Extracting figures from scanned electronic theses and dissertations

M.S. thesis defense

Aug. 6, 2020
Virginia Tech
Blacksburg, VA 24061

Student:
Sampanna Kahu (Dept. of Electrical and Computer Engineering)

Committee:
Dr. Edward A. Fox, Chair (Professor, Dept. of Computer Science, and by courtesy, ECE)
Dr. A. Lynn Abbott (Professor, Dept. of Electrical and Computer Engineering)
Dr. William Diehl (Assistant Professor, Dept. of Electrical and Computer Engineering)

Slide count: 57

Outline

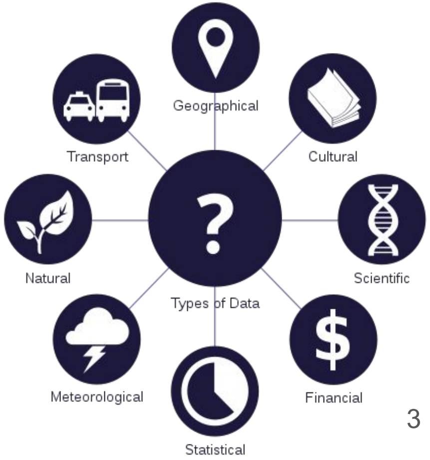
Introduction

Extracting figures* from scanned electronic theses and dissertations.

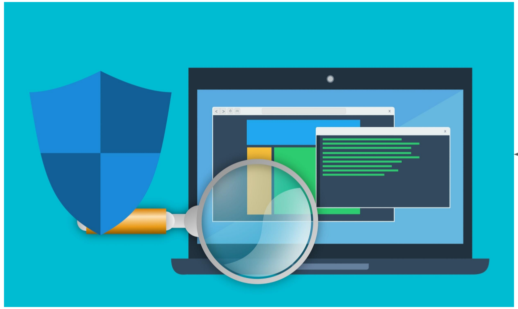
- Vast majority of published research is in PDF.
- Downstream tasks rely on accurate figure extraction

Publish

Figure extraction*



Scientific search and information extraction



*Refers to both figure and table extraction in the rest of the presentation.

Image sources:
[1] https://pxhere.com/en/photo/1451109
[2] https://pixabay.com/illustrations/pdf-logo-adobe-filetype-mime-type-3383632/
[3] https://commons.wikimedia.org/wiki/File:Exclamation_Circle_Red.svg
[4] https://commons.wikimedia.org/wiki/File:Data_types_-_en.svg
[5] https://pxhere.com/en/photo/1565521

Introduction

<u>Extracting figures</u> from scanned electronic theses and dissertations.

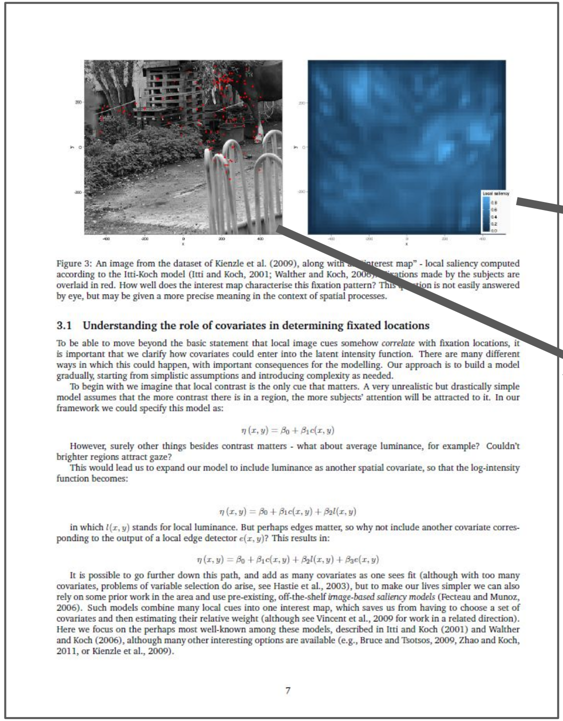


Figure 3: An image from the dataset of Kienzle et al. (2009), along with an “interest map” - local saliency computed according to the Itti-Koch model (Itti and Koch, 2001; Walther and Koch, 2006). Fixations made by the subjects are overlaid in red. How well does the interest map characterise this fixation pattern? This question is not easily answered by eye, but may be given a more precise meaning in the context of spatial processes.

3.1 Understanding the role of covariates in determining fixated locations

To be able to move beyond the basic statement that local image cues somehow *correlate* with fixation locations, it is important that we clarify how covariates could enter into the latent intensity function. There are many different ways in which this could happen, with important consequences for the modelling. Our approach is to build a model gradually, starting from simplistic assumptions and introducing complexity as needed.

To begin with we imagine that local contrast is the only cue that matters. A very unrealistic but drastically simple model assumes that the more contrast there is in a region, the more subjects' attention will be attracted to it. In our framework we could specify this model as:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y)$$

However, surely other things besides contrast matters - what about average luminance, for example? Couldn't brighter regions attract gaze?

This would lead us to expand our model to include luminance as another spatial covariate, so that the log-intensity function becomes:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y)$$

in which $l(x, y)$ stands for local luminance. But perhaps edges matter, so why not include another covariate corresponding to the output of a local edge detector $e(x, y)$? This results in:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y) + \beta_3 e(x, y)$$

It is possible to go further down this path, and add as many covariates as one sees fit (although with too many covariates, problems of variable selection do arise, see Hastie et al., 2003), but to make our lives simpler we can also rely on some prior work in the area and use pre-existing, off-the-shelf *image-based saliency models* (Fecteau and Munoz, 2006). Such models combine many local cues into one interest map, which saves us from having to choose a set of covariates and then estimating their relative weight (although see Vincent et al., 2009 for work in a related direction). Here we focus on the perhaps most well-known among these models, described in Itti and Koch (2001) and Walther and Koch (2006), although many other interesting options are available (e.g., Bruce and Tsotsos, 2009, Zhao and Koch, 2011, or Kienzle et al., 2009).

7

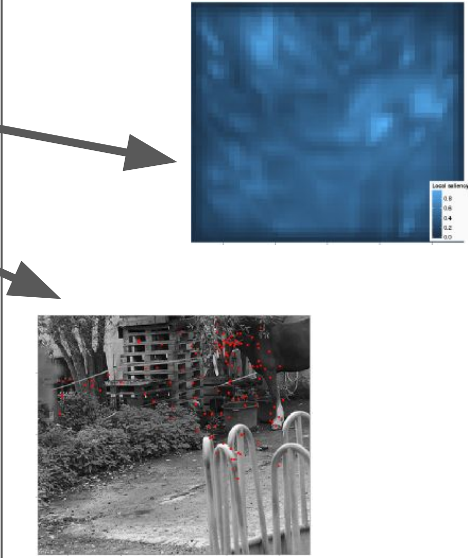


Image sources:
[1] Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann.2012. Modelling fixation locations using spatial point processes. (2012). arXiv:stat.AP/1207.2370 http://arxiv.org/abs/1207.23701 (Page no. 7)

Introduction

Extracting figures from <u>scanned</u> electronic theses and dissertations.

Born digital PDF files:

Contain the complete description to render its elements (text, fonts, vector graphics, raster images, etc.)

Scanned PDF files:

Originally handwritten or typed using a typewriter.

Later digitized using scanning devices.

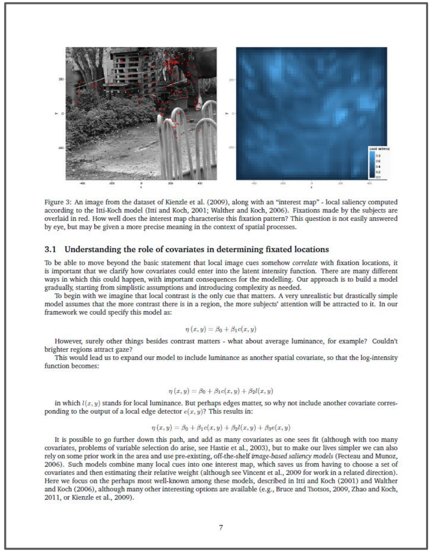


Figure 3: An image from the dataset of Kienzle et al. (2009), along with an “interest map” - local saliency computed according to the Itti-Koch model (Itti and Koch, 2001; Walther and Koch, 2006). Fixations made by the subjects are overlaid in red. How well does the interest map characterise this fixation pattern? This question is not easily answered by eye, but may be given a more precise meaning in the context of spatial processes.

3.1 Understanding the role of covariates in determining fixated locations

To be able to move beyond the basic statement that local image cues somehow *correlate* with fixation locations, it is important that we clarify how covariates could enter into the latent intensity function. There are many different ways in which this could happen, with important consequences for the modelling. Our approach is to build a model gradually, starting from simplistic assumptions and introducing complexity as needed.

To begin with we imagine that local contrast is the only cue that matters. A very unrealistic but drastically simple model assumes that the more contrast there is in a region, the more subjects' attention will be attracted to it. In our framework we could specify this model as:

$$\eta(x,y) = \beta_0 + \beta_1 c(x,y)$$

However, surely other things besides contrast matters - what about average luminance, for example? Couldn't brighter regions attract gaze?

This would lead us to expand our model to include luminance as another spatial covariate, so that the log-intensity function becomes:

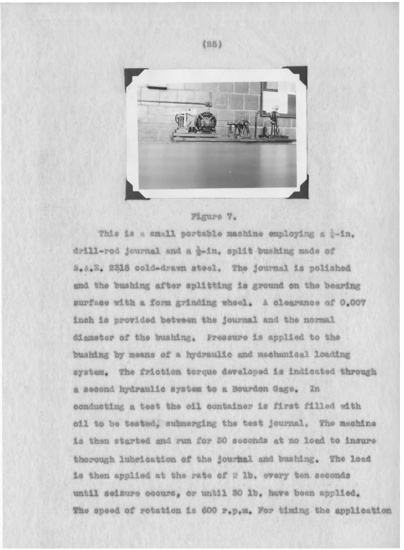
$$\eta(x,y) = \beta_0 + \beta_1 c(x,y) + \beta_2 l(x,y)$$

in which $l(x,y)$ stands for local luminance. But perhaps edges matter, so why not include another covariate corresponding to the output of a local edge detector $e(x,y)$? This results in:

$$\eta(x,y) = \beta_0 + \beta_1 c(x,y) + \beta_2 l(x,y) + \beta_3 e(x,y)$$

It is possible to go further down this path, and add as many covariates as one sees fit (although with too many covariates, problems of variable selection do arise, see Hastie et al., 2003), but to make our lives simpler we can also rely on some prior work in the area and use pre-existing, off-the-shelf *image-based saliency models* (Fecteau and Munoz, 2006). Such models combine many local cues into one interest map, which saves us from having to choose a set of covariates and then estimating their relative weight (although see Vincent et al., 2009 for work in a related direction). Here we focus on the perhaps most well-known among these models, described in Itti and Koch (2001) and Walther and Koch (2006), although many other interesting options are available (e.g., Bruce and Tsotsos, 2009, Zhao and Koch, 2011, or Kienzle et al., 2009).

7



(25)

Figure 7.

This is a small portable machine employing a $\frac{1}{2}$-in. drill-rod journal and a $\frac{1}{2}$-in. split bushing made of S.A.E. 2315 cold-drawn steel. The journal is polished and the bushing after splitting is ground on the bearing surface with a form grinding wheel. A clearance of 0.007 inch is provided between the journal and the normal diameter of the bushing. Pressure is applied to the bushing by means of a hydraulic and mechanical loading system. The friction torque developed is indicated through a second hydraulic system to a Bourdon Gage. In conducting a test the oil container is first filled with oil to be tested, submerging the test journal. The machine is then started and run for 30 seconds at no load to insure thorough lubrication of the journal and bushing. The load is then applied at the rate of 2 lb. every ten seconds until seizure occurs, or until 30 lb. have been applied. The speed of rotation is 600 r.p.m. For timing the application

Image sources:
[1] Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann.2012. Modelling fixation locations using spatial point processes. (2012). arXiv:stat.AP/1207.2370 http://arxiv.org/abs/1207.23701 (Page no. 7)
[2] Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159
[3] https://www.123rf.com/photo_78921823_modern-desktop-pc-computer-isolated-.html
[4] https://www.brandeps.com/logo/M/Microsoft-Word-01
[5] https://i.stack.imgur.com/zHFFO.png
[6] https://www.vectorstock.com/royalty-free-vector/the-old-portable-typewriter-vector-22613165
[7] https://images.techhive.com/images/article/2016/01/flatbed-scanner-stock-100636615-large.jpg

Introduction

Extracting figures from scanned electronic theses and dissertations.

This work focuses on ETDs, which are longer, book-length documents.

But can possibly be extended to other scientific documents.

ETD=Electronic Theses and Dissertations

Outline

1. Introduction
2. **Research questions**
3. Related work
4. Methodology
    a. Data augmentation
    b. Training at scale
    c. Gold standard
5. Experiments
    a. Experiments, results/discussion, answers to research questions.
6. Conclusions
7. Future work

Research questions

RQ1: How well can existing methods perform figure extraction from scanned ETDs?

RQ2: Can this performance be improved by using simple data augmentation techniques and weight initialization from the original pre-trained model?

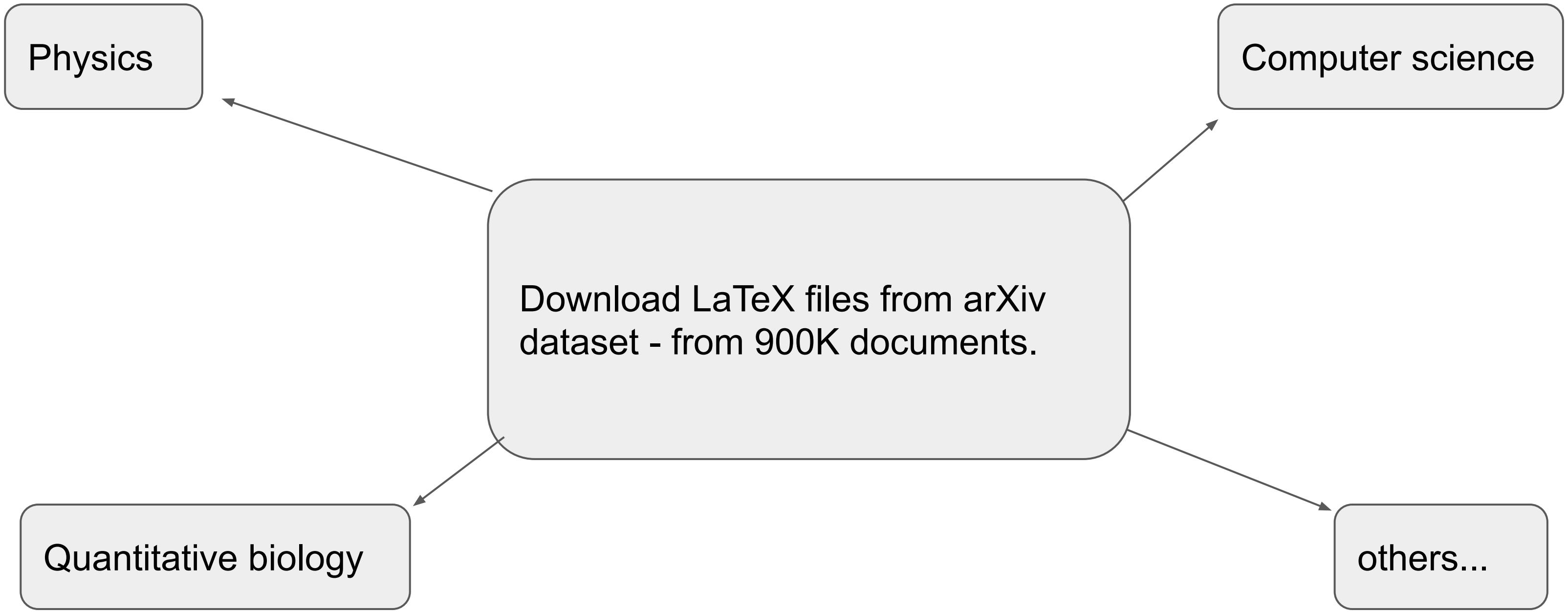
RQ3: Can this performance be improved by training on manually labelled data?

RQ4: Can this performance be improved by using transfer learning techniques?

Outline

1. Introduction
2. Research questions
3. **Related work**
4. Methodology
    a. Data augmentation
    b. Training at scale
    c. Gold standard
5. Experiments
    a. Experiments, results/discussion, answers to research questions.
6. Conclusions
7. Future work

Related work - Deepfigures

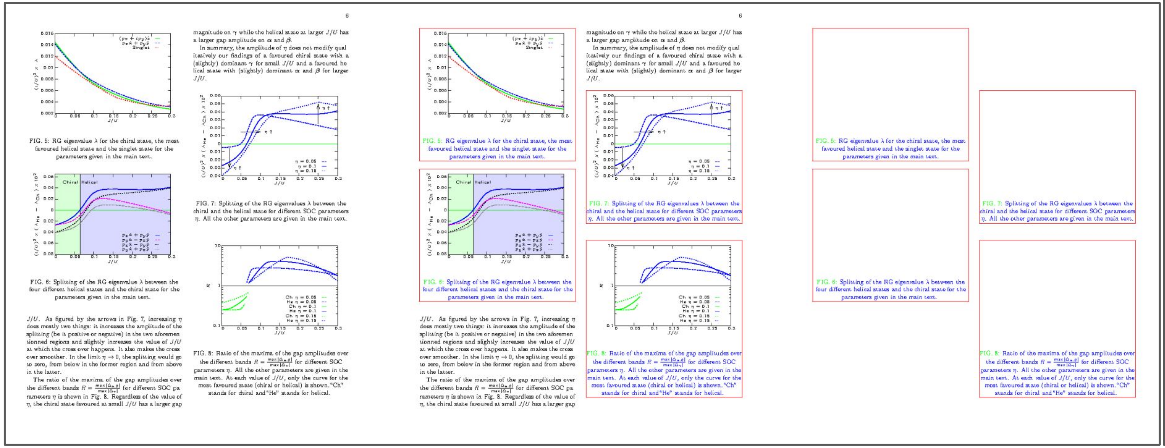


Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRR abs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445

```
\usepackage{color}
\usepackage{floatrow}
\usepackage{tcolorbox}

\DeclareColorBox{figurecolorbox}{\fcolorbox{red}{white}}
\DeclareColorBox{tablecolorbox}{\fcolorbox{yellow}{white}}

\floatsetup[figure]{framestyle=colorbox,
    colorframeset=figurecolorbox, framearound=all,
    frameset={\fboxrule1pt\fboxsep0pt}}
\floatsetup[table]{framestyle=colorbox,
    colorframeset=tablecolorbox, framearound=all,
    frameset={\fboxrule1pt\fboxsep0pt}}
```



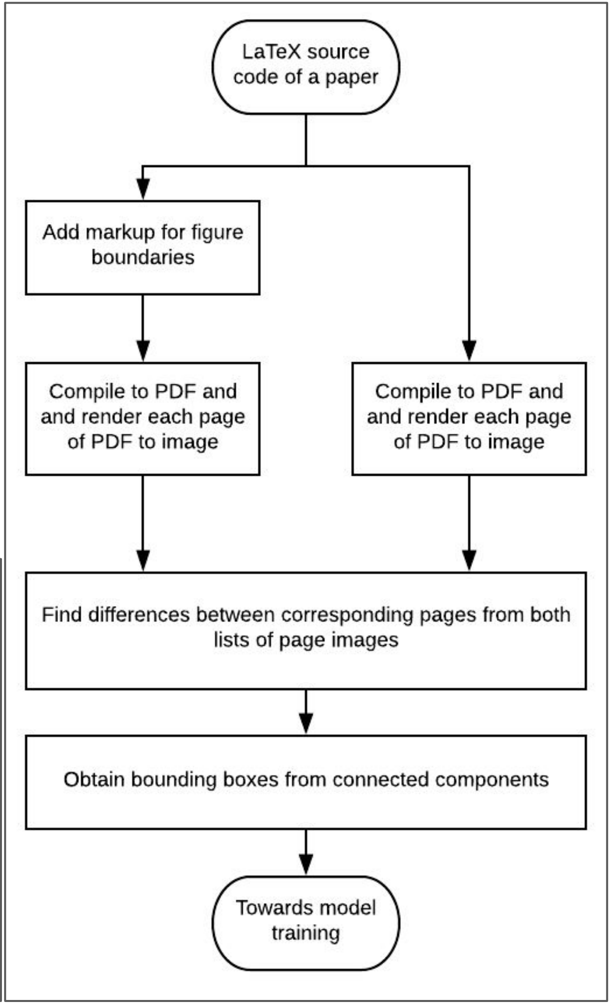
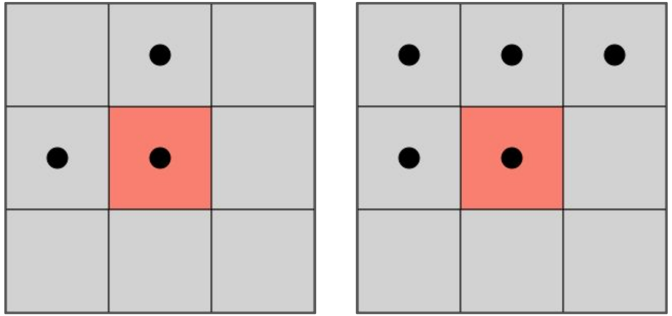


Image sources:
[1] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRR abs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445

Related work - Deepfigures

Connected Component Labelling Algorithm: Assume that region pixels have the value 0 (black) and that background pixels have the value 255 (white).

1. Scan the image to find an unlabeled 0 (pixel and assign it a new label L.
2. Recursively assign a label L to all of its 0 neighbors.
3. Stop if there are no more unlabeled 0 pixels.
4. Go to step 1.



4-connectivity 8-connectivity

[1] https://en.wikipedia.org/wiki/Connected-component_labeling
[2] https://www.cse.unr.edu/~bebis/CS791E/Notes/ConnectedComponents.pdf

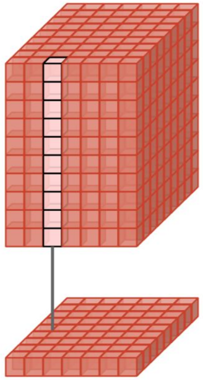
Related work - Deepfigures

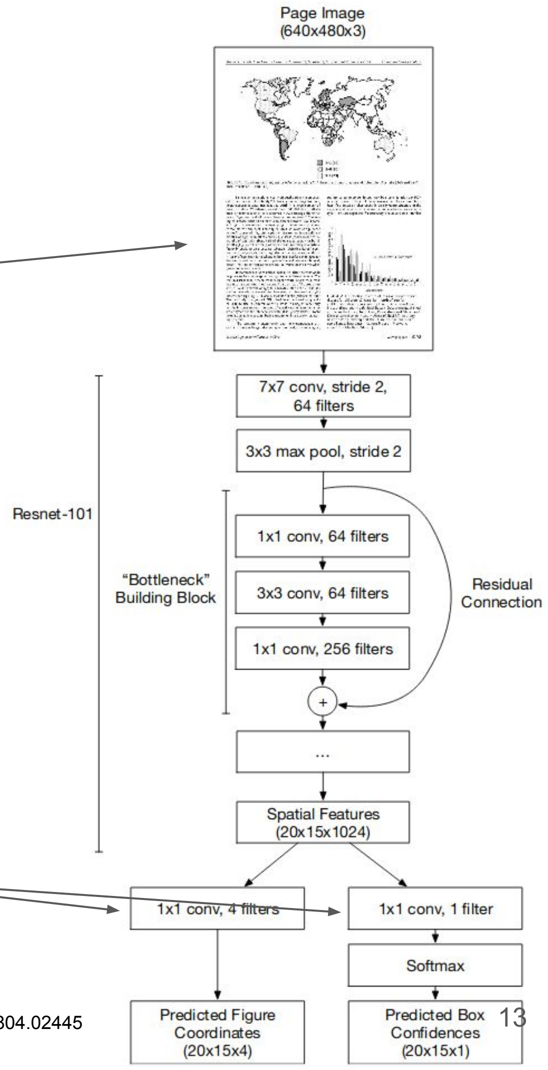
The input page of the PDF
(converted to an image 640x480)

The Resnet-101 model

Model	CS-Large	PubMed
Deepfigures	0.849	0.806

F1-scores of Deepfigures on manually labelled datasets[1] of born-digital documents.





[1] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRR abs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445

Outline

1. Introduction
2. Research questions
3. Related work
4. **Methodology**
    a. Data augmentation
    b. Training at scale
    c. Gold standard
5. Experiments
    a. Experiments, results/discussion, answers to research questions.
6. Conclusions
7. Future work

Feature distributions (Born-digital vs. scanned)

Features:

Font, line spacing, content layout, scanner noise, etc.

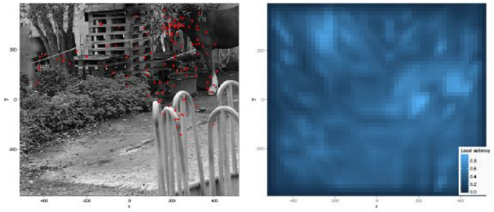


Figure 3: An image from the dataset of Kienzle et al. (2009), along with an "interest map" - local saliency computed according to the Itti-Koch model (Itti and Koch, 2001; Walther and Koch, 2006). Fixations made by the subjects are overlaid in red. How well does the interest map characterise this fixation pattern? This question is not easily answered by eye, but may be given a more precise meaning in the context of spatial processes.

3.1 Understanding the role of covariates in determining fixated locations

To be able to move beyond the basic statement that local image cues somehow *correlate* with fixation locations, it is important that we clarify how covariates could enter into the latent intensity function. There are many different ways in which this could happen, with important consequences for the modelling. Our approach is to build a model gradually, starting from simplistic assumptions and introducing complexity as needed.

To begin with we imagine that local contrast is the only cue that matters. A very unrealistic but drastically simple model assumes that the more contrast there is in a region, the more subjects' attention will be attracted to it. In our framework we could specify this model as:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y)$$

However, surely other things besides contrast matters - what about average luminance, for example? Couldn't brighter regions attract gaze?

This would lead us to expand our model to include luminance as another spatial covariate, so that the log-intensity function becomes:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y)$$

in which $l(x, y)$ stands for local luminance. But perhaps edges matter, so why not include another covariate corresponding to the output of a local edge detector $e(x, y)$? This results in:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y) + \beta_3 e(x, y)$$

It is possible to go further down this path, and add as many covariates as one sees fit (although with too many covariates, problems of variable selection do arise, see Hastie et al., 2003), but to make our lives simpler we can also rely on some prior work in the area and use pre-existing, off-the-shelf *image-based saliency models* (Fecteau and Munoz, 2006). Such models combine many local cues into one interest map, which saves us from having to choose a set of covariates and then estimating their relative weight (although see Vincent et al., 2009 for work in a related direction). Here we focus on the perhaps most well-known among these models, described in Itti and Koch (2001) and Walther and Koch (2006), although many other interesting options are available (e.g., Bruce and Tsotsos, 2009, Zhao and Koch, 2011, or Kienzle et al., 2009).

7

(25)

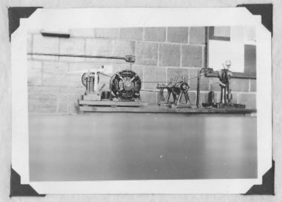


Figure 7.

This is a small portable machine employing a ½-in. drill-rod journal and a ½-in. split bushing made of S.A.E. 2315 cold-drawn steel. The journal is polished and the bushing after splitting is ground on the bearing surface with a form grinding wheel. A clearance of 0.007 inch is provided between the journal and the normal diameter of the bushing. Pressure is applied to the bushing by means of a hydraulic and mechanical loading system. The friction torque developed is indicated through a second hydraulic system to a Bourdon Gage. In conducting a test the oil container is first filled with oil to be tested, submerging the test journal. The machine is then started and run for 30 seconds at no load to insure thorough lubrication of the journal and bushing. The load is then applied at the rate of 2 lb. every ten seconds until seizure occurs, or until 30 lb. have been applied. The speed of rotation is 600 r.p.m. For timing the application

Goal:

To make the first document look like the second.

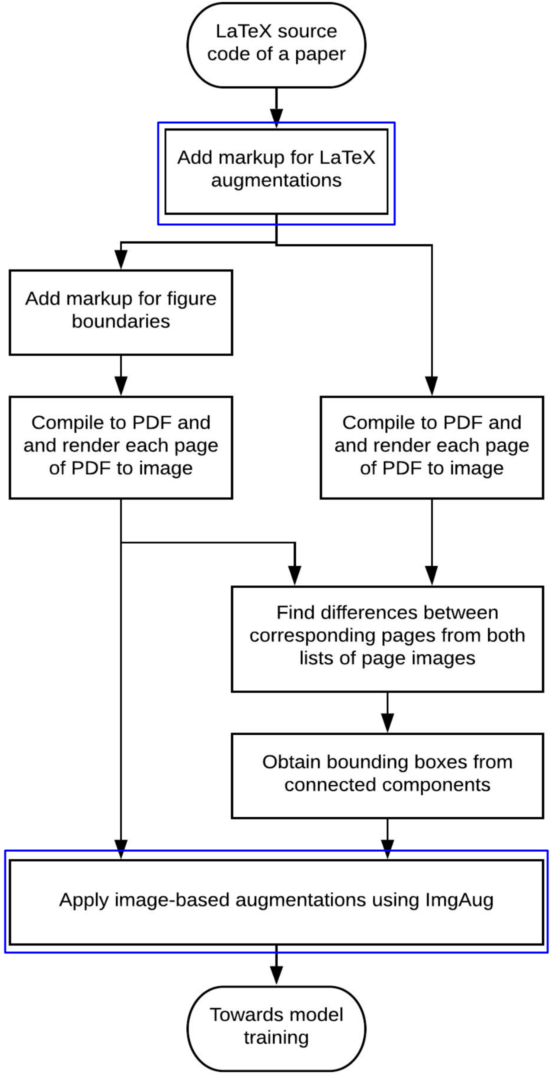
Image sources:

[1] Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann.2012. Modelling fixation locations using spatial point processes. (2012). arXiv:stat.AP/1207.2370 http://arxiv.org/abs/1207.23701 (Page no. 7)

[2] Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

Data augmentation

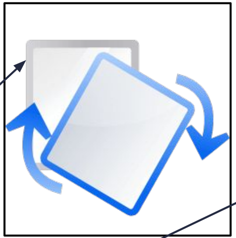
The overall pipeline along with the data augmentation steps.

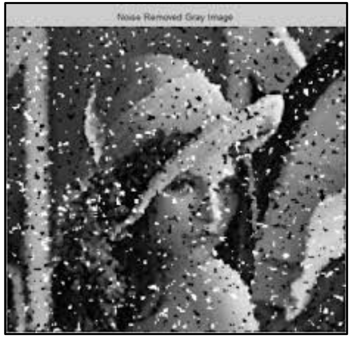


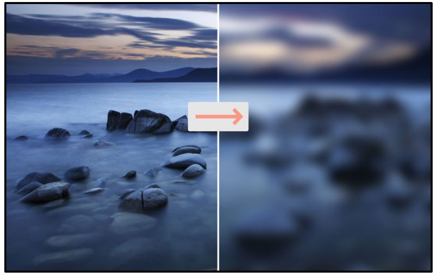
Data augmentation

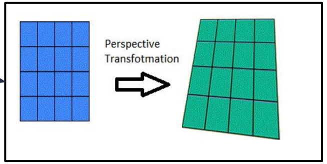
Image-based data augmentations

1. Random affine rotation (limited to +/- 5 degrees)
2. Additive Gaussian noise
3. Salt-and-pepper noise
4. Gaussian blur
5. Linear contrast
6. Perspective transform









Low Contrast Image High Contrast Image

Data augmentation

LaTeX-based data augmentations

1. Following line was replaced:
   ```
   \documentclass[sigconf]{acmart}
   \documentclass[sigconf,12pt]{acmart}
   ```
2. Following code added at the beginning:
   ```
   \renewcommand\ttdefault{cmvtt}
   \renewcommand{\familydefault}{\ttdefault}
   \linespread{1.5}
   ```

Data augmentation

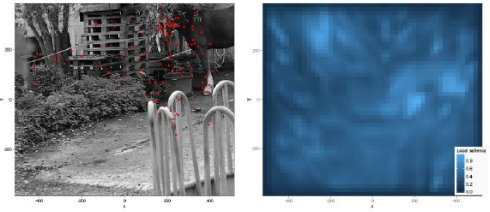


Figure 3: An image from the dataset of Kienzle et al. (2009), along with an "interest map" - local saliency computed according to the Itti-Koch model (Itti and Koch, 2001; Walther and Koch, 2006). Fixations made by the subjects are overlaid in red. How well does the interest map characterise this fixation pattern? This question is not easily answered by eye, but may be given a more precise meaning in the context of spatial processes.

3.1 Understanding the role of covariates in determining fixated locations

To be able to move beyond the basic statement that local image cues somehow *correlate* with fixation locations, is important that we clarify how covariates could enter into the latent intensity function. There are many differe ways in which this could happen, with important consequences for the modelling. Our approach is to build a model gradually, starting from simplistic assumptions and introducing complexity as needed.

To begin with we imagine that local contrast is the only cue that matters. A very unrealistic but drastically simple model assumes that the more contrast there is in a region, the more subjects' attention will be attracted to it. In our framework we could specify this model as:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y)$$

However, surely other things besides contrast matters - what about average luminance, for example? Couldn't brighter regions attract gaze?

This would lead us to expand our model to include luminance as another spatial covariate, so that the log-intensity function becomes:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y)$$

in which $l(x, y)$ stands for local luminance. But perhaps edges matter, so why not include another covariate corresponding to the output of a local edge detector $e(x, y)$? This results in:

$$\eta(x, y) = \beta_0 + \beta_1 c(x, y) + \beta_2 l(x, y) + \beta_3 e(x, y)$$

It is possible to go further down this path, and add as many covariates as one sees fit (although with too many covariates, problems of variable selection do arise, see Hastie et al., 2003), but to make our lives simpler we can also rely on some prior work in the area and use pre-existing, off-the-shelf *image-based saliency models* (Fecteau and Munoz, 2006). Such models combine many local cues into one interest map, which saves us from having to choose a set of covariates and then estimating their relative weight (although see Vincent et al., 2009 for work in a related direction). Here we focus on the perhaps most well-known among these models, described in Itti and Koch (2001) and Walther and Koch (2006), although many other interesting options are available (e.g., Bruce and Tsotsos, 2009, Zhao and Koch, 2011, or Kienzle et al., 2009).

7

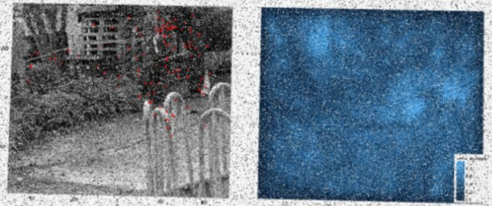


Figure 3. An image from the dataset of ?, along with an "interest map" - local saliency computed according to the Itti-Koch model (??). Fixations made by the subjects are overlaid in red. How well does the interest map characterise this fixation pattern? This question is not easily answered by eye, but may be given a more precise meaning in the context of spatial processes.

(although see ? for work in a related direction). Here we focus on the perhaps most well-known among these models, described in ? and ?, although many other interesting options are available (e.g., ?, ?, or ?).

The model computes several feature maps (orientation, contrast, etc.) according to physiologically plausible mechanisms, and combines them into one master map which aims to predict what the interesting features in image i are. For a given image i we can obtain the interest map $m_i(x, y)$ and use that as the unique covariate in a point process:

$$\eta_i(x, y) = \alpha_i + \beta_i m_i(x, y) \quad (3)$$

This last equation will be the starting point of our modelling. We have changed the notation somewhat to reflect some of the adjustments we need to make in order to learn anything from applying model to data. To summarise:

- $\eta_i(x, y)$ denotes the log-intensity function for image i, which depends on the spatial covariate $m_i(x, y)$ that corresponds to the interest map given by the low-level saliency of ?
- β_i is an image-specific coefficient that measures to what extent spatial intensity can be

13

(25)

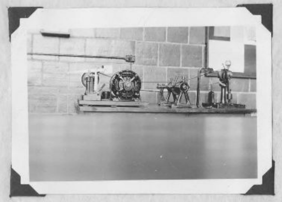


Figure 7.

This is a small portable machine employing a ½-in. drill-rod journal and a ½-in. split bushing made of S.A.E. 2315 cold-drawn steel. The journal is polished and the bushing after splitting is ground on the bearing surface with a form grinding wheel. A clearance of 0.007 inch is provided between the journal and the normal diameter of the bushing. Pressure is applied to the bushing by means of a hydraulic and mechanical loading system. The friction torque developed is indicated through a second hydraulic system to a Bourdon Gage. In conducting a test the oil container is first filled with oil to be tested, submerging the test journal. The machine is then started and run for 30 seconds at no load to insure thorough lubrication of the journal and bushing. The load is then applied at the rate of 2 lb. every ten seconds until seizure occurs, or until 30 lb. have been applied. The speed of rotation is 600 r.p.m. For timing the application

Image sources:

[1] Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann.2012. Modelling fixation locations using spatial point processes. (2012). arXiv:stat.AP/1207.2370 http://arxiv.org/abs/1207.23701 (Page no. 7)

[2] Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

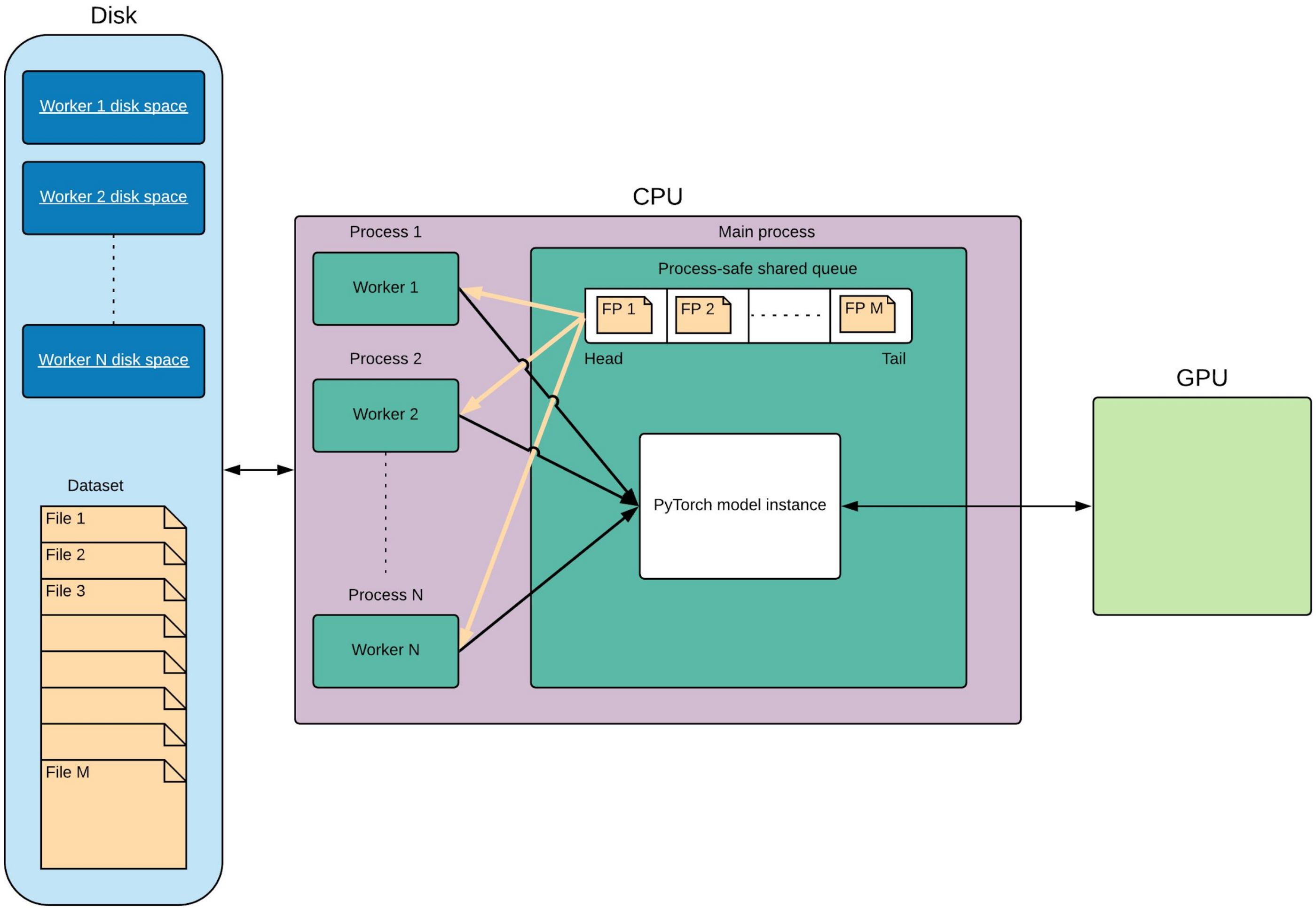
Training at scale

We primarily use the arXiv dataset:

- Compressed size on disk: 1.3 TB.
- Divided into 2600 files of 500MB each.
- Each 500MB-file contains several hundred scientific documents along with their LaTeX source code.

Training deep learning models on a huge dataset is logistically challenging:

- Not feasible to unzip and compile all LaTeX into PDFs before training.
- Computationally expensive to do using a single thread.


Disk
Worker 1 disk space
Worker 2 disk space
Worker N disk space
Dataset
File 1
File 2
File 3
File M
CPU
Process 1
Worker 1
Process 2
Worker 2
Process N
Worker N
Main process
Process-safe shared queue
FP 1
FP 2
FP M
Head
Tail
PyTorch model instance
GPU

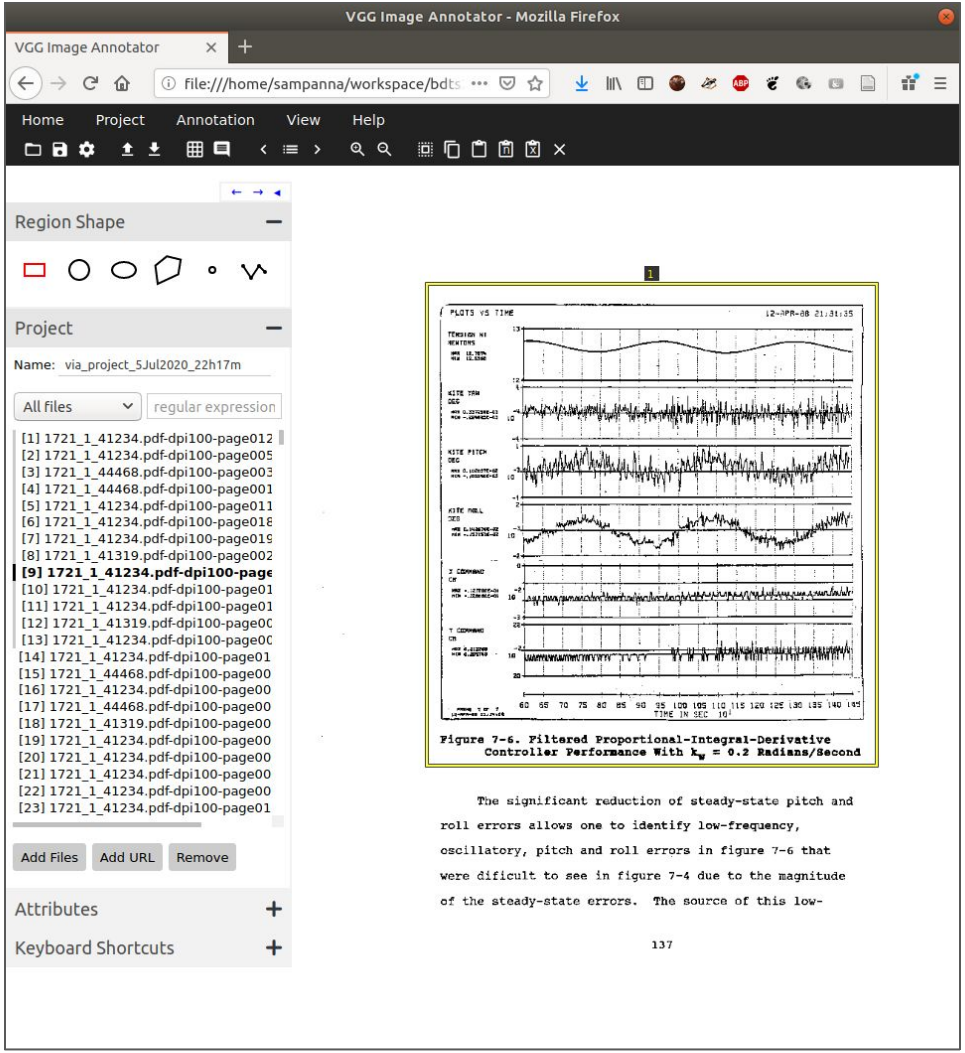
Gold standard dataset

1. Since the labels of the arXiv dataset are self-generated, labels of the augmented dataset cannot be considered as ground-truth.
2. Thus, an evaluation set labelled manually needs to be created.

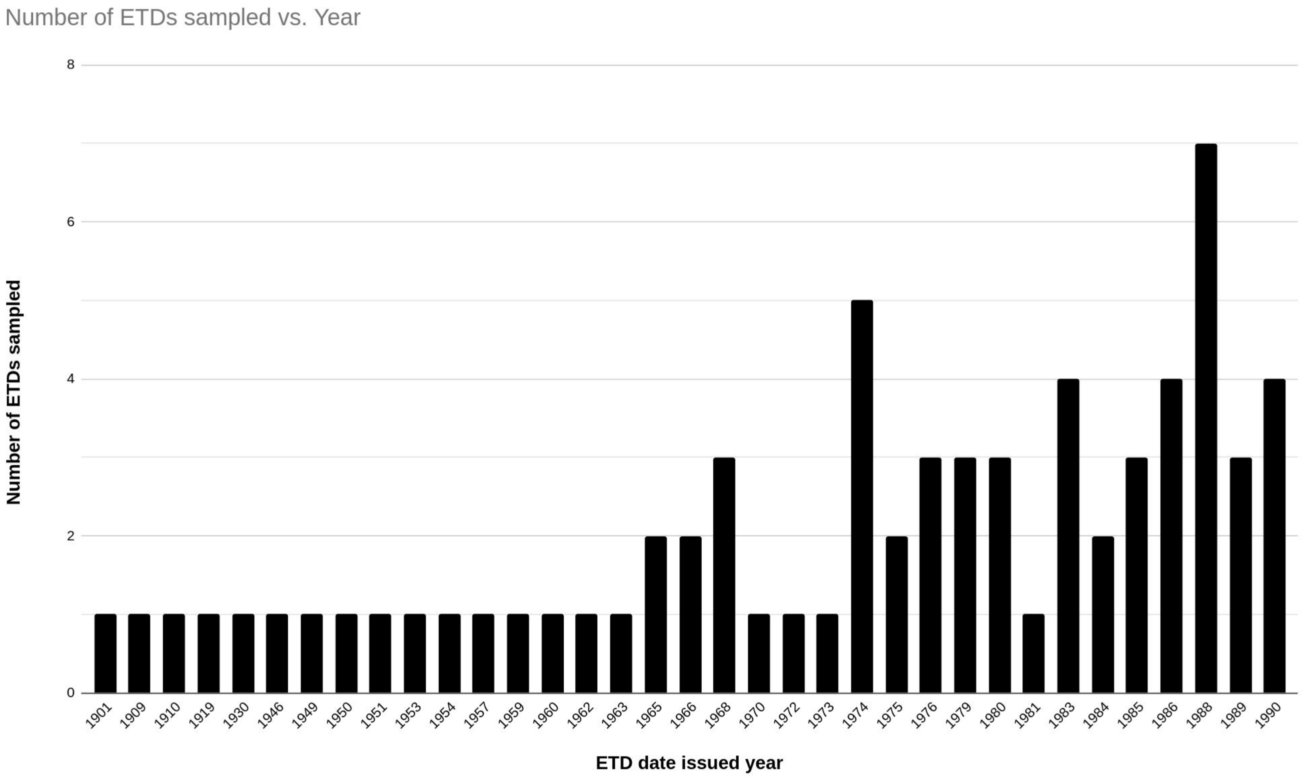
Val set (50%) (used for selecting the best model out of all the model checkpoints)	Test set (50%) (used for reporting the performance of the chosen best model)

Gold standard
(70 ETDs, 10.1K pages, manually labelled, 3.3K figures)

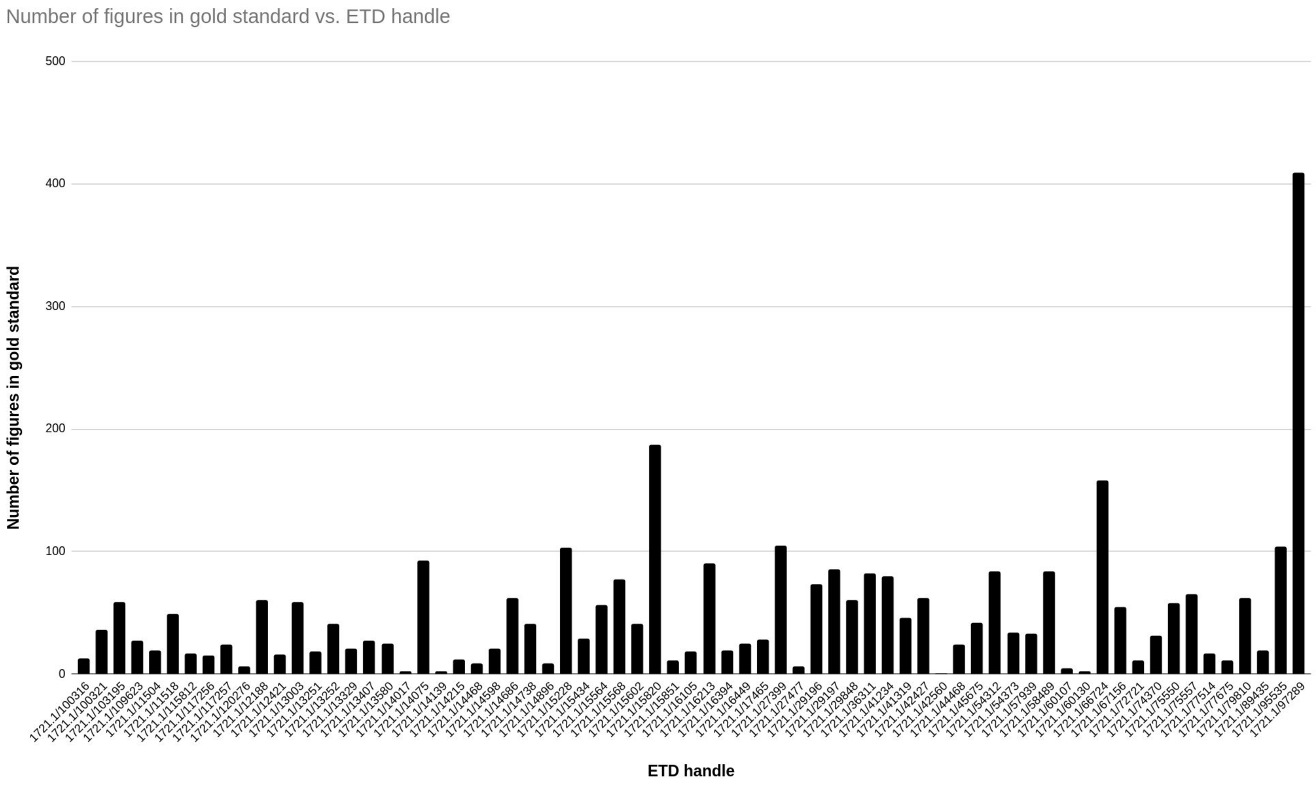
Gold standard dataset



Gold standard dataset



Gold standard dataset

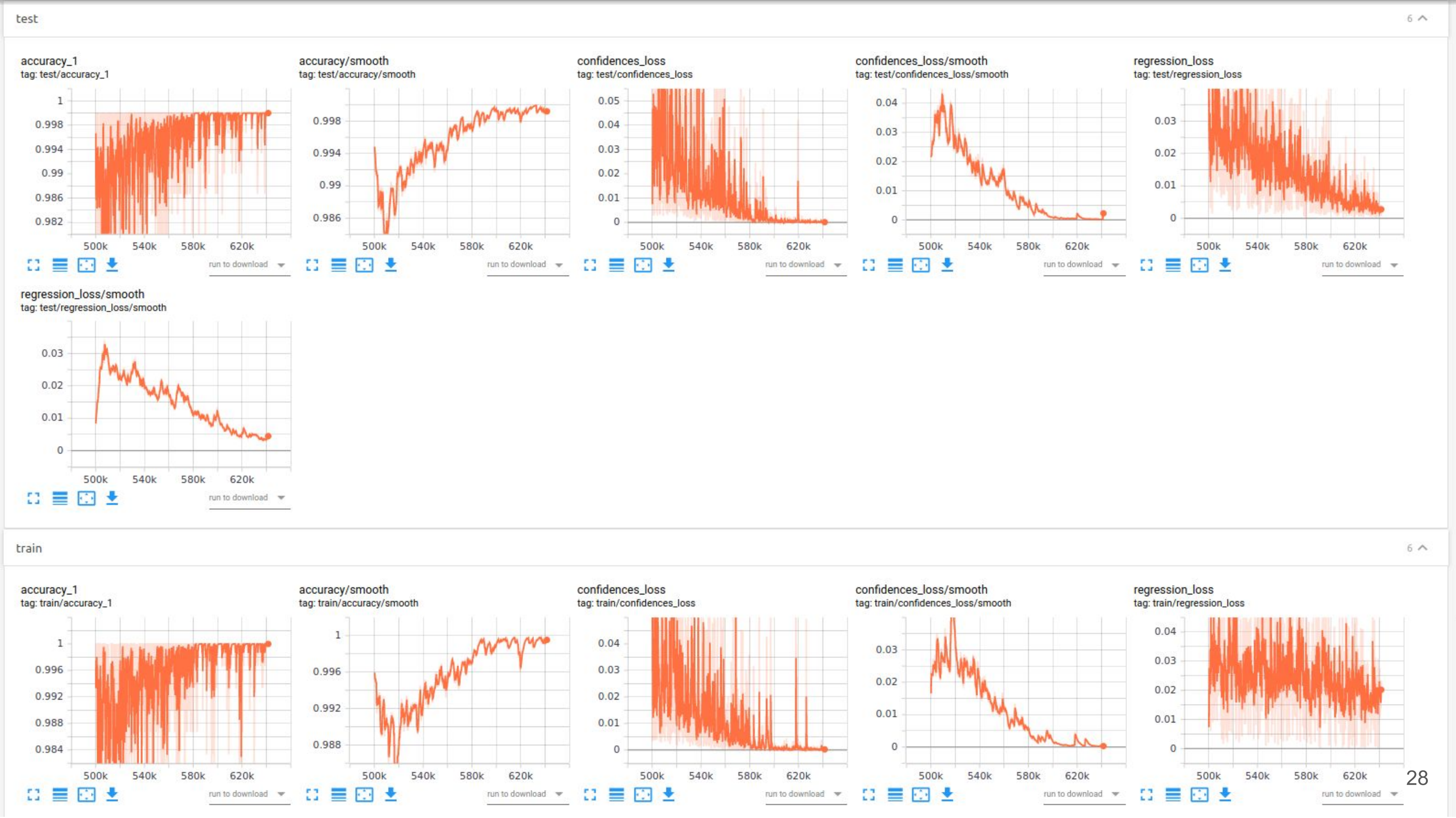


Outline

Exp1: Proof of concept

1. Experimental setup
   a. **Model**: Deepfigures
   b. **Weight initialization**: Pre-trained weights from Deepfigures
   c. **Data**: About 2% of the original arXiv data with all transformations applied. Evaluated on an old scanned ETD from VTechWorks [1].
   d. **Duration**: About 100K training steps.
   e. **Batch size**: 1

[1] Walter Douglas Chiles. Effect of service on automobile crankcase oils. PhD thesis, Virginia Agricultural and Mechanical College and Polytechnic Institute, 1935. URL http://hdl.handle.net/10919/56159.


test
6
accuracy_1
tag: test/accuracy_1
accuracy/smooth
tag: test/accuracy/smooth
confidences_loss
tag: test/confidences_loss
confidences_loss/smooth
tag: test/confidences_loss/smooth
regression_loss
tag: test/regression_loss
regression_loss/smooth
tag: test/regression_loss/smooth
500k 540k 580k 620k
run to download
train
6
accuracy_1
tag: train/accuracy_1
accuracy/smooth
tag: train/accuracy/smooth
confidences_loss
tag: train/confidences_loss
confidences_loss/smooth
tag: train/confidences_loss/smooth
regression_loss
tag: train/regression_loss

Exp1: Proof of concept

Model	TPs	FPs	FNs	Precision	Recall	F1
Deepfigures	0	29	26	0	0	0
Ours (image-based transformations)	7	16	15	0.30	0.318	0.309
Ours (all transformations)	10	15	16	0.4	0.385	0.392

TP=True Positive, FP=False Positive, FN=False Negative

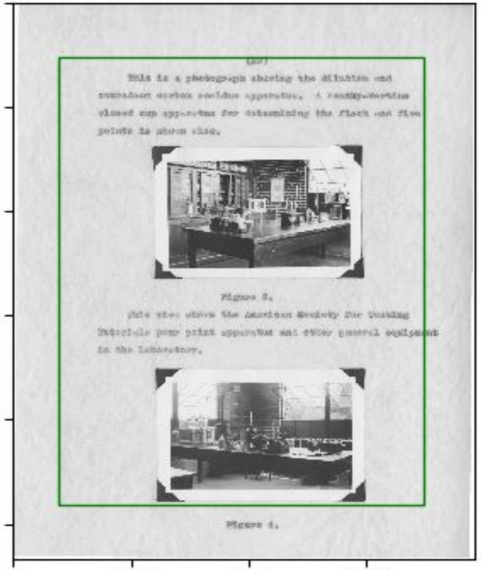
Observations:

1. Our models have a higher F1-score.
2. More augmentations result in a higher F1-score.

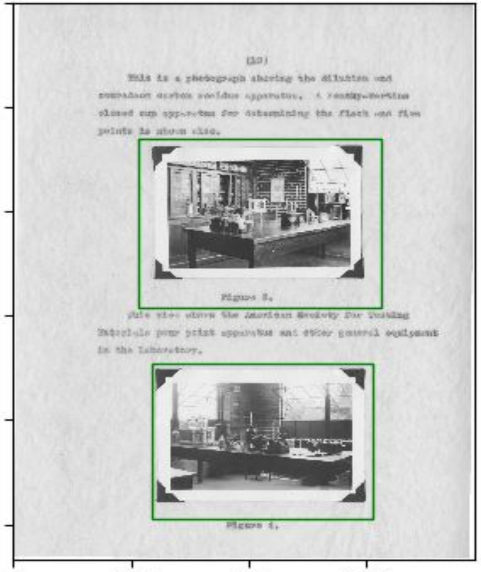
However:

1. This evaluation is only on a single ETD.
2. The TPs, FPs, FNs were manually calculated. We need better metrics.

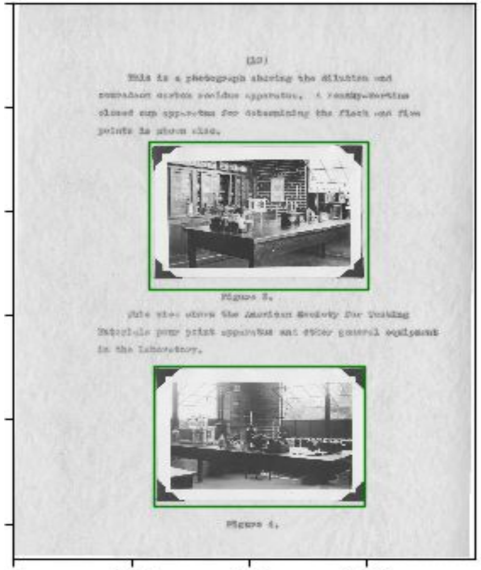
Exp1: Proof of concept



Original model



(Ours) Model trained on image-based transformations



(Ours) Model trained on all transformations

ETD source: Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

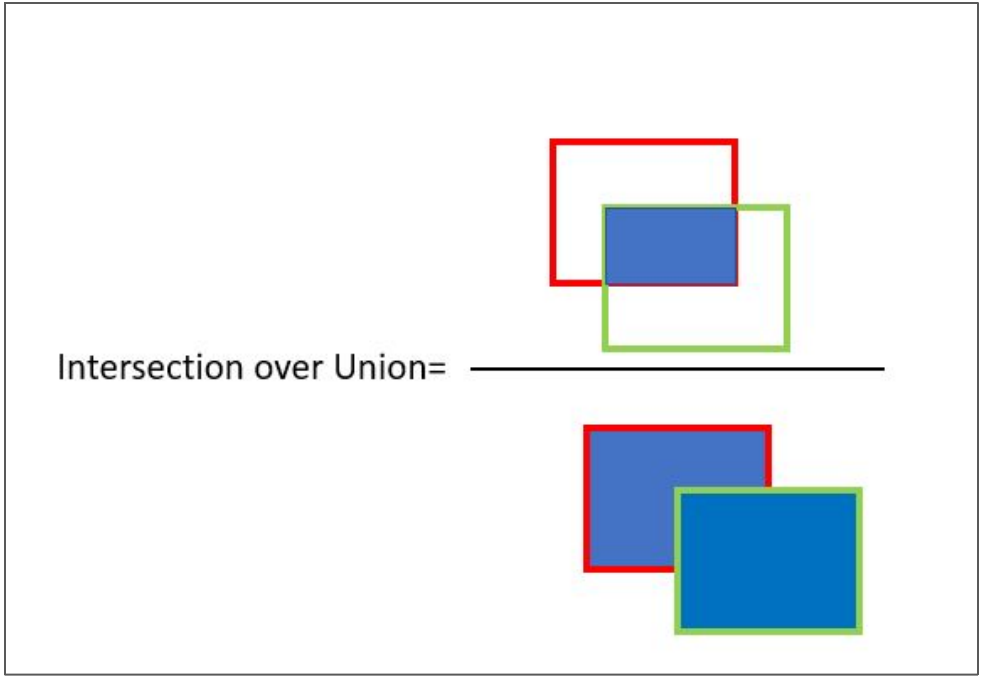
Exp2: Evaluate Deepfigures on gold standard

Better metric for TPs, FPs and FNs.

True Positive: IOU >= 0.8

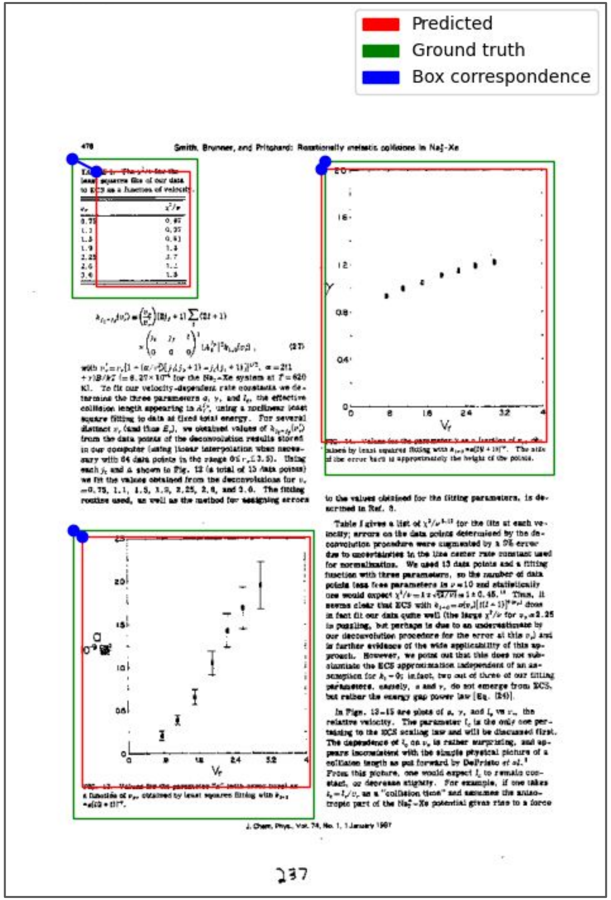
False Positive: IOU < 0.8

False Negative: Ground truth exists but prediction is missing.

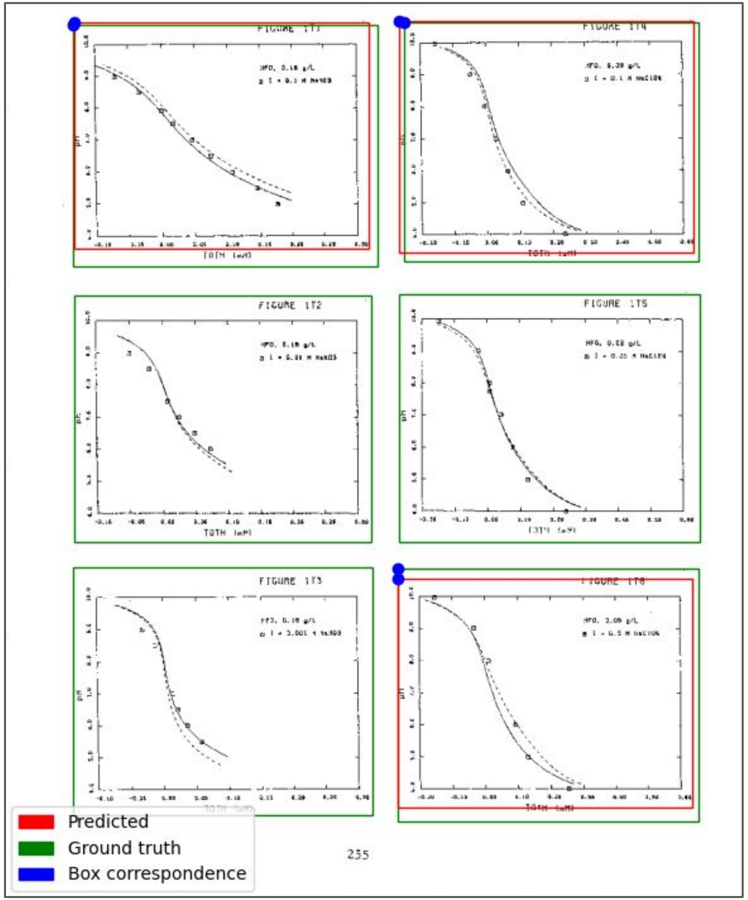


IOU=Intersection Over Union

Image sources:
[1] https://www.pyimagesearch.com/wp-content/uploads/2016/09/iou_equation.png



A page from a scanned ETD[1] showing the predictions, ground truth and box correspondences.



Another page from a scanned ETD[2] showing the predictions, ground truth and box correspondences.

Image sources:

[1] https://dspace.mit.edu/handle/1721.1/15568

[2] https://dspace.mit.edu/handle/1721.1/97289

Exp2: Evaluate Deepfigures on gold standard

1. Experimental setup:
    a. **Model**: Deepfigures.
    b. **Weight initialization**: Pre-trained weights from Deepfigures.
    c. **Data**: Gold standard dataset. Used for evaluation.
    d. **IOU thresh**: 0.8.

Exp2: Evaluate Deepfigures on gold standard

Model	TPs	FPs	FNs	Precision	Recall	F1
Deepfigures	1005	1227	1143	0.450	0.468	0.459

1. Results:
   a. F1-score of 0.459.
   b. This is the true performance of Deepfigures on scanned ETDs for figure extraction.
   c. RQ1 answered: Existing methods for figure extraction from scanned ETDs perform with an F1-score of 0.459.

Exp3: Ablation studies

1. 2^n combinations are possible for n transformations. (Enabled or disabled)
2. Leave-one-out ablation study:
   a. train n separate models, where, for the n-th model, the n-th transformation is disabled.
3. Doesn't guarantee an optimal combination.
4. Could give a general idea.

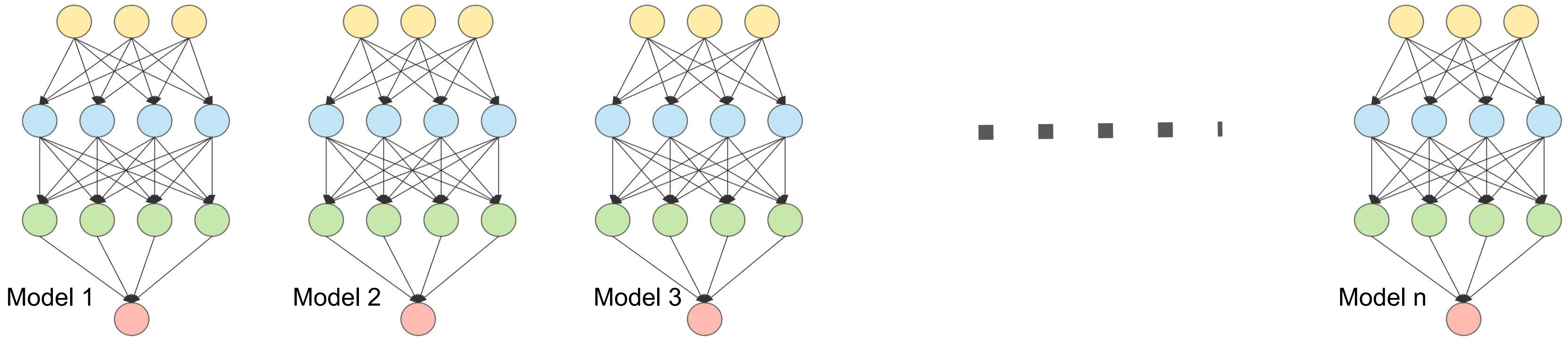
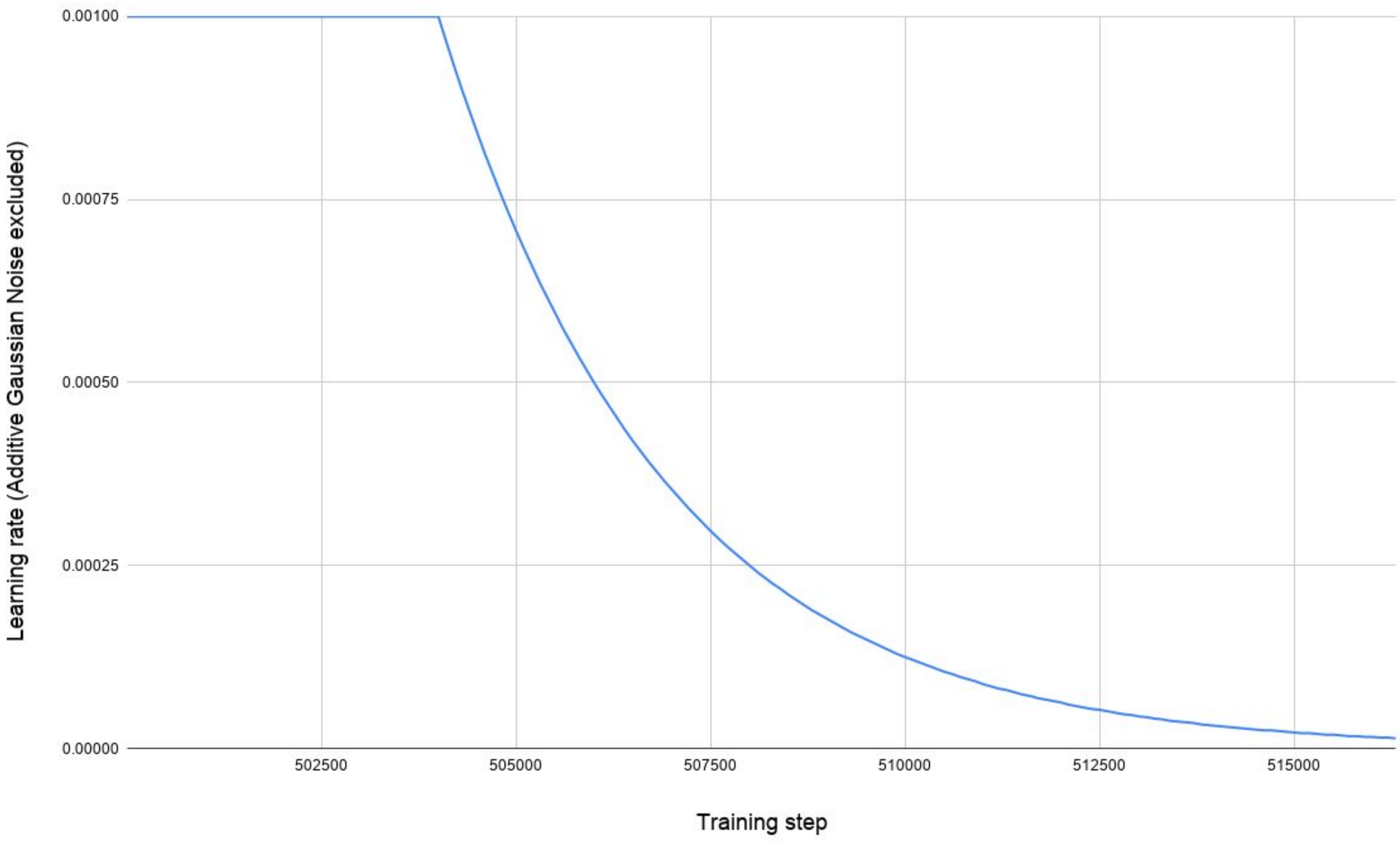
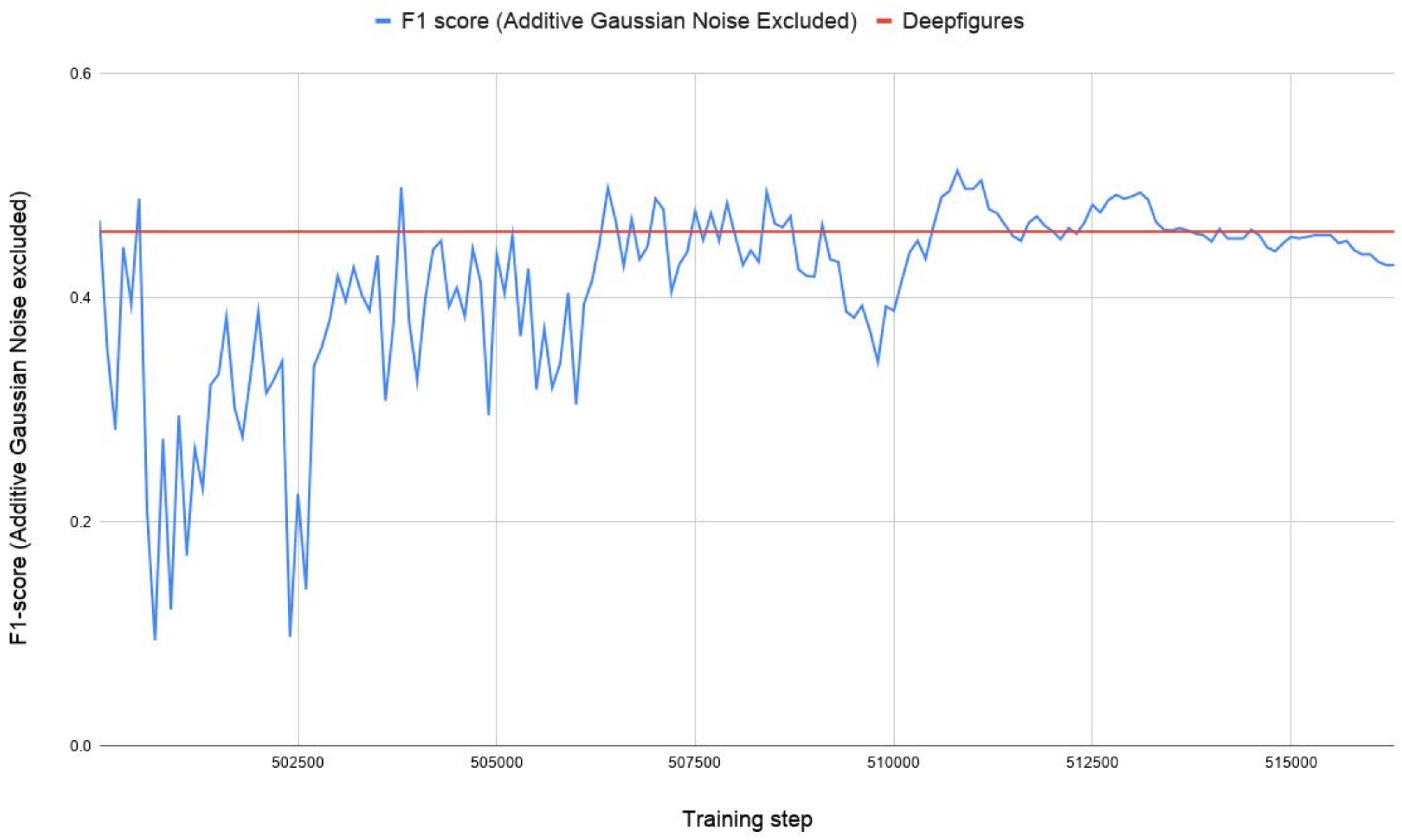


Image sources:
[1] https://www.kdnuggets.com/2019/02/neural-networks-intuition.html

Exp3: Ablation studies

1. Experimental setup
    a. **Model**: Deepfigures.
    b. **Weight initialization**: Pre-trained weights from Deepfigures
    c. **Data**: Trained on the entire arXiv dataset with n-th transformation disabled. Evaluated on the gold standard (val split for choosing the best model, test split for reporting performance).
    d. **Duration**: 24-hours. 16K steps.
    e. **Batch size**: 1


0.00100
0.00075
0.00050
0.00025
0.00000
502500
505000
507500
510000
512500
515000
Training step
Learning rate (Additive Gaussian Noise excluded)


F1 score (Additive Gaussian Noise Excluded)
Deepfigures
0.6
0.4
0.2
0.0
F1-score (Additive Gaussian Noise excluded)
502500
505000
507500
510000
512500
515000
Training step

Exp3: Ablation studies

Model	TPs	FPs	FNs	Precision	Recall	F1
Deepfigures	1005	1227	1143	0.450	0.468	0.459
Ours (All enabled)	619	506	569	0.550	0.521	0.535
Ours (Additive Gaussian Noise)	561	465	668	0.547	0.456	0.498
Ours (Affine)	577	530	587	0.521	0.496	0.508
Ours (Gaussian Blur)	506	619	569	0.450	0.471	0.460
Ours (Linear Contrast)	630	498	566	0.559	0.527	0.542
Ours (Perspective Transform)	597	539	558	0.526	0.517	0.521
Ours (Salt and Pepper)	686	509	499	0.574	0.579	0.576
Ours (Line spacing 1.5)	614	737	343	0.454	0.642	0.532
Ours (Typewriter font)	566	476	652	0.543	0.465	0.501

Exp3: Ablation studies

1. Results:
   a. F1-scores of almost all our models are higher than the original Deepfigures model. This supports to answer RQ2 positively.
   b. Our model with Gaussian Blur disabled has F1-score close to the original Deepfigures model. Indicates that Gaussian Blur could be the most `helpful’ transform.
2. However:
   a. Since this is only a single set of observation, we conduct the next experiment to investigate further.

Exp4: Ablation studies - longer training

1. Experimental setup
   a. **Model**: Deepfigures.
   b. **Weight initialization**: Pre-trained weights from Deepfigures
   c. **Data**: Trained on the entire arXiv dataset with n-th transformation disabled. Evaluated on the gold standard (val split for choosing the best model, test split for reporting performance).
   d. **Duration**: 72-hours.
   e. **Batch size**: 1

Exp4: Ablation studies - longer training

Model	TPs	FPs	FNs	Precision	Recall	F1
Deepfigures	1005	1227	1143	0.450	0.468	0.459
Ours (All enabled)	604	482	608	0.556	0.498	0.526
Ours (Additive Gaussian Noise)	613	448	633	0.578	0.492	0.531
Ours (Affine)	589	407	698	0.591	0.457	0.516
Ours (Gaussian Blur)	642	510	542	0.557	0.542	0.550
Ours (Linear Contrast)	602	460	632	0.567	0.488	0.524
Ours (Perspective Transform)	560	739	395	0.431	0.586	0.497
Ours (Salt and Pepper)	625	503	566	0.554	0.525	0.539
Ours (Line spacing 1.5)	705	594	395	0.542	0.641	0.588
Ours (Typewriter font)	641	386	667	0.624	0.490	0.549

Exp5: Training Deepfigures on the gold standard

1. Experimental setup
    a. **Model**: Deepfigures.
    b. **Weight initialization**: Pre-trained weights from Deepfigures
    c. **Data**: Gold standard dataset. 80-20 random train-test split.
    d. **Duration**: 2 hours
    e. **Batch size**: 1

Test accuracy vs. Training Step
1.00
0.99
0.98
0.97
0.96
Test accuracy
502000
504000
506000
508000
510000
Training Step

Exp6: Training Deepfigures on the gold standard (some layers training)

1. Experimental setup
   a. **Model**: Deepfigures.
   b. **Weight initialization**: Pre-trained weights from Deepfigures
   c. **Data**: Gold standard dataset. 80-20 random train-test split.
   d. **Duration**: 2 hours
   e. **Batch size**: 1
   f. **Method**: Freeze the weights of the ResNet backbone. Train only the last FC layers.

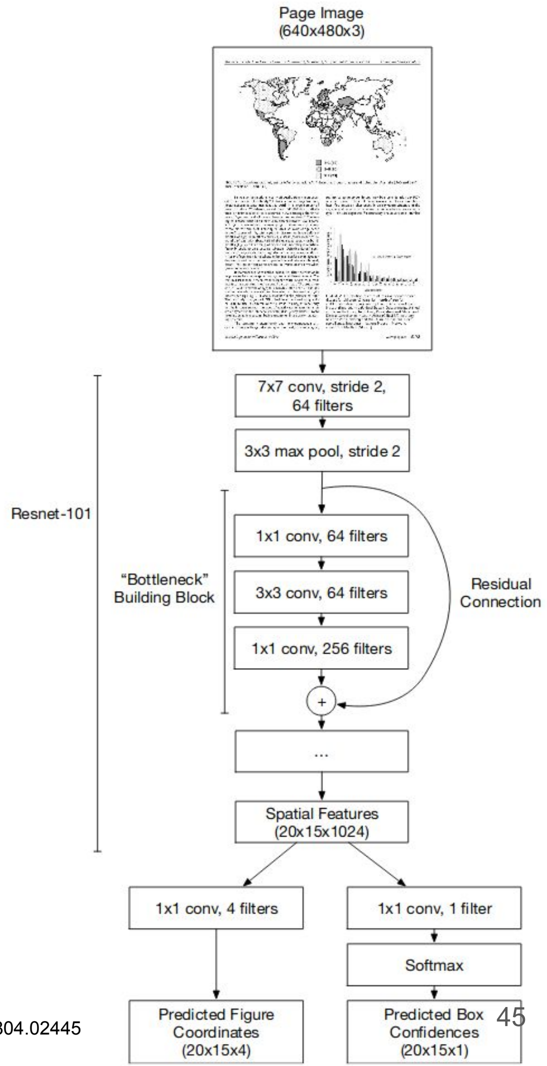
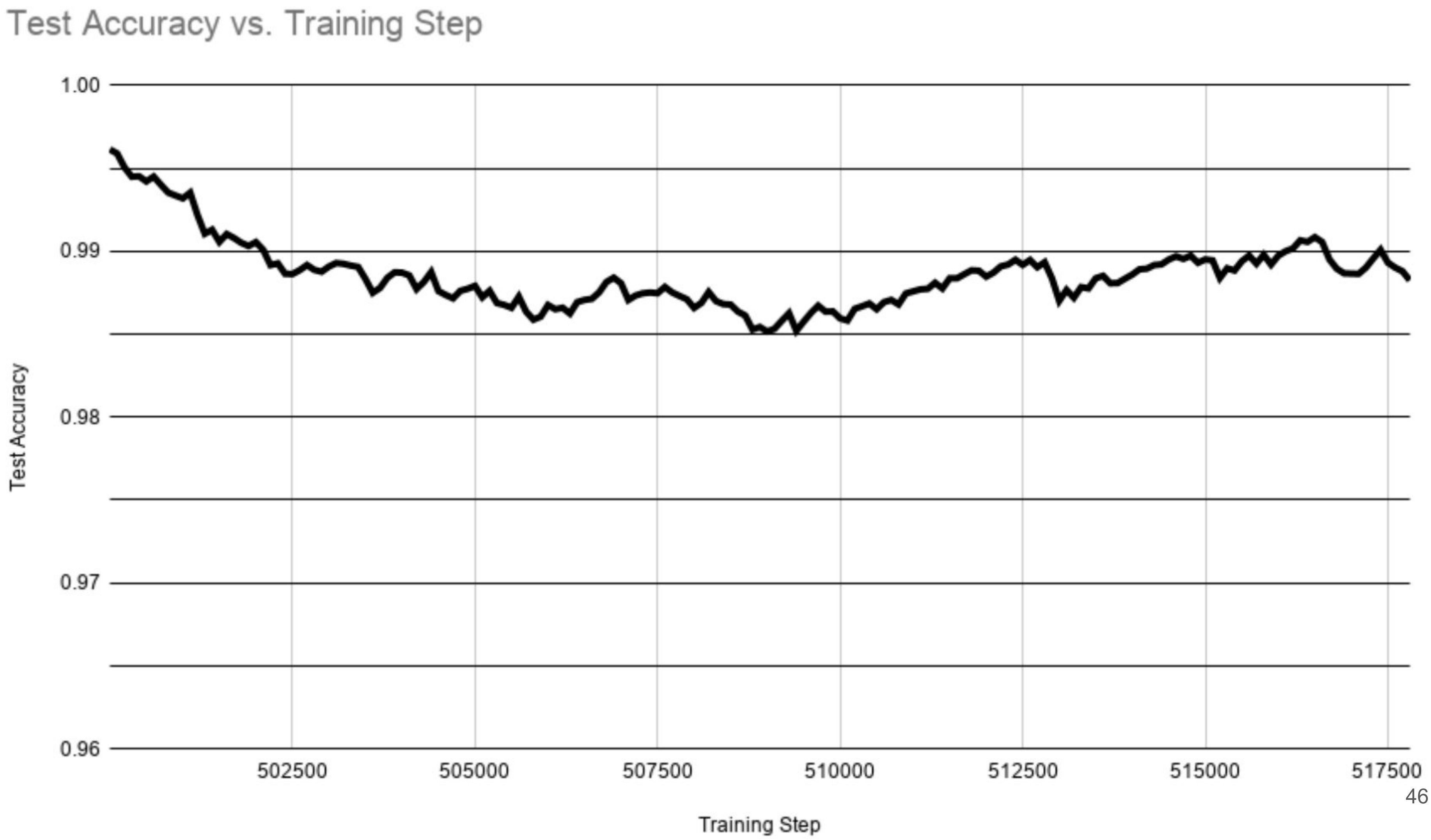
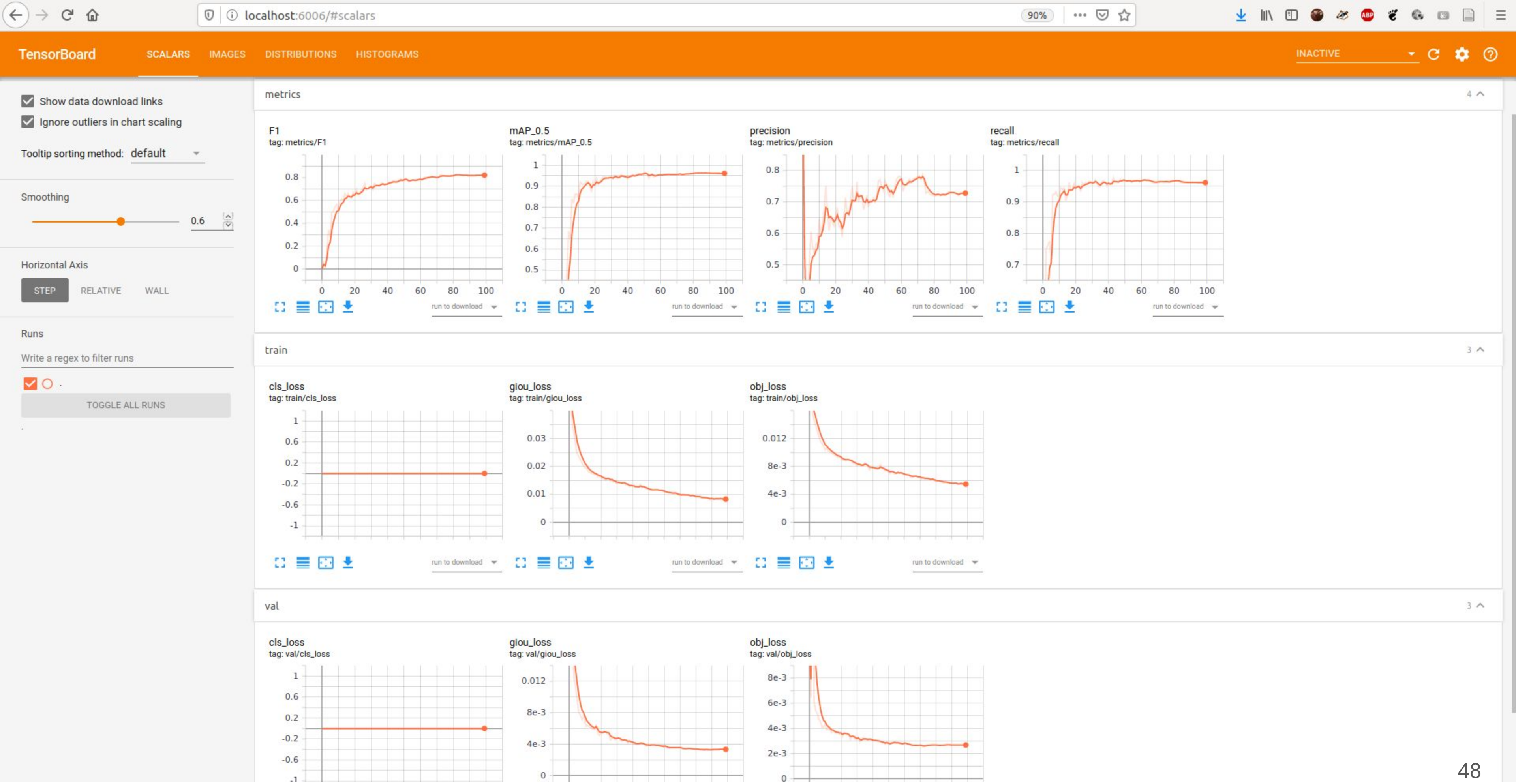


Image sources:
[1] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRR abs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445


Test Accuracy vs. Training Step
1.00
0.99
0.98
0.97
0.96
Test Accuracy
502500
505000
507500
510000
512500
515000
517500
Training Step

Exp7: Training YOLOv5 on the gold standard

1. Experimental setup
    a. **Model**: YOLOv5.
        i. You Only Look Once. A popular objector detection model.
        ii. Released May 2020
        iii. Outperforms all previous YOLO versions.
    b. **Weight initialization**: Random.
    c. **Data**: Gold standard dataset. K-fold cross validation (K=8).
    d. **Duration**: ~30 hours. 100 epochs.
    e. **Batch size**: 8


localhost:6006/#scalars
TensorBoard
SCALARS
IMAGES
DISTRIBUTIONS
HISTOGRAMS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
TOGGLE ALL RUNS
metrics
F1
tag: metrics/F1
mAP_0.5
tag: metrics/mAP_0.5
precision
tag: metrics/precision
recall
tag: metrics/recall
train
cls_loss
tag: train/cls_loss
giou_loss
tag: train/giou_loss
obj_loss
tag: train/obj_loss
val
cls_loss
tag: val/cls_loss
giou_loss
tag: val/giou_loss
obj_loss
tag: val/obj_loss
run to download

Exp7: Training YOLOv5 on the gold standard

Fold ID	Mean IOU	TP	FP	FN	Prec	Recall	F1
0	0.6282731795	298	100	45	0.7487437186	0.8688046647	**0.8043184885**
1	0.6511308301	262	39	57	0.8704318937	0.8213166144	**0.8451612903**
2	0.5854548651	381	127	170	0.75	0.6914700544	**0.7195467422**
3	0.8130208492	282	22	8	0.9276315789	0.9724137931	**0.9494949495**
4	0.7886945932	358	46	24	0.8861386139	0.9371727749	**0.9109414758**
5	0.7838215799	457	58	32	0.8873786408	0.9345603272	**0.9103585657**
6	0.7312581412	209	51	26	0.8038461538	0.8893617021	**0.8444444444**
7	0.6920578815	261	43	19	0.8585526316	0.9321428571	**0.8938356164**
Mean	0.70921399	313.5	60.75	47.625	0.8415904039	0.8809053485	**0.8597626966**
Std. dev.	0.0833932069	79.90172535	34.92747588	51.72713435	0.0665879393	0.08999876353	**0.07326514907**

- 9 -

(lyophilic), a high pressure has to be mantained on the gas side of the electrode if penetration is to be kept at a minimum. (Figure 2.4).

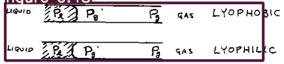


Figure 2.4. Gas-solution interface in capillaries.

Of the electrodes used presently, it appears that metallic electrodes are lyophilic, while carbon electrodes are lyophobic. For either case the pressure difference between the gas phase and the liquid phase can be derived directly from Laplace's relationship(2), and is given by,

$$P_1 - P_g'' = -\frac{2\gamma_{LA}}{r}\cos\theta \qquad (1)$$

where γ_{LA} is the liquid gas interfacial tension, r is the radius of the capillary, and θ is the contact angle associated with the gas-liquid-solid interface. For lyophobic surfaces, θ is between 90° and 180°, so that $\cos\theta$ is negative and the pressure in the gas phase is less than in the liquid phase. For the lyophilic case, clearly, the pressure would be higher in the gas phase.

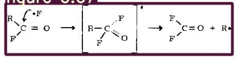
Gas transport within the electrodes of the type described above can occur through the following mechanisms:

1) Viscous flow under a pressure gradient in the gas filled pores.

-18-

The renewal of the reactant surface during later stages of fluorination for 1,2-diethoxyethane is better than octanoyl fluoride because the boiling point is so much lower.

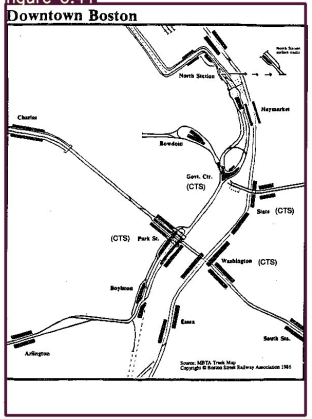
Another large problem causing the small yields of the acid fluorides is decarbonylation. The carboxylic acid fluoride group is flat and thus open to relatively unhindered attack from above or below the plane of the group. Thus, more fragmentations occur at this site in the molecule than would occur for a carbon atom in the sp^3 state.

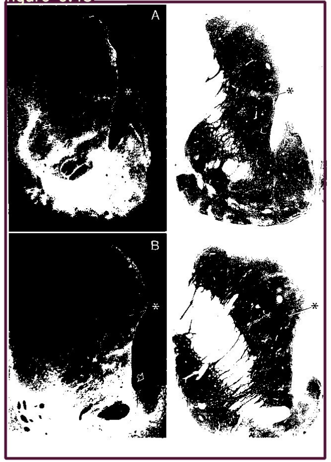


The carbon-fluorine bond is approximately 110-118 kcal/mole compared to 86 kcal/mole for the carbon-carbon bond. Thus, the carbon-carbon bond breaks and a alkyl radical and carbonyl fluoride are produced.

-27-

Figure 3.2 Downtown Subway System





- 8 -

2.4. Transport Process in Porous Electrodes. The need for large surface area of reaction to obtain high current densities has made the use of porous electrodes necessary. Sintered porous metal electrodes and porous carbon electrodes are the most widely used. They consist of macropores (10^{-3} to 10^{-5} cm.), where the transport processes take place, and smaller pores or microcapillaries which provide extensive reaction area.

The reaction zone is usually wihtin the electrode. Furthermore, since the diffusion of ions in liquid is much slower than diffusion in gas, then, assuming that transport of reactants and products takes place mainly by diffusion, it is desired to keep the reaction zone close to the electrolyte side; otherwise electrolyte concentration polarization would be encountered at low current densities. (Figure 2.3).

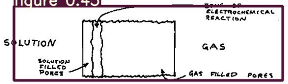


Figure 2.3. Model of porous electrode.

To avoid penetration of electrolyte into the electrode pores, the wetting properties of the electrolyte and electrode must be of such nature that low wetting prevails (lyophobic). If the electrode surface allows good wetting

-26-

Figure 3.1 MBTA Subway System

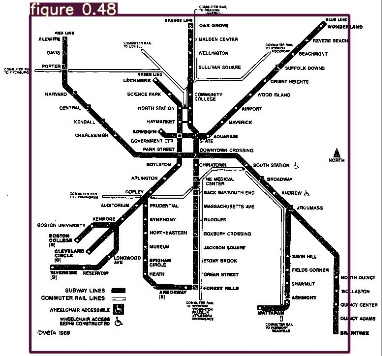


Figure A.5 DWELL TIME vs. MAXASLS

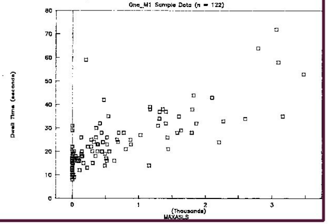
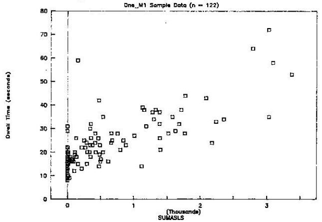


Figure A.6 DWELL TIME vs. SUMASLS



Direct Vision:

1) 12 ft, force feedback
2) 12 ft, no force feedback
3) 8 ft, no force feedback
4) 4 ft, no force feedback
5) 4 ft, force feedback
6) 8 ft, force feedback

Video Monitor:

1) 30 fps, force feedback
2) 5 fps, force feedback
3) 3 fps, force feedback
4) 3fps, no force feedback
5) 5 fps, no force feedback
6) 30 fps, no force feedback

The balanced latin square technique yielded the following ordering assignments for each subject:

Subj. #1	Subj. #2	Subj. #3	Subj. #4	Subj. #5	Subj. #6
1	2	3	4	5	6
2	3	4	5	6	1
6	1	2	3	4	5
3	4	5	6	1	2
5	6	1	2	3	4
4	5	6	1	2	3

Since each task was performed an equal number of times going to the left and an equal number of times going to the right, effects due to direction were also analyzed.

60

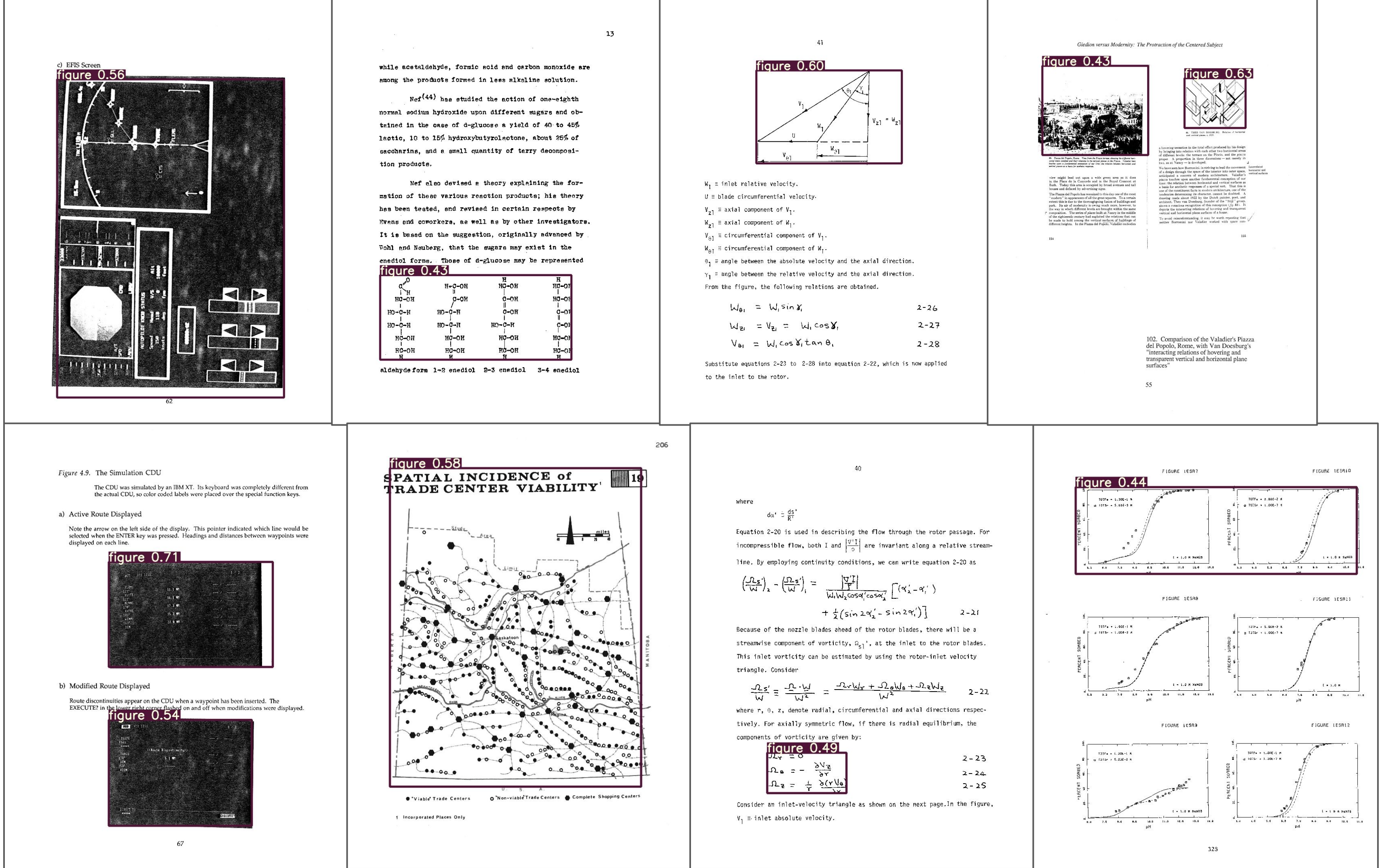


figure 0.56
figure 0.43
figure 0.60
figure 0.43
figure 0.63
figure 0.71
figure 0.54
figure 0.58
figure 0.49
figure 0.44

Answers to research questions

RQ1: How well can existing methods perform figure extraction from scanned ETDs?
Ans: Deepfigures is able to extract figures from scanned ETDs with an F1-score of 0.459.

RQ2: Can this performance be improved by using simple data augmentation techniques and weight initialization from the original pre-trained model?
Ans: Yes. In general, the Deepfigures model trained on born-digital ETDs performs better when further trained on augmented born-digital ETDs.

RQ3: Can this performance be improved by training on manually labelled data?
Ans: Using the Deepfigures model architecture and weight initialization, we did not see an improvement. However, by training YOLOv5 with random weight initialization on manually labelled data (gold standard), the performance improved to an F1-score of 0.859 (std. dev. 0.07).

RQ4: Can this performance be improved by using transfer learning techniques?
Ans: Using transfer learning techniques on the pre-trained Deepfigures model, we did not see any improvement.

Outline

1. Introduction
2. Research questions
3. Related work
4. Methodology
    a. Data augmentation
    b. Training at scale
    c. Gold standard
5. Experiments
    a. Experiments, results/discussion, answers to research questions.
6. **Conclusions**
7. Future work

Conclusions

1. In this thesis, we focus on extracting figures from scanned ETDs.
2. We describe the research problem, formulate RQs, and review related work.
3. We propose LaTeX and image-based transformations.
4. We describe our system to apply these transformations at scale.
5. We curate a gold standard dataset for evaluation.
6. Finally, we describe the various experiments we conducted.

Outline

1. Introduction
2. Research questions
3. Related work
4. Methodology
    a. Data augmentation
    b. Training at scale
    c. Gold standard
5. Experiments
    a. Experiments, results/discussion, answers to research questions.
6. Conclusions
7. **Future work**

Future work

1. Hyper-parameter tuning.
2. Visual similarity metric for choosing transformations.
3. More ablation studies.
4. Pre-training for unsupervised visual representation learning, and then fine-tuning using these visual representations.

Thank you.

Questions are welcome.

Email: sampanna@vt.edu