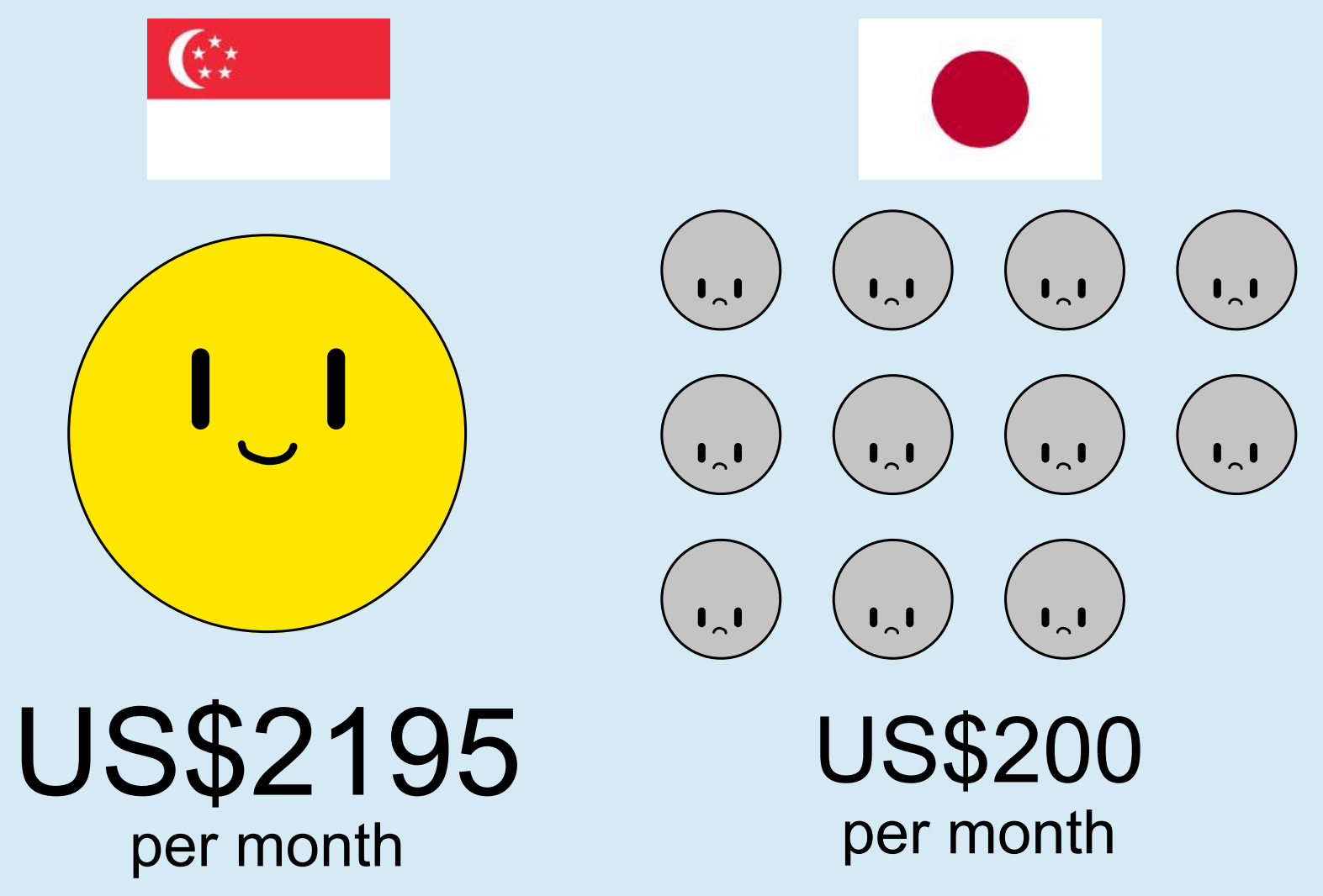
Manga Layout Analysis via Deep Learning

Nyx Audrey Angelo Iskandar
Raffles Institution, Singapore

Introduction

Motivation

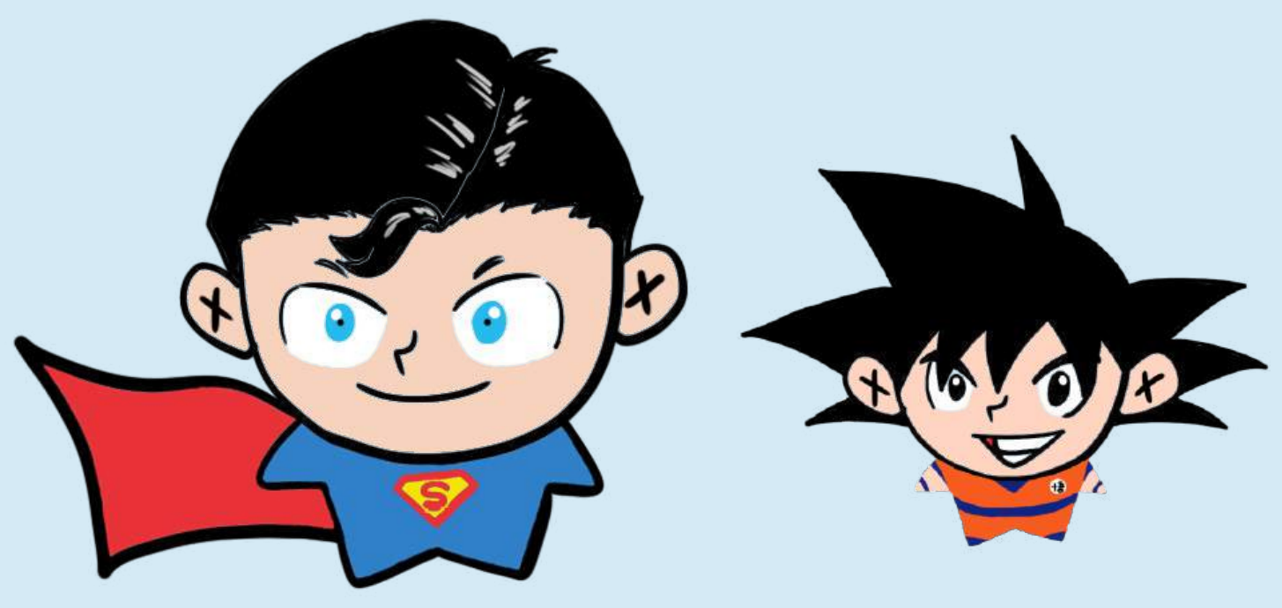


Japanese animators are **overworked** and **underpaid**

- Extremely low pay -- Singaporean animators earn 11 times more than them
- Outrageously long working hours (12-18 hours daily, 400-600 hours monthly) -- violate Japanese labour regulations
- **Manually adapting manga into anime and other media is an extremely time-consuming and draining process -- one anime episode takes two years to create!**

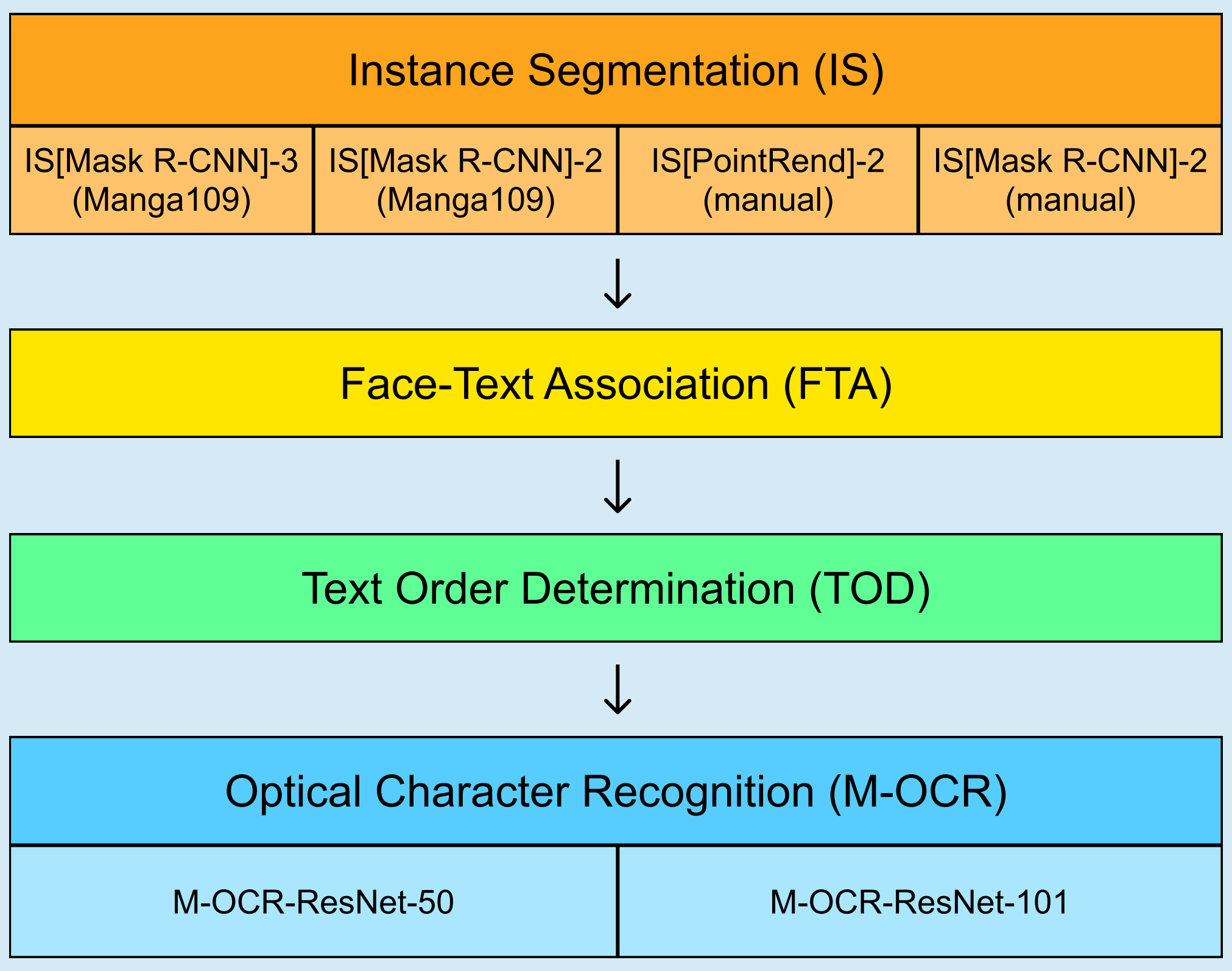
Limited work has been done on manga-based document layout analysis

- Most existing research was done on Western comic books
- Existing models only focus on particular components of Manga Layout Analysis -- either only one-class instance segmentation or only optical character recognition



Aim

- **Automate** the preprocessing stage of adapting manga
- **Innovate** an **integrated** solution to **manga**-based document layout analysis research



Instance Segmentation (IS)

IS[Mask R-CNN]-3 (Manga109) | IS[Mask R-CNN]-2 (Manga109) | IS[PointRend]-2 (manual) | IS[Mask R-CNN]-2 (manual)

↓

Face-Text Association (FTA)

↓

Text Order Determination (TOD)

↓

Optical Character Recognition (M-OCR)

M-OCR-ResNet-50 | M-OCR-ResNet-101

Methodology

Instance Segmentation

IS[Mask R-CNN]-3 (Manga109)

A Mask R-CNN model performing instance segmentation on text, frames, and faces

Dataset: Manga109*

Number of training rounds: 4

Round	Learning Rate	Number of Epochs	Layers	Steps per epoch
1	0.001	10	heads	131
2	0.0001	40	all	131
3	0.001	1	all	1000
4	0.001	1	all	2524

Specifications:

- 1 CUDA-enabled GPU
- ResNet-101 backbone
- Backbone strides for each FPN Pyramid layer = [4, 8, 16, 32, 64]
- Batch size = 2
- Number of images per GPU = 2
- Learning momentum = 0.9
- Weight decay = 0.0001
- Pool size = 7
- Loss weights (value = 1.0) = rpn_class_loss, rpn_bbox_loss, mrcnn_class_loss, mrcnn_bbox_loss, mrcnn_mask_loss
- Image shape = [1024, 1024, 3] (the minimum and maximum image dimensions are 800 and 1024 respectively)
- Mask shape = [28, 28]
- Number of classes = 4 (background + text + frame + face)
- Image meta size = 16

IS[Mask R-CNN]-2 (Manga109)

A Mask R-CNN model performing instance segmentation for text and frames

Specifications:

- Number of classes = 3 (background + text + frame)
- Image meta size = 15

Other details are the same as those for IS[Mask R-CNN]-3 (Manga109)

IS[Mask R-CNN]-2 (manual)

A Mask R-CNN model performing instance segmentation for text and frames

Dataset: Manual**

Other details are the same as those for IS[Mask R-CNN]-2 (Manga109)

IS[PointRend]-2 (manual)

A PointRend model performing instance segmentation for text and frames

Dataset: Manual**

Number of training rounds: 1

Specifications:

- 1 CUDA-enabled GPU
- ResNet-101 backbone
- Learning rate = 0.0005
- Learning momentum = 0.9
- Maximum iteration count = 2500
- Weight decay = 0.0001
- Number of workers = 2
- Images per batch = 2
- Number of classes = 2 (text + frame)

*This dataset contains the original annotations of the Manga109 dataset whose format was modified slightly for training compatibility; originally, one XML file contained annotations for one book, but after modification, one XML file contains annotations for one image (two pages of one book)

**This dataset is a COCO-like dataset containing Manga109 images which were manually annotated

Face-Text Association

Inputs: Output vertices of IS[Mask R-CNN]-3 (Manga109), image from manga

Methodology:

- Draw a box around each region of interest (text, frame, face)
- Calculate centres of text and faces using the formula below
  x_centre = (x1 + x2) / 2
  y_centre = (y1 + y2) / 2
- Determine which text and faces lie inside each frame -- whether centres of text and faces lie within frame
- Find face-text pairs by associating each face to the text nearest to it in a frame
- Draw a line from the centre of each face to the centre of the associated text

Text Order Determination

Inputs: Output vertices of IS[Mask R-CNN]-3 (Manga109), image from manga

Methodology:

- Draw box around each text
- Find centres of each text and frame
- Order frames -- topmost and rightmost frame is the first frame
- Order text
- Repeat
- Display text order visually by writing numbers in the speech bubbles of the text -- 1 represents the text to be read first, 2 for the next text, and so on and so forth

Optical Character Recognition

M-OCR-ResNet-50

A convolutional neural network-based optical character recognition model built on ResNet-50

M-OCR-ResNet-101

A convolutional neural network-based optical character recognition model built on ResNet-101

Datasets:

- Kuzushiji-49 (48 hiragana characters, 1 Hiragana iteration mark)
- A subset of Kuzushiji-Kanji (50 kanji characters)

Number of training rounds: 5

Round	Learning Rate	Number of Epochs
1	0.01	30
2	0.01	50
3	0.001	50
4	0.005	100
5	0.005	50

Specifications:

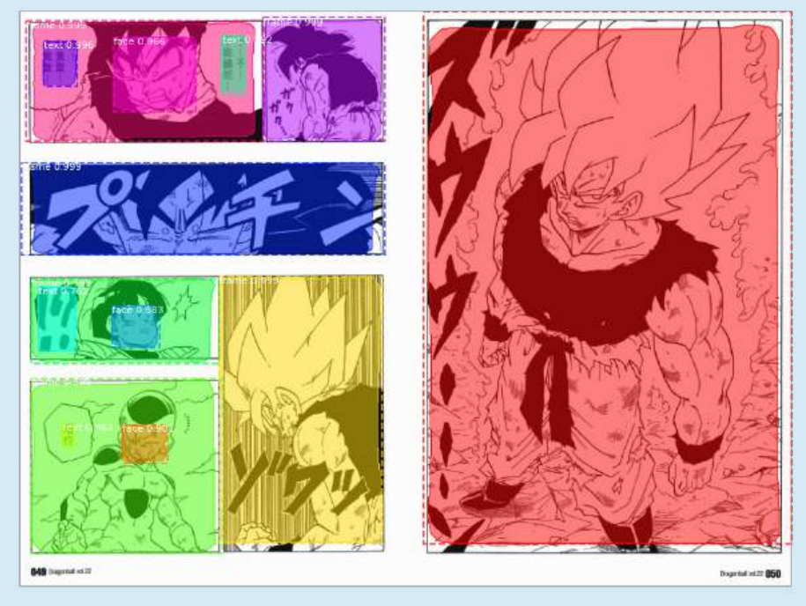
- 1 GPU
- Batch size = 128
- Number of classes = 99 (hiragana + kanji)
- Steps per epoch = 1839
- Image size = (32, 32, 1)
- Binarized image labels
- Image augmentation
  - Rotation of range 10
  - Zoom of range 0.05
  - Width-shift of range 0.1
  - Height-shift of range 0.1
  - Shearing of range 0.15

Results

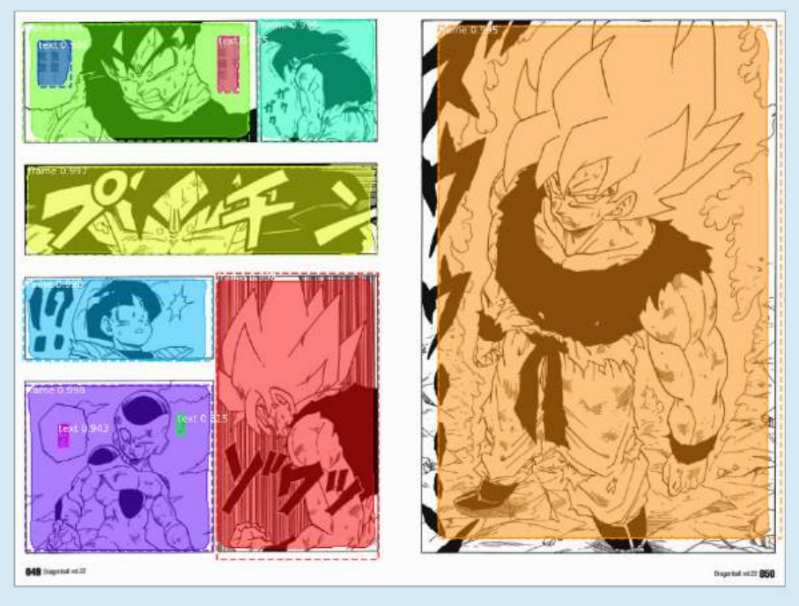
Instance Segmentation

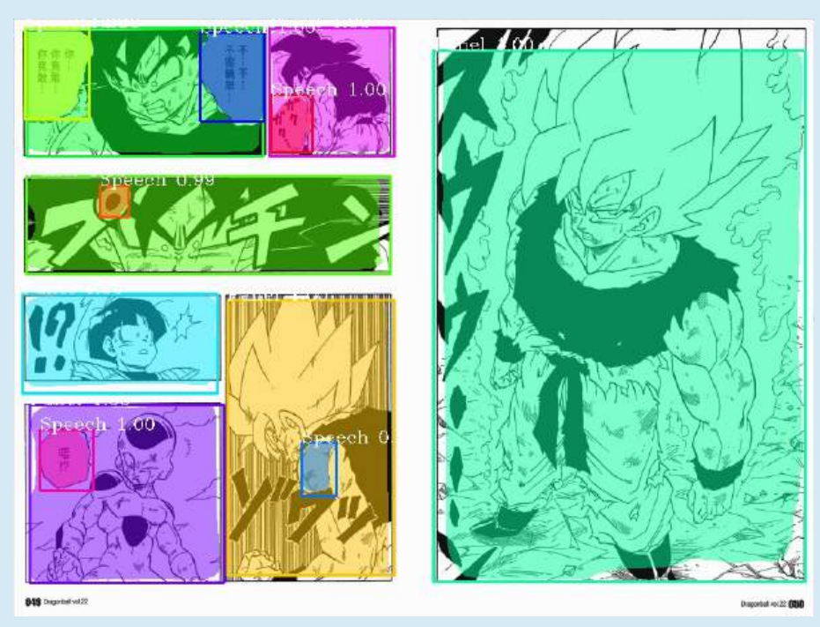
Name	mAP (IoU=0.5)	mAP (IoU=0.75)
IS[Mask R-CNN]-3 (Manga109)	0.90	0.80
IS[Mask R-CNN]-2 (Manga109)	0.93	0.86
IS[PointRend]-2 (manual)	0.95	0.92
IS[Mask R-CNN]-2 (manual)	0.93	0.85

IS[Mask R-CNN]-3 (Manga109)

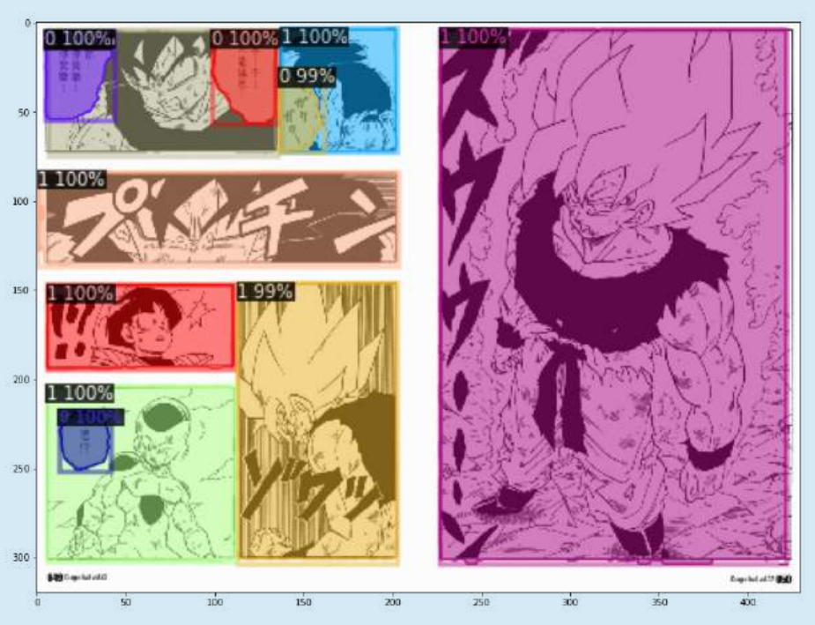


IS[Mask R-CNN]-2 (Manga109)

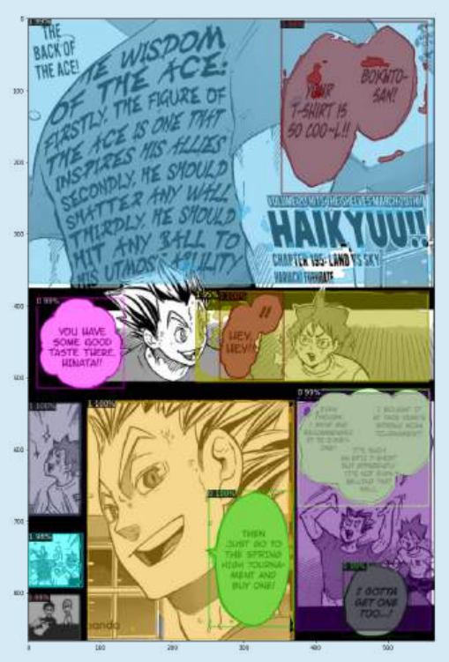




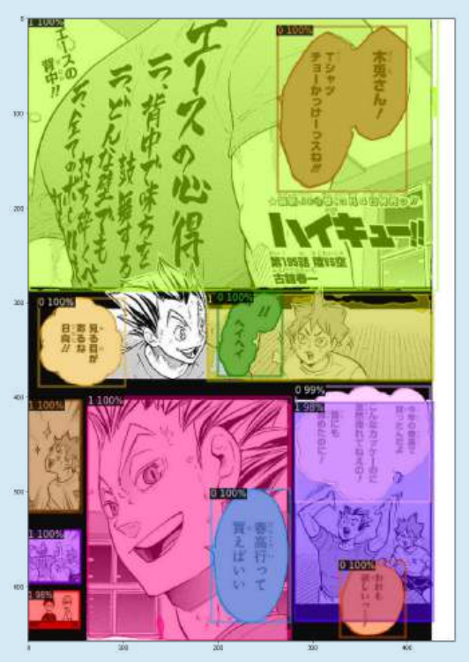
IS[Mask R-CNN]-2 (manual)



IS[PointRend]-2 (manual)



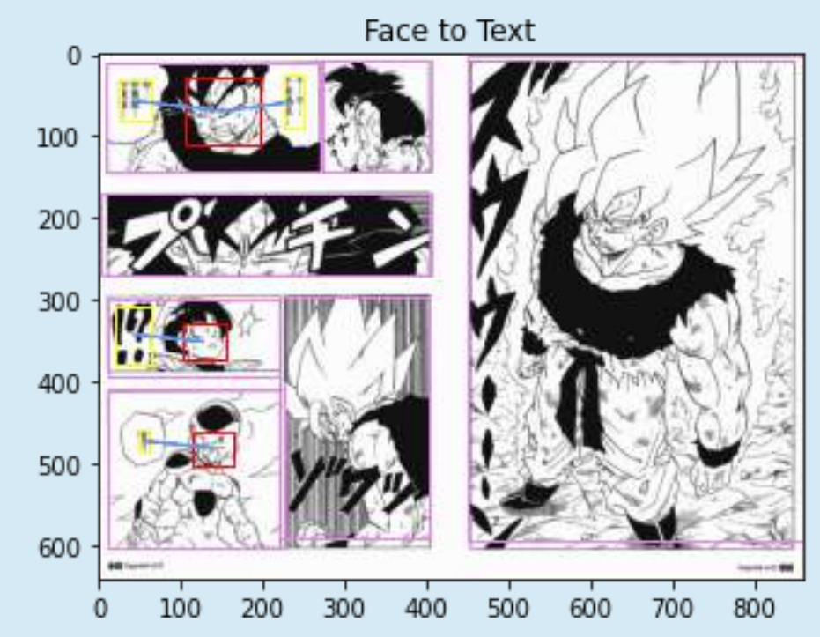
English



Japanese

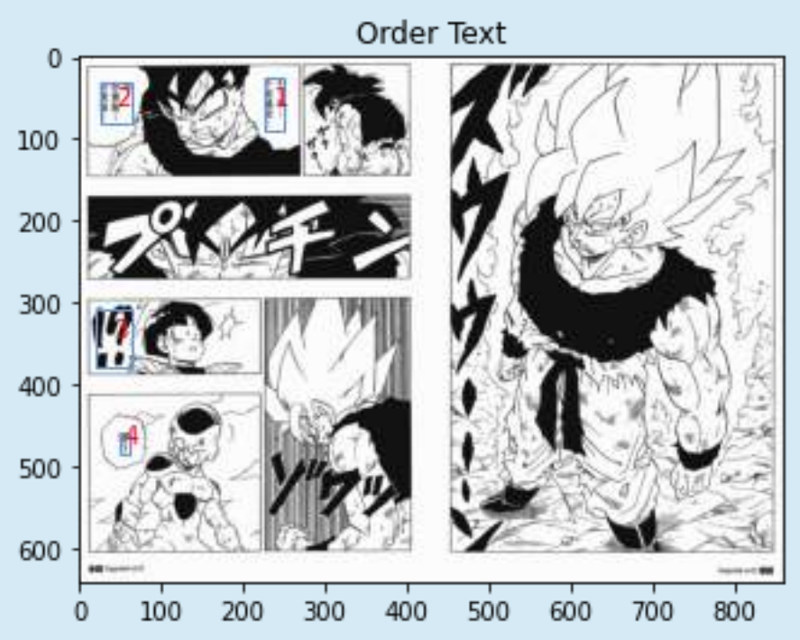
- **IS[PointRend]-2 (manual)** is the **most accurate** IS -- highest mAP scores
- **IS[Mask R-CNN]-3 (Manga109)** is the **most comprehensive** IS -- able to segment three classes
- Results for **English** manga are **comparable** to those for **Japanese** manga

Face-Text Association



- Successfully associates each segmented face and text
- Pairs each face with the text nearest to it where the face is in a frame with multiple faces and text

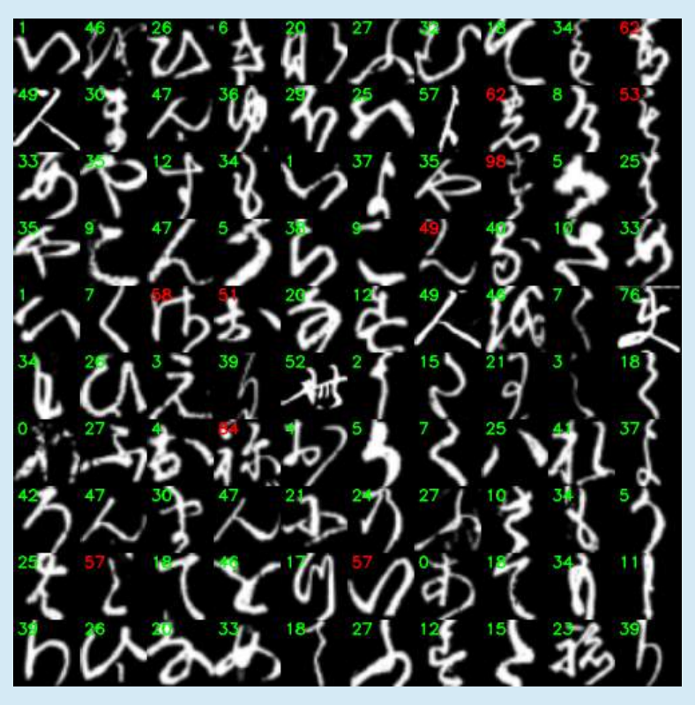
Text Order Determination



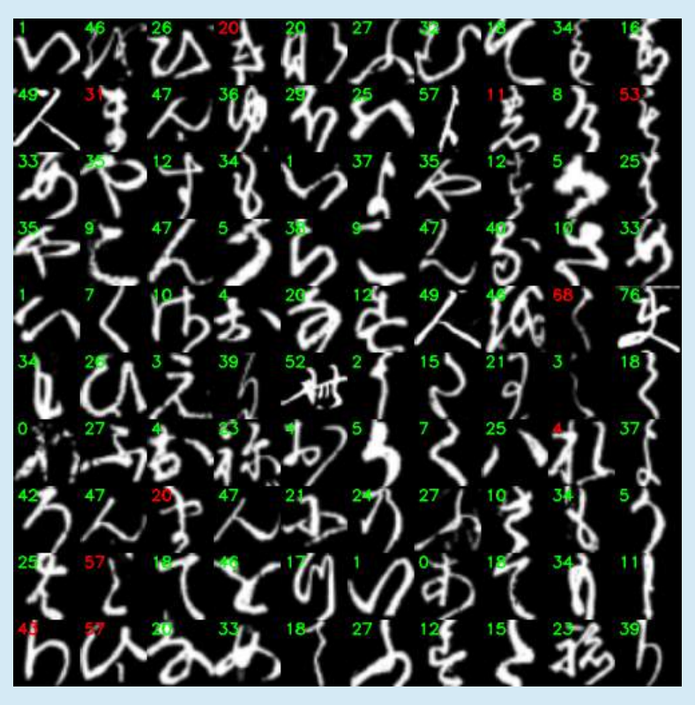
- Successfully orders text as per the reading order of manga (from top to bottom and from right to left)
- Able to distinguish between text in speech bubbles and onomatopoeia and text outside of frames -- only orders text within speech bubbles

Optical Character Recognition

Name	F1 Score (Accuracy)	F1 Score (Macro Average)	F1 Score (Weighted Average)
M-OCR-ResNet-50	0.87	0.79	0.89
M-OCR-ResNet-101	0.89	0.81	0.90



M-OCR-ResNet-50



M-OCR-ResNet-101

- **M-OCR-ResNet-101** is the **more accurate** M-OCR -- higher F1 Score
- **M-OCR-ResNet-50** is the **faster** M-OCR -- higher training and execution speeds

Conclusion

Manga Layout Analysis via Deep Learning has innovated an **integrated** system of instance segmentation, novel algorithms, and optical character recognition

- **Comparable performance** to state-of-the-art models
- **More comprehensive** than existing research/ solutions

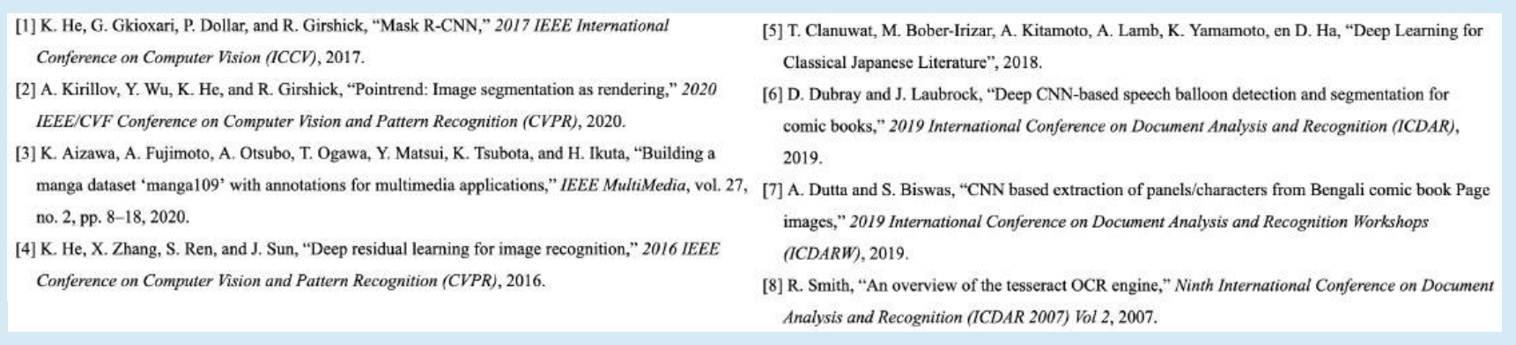
Future Work

- Identify the name of the character whose face is segmented
- Identify the emotion conveyed through the shape of a speech bubble
- Develop an MLA mobile or web application

Acknowledgements

I would like to thank my research mentor, Dr Loke Yuan Ren, as well as my teacher-mentors, Mr Harapan Ong, Dr Tan Guoxian, and Dr Lee Zhiying, for their invaluable guidance and feedback through the course of this project

References



[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] K. Aizawa, A. Fujimoto, A. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota, and H. Ikuta, "Building a manga dataset 'manga109' with annotations for multimedia applications," *IEEE MultiMedia*, vol. 27, no. 2, pp. 8–18, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, en D. Ha, "Deep Learning for Classical Japanese Literature", 2018.

[6] D. Dubray and J. Laubrock, "Deep CNN-based speech balloon detection and segmentation for comic books," *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[7] A. Dutta and S. Biswas, "CNN based extraction of panels/characters from Bengali comic book Page images," *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019.

[8] R. Smith, "An overview of the tesseract OCR engine," *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 2007.