Synthetic Data for Text Localisation in Natural Images
Supplementary Material

Ankush Gupta Andrea Vedaldi Andrew Zisserman
University of Oxford
{ankush,vedaldi,az}@robots.ox.ac.uk

We highlight some components of our synthetic text dataset — *SynthText in the Wild*, and show some sample images from the dataset. Next, we compare the detection results from the "FCRNall multi-flit" method and Jaderberg *et al.* [1] on ICDAR 2013 and Street View Text (SVT) datasets.

Contents

1 Variation in Fonts, Colors and Sizes

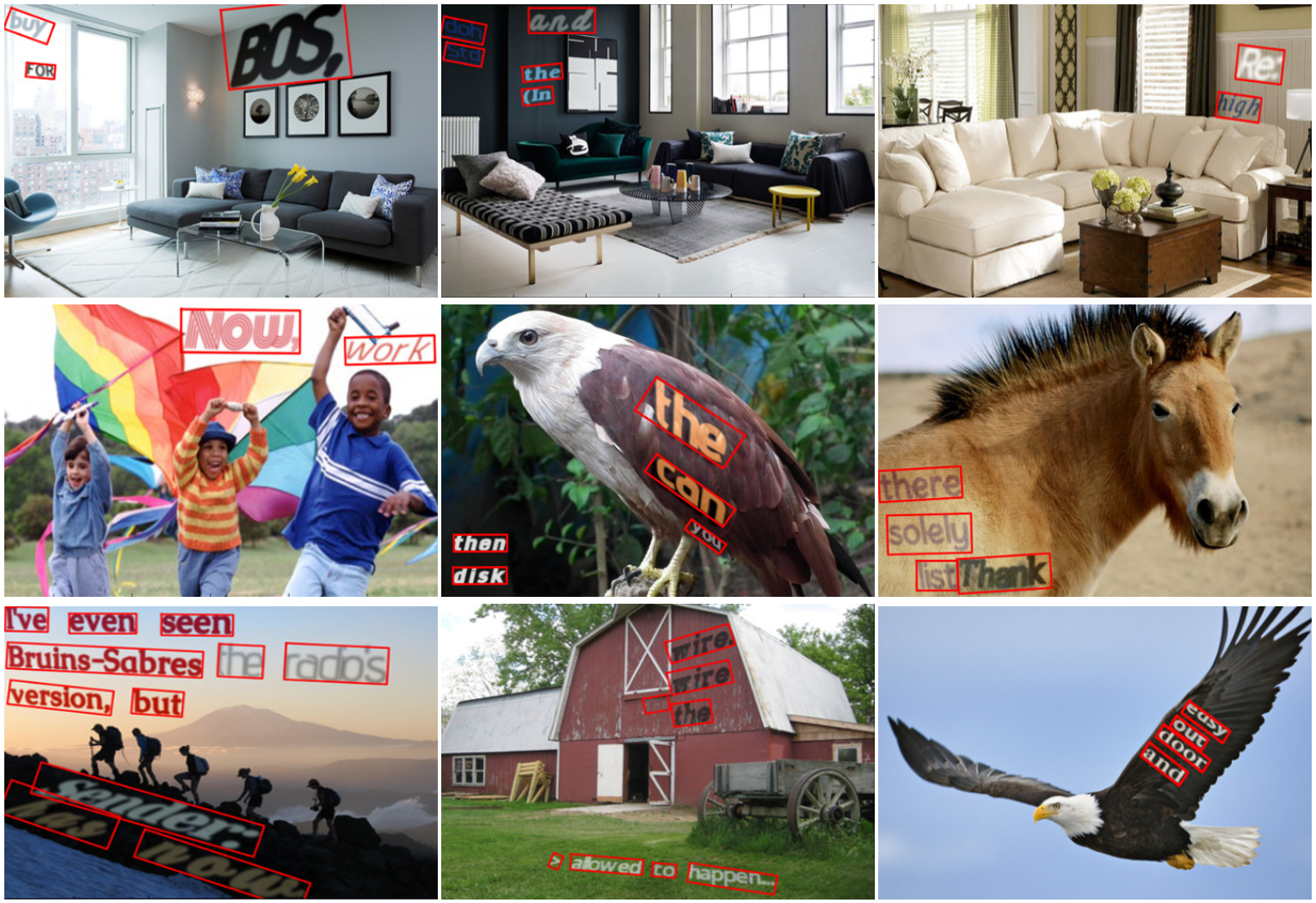
The following images show synthetic text renderings for the same text – “vamos!”.
Along the rows, the text is rendered in approximately the same location and against the same background image but in different fonts, colors and sizes.

2 Poisson Editing vs. Alpha Blending

Comparison between simple alpha blending **(bottom row)** and Poisson Editing [2] **(top row)**.
NOTE: Poisson Editing preserves local illumination gradient and texture details.

3 SynthText in the Wild

Sample images from our synthetic text dataset (continued on the next page).
These images show text instances in various fonts, colors, sizes, with borders and shadows, against different backgrounds, and transformed according to the local geometry and constrained to local contiguous regions of color and text. Ground-truth word bounding-boxes are marked in red.




Sincerely
THEN
know
it
APR
93
Internet:
Newsgroups:
From:
SOME
Does
his
From:
THR
Just
the
for
you
Call
very
one
why?
(that
The
world
dealer
mark
>Public
>Space
Far
player?
week,

4 ICDAR 2013 Detections

Example detections on the ICDAR 2013 dataset from "FCRNall + multi-flit" **(top row)** and those from Jaderberg *et al.* [1] **(bottom row)**.

Precision, recall and F-measure values **(P/R/F)** are indicated at the top of each image.

FCRNall multi-filt

1 / 1 / 1 | 1 / 1 / 1 | 1 / 0.88 / 0.94 | 1 / 1 / 1 | 1 / 1 / 1

Jaderberg *et al.*

1 / 0.33 / 0.50 | 1 / 0.33 / 0.50 | 1 / 0.28 / 0.44 | 1 / 0.75 / 0.86 | 1 / 0.66 / 0.80

5 Street View Text (SVT) Detections

Example detections on the Street View Text (SVT) dataset from "FCRNall + multi-flit" **(top row)** and those from Jaderberg *et al.* [1] **(bottom row)**.

Precision, recall and F-measure values **(P/R/F)** are indicated at the top of each image: both the methods have a precision of 1 on these images (except in one case due to missing ground-truth annotation).

References

[1] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *IJCV*, 2015. 1, 6, 7

[2] P. Perez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. 3