MangaNet: Building an Object Detection System for Mangas

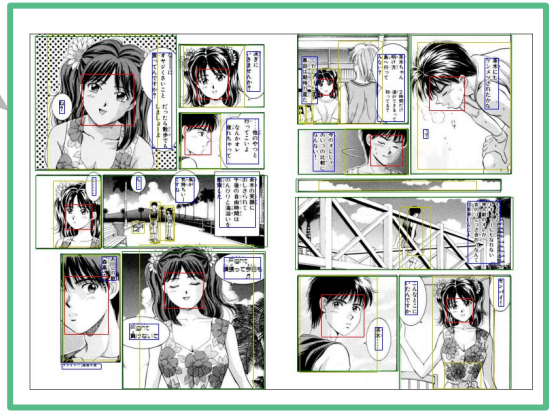
Robert Yang and Tai Vu

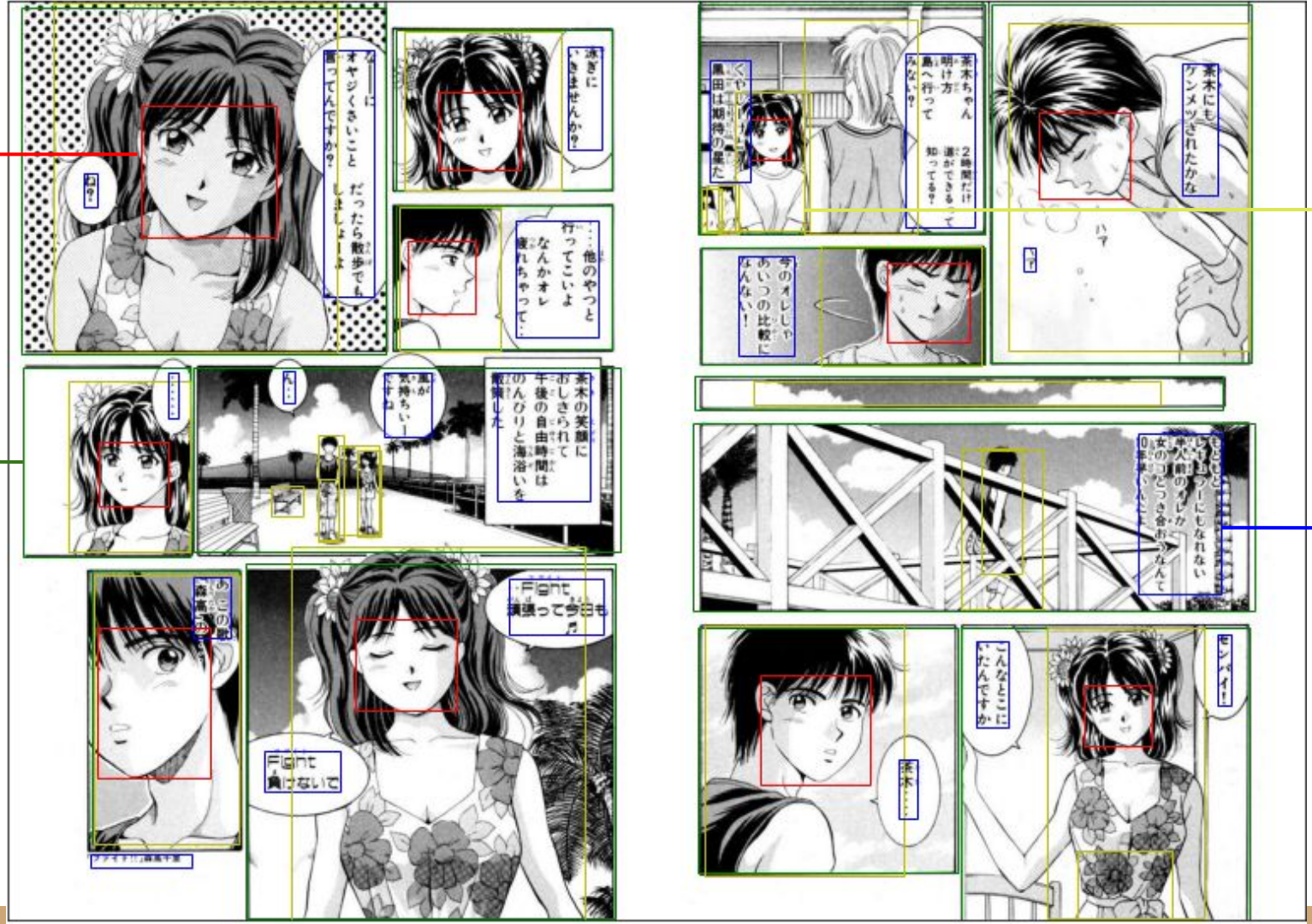
Problem Statement

Model

Frame
Text
Body
Face

Let’s zoom in




Face
Frame
Body
Text

Technical Challenges

Overlapping objects

Variation in characters' face, body, etc.

Large areas of whitespace

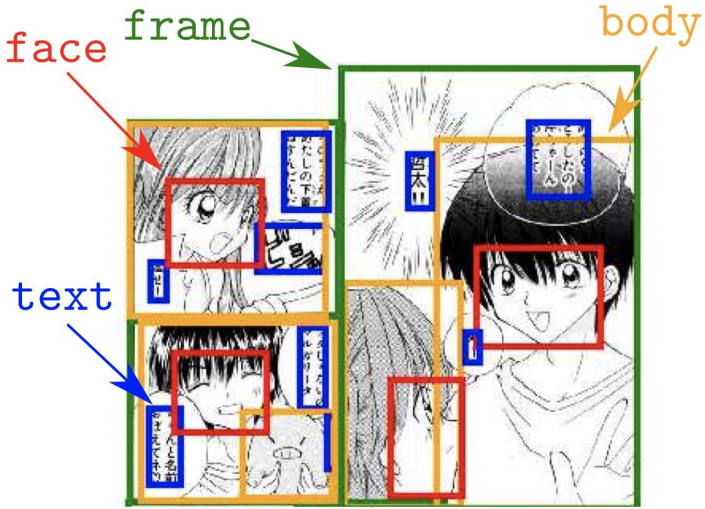
Grayscale images lack crucial color information

Related Works

- Object detection approaches: R-CNN, Fast R-CNN, Faster R-CNN, SSD, YOLO, RetinaNet
- Rigaud et al. (2011): connected component labeling + k-means clustering => classifying texts & frames
- Sun et al. (2013): SIFT descriptors + local feature matching => identifying characters
- Ogawa et al. (2018): SSD + YOLOv2 => localizing objects

Dataset

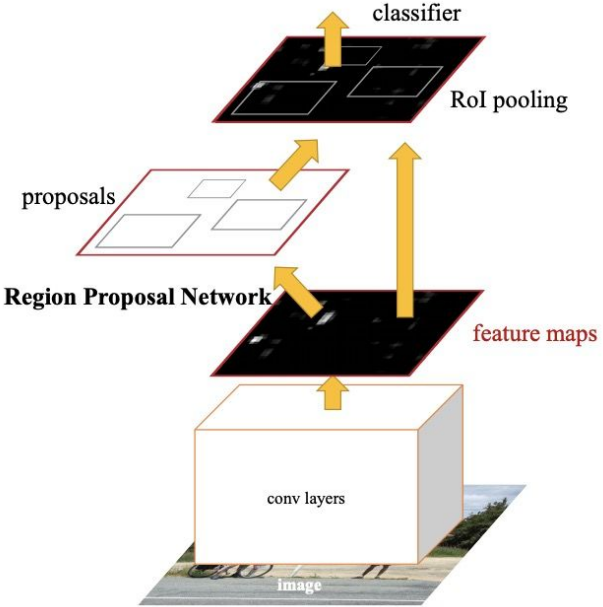
- Manga109 dataset
- 10130 examples
- Classes: frame, text, face, body
- 80% training, 10% validation, 10% testing



Ogawa, Toru, et al. "Object detection for comics using manga109 annotations." arXiv preprint arXiv:1803.08670 (2018).

Approach: Faster R-CNN

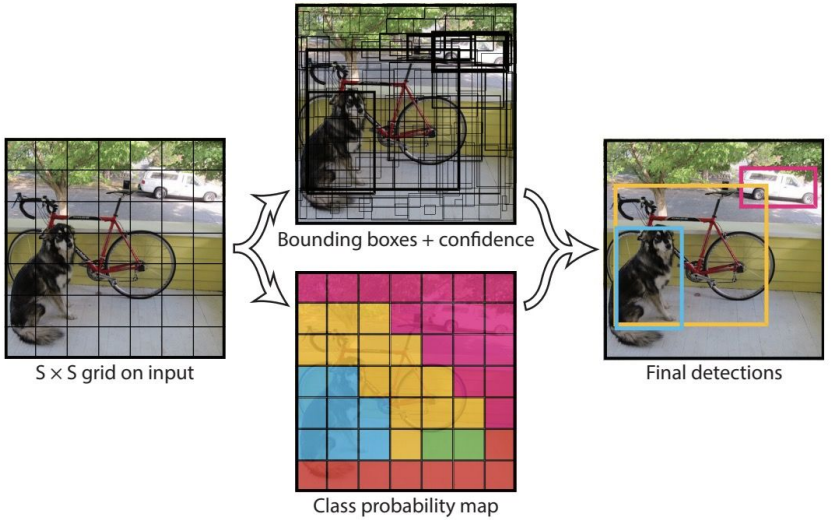
- Loss function:
  - Classification loss (cross entropy)
  - Bounding box regression loss (L1)



Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." arXiv preprint arXiv:1506.01497 (2015).

Approach: YOLOv3

- Loss function:
  - Classification loss (L2)
  - Bounding box regression loss (L2)

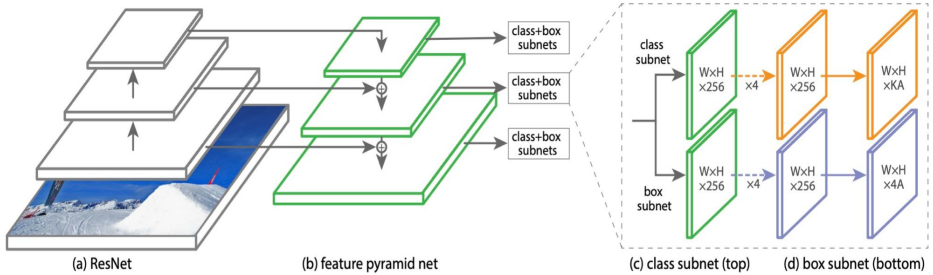


Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

Approach: RetinaNet

- Loss function:
  - Focal loss

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1-p_\mathrm{t})^\gamma \log(p_\mathrm{t}).$$



Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

Experiments: Neural Style Transfer as a featurizer

Overlapping objects

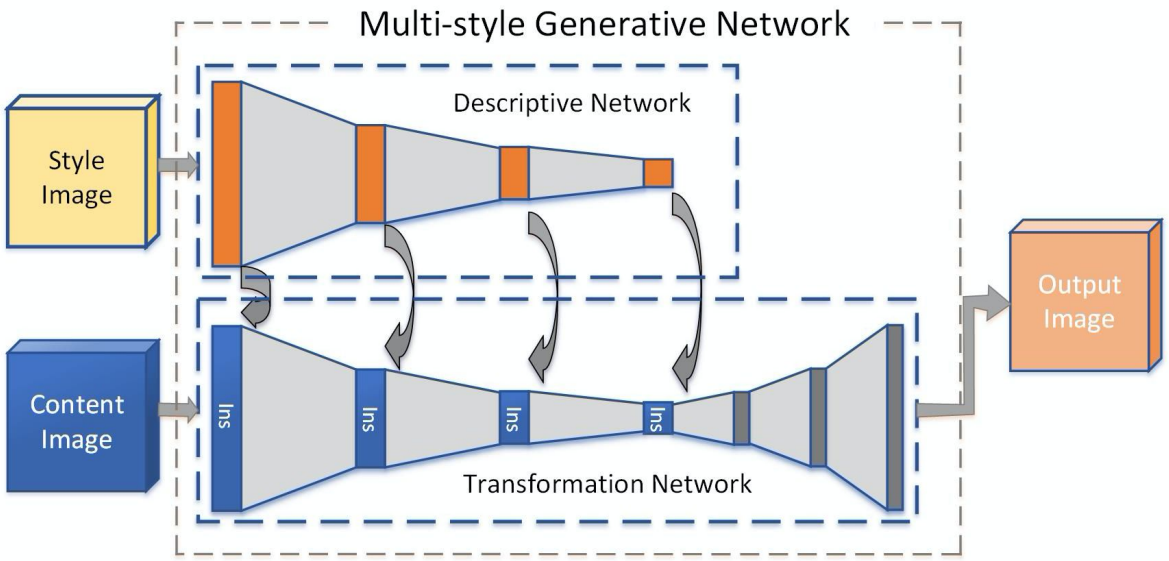
Variation in characters' face, body, etc.

Large areas of whitespace

Grayscale images lack crucial color information

Experiments: Neural Style Transfer as a featurizer

We used MSG-Net (Zhang and Dana 2018), an advanced version of style transfer



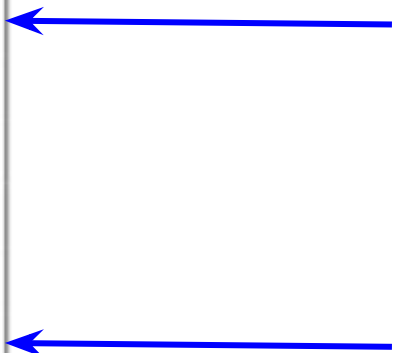
Source: Multi-style Generative Network for Real-time Transfer, Zhang and Dana, 2018

Experiments

- Data augmentation: pixel removal, affine transformation, brightness change, horizontal flipping
- Hyperparameters:
  - Faster R-CNN, RetinaNet: learning rate 0.0001, Adam, batch size 8
  - YOLOv3: learning rate 0.001, Adam, batch size 16
- Metric: Mean Average Precision (mAP)

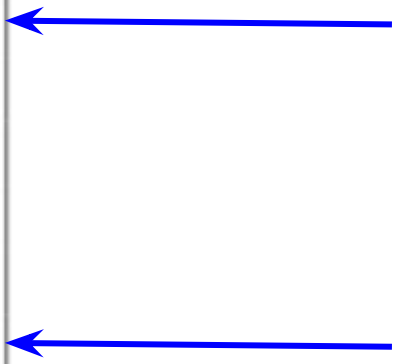
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



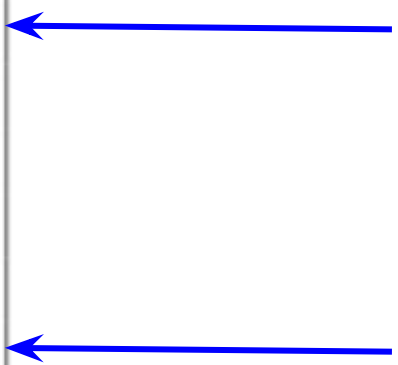
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



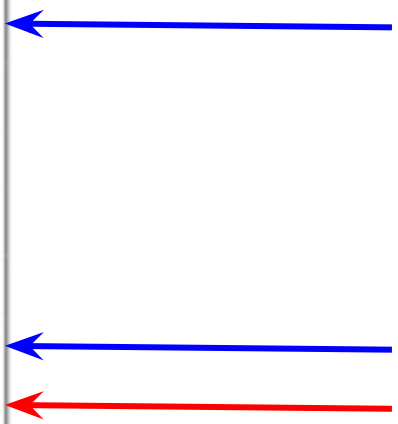
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



Results

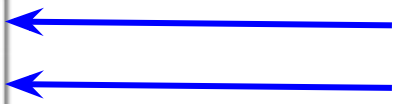
Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**

Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**

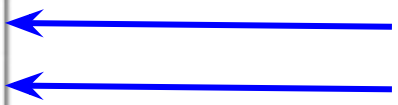
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



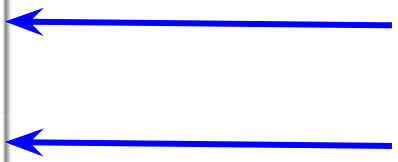
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**



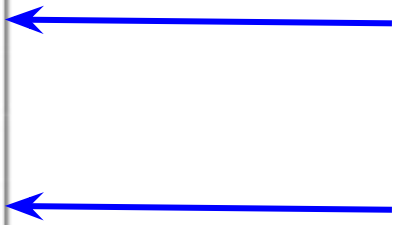
Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**

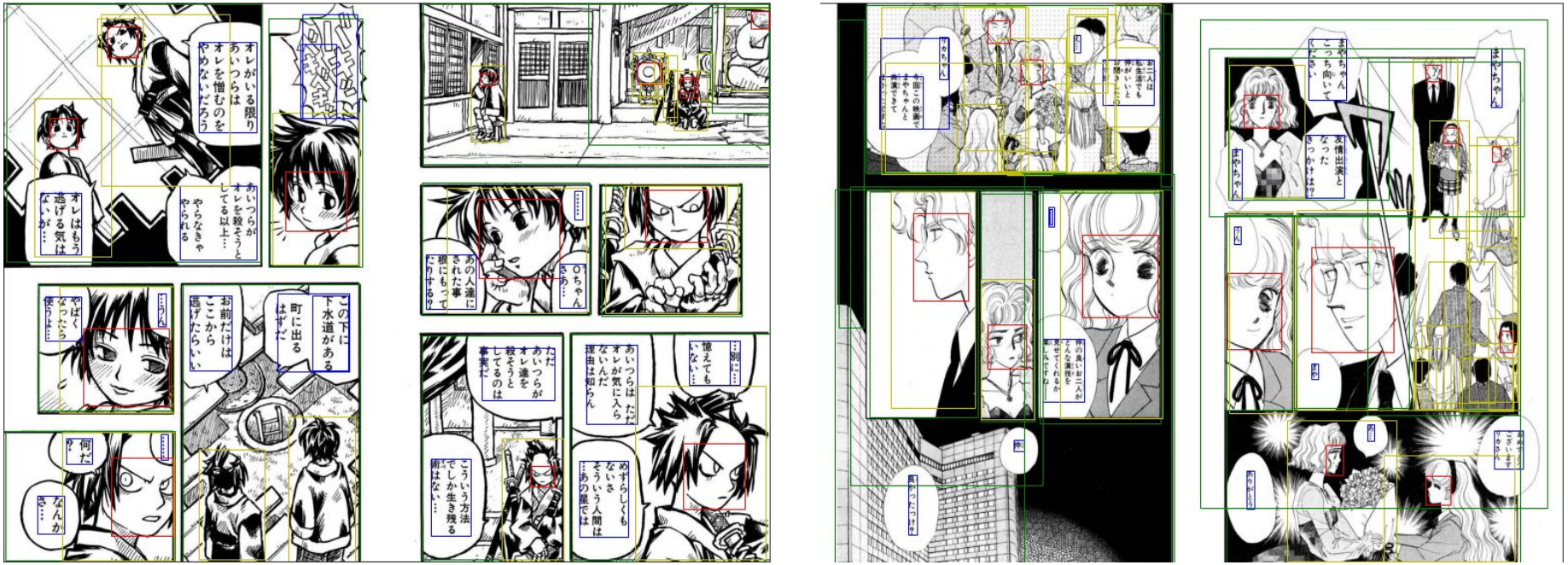


Results

Model	mAP
YOLOv2 (baseline) [19]	59.7
Faster R-CNN	57.4
Faster R-CNN (TL)	62.0
Faster R-CNN (TL + DA)	64.5
Faster R-CNN (TL + NST1)	55.4
Faster R-CNN (TL + NST2)	56.4
RetinaNet	57.4
RetinaNet (TL)	60.8
RetinaNet (TL + DA)	63.5
YOLOv3 (TL + DA)	**71.0**

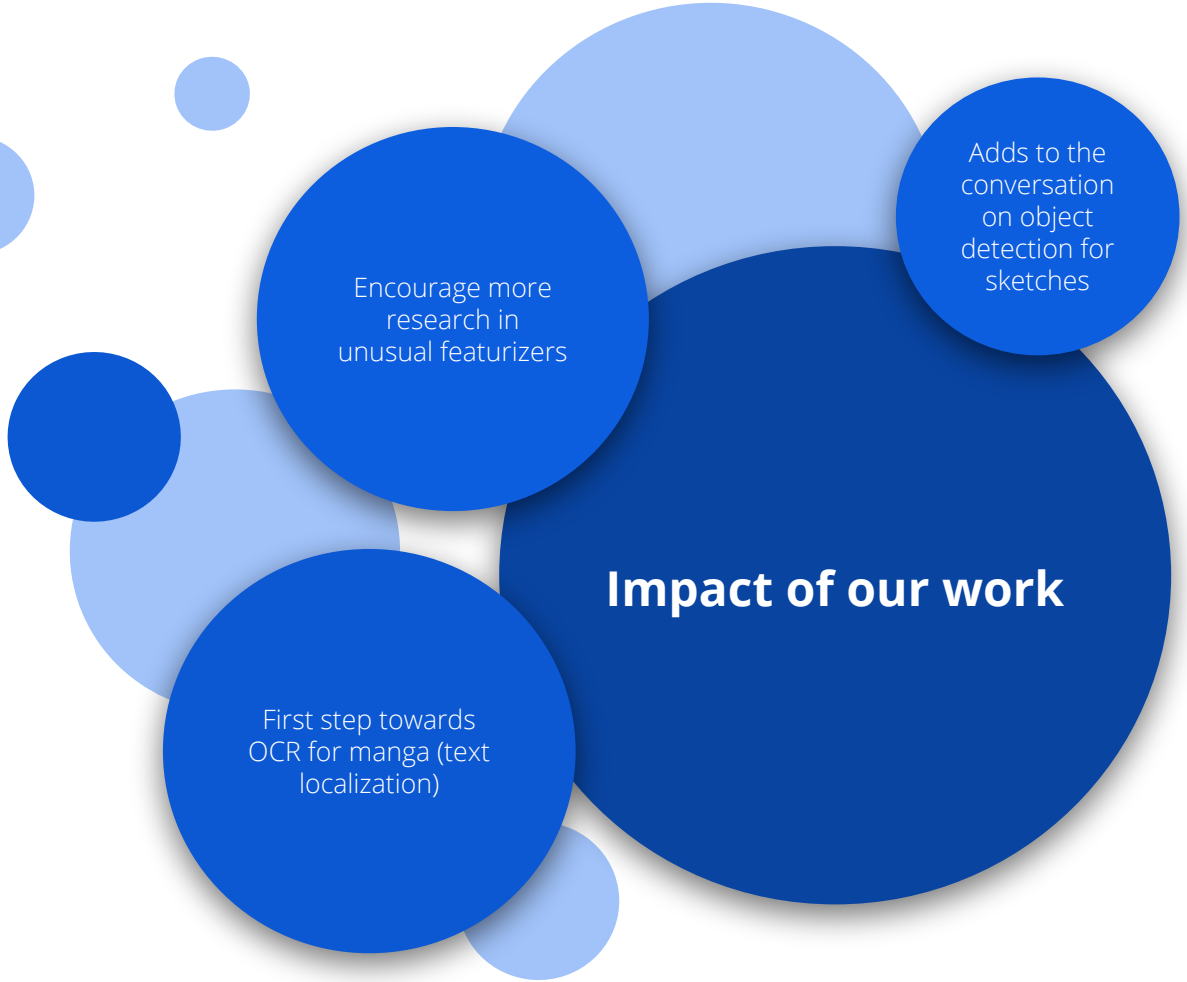


Qualitative Results



Labels: Frame, Text, Face, Body

Broader Impact



Broader Impact