and semantic segmentation, respectively. To infer with multiple names, we follow prior work Yu et al. (2023) to choose the highest logit prediction per class.

In Tab. 2, we demonstrate that upgraded names significantly boost the performance of pre-trained open-vocabulary models. This shows that open-vocabulary performances of these pre-trained models were underestimated due to inappropriate old class names. Particularly, with the FC-CLIP model, our approach achieves $\sim$16% PQ improvement on both ADE20K and Cityscapes benchmarks.

4 Conclusion

This paper proposes the task of data upgrading, aimed at enhancing data annotation quality to increase dataset utility. We concentrate on multimodal data upgrading and present a framework with examples and case studies to address this challenge. We hope our work brings attention to a novel issue in data quality research and inspires future studies.

5 LIMITATIONS

Our work is only a preliminary study on the data upgrading problem and our case study is also far from complete. We first note that there exist many other possibilities for both the candidate generation step and the cross-modality matching step in our case study. For example, some retrieval-based methods for candidate generation may be considered (Parashar et al., 2024). Also, the training of the cross-modality matching function is supervised learning, but it can also be semi-supervised (e.g., using COCO and a large-scale unlabeled dataset) or even self-supervised. For different learning paradigms, other cross-modality matching objectives such as contrastive losses (Radford et al., 2021) or masked predictions (Lu et al., 2022; Wang et al., 2022) should also be explored or combined with the current segmentation-based objective. These losses are also generally applicable to cross-modality matching in many other modalities.

In addition, we also note that our evaluation of data upgrading is very preliminary. Even in our case study, we only use two ways to validate the quality of our upgraded names, i.e., human evaluation and open-vocabulary segmentation. In practice, we think multiple downstream tasks should be considered to fully understand the quality of the upgraded datasets. We also encourage more research in studying the value of data upgrading beyond downstream tasks, e.g., dataset coverage analysis and dataset monitoring.

6 SOCIETAL IMPACT

This work proposes the task of multimodal dataset upgrading which aims to improve the quality of current datasets by improving their correctness and granularity. The data upgrading process is a potential way to address the fairness and bias issues in the current datasets. For example, in our case study of class name upgrading, we will be able to have more detailed understanding of the dataset subclasses (e.g., we will know that the "person" class is composed of "man, woman, boy, girl,...") through the cross-modality matching and name selection process. This gives us tools to further filter and process the datasets to control both the quality and societal impacts of the datasets. In addition, the data upgrading paradigm aims to develop an automatic tool to improve the current publicly available large-scale datasets. This will contribute to the development of stronger open-source models and benefit researchers and users who do not have access to the models for commercial uses.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *arXiv preprint arXiv:2306.00917*, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. June 2016.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. 2022.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

R OpenAI. Gpt-4 technical report. *arXiv*, 2023.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. 2023a.

Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. 2023b.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. 2017.