Journal of Machine Learning for Biomedical Imaging. 2022:025. pp 1-18 Submitted 02/2022; Published 08/2022

Compound Figure Separation of Biomedical Images: Mining Large Datasets for Self-supervised Learning

Tianyuan Yao tianyuan.yao@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Chang Qu chang.qu@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Jun Long junlong@csu.edu.cn
Central South University, Big Data Institute, Changsha, Hunan, China 410083

Quan Liu quan.liu@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Ruining Deng r.deng@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Yuanhan Tian yuanhan.tian@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Jiachen Xu jiachen.xu@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Aadarsh Jha aadarsh.jha@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Zuhayr Asad zuhayr.asad@vanderbilt.edu
Vanderbilt University, Department of Computer Science, Nashville, TN, USA 37215

Shunxing Bao shunxing.bao@vanderbilt.edu
Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, TN, USA 37215

Mengyang Zhao mengyang.zhao@dartmouth.edu
Dartmouth College, Hanover, NH, USA 03755

Agnes B. Fogo agnes.fogo@vumc.org
Vanderbilt University Medical Center, Department of Pathology, Nashville, TN, USA 37215

Bennett A. Landman bennett.landman@vanderbilt.edu
Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, TN, USA 37215

Haichun Yang haichun.yang@vumc.org
Vanderbilt University Medical Center, Department of Pathology, Nashville, TN, USA 37215

Catie Chang catie.chang@vanderbilt.edu
Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, TN, USA 37215

Yuankai Huo yuankai.huo@vanderbilt.edu
Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, TN, USA 37215

Abstract

With the rapid development of self-supervised learning (e.g., contrastive learning), the importance of having large-scale images (even without annotations) for training a more generalizable AI model has been widely recognized in medical image analysis. However, collecting large-scale task-specific unannotated data at scale can be challenging for individual labs. Existing online resources, such as digital books, publications, and search engines, provide a new resource for obtaining large-scale images. However, published images in healthcare (e.g., radiology and pathology) consist of a considerable amount of compound figures with

https://www.melba-journal.org/papers/2022:025.html.

subplots. In order to extract and separate compound figures into usable individual images for downstream learning, we propose a simple compound figure separation (SimCFS) framework without using the traditionally required detection bounding box annotations, with a new loss function and a hard case simulation. Our technical contribution is four-fold: (1) we introduce a simulation-based training framework that minimizes the need for resource extensive bounding box annotations; (2) we propose a new side loss that is optimized for compound figure separation; (3) we propose an intra-class image augmentation method to simulate hard cases; and (4) to the best of our knowledge, this is the first study that evaluates the efficacy of leveraging self-supervised learning with compound image separation. From the results, the proposed SimCFS achieved state-of-the-art performance on the ImageCLEF 2016 Compound Figure Separation Database. The pretrained self-supervised learning model using large-scale mined figures improved the accuracy of downstream image classification tasks with a contrastive learning algorithm. The source code of SimCFS is made publicly available at `https://github.com/hrlblab/ImageSeperation`.

Keywords: Compound figures, Biomedical data, Deep learning, Contrastive learning, Self-supervised learning

1. Introduction

Self-supervised learning algorithms (e.g., contrastive learning) allow deep learning models to learn effective image representations from large-scale unlabeled data (Celebi and Aydin, 2016; Sathya and Abraham, 2013; Chen et al., 2020). Thus, the important role of having large-scale images (even without annotations) for training a more generalizable AI model has been widely recognized in medical image analysis. Even unannotated medical images can be difficult to obtain at scale for individual labs (Zhang et al., 2017). Fortunately, online resources (e.g., NIH Open-i® (Demner-Fushman et al., 2012) search engine, academic images released by journals) have provided a cost-effective and scalable way of obtaining large-scale images. However, the images from such resources consist of a considerably large amount of compound figures with subplots that cannot be directly used by modern self-supervised learning approaches (Fig 1). To make the data useful, we need to extract individual subplots from the compound figure, with compound figure separation algorithms (Lee and Howe, 2015b).

Recent contrastive learning methods have demonstrated advantages in pretraining a more generalizable deep learning model using large-scale unannotated individual images. However, the web-mined images from medical literature and search engines are not necessarily single images that can be directly used for contrastive learning. Therefore, the proposed SimCFS framework can be used to separate such compound images into individual images as unannotated training data for self-supervised learning.

Various compound figure separation approaches have been developed (Davila et al., 2020; Lee and Howe, 2015a; Apostolova et al., 2013; Tsutsui and Crandall, 2017; Shi et al., 2019; Jiang et al., 2021; Huang et al., 2005), especially with recent advances in deep learning. However, previous approaches typically required resource extensive bounding box annotation to form the problem as a training detection task. In this paper, we propose a simple compound figure separation (SimCFS) framework that minimizes the need for bounding box annotations in compound figure separation. Briefly, the contribution of this study is four-fold:

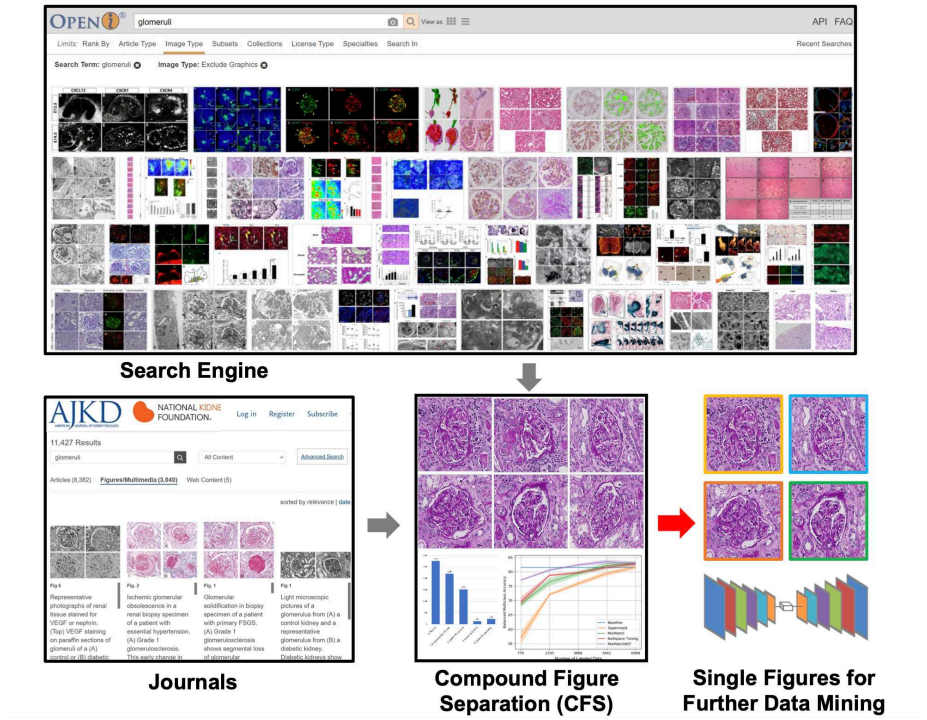

Figure 1: **Value of compound figure separation.** This figure shows the hurdle (red arrow) of training self-supervised machine learning algorithms directly using large-scale biomedical image data from biomedical image databases (e.g., NIH OpenI) and academic journals (e.g., AJKD). When searching desired tissues (e.g., search "glomeruli"), a large amount of data are compound figures. Such data would advance medical image research via recent self-supervised learning algorithms, such as self-supervised learning, contrasting learning, and auto encoder networks Huo et al. (2021)

• We introduce a simulation-based training framework that minimizes the need of resource extensive bounding box annotations.

• We propose a new Side loss, which is an optimized detection loss for figure separation.

• We propose an intra-class image augmentation method to mimic the hard cases of compound images without clear boundaries.

• To the best of our knowledge, this is the first study that evaluates the efficacy of leveraging self-supervised learning with compound image separation.

We apply our technique to conduct compound figure separation for renal pathology (in-house data) as well as on the ImageCLEF 2016 Compound Figure Separation Database (publicly available). Glomerular phenotyping (Koziell et al., 2002) is a fundamental task for efficient diagnosis and quantitative evaluations in renal pathology. Recently, deep learning techniques have played increasingly important roles in renal pathology to reduce clinical working load of pathologists and enable large-scale population based research (Gadermayr

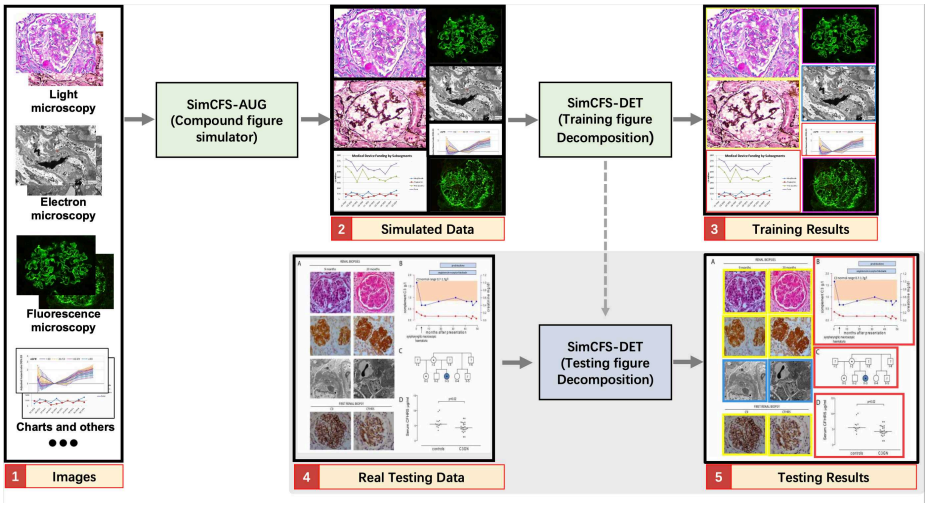

Figure 2: **The overall workflow of the proposed simple compound figure separation (SimCFS) workflow.** In the training stage, SimCFS only requires individual images from different categories. The pseudo compound figures are generated from the proposed augmentation simulator (SimCFS-AUG). Then, a detection network (SimCFS-DET) is trained to perform compound figure separation. In the testing stage (the gray panel), only the trained SimCFS-DET is required for separating the images.

et al., 2017; Bueno et al., 2020; Govind et al., 2018; Kannan et al., 2019; Ginley et al., 2019). Due to the lack of a publicly available dataset for renal pathology, it is appealing to extract large-scale glomerular images from public databases (e.g., NIH Open-i® search engine) for downstream self-supervised or semi-supervised learning (Huo et al., 2021). Meanwhile, the Image-CLEF 2016 dataset consists of various types of organs, and resources of large-scale medical images, which is arguably the most widely used testbed for compound image separation tasks. Both cohorts are used to evaluate the performance of different methods.

This work is extended from our conference paper (Yao et al., 2021) with the new efforts listed below: (1) we included more technical and evaluation details for the proposed method; (2) More comprehensive literature review and related work have been introduced; (3) We performed more rigorous evaluation (five-fold cross-validation) during the evaluation stages; (4) We conducted more comprehensive evaluation with more baseline compound image generation and separation methods (e.g., Tsutsui and Crandall (2017)); (5) We evaluated the efficacy of leveraging self-supervised learning with compound image separation by evaluating with both supervised and semi-supervised methods; (6) Our web mined glomerular dataset (20,000 images), as well as the source code of SimCFS, are released to public in the paper.

2. Related Work

2.1 Compound Figure Separation

In biomedical articles, about 40-60% of figures are multi-panel (Kalpathy-Cramer et al., 2015). Several methods have been proposed in the document analysis community that envolve, extracting figures and their semantic information. For example, Huang et al. (2005) presented their recognition results of textual and graphical information in literary figures. Davila et al. (2020) presented a survey of approaches of several data mining pipelines for future research.

2.1.1 Traditional vision approaches

In order to collect scientific data massively and automatically, various approaches have been proposed in the prior arts(Li et al., 2017b,a; Lee and Howe, 2015b). For example, Lee and Howe (2015a) proposed an SVM-based binary classifier to distinguish completed charts from visual markers, such as labels, legend, and ticks. Apostolova et al. (2013) proposed a figure separation method via a capital index. These traditional computer vision approaches were commonly performed on the figure's grid-based layout. Thus, the separation was usually accomplished by simple horizontal and vertical cuts based on the image boundary information.

2.1.2 Deep learning Methods

In the past few years, deep learning based algorithms, especially convolutional neural networks (CNNs), have provided considerably superior performance in extracting and separating subplots from from compound images. Tsutsui and Crandall (2017) proposed a CNN based approach that treated compound figure segmentation as an object localization problem by estimating the bounding boxes of subplots. This was one of the earliest deep learning-based approaches to achieve compound figure separation via a deep convolutional neural network. Tsutsui et al. applied the You Only Look Once (YOLO) Version 2, a CNN based detection network, which utilized a single convolutional network to predict bounding boxes and class probabilities simultaneously. They also implemented training on artificially constructed datasets and reported superior performances on ImageCLEF dataset (García Seco de Herrera et al., 2016). Shi et al. (2019) developed a multi-branch output CNN to predict the irregular panel layouts and provided augmented data to drive learning. Their network separated compound figures of different sizes of structures with better accuracy.

More recently, anchor-based approaches have attracted great attentions in the object detection field due to their concise network architectures and high computational efficiency. The introducing of anchor has prior knowledge to object distribution which is also closer to the compound figure situation. YOLOv4 was used by Jiang et al. (2021) to achieve a superior detection performance. They combined a traditional vision method with high performance of deep learning networks by detecting the sub-figure label and then optimizing the feature selection process in the sub-figure detection. Now, YOLO has been updated to V5, which inherited the advantages of YOLOv4 (Bochkovskiy et al., 2020). YOLOv5 integrated spatial pyramid pooling with new data enhancement methods like Mosaic training,

balanced model size and detection speed which achieved faster detection speed and higher accuracy.

2.2 Self-supervised learning method

Supervised learning refers the usage of a set of input variables to predict the value of a labeled output variable. It requires labeled data (like an answer key that the model can use to evaluate its performance). Conversely, self-supervised learning (Celebi and Aydin, 2016) refers to inferring underlying patterns from an unlabeled dataset without any reference to labeled outcomes or predictions.

Recently, a new family of self-supervised representation learning, called contrastive learning, shows its superior performance in various vision tasks (Wu et al., 2018; Noroozi and Favaro, 2016; Zhuang et al., 2019; Hjelm et al., 2018). Learning from large-scale unlabeled data, contrastive learning can learn discriminative features for downstream tasks. SimCLR (Chen et al., 2020) maximizes the similarity between images in the same category and repels the representations of different category images. Wu et al. (2018) uses an offline dictionary to store all data representation and randomly selects training data to maximize negative pairs. MoCo (He et al., 2020) introduces a momentum design to maintain a negative sample pool instead of an offline dictionary. Such works demand a large batch size in order to include sufficient negative samples. To eliminate the needs of negative samples, BYOL (Grill et al., 2020) was proposed to train a model with an asynchronous momentum encoder. Recently, SimSiam (Chen and He, 2020) was proposed to further eliminate the momentum encoder in BYOL, allowing for less GPU memory consumption.

3. Methods

The overall framework of SimCFS is presented in Fig. 2. The training stage of SimCFS contains two major steps: (1) compound figure simulation, and (2) sub-figure detection. In the training stage, the SimCFS network can be trained with either a binary (background and sub-figure) or multi-class setting. The purpose of the compound figure simulation is to achieve collecting large-scale training compound images in an annotation free manner. In the testing stage, only the detection network is needed, where the output will be the bounding boxes of the sub-figures which shall enable us to crop those images in a fully automatic manner. The binary setting detector can serve as a compound figure separator, while the multi-class detector can be used for web image mining for images of concerned categories.

3.1 Anchor-based detection

YOLOv5, the latest version in the YOLO family (Bochkovskiy et al., 2020), is employed as the backbone network for sub-figure detection. The rationale for choosing YOLOv5 is that the sub-figures in compound figures are typically located in horizontal or vertical orders. Herein, the grid-based design with anchor boxes is well adaptable to our application. A new Side loss is introduced to the detection network that further optimizes the performance of compound figure separation.

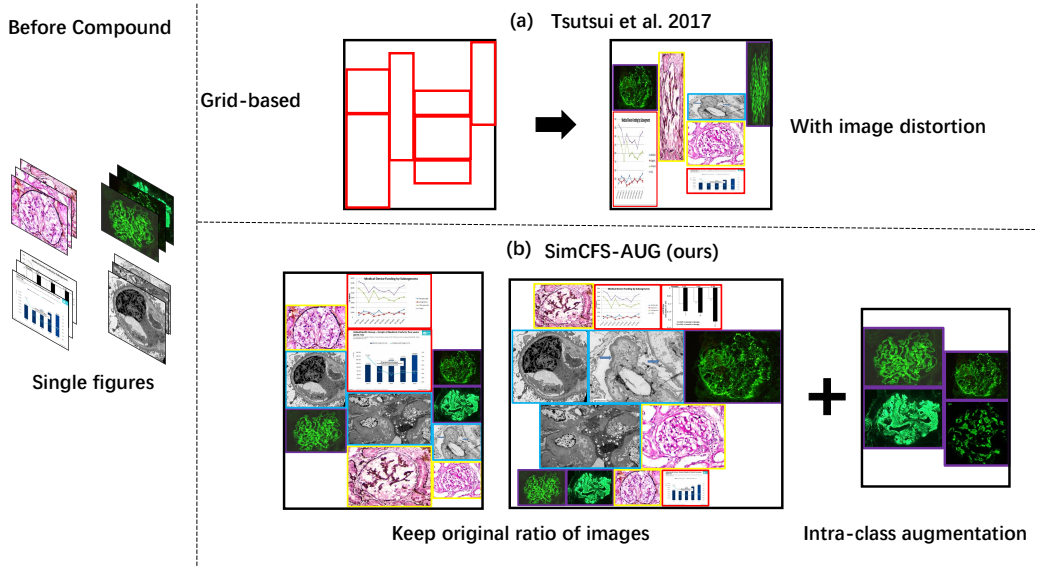

Figure 3: **Compound figure simulation.** (a) The upper panel shows the previously proposed compound figure synthesis strategy. It first generates the figure grids and then fills with images that have undergone image distortion, which is unusual in real compound figures. (b) The lower panel presents the proposed SimCFS-AUG compound figure simulator. It keeps the original ratio of individual images in an adaptive manner. Beyond this step of keeping original ratios, an intra-class augmentation is introduced to simulate the hard cases in which the boundaries are not explicitly visible between similar subplots. (Bounding boxes are displayed for visualization and are not actually visible in the training data)

3.2 Compound figure simulation

Our goal is to only utilize individual images, which are non-compound images with weak classification labels in training a compound image separation method. In previous studies, the same task typically requires stronger bounding box annotations of subplots using real compound figures. In compound figure separation tasks, a unique advantage is that the sub-figures are not overlapped. Moreover, their spatial distributions are more ordered as compared with natural images in object detection. Therefore, we propose to directly simulate compound figures from individual images as the training data for the downstream sub-figure detection.

Tsutsui and Crandall (2017) proposed a compound figure synthesis approach (Fig. 3). The method first randomly samples a number of rows and generates random heights for each row. Then a random number of single figures fills the empty template. However, the single figures are naively resized to fit the template, with large distortion (Fig. 3).

Inspired by prior arts (Tsutsui and Crandall, 2017), we propose a simple augmentation strategy that is specific to compound figure separation data, called SimCFS-AUG, to perform compound figure simulation. The inputs of the simulator are single images with

Algorithm 1 Compound figure simulation

Input:

Single images X_i in k classes

Set of training input indices with known labels $L_1, L_2, ..., L_k$

Output:

Compound figure $\overline{C}_j$

Annotation file A_j

for each pseudo compound figure $\overline{C}_j$ **do**
  Stage 1: Space initialize ▷ Multi real world case simulation
  Layout $\leftarrow$ *row-restricted or column-restricted*
  Classes $\leftarrow$ *multi or intra* ▷ Add intra-class augmentation
  Number of rows/columns $\leftarrow n \in [2, 5]$
  if layout is row-restricted **then** ▷ Keep close to real world aspect ratio
    $WidthW_{\overline{C}_j} \leftarrow 640, HeightH_{\overline{C}_j} \leftarrow \sum_{p=1}^{n} H_p \quad while \quad \frac{3}{4} \leq aspect\ ratio \leq \frac{4}{3}$
    ▷ Each row's height $H_1, ...H_p$ should be in certain range
  else if layout is column-restricted **then**
    $HeightH_{\overline{C}_j} \leftarrow 640, WidthW_{\overline{C}_j} \leftarrow \sum_{q=1}^{n} W_q \quad while \quad \frac{3}{4} \leq aspect\ ratio \leq \frac{4}{3}$
    ▷ Each column's width $W_1, ...W_q$ should be in certain range
  Stage 2: Fit in preset space
  for row/column in n **do**
    if Classes is multi **then**
      Create ImagePool I, for images X_i in I, i $\in L_1, L_2, ..., L_k$
    else if Classes is intra **then**
      Create ImagePool I, for images X_i in I, i $\in L_m, m \in [1, k]$
    Random fill in resized images from ImagePool (keeping original ratio)
    Save resized w_i', h_i', center position x_i, y_i to A_j
  Stage 3: Output: compound figure $\overline{C}_j$, annotation A_j

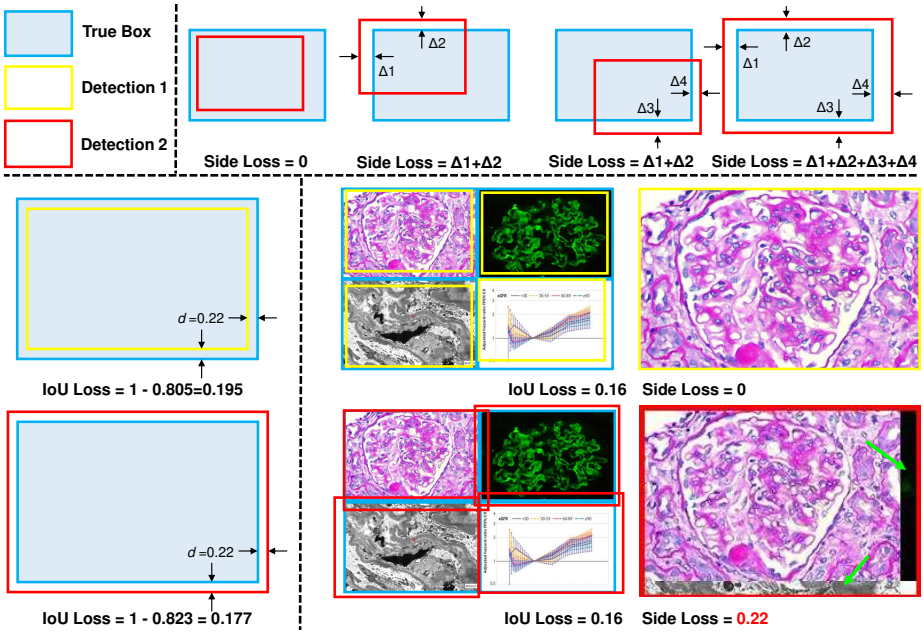

Figure 4: **Proposed Side loss for figure separation**. The upper panel shows the principle of side loss, in which penalties only apply when vertices of detected bounding boxes are outside of true box regions. The lower left panel shows the bias of current IoU loss towards over detection. When an under detection case (yellow box) and an over detection case (red box) have the same margins (d), from predicted to true boxes, the over detection has the smaller loss (larger IoU). The lower right panel shows the under detection and over detection examples of the compound figure separation, with the same IoU loss. Side loss is proposed to break IoU loss, given the results in the yellow boxes are less contaminated by nearby figures than the results in the red boxes (green arrows).

specified classes. Two groups are generated when simulating compound figures; these groups are row-restricted and column-restricted. The length of each row or column is randomly generated within a certain range. Then, images from our database are randomly selected and concatenated together to fit in the preset space. As opposed to previous studies, the original ratio of individual images is kept within our SimCFS-AUG simulator so as to mimic more realistic common compound images without distortion in individual images.

3.3 Side loss for compound figure separation

For object detection on natural images, there is no specific preference between over detection and under detection as objects can be randomly located and even overlapped. In medical compound images, however,objects are typically closely attached to each other without overlapping. In this case, over detection would introduce undesired pixels from the nearby plots (Fig. 4), which are not ideal for downstream deep learning tasks. Unfortunately, over

detection is often encouraged by the current Intersection Over Union (IoU) loss in object detection (Fig. 4), as compared with under detection.

In the SimCFS-DET network, we introduce a simple side loss, which will penalize over detection. We define a predicted bounding box as B^p and a ground truth box as B^g, with coordinates: $B^p = (x_1^p, y_1^p, x_2^p, y_2^p)$, $B^g = (x_1^g, y_1^g, x_2^g, y_2^g)$. The over detection penalty of vertices for each box is computed as:

$$\begin{aligned} x_1^{\mathcal{I}} &= \max(0, x_1^g - x_1^p), y_1^{\mathcal{I}} = \max(0, y_1^g - y_1^p) \\ x_2^{\mathcal{I}} &= \max(0, x_2^p - x_2^g), y_2^{\mathcal{I}} = \max(0, y_2^p - y_2^g) \end{aligned} \tag{1}$$

Then, the Side loss is defined as:

$$\mathcal{L}_{side} = x_1^{\mathcal{I}} + y_1^{\mathcal{I}} + x_2^{\mathcal{I}} + y_2^{\mathcal{I}} \tag{2}$$

The side loss is combined with canonical loss functions in YOLOv5, including bounding box loss (L_{box}), object probability loss (L_{obj}), and classification loss (L_{cls}).
$\mathcal{L}_{total} = \lambda_1 L_{box} + \lambda_2 L_{obj} + \lambda_3 L_{cls} + \lambda_4 L_{side}$,where λ_1, λ_2, λ_3, λ_4 are constant weights to balance the four loss functions. Following YOLOv5's implementation [1], the parameters were set as $\lambda_1 = box \times (3/nl)$, $\lambda_2 = obj \times (imgsize/640)^2 \times (3/nl)$, $\lambda_3 = (cls \times num\_cls/80) \times (3/nl)$, where num_cls was the number of classes, nl was the number of layers, and $imgsize$ was the image size.The λ_4 of the Side loss was empirically set to $\lambda_1/30$ across all experiments as the Side loss and Box loss are all based on the coordinates.

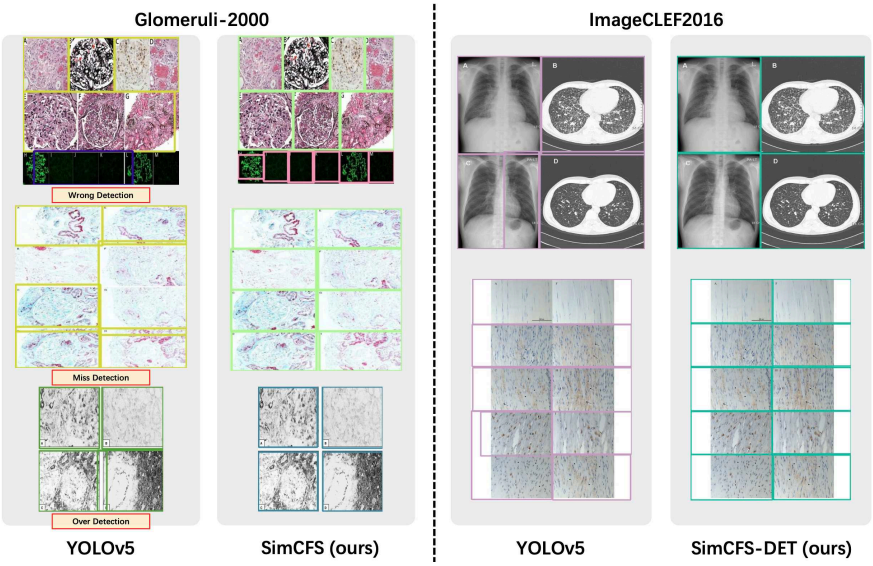

Figure 5: **Qualitative Results.** This figure shows the qualitative results of comparing proposed SimCFS approach with the YOLOv5 benchmark.

1. https://github.com/ultralytics/yolov5

4. Experimental Design

4.1 Data

We collected two in-house datasets for evaluating the performance of different compound figure separation strategies. One compound figure dataset (called Glomeruli-2000) consisted of 917 training and 917 testing real figure plots from the American Journal of Kidney Diseases (AJKD), with keywords "glomerular OR glomeruli OR glomerulus". Each figure was annotated manually with four classes, including glomeruli from (1) light microscopy, (2) fluorescence microscopy, (3) electron microscopy, and (4) charts/plots.

To obtain individual images to simulate compound figures, we downloaded 5,663 single individual images from online resources. Briefly, we obtained 1,037 images from Twitter, and obtained 4,626 images from Google search, with five classes, including individual images from (1) glomeruli with light microscopy, (2) glomeruli with fluorescence microscopy, (3) glomeruli with electron microscopy, (4) charts/plots, and (5) others. The individual images were combined using the SimCFS-AUG simulator in order to generate 7,000 pseudo training images. 2,000 of the pseudo images (with multiple sub-figures) were simulated using intra-class augmentation. In addition, 2,947 individual images were further employed as training data. The implementation of SimCFS-DET was based on YOLOv5 with PyTorch implementations. Google Colab was used to perform all experiments in this study.

4.2 Implement Details

In the experiment setting, the parameters are empirically chosen. We set the learning rate to 0.01, weight decay to 0.0005 and momentum to 0.937. The input image size was set to 640, *box* to 0.5, *obj* to 1, *cls* to 0.5, and the number of layers to 3. For our in-house datasets, we trained 50 epochs using a batch size of 64. For the imageCLEF2016 dataset (García Seco de Herrera et al., 2016), we trained 50 epochs using a smaller batch size of 8.

4.3 Evaluation Metrics

Mean average precision was the primary metric used to evaluate detection performance. For a given threshold IOU, average precision was obtained by calculating the area under the 101-point interpolated precision-recall curve. Then, mean average precision (AP) is the mean of the average precision for IOU thresholds from 0.5 to 0.95 with a step size of 0.05. AP_{50} is the average precision with an IOU threshold at 0.5. AP_{75} is the average precision with an IOU threshold at 0.75. AP_S is the mean average precision for small objects (area less than 32^2). AP_M is the mean average precision for medium objects (area between 32^2 and 96^2). Since no objects contained an area greater than 96^2, the large mean average precision (AP_L) was not utilized.

5. Results

5.1 Ablation Study

In this ablation study, we evaluate the image separation performance via 917 real compound images with manual box annotations as testing data in 1 and Fig. 5. For training, we assessed the performance of using 917 real compound training images ("Real Training

Table 1: The ablation study with different types of training data.

Method	Training Data	SL	AUG	All	Light	Fluo.	Elec.	Chart
YOLOv5	R			69.8	77.1	71.3	73.4	57.4
SimCFS-DET (ours)	R	✓		<u>79.2</u>	<u>86.1</u>	**80.9**	84.2	**65.8**
YOLOv5	$\bar{S}$			63.8	76.4	60.1	72.5	46.8
YOLOv5	S			66.4	79.3	62.1	76.1	48.0
YOLOv5	S		✓	71.4	82.8	72.1	75.3	47.1
SimCFS (ours)	$\bar{S}$	✓		68.9	77.1	66.8	82.5	49.1
SimCFS (ours)	S	✓		69.4	77.6	67.1	<u>84.1</u>	48.8
SimCFS (ours)	S	✓	✓	**80.3**	**89.9**	<u>78.7</u>	**87.4**	<u>58.8</u>

*The best and second best performances are denoted by **bold** and <u>underline</u>.
*For training data, R is using real compound figure while S is using simulated images, $\bar{S}$ is using Tsutsui and Crandall (2017) grid-based synthetic method.
*SL is the side loss, AUG is the intra-class self-augmentation.
*ALL is the Overall $\text{mAP}_{0.5:.95}$, which is reported for all concerned classes, (light, fluorescence, and electron microscopy).

Table 2: The results on ImageCLEF2016 dataset.

Method	Backbone	$\text{mAP}_{0.5}$	$\text{mAP}_{0.5:.95}$
Tsutsui and Crandall (2017)	YOLOv2	69.8	-
Tsutsui and Crandall (2017)	Transfer	77.3	-
Zou et al. (2020)	ResNet152	78.4	-
Zou et al. (2020)	VGG19	81.1	-
YOLOv5 (Bochkovskiy et al., 2020)	YOLOv5	85.3	69.5
SimCFS-DET (ours)	YOLOv5	88.9	71.2
SimCFS-DET esemble (ours)	YOLOv5	**90.3**	**71.5**

Images"), as well as the performance when only using simulated training images ("Simulated Training Images").

From the result, the proposed Side loss consistently improves the detection performance by a decent margin. The proposed compound image simulation method (with intra-class self-augmentation) achieves superior performance as compared to the benchmarks.

5.2 Comparison with State-of-the-art

We also compare CFS-DET with the state-of-the-art approaches including Tsutsui and Crandall (2017) and Zou et al. (2020) using the ImageCLEF2016 dataset (García Seco de Herrera et al., 2016). ImageCLEF2016 is the commonly accepted benchmark for compound figure separation, including total 8,397 annotated multi-panel figures (6,783 figures for training and 1,614 figures for testing). Table 2 shows the results of the ImageCLEF2016 dataset. The proposed CFS-DET approach consistently outperforms other methods by considering

evaluation metrics. Additionally, we applied five-fold cross validation to our model training using weighted boxes fusion as proposed by (Solovyev et al., 2021). To merge the bounding boxes results from the five predictions, the proposed method used the confidence scores of all of the proposed bounding boxes in order to construct the average boxes. Eventually, when combining SimCFS with the weighted boxes fusion (SimCFS-DET ensemble), the performance was further improved.

5.3 Application on Contrastive Learning

We demonstrate the application of our SimCFS framework and how it helps to provide massive biomedical image data and benefits further data analysis with self-supervised representation learning.

In this study, self-supervised contrastive learning was employed as an example downstream task for our SimCFS compound image separation approach. We demonstrate how our approach helps to provide massive biomedical image data and benefits further data analysis with self-supervised representation learning. To evaluate the performance of introducing separated images, a semi-supervised method was evaluated beyond the supervised benchmark to present the performance of using the same set of unannotated images as the contrastive learning approach.(Table 3) Specifically, the stain and imaging modality classification task is employed to evaluate the performance of different approaches.

5.3.1 Data

We first collected 10,000 compound figures with the keywords 'glomerular OR glomeruli OR glomerulus'. Then we used our SimCFS network to process all compound images to get more than 20,000 glomeruli pathologies obtained by different microscopy or in different stains with a confidence threshold of 0.7.

Other in-house data are 3,000 manually annotated glomeruli pathologies with seven classes, including glomeruli from (1) electron microscopy, (2) fluorescence microscopy, and light microscopy with different stains of (3) PAS, (4) silver, (5) H&E, (6) Masson and (7) other.

5.3.2 Approach

We used the SimSiam network (Chen et al., 2020) as the baseline method of contrastive learning. 20,000 glomeruli pathologies were used to train the SimSiam network. Two random augmentations from the same image were used as training data. In all of our self-supervised pre-training, images for model training were resized to 224×224 pixels. We used the momentum SGD as the optimizer. The weight decay was set to 0.0005. The base learning rate was $lr = 0.05$ and the batch size equals 64. The learning rate was $lr \times \text{BatchSize}/256$, which followed a cosine decay schedule (Loshchilov and Hutter, 2017).

To apply the self-supervised pre-training networks, we froze the pretrained ResNet-50 model by adding one extra linear layer which followed the global average pooling layer. When finetuning with the 3,000 manually annotated glomeruli data, only the extra linear layer was trained.To prevent model over-fitting, we applied 5-fold cross validation by dividing our data into 5 folders, using four of the five folders as training data and the other folder as validation. We used the SGD optimizer to train linear classifier with a based (initial)

Table 3: Classification performance.

Methods	Unlabeled Images	labeled Images	F1 Score	Balanced Acc
Supervised method:				
Random Int	-	2.3k	0.845	0.843
ImageNet Int	-	2.3k	0.888	0.883
Semi-supervised method:				
Temporal Ensembling	20k	2.3k	0.892	0.885
Self-supervised method:				
Simsiam	-	2.3k	0.891	0.893
Simsiam w.SimCFS	20k	2.3k	**0.900**	**0.904**

*For the supervised method, we trained the entire ResNet-50 (random initialized and ImageNet pretrained) from scratch with fully supervised learning.

learning rate lr=30, weight decay=0, momentum=0.9, and batch size=64 (follows Chen and He (2020)). We trained linear classifiers for 100 epochs and selected the best model based on the validation set.

5.3.3 Results

Fine-tuning our pretrained SimSiam (Backbone:ResNet-50) on 2.3K labeled images is significantly better then training from scratch. Interestingly, our model also outperformed ResNet-50 models pretrained on ImageNet. Table 3 shows the results.

6. Discussion

In this study, we develop a new compound image separation framework with the ultimate goal to advance downstream machine learning tasks. The recent contrastive learning methods demonstrated their advantages of pretraining a more generalizable deep learning model using large-scale unannotated individual images. However, the web-mined images from medical literatures and search engines are not necessarily single images that can be directly used for contrastive learning. Therefore, the proposed SimCFS can be used to separate such compound images into individual images as unannotated training data for self-supervised learning.

The YOLO method was employed since it was a broadly used anchor-based backbone in previous compound image separation algorithms. However, our framework is an open framework, where the YOLO method can be replaced by other object detection backbones (e.g., anchor-free methods) and even with an even better performance.

The new application, through the optimization of both Side loss function and hard case simulation, proposes to improve the accuracy of image separation. Our proposed Side loss is designed based on the knowledge that there is no overlapping case in compound figures. By adding a penalty for the overestimated bounding box, the predictions are less overlapped as compared to the true box regions.

Secondly, with our compound figure simulation method, SimCFS can be trained with only synthetic compound figures which are generated by only a small quantity of annotated individual images. At the beginning of our experiment, when we synthesized row-restricted and column-restricted compound figures using images from all classes, the results were not as good as the real compound image data. To overcome such issues, we proposed the intra-class image augmentation method. By simulating those hard cases and adding the new intra-class compound figures to our previous synthesized data, the performance of the simulated training data has outperformed the real data by its large quantity and various simulated cases.

Recent advances in computer vision are due, to a large extent, to the growing size of annotated training data. However, one key limitation to the SimCFS network is that the ImageCLEF Medical dataset , the largest available dataset for compound figure separation, has only 7,000 images for training, which is much smaller than most modern object detection datasets. An important goal for this community could be to build up a much larger size dataset with multi-classes annotations like MRI, pathology, and charts etc. In this study, we assessed the promising application of SimCFS, which is to create large-scale unlabeled images for downstream contrastive learning. Using NIH OpenI, tens of thousands of free biomedical data can be achieved by searching the desired tissue types. The self-supervised learning strategy achieved better accuracy than the fully supervised approach with ImageNet initialization.

Several potential improvements for our SimCFS framework are as follows. First, we could further introduce image synthesis approaches to the proposed pipeline to obtain more unique images. Furthermore, we can perform textual contents extractions for captions, notes and labels while separating figures. These data in multi-forms could benefit further data mining research.

7. Conclusion

In this paper, we introduced the SimCFS framework to extract images of interests from large-scale compounded figures with weak classification labels. The pseudo training data were built using the proposed SimCFS-AUG simulator. The anchor-based SimCFS-DET detection achieved state-of-the-art performance by introducing a simple side loss. Additionally, our SimCFS framework provided cost-efficient and large-scale unannotated images to train un-/self-supervised representation learning methods (e.g., SimSiam). It achieved better performance than ImageNet's supervised pre-trained counterparts in classification tasks.

Acknowledgments

This work was supported in part by NIH NIDDK DK56942(ABF) and NSF CAREER 1452485 (Landman).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Emilia Apostolova, Daekeun You, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, and George R Thoma. Image retrieval from scientific publications: Text and image content processing to separate multipanel figures. *Journal of the American Society for Information Science and Technology*, 64(5):893–908, 2013.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Gloria Bueno, M Milagro Fernandez-Carrobles, Lucia Gonzalez-Lopez, and Oscar Deniz. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Computer Methods and Programs in Biomedicine*, 184:105273, 2020.

M Emre Celebi and Kemal Aydin. *Unsupervised learning algorithms*. Springer, 2016.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Kenny Davila, Srirangaraj Setlur, David Doermann, Urala Kota Bhargava, and Venu Govindaraju. Chart mining: a survey of methods for automated chart analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

Michael Gadermayr, Ann-Kathrin Dombrowski, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. Cnn cascades for segmenting whole slide images of the kidney. *arXiv preprint arXiv:1708.00251*, 2017.

Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, and Henning Müller. Overview of the ImageCLEF 2016 medical task. In *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, September 2016.

Brandon Ginley, Brendon Lutnick, Kuang-Yu Jen, Agnes B Fogo, Sanjay Jain, Avi Rosenberg, Vighnesh Walavalkar, Gregory Wilding, John E Tomaszewski, Rabi Yacoub, et al. Computational segmentation and classification of diabetic glomerulosclerosis. *Journal of the American Society of Nephrology*, 30(10):1953–1967, 2019.

Darshana Govind, Brandon Ginley, Brendon Lutnick, John E Tomaszewski, and Pinaki Sarder. Glomerular detection and segmentation from multimodal microscopy images using a butterworth band-pass filter. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058114. International Society for Optics and Photonics, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weihua Huang, Chew Lim Tan, and Wee Kheng Leow. Associating text and graphics for scientific chart understanding. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 580–584. IEEE, 2005.

Yuankai Huo, Ruining Deng, Quan Liu, Agnes B Fogo, and Haichun Yang. Ai applications in renal pathology. *Kidney International*, 2021.

Weixin Jiang, Eric Schwenker, Trevor Spreadbury, Nicola Ferrier, Maria KY Chan, and Oliver Cossairt. A two-stage framework for compound figure separation. *arXiv preprint arXiv:2101.09903*, 2021.

Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at imageclef 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015.

Shruti Kannan, Laura A Morgan, Benjamin Liang, McKenzie G Cheung, Christopher Q Lin, Dan Mun, Ralph G Nader, Mostafa E Belghasem, Joel M Henderson, Jean M Francis, et al. Segmentation of glomeruli within trichrome images using deep learning. *Kidney international reports*, 4(7):955–962, 2019.

Ania Koziell, Victor Grech, Sagair Hussain, Gary Lee, Ulla Lenkkeri, Karl Tryggvason, and Peter Scambler. Genotype/phenotype correlations of nphs1 and nphs2 mutations in nephrotic syndrome advocate a functional inter-relationship in glomerular filtration. *Human molecular genetics*, 11(4):379–388, 2002.

Po-Shen Lee and Bill Howe. Detecting and dismantling composite visualizations in the scientific literature. In *International Conference on Pattern Recognition Applications and Methods*, pages 247–266. Springer, 2015a.

Po-Shen Lee and Bill Howe. Dismantling composite visualizations in the scientific literature. In *ICPRAM (2)*, pages 79–91. Citeseer, 2015b.

Pengyuan Li, Xiangying Jiang, Chandra Kambhamettu, and Hagit Shatkay. Segmenting compound biomedical figures into their constituent panels. In Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 199–210, Cham, 2017a. Springer International Publishing. ISBN 978-3-319-65813-1.

Pengyuan Li, Xiangying Jiang, Chandra Kambhamettu, and Hagit Shatkay. Compound image segmentation of published biomedical figures. *Bioinformatics*, 34(7):1192–1199, 10 2017b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx611. URL `https://doi.org/10.1093/bioinformatics/btx611`.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

Ramadass Sathya and Annamma Abraham. Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2):34–38, 2013.

Xiangyang Shi, Yue Wu, Huaigu Cao, Gully Burns, and Prem Natarajan. Layout-aware subfigure decomposition for complex figures in the biomedical literature. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1343–1347. IEEE, 2019.

Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, pages 1–6, 2021.

Satoshi Tsutsui and David J Crandall. 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, 2017.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.

Tianyuan Yao, Chang Qu, Quan Liu, Ruining Deng, Yuanhan Tian, Jiachen Xu, Aadarsh Jha, Shunxing Bao, Mengyang Zhao, Agnes B Fogo, et al. Compound figure separation of biomedical images with side loss. *arXiv preprint arXiv:2107.08650*, 2021.

Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.

Jie Zou, George Thoma, and Sameer Antani. Unified deep neural network for segmentation and labeling of multipanel biomedical figures. *Journal of the Association for Information Science and Technology*, 71(11):1327–1340, 2020.