# Medical Vision-Language Pre-Training for Brain Abnormalities

**Masoud Monajatipoor, Zi-Yi Dou, Aichi Chien, Nanyun Peng, Kai-Wei Chang**
UCLA
{monajati, zdou, aichi}@ucla.edu
{violetpeng, kwchang}@cs.ucla.edu

**Abstract**

Vision-language models have become increasingly powerful for tasks that require an understanding of both visual and linguistic elements, bridging the gap between these modalities. In the context of multimodal clinical AI, there is a growing need for models that possess domain-specific knowledge, as existing models often lack the expertise required for medical applications. In this paper, we take *brain abnormalities* as an example to demonstrate how to automatically collect medical image-text aligned data for pretraining from public resources such as PubMed. In particular, we present a pipeline that streamlines the pre-training process by initially collecting a large brain image-text dataset from case reports and published journals and subsequently constructing a high-performance vision-language model tailored to specific medical tasks. We also investigate the unique challenge of mapping subfigures to subcaptions in the medical domain. We evaluated the resulting model with quantitative and qualitative intrinsic evaluations. The resulting dataset and our code can be found here https://github.com/masoud-monajati/MedVL_pretraining_pipeline

**Keywords:** vision-language, pre-training, medical

## 1. Introduction

Readily available open-access datasets in the broad domain (Lin et al., 2014; Sharma et al., 2018) have enabled the development of vision-language (VL) models. Typically, researchers pre-train VL models on large image-text data and then fine-tune them for specific downstream tasks. Such a pre-training/fine-tuning paradigm is highly effective for tasks with limited data and therefore is a standard approach for various downstream applications.

On the other hand, the medical domain presents unique challenges due to the scarcity and complexity of its data (Brigato and Iocchi, 2021). Pre-trained models, trained on general domain datasets, often exhibit reduced effectiveness when applied to a medical domain with limited data, as observed in the cases of chest x-ray analysis (Wu et al., 2023; Monajatipoor et al., 2022) and Alzheimer's disease detection (Chen et al., 2023). Furthermore, domain-specific data suitable for pre-training is notably limited. Although several domain-specific medical Vision-Language (VL) datasets do exist, such as MIMIC (Johnson et al., 2019), CheXNet (Rajpurkar et al., 2017), and datasets for Alzheimer's disease (Petersen et al., 2010), their creation requires substantial manual curation, involving significant labor and time. Besides, they are presented for specific domains and their knowledge is not transferable to other medical domains.

In this paper, we take the brain disease domain as an example and propose an automatic pipeline for extracting image-text pairs for pre-training a VL model for specific medical domains. The pipeline first collects raw image-text pairs from medical sources like PubMed and prepares aligned image-text pairs which can be leveraged for VL pre-training. There are two main challenges with the data we collected. First, medical data in both image and text form inherently possess a complex nature (Han et al., 2018). Moreover, image-caption pairs sourced from PubMed literature and journals often incorporate subfigures and subscaptions, introducing additional intricacies that can pose formidable challenges for the pre-training of VL models. Therefore, there is a unique challenge in medical VL pre-training where we care about the fine-grained alignment between subfigures/subcaptions. Our pipeline is equipped with subfigure/subcaption aligning designed to enhance multimodal learning. The letters in subfigures and subcaptions are often referred to as "subfigure labels" or "subcaption labels." are used to identify the alignment between subfigures and subcaptions.

Getting the aligned image-text pairs from our pipeline, we pre-train a VL model, BLIP (Li et al., 2022), on both the original collected data and processed data to analyze the effectiveness of our pipeline. We consider quantitative and qualitative intrinsic evaluations including image-text retrieval and attention visualization. We have observed that the model pre-trained on the processed data has a better multimodal understanding. We believe that our pipeline can be used for other domain-specific medical applications such as Prostate Cancer Diagnosis or Alzheimer's Disease Prediction.

Our work has three main contributions. We identified the challenge in medical VL pre-training by taking brain disease as an example. We built a pipeline that collects domain-specific medical image-text pairs followed by a module for matching subfigures/subcaptions. We release the dataset which

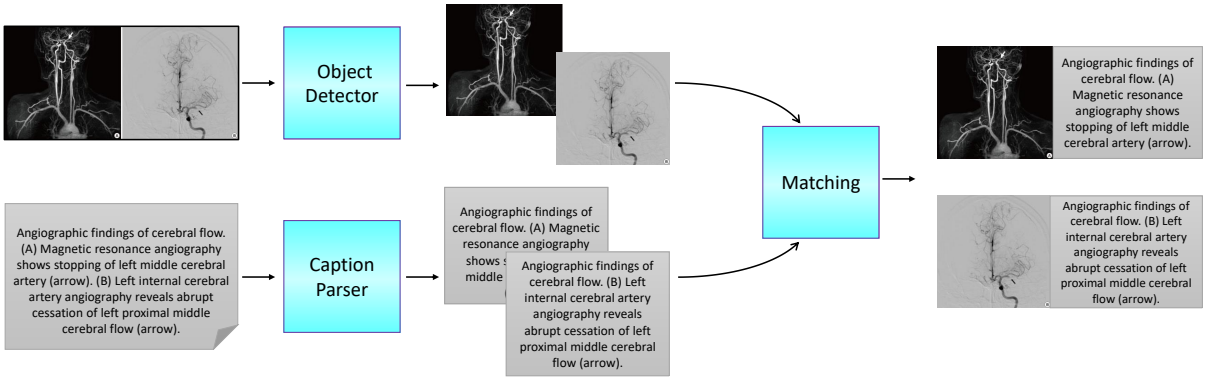


Figure 1: Our main pipeline for matching subfigures/subcaptions. The object detector outputs subfigures while the caption parser parses the caption into subcaptions simultaneously. Then, the module called 'matching' provides us the subfigure/subcaption pairs for the pre-training

is suitable for pre-training a VL model for brain diseases upon the acceptance of our work. Lastly, we pre-trained a VL model (BLIP) (Li et al., 2022) on our dataset, and with quantitative and qualitative analysis, we demonstrate that our pre-trained dataset and model are useful for building VL models in the medical domain.

## 2. Brain Image-Caption for VL Pre-Training

In this section, we explain the details of our pipeline for processing the data collected from PubMed and pre-training the vision-language (VL) model.

### 2.1. Data collection and processing

**Image-caption pairs.** Out of 9,371 journal papers from 1,021 various journal titles like "Brain tumor research and treatment" and "BMC Medical Imaging" dated from 1937-2018, We collected image-caption pairs from PubMed and filtered them by only scraping "case reports" with both types of image and text data. Pairs are additionally filtered by the keyword "brain" that appears in the corresponding caption assuming that both the image and caption are brain-related data. We collected 22,795 data initially and pre-trained the BLIP model on image-caption pairs as our baseline. Other data from the medical domain can be collected following our procedure to build another domain-specific VL model.

**Fine-grained alignment.** Due to the presence of subfigures and subcaptions in a significant portion of collected images ( 43 percent of data ), there is a unique challenge in their usage for the pre-training process and following the pre-training schema of existing VL models (taking the entire image and entire caption as image-text pairs) could lead to a low-performance model. This is because the model may struggle to match subcaptions and subfigures, which is not the issue in general domain VL datasets. Additionally, in numerous images within the dataset, subfigures exhibit significant distribution variations, possibly stemming from diverse sources such as MRI scans, surgical cameras, or simulation visualizations (as exemplified in Fig. 2). Presenting the complete figure and caption to the model could potentially lead to misguidance, as it may focus on image portions less pertinent to the caption and subcaptions. Moreover, the complexity of medical terms in clinical notes and abnormal regions in images poses a significant challenge to the model's understanding. To overcome these challenges, we developed a pipeline shown in Figure 1 where we used an object detection model (Tsutsui and Crandall, 2017) to identify subfigures in the images while simultaneously leveraging an NLP tool (Te-Lin et al., 2021) to parse captions into subcaptions. With these two steps, we were able to convert the collected image-caption pairs into 39,535 subfigure/subcaption pairs. However, matching subfigures and subcaptions was another challenge in training the VL model. We utilized an OCR tool[1] to detect the small characters in the subfigures commonly referred to as "subfigures labels" that are most likely the letters we can use to match with subcaptions. If the detected text does not match with any of the "subcaption labels" or does match with two or more of them, we pair the subfigure with the entire caption. Then aligned subfigures/subcaptions pairs are used for pre-training the model.

[1] google OCR: https://cloud.google.com/vision/docs/ocr

| | training schema | val | | | | test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | i2t@1 | i2t@10 | t2i@1 | t2i@10 | i2t@1 | i2t@10 | t2i@1 | t2i@10 |
| Only PT | raw data | 18.04 | 48.36 | 21.51 | 51.98 | 17.37 | 47.65 | 19.58 | 39.86 |
| | processed data | **36.9** | **69.88** | **38.4** | **69.04** | **36.94** | **69.37** | **37.40** | **69.58** |
| PT + FT | raw data | 24.6 | 60.47 | 27.13 | 59.14 | 24.45 | 59.4 | 26.1 | 58.7 |
| | processed data | **38.22** | **72.09** | **39.98** | **69.56** | **36.52** | **72.62** | **36.38** | **70.1** |

Table 1: Image-text retrieval results for the BLIP model were pre-trained/Fine-tuned (PT means only pre-training and PT + FT means pre-training plus fine-tuning) with two different schemes. One with following existing VL models and the other with following our pipeline.

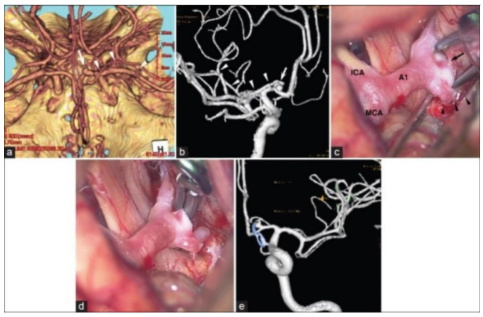

Figure 2: An example data from the dataset where each subfigure comes from a different source and Providing the model with the entire figure and its accompanying caption could potentially result in misdirection, causing it to emphasize image areas that are less relevant to both the caption and sub-captions.

## 3. Pre-Training and Evaluation

We trained the BLIP model on the processed data with a learning rate of 1e-5 for 60 epochs, using Adam as the optimization algorithm and Masked Language Modeling (MLM) (Devlin et al., 2018), Image-text matching (ITM) (Chen et al., 2020), and Image-text contrastive (ITC) (Radford et al., 2021) as the pre-training objectives. We evaluated the pre-trained model on qualitative and quantitative intrinsic tasks including image-text retrieval. Image-text retrieval evaluates the model's ability to retrieve the corresponding text/image from a pool of data given an image/text. We also conducted qualitative analysis by visualizing the attention map of some important medical terms and their relationship with abnormal regions.

## 4. Results

### 4.1. Quantitaive evaluation

Our pre-trained model for image-text matching is evaluated to assess its ability to match an image to its corresponding caption. To perform this evaluation, we evaluate the model performance for image-text retrieval tasks as an intrinsic evaluation. In this task, we test how well our model can retrieve the corresponding caption given an image and a pool of all captions, and vice versa. We evaluate our model's performance in two settings: '@1', where we determine the percentage of val/test dataset where the corresponding image or caption is the top 1 retrieved data; and '@10', where we determine the percentage of data where the corresponding image or caption is among the top 10 retrieved data. Our results which are summarized in Table 1validate our pre-trained model's capability for VL modeling. Additionally, we conducted pre-training experiments with varying proportions of the available data, ranging from 25 to 100 percent, and assessed the impact on our image-text retrieval performance. As depicted in Figure 3, our model exhibited a steeper performance curve, indicating its increased learning capacity as more data became available. In contrast, the baseline performance appeared to saturate and reach a plateau. This observation underscores the effectiveness of our pre-training pipeline in enhancing a model's learning capabilities.

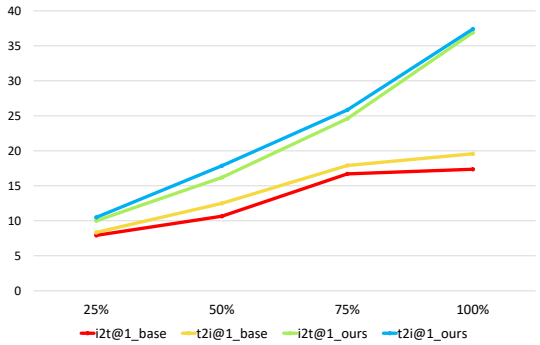


Figure 3: The plot visually demonstrates the dynamics of image-text retrieval performance across varying proportions of pre-training data. Our model results reveal a notable increase in learning capacity as more data becomes available, while the baseline exhibits characteristics that suggest a form of saturation.

### 4.2. Qualitative analysis

As part of our qualitative analysis, we employ attention maps to visualize the model's focus on some important medical terms, such as 'aneurysm'—representing abnormal blood vessel outpouching—within the associated images and 'cerebral artery' a condition where there is a deviation from the normal, healthy state of arteries. In Fig. 4 and 5, we present the model attention map for both the baseline and our pre-trained model concerning the terms 'cerebral artery' and 'aneurysm' respectively where 'aneurysm' is divided into subtokens. The upper heatmaps correspond to the baseline, while the lower ones depict the outputs of our pre-trained model. The attention maps demonstrate that our model emphasizes the relevant areas.

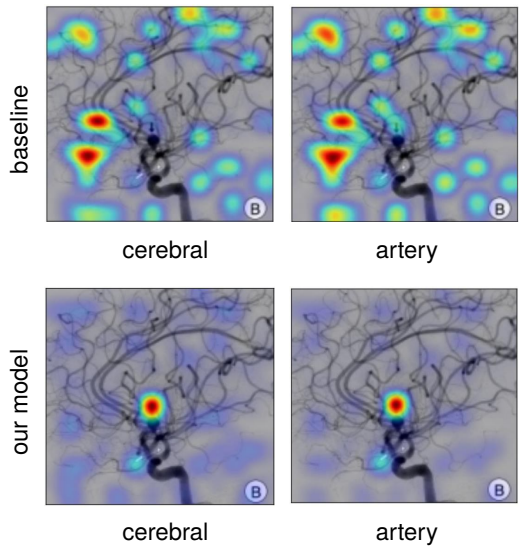


Figure 4: The visualization of the attention of the "cerebral artery" on the corresponding image. It is evident that ours highlighted the abnormal region more specifically.

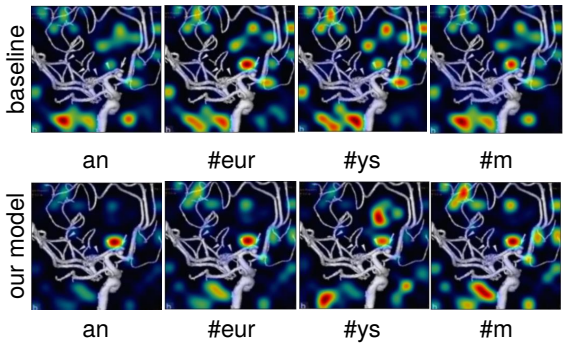


Figure 5: The visualization of the attention of the "aneurysm" on the corresponding image. It is evident that ours highlighted the abnormal region more specifically.

## 5. Related Work

**Vision-Language Learning.** Large-scale VL pre-training has demonstrates impressive performance for various VL tasks (Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2019; Li et al., 2019). Early works in this direction use off-the-shelf object detectors to extract objects (Anderson et al., 2018) and feed the region features into a modality fusion module (Su et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020; Zhang et al., 2021). End-to-end training methods have been proposed to be efficient and effective alternatives to object detector-based models (Kim et al., 2021; Li et al., 2021; Dou et al., 2022) because of their ability to unleash the power of vision encoders. While state-of-the-art VL models have achieved impressive performance, many of them are mainly focused on general-domain VL tasks such as visual question answering and image captioning, and it is unclear whether the pre-trained representations can be helpful for medical tasks.

**Medical-Domain VL Learning.** Multi-modal learning in the medical domain is of great significance due to the presence of medical images, text notes, and electronic health records. While there are a few medical-domain datasets and models being proposed (Eslami et al., 2023; Liu et al., 2021), many of them are focused on CLIP-style models (Radford et al., 2021) trained with contrastive objectives, neglecting to investigate more effective VL training methods in a large-scale setting. In addition to this line of work, Zhang et al. (2022) introduce the mmFormer, a novel Transformer-based approach, for accurate brain tumor segmentation from incomplete MRI modalities, achieving significant improvements in segmentation performance compared to state-of-the-art methods on the BraTS 2018 dataset. Lin et al. (2023) present PMC-OA, a large-scale biomedical dataset with 1.6M image-caption pairs from PubMedCentral's OpenAccess subset. Using this dataset, the PMC-CLIP model achieves state-of-the-art results in image-text retrieval and classification tasks, addressing data scarcity issues in the biomedical domain. In this work, we adapt the advanced techniques in the general domain VL learning to the medical domain and demonstrate its effectiveness.

## 6. Conclusion

Our study is centered on the creation of a pre-training dataset for domain-specific medical vision language models, utilizing a VL dataset sourced

from PubMed and a pipeline for VL modeling. Our approach has been evaluated through intrinsic task assessments, demonstrating its effectiveness in constructing data-efficient VL models for the medical field. Our model has surpassed the performance of standard VL training methods, further validating the efficacy of our pipeline. Our proposed pipeline helps build powerful multimodal AI models which could be used for various tasks like diagnosis, drug discovery, and information extraction.

## 7. Aknowledgment

This research has received support from Optum and was partially funded by the National Institutes of Health (NIH) under grant number R01HL152270. We extend our gratitude to Joel Stremmel and his team at OptumLabs, as well as to the reviewers whose valuable feedback significantly contributed to the enhancement of our work.

## 8. References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Lorenzo Brigato and Luca Iocchi. 2021. A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2490–2497. IEEE.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. 2023. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *arXiv preprint arXiv:2305.10799*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *ECCV*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

Yoseob Han, Jaejun Yoo, Hak Hee Kim, Hee Jung Shin, Kyunghyun Sung, and Jong Chul Ye. 2018. Deep learning with domain adaptation for accelerated projection-reconstruction mr. *Magnetic resonance in medicine*, 80(3):1189–1205.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021. Contrastive pre-training and representation distillation for medical visual question answering

based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li, C-C Jay Kuo, Aichi Chien, and Kai-Wei Chang. 2022. Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 725–734. Springer.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. 2010. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209.

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

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*, pages 2556–2565.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Wu Te-Lin, Shikhar Singh, Sayan Paul, Gully Burns, and Nanyun Peng. 2021. Melinda: A multimodal dataset for biomedical experiment method classification. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Satoshi Tsutsui and David J Crandall. 2017. A data driven approach for compound figure separation using convolutional neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 533–540. IEEE.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *CVPR*.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer.