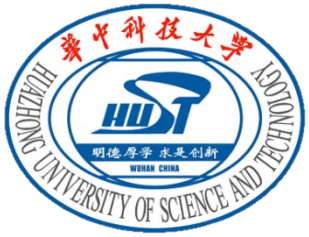


華中科技大學

Huazhong University of Science & Technology

DeepFlux & TextField for Skeleton and Text Detection in the Wild

Yongchao Xu
Huazhong University of Science and Technology (HUST)

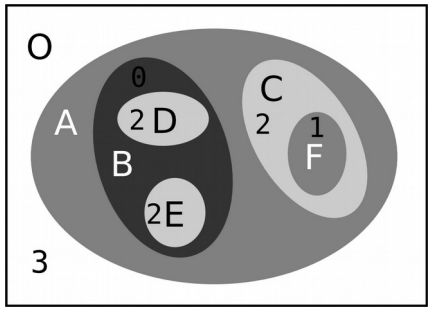
Joint work with Xiang Bai, Stavros Tsogkas, Sven Dickinson, Kaleem Siddiqi

Outline

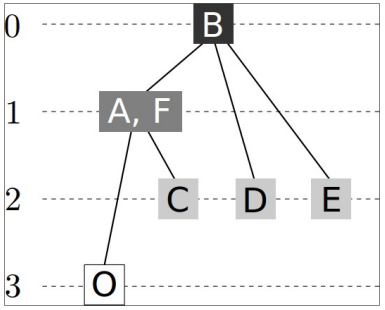
- Context
- Flux Representation
- DeepFlux for Skeletonization in the Wild
- TextField for Irregular Scene Text Detection
- Conclusion

Context

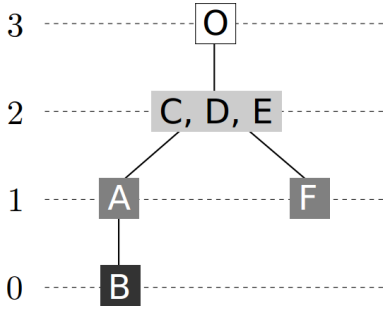
Image Representation



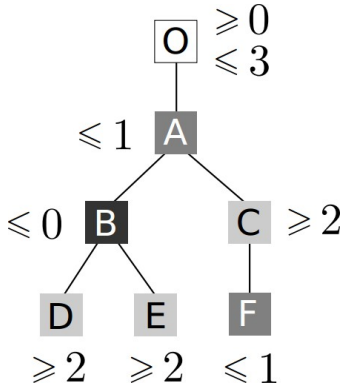
Image



Max-tree
[Salembier *et al.*, ITIP, 1998]



Min-tree
[Salembier *et al.*, ITIP, 1998]



Tree of shapes
[Monasse & Guichard, ITIP, 2000]

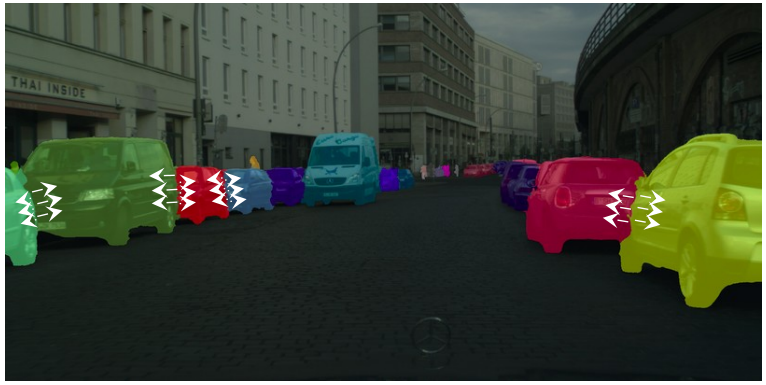
Pixel Relationship → Regions → Trees

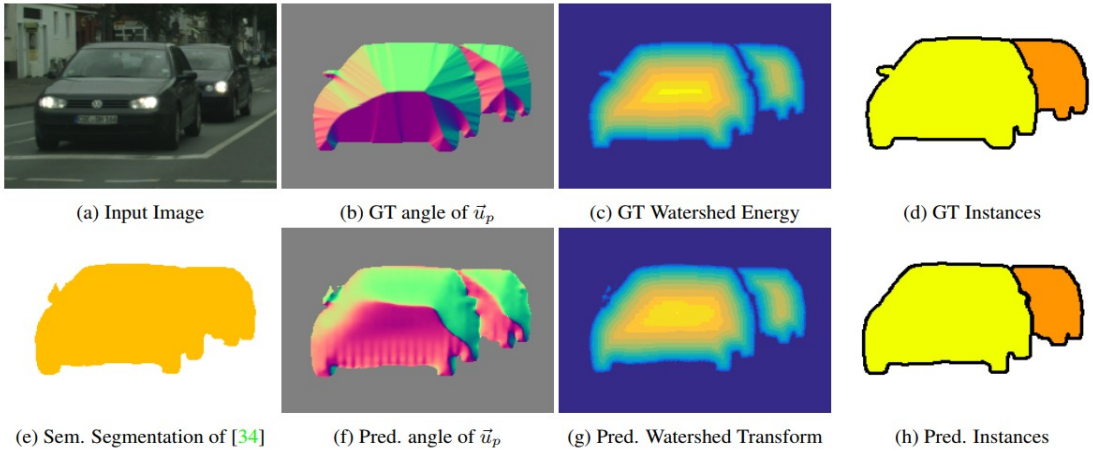
Context

Instance Segmentation

Semantic segmentation → (pixel relationship) → Instance segmentation

relationship between pixels and instance-level boundary





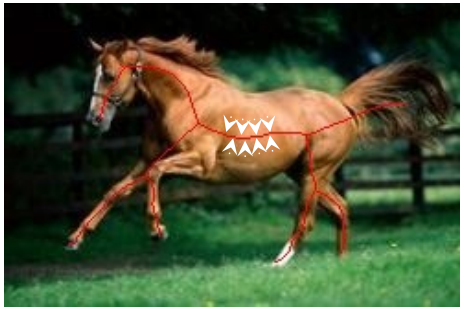
(a) Input Image (b) GT angle of $\vec{u}_p$ (c) GT Watershed Energy (d) GT Instances

(e) Sem. Segmentation of [34] (f) Pred. angle of $\vec{u}_p$ (g) Pred. Watershed Transform (h) Pred. Instances

Deep Watershed Transform for Instance Segmentation [Bai *et al.*, CVPR, 2017]

Context

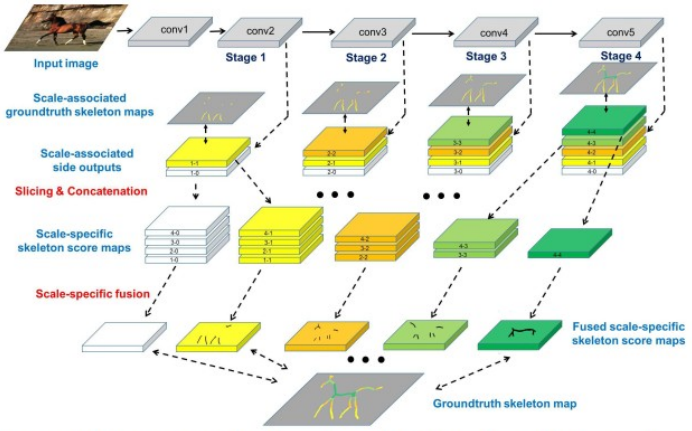
Skeleton Detection



one kind of pixel relationship

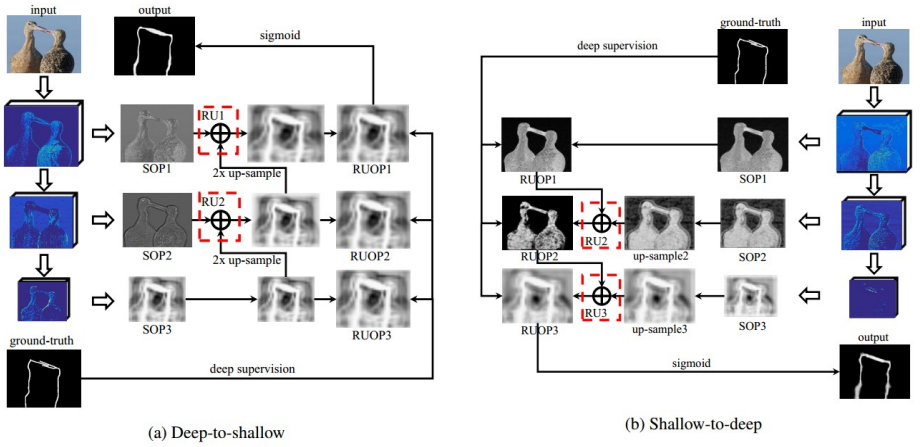
relationship between pixels and skeleton

Previous methods for skeleton detection:



FSDS

[Shen *et al.*, CVPR, 2016]



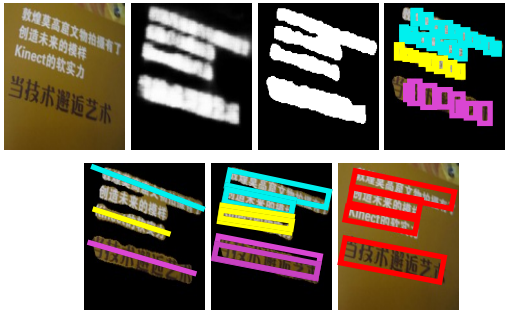
SRN

[Ke *et al.*, CVPR, 2017]

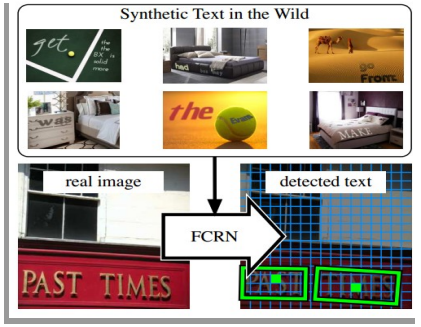
Context

Scene Text Detection

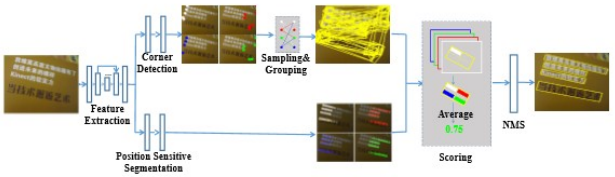
Pixel relationship for irregular scene text



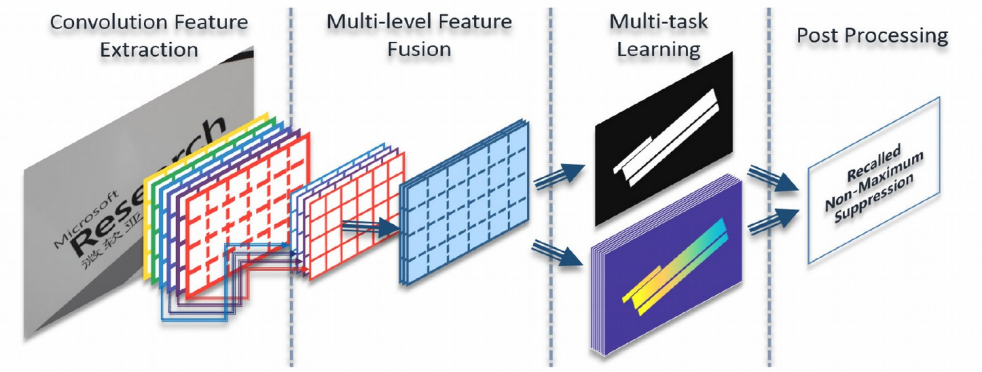
Segmentation-based method
[Zhang *et al.*, CVPR, 2016]



Proposal-based method
[Gupta *et al.*, CVPR, 2016]



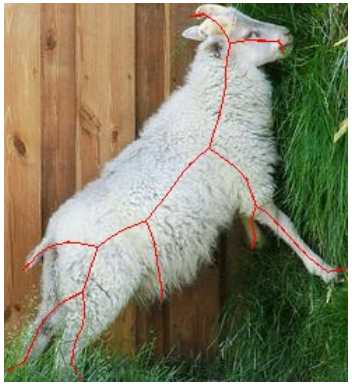
Part-based method
[Lyu *et al.*, CVPR, 2018]



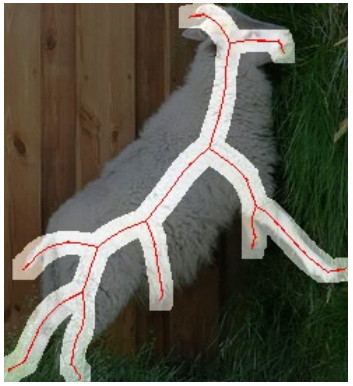
Hybrid method
[He *et al.*, ICCV, 2017]

Flux Representation

Flux Representation (attracted by skeleton)



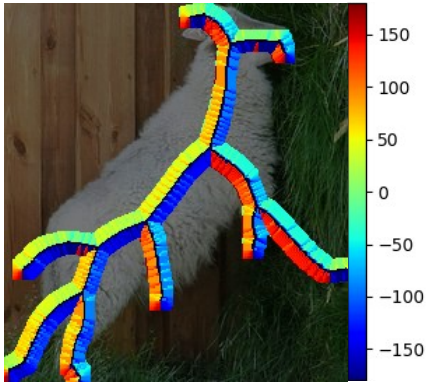
Skeleton Annotation



Dilated Skeleton Mask

compute flux

visualize



$$F(p) = \begin{cases} \overrightarrow{pN_p} / \left|\overrightarrow{pN_p}\right|, & p \in R_c \\ (0,0), & p \in R_s \cup R_b, \end{cases}$$

- unit vectors **pointing to** the object skeleton.

Flux Representation

Flux Representation (repulsed by boundary)

Bbox Annotation

Text Mask

compute flux

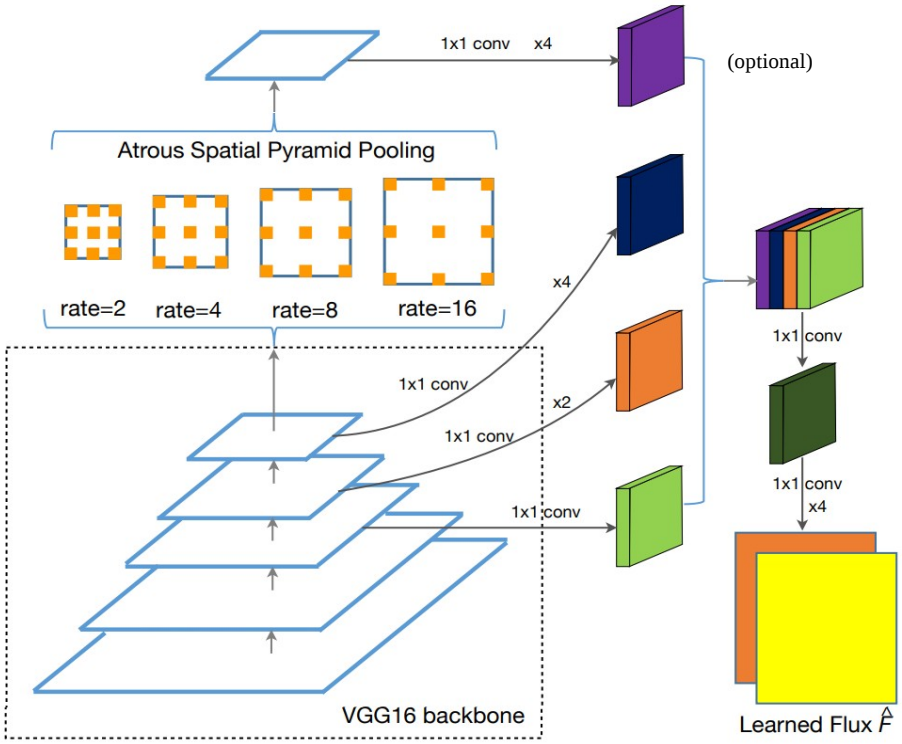
visualize

$$F(p) = \begin{cases} \overrightarrow{N_p p} / \left| \overrightarrow{N_p p} \right|, & p \in R_t \\ (0,0), & p \notin R_t, \end{cases}$$

- unit vectors **pointing away from** the bbox boundary.

Flux Representation

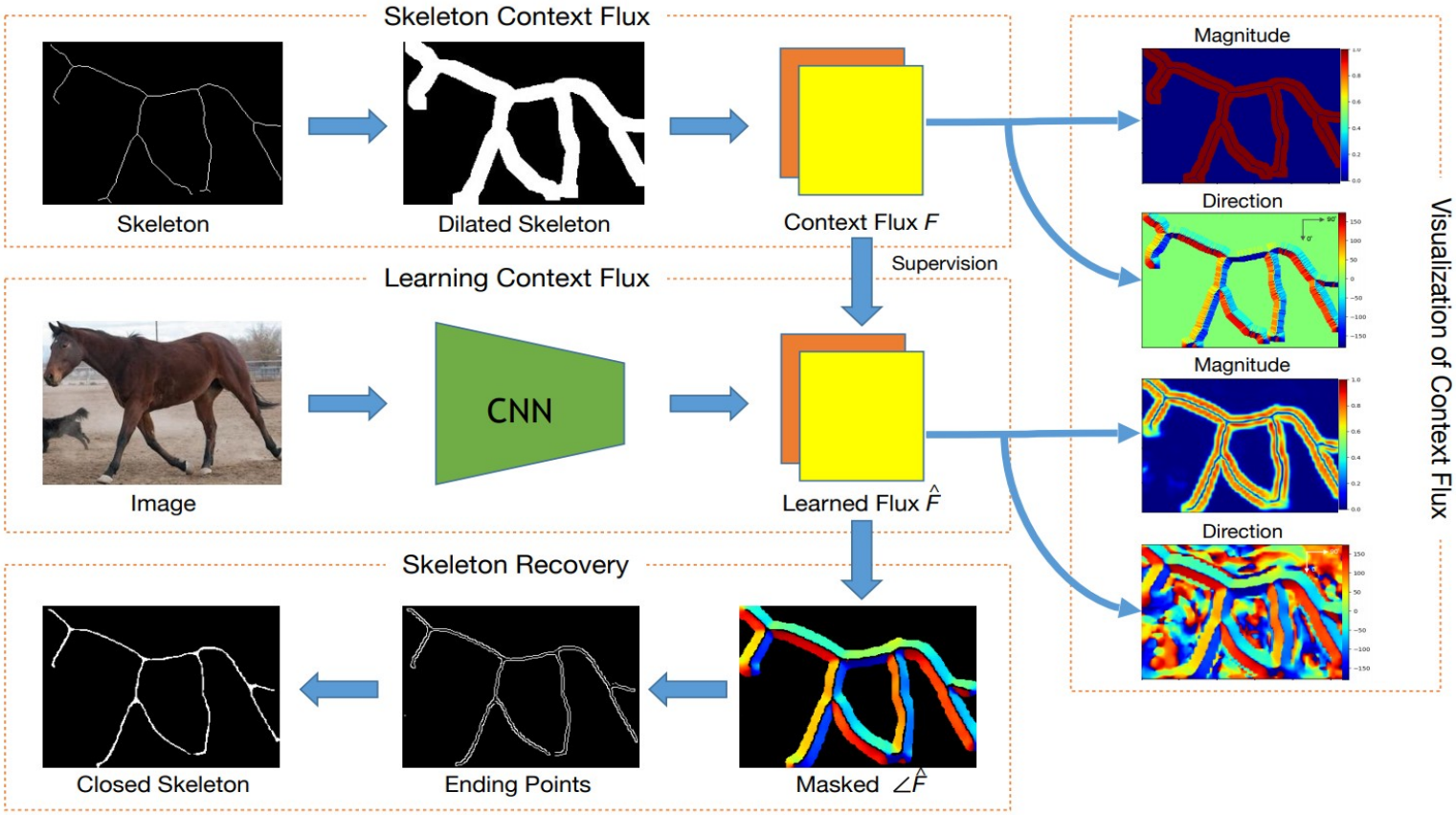
Learning Flux Representation



- VGG16 backbone
- ASPP module (optional)
- feature fusion

DeepFlux for Skeletons in the Wild

How to use Flux Representation to detect skeleton ?



Pixel Relatic Flux → Regions → Trees → (search) Skeleton

DeepFlux for Skeletons in the Wild

Binary Skeleton vs. Flux Representation

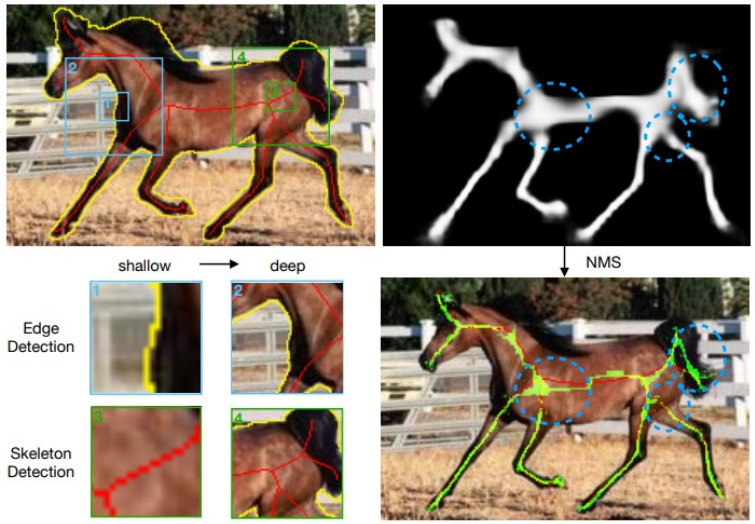
Previous CNN-based methods :

skeleton detection ⟺ edge detection

⇓

binary pixel classification task with NMS

- poor localization
- poor connectedness



(a) Previous CNN-based skeleton detections rely on NMS.

With flux representation :

skeleton detection ⇒ regression task

- encode the relative position of skeleton
- can accurately recover the object skeleton from the learned flux
- large receptive fields

(b) Flux provides an alternative way for accurately detecting skeletons.

DeepFlux for Skeletons in the Wild

Qualitative Results

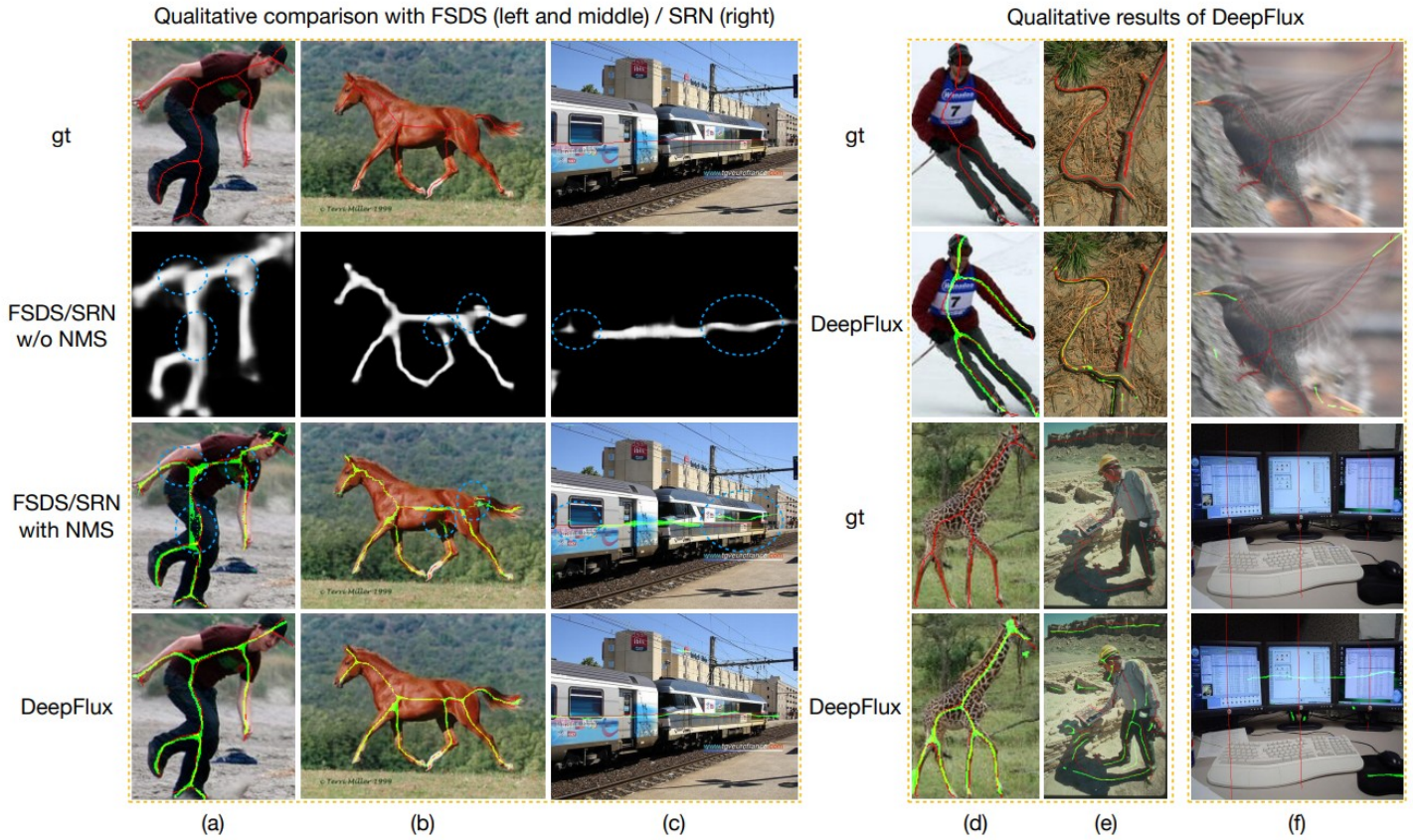


Figure 6. Qualitative results on SK-LARGE, WH-SYMMAX, and SYM-PASCAL (a-c), SK506 (d), SYMMAX300 (e), and two failure cases (f). Red: GT; Green: detected skeleton; Yellow: detected skeleton and GT overlap. DeepFlux fails to detect the skeleton on the bird body due the severe blurring. In the second failure example DeepFlux detects a symmetry axis not annotated in the ground truth.

DeepFlux for Skeletons in the Wild

Quantitative Results

Methods	SK-LARGE	SK506	WH-SYMMAX	SYM-PASCAL	SYMMAX300
MIL [45]	0.353	0.392	0.365	0.174	0.362
HED [47]	0.497	0.541	0.732	0.369	0.427
RCF [25]	0.626	0.613	0.751	0.392	-
FSDS* [35]	0.633	0.623	0.769	0.418	0.467
LMSDS* [34]	0.649	0.621	0.779	-	-
SRN [17]	0.678	0.632	0.780	0.443	0.446
LSN [22]	0.668	0.633	0.797	0.425	0.480
Hi-Fi* [51]	0.724	0.681	0.805	0.454	-
DeepFlux (Ours)	**0.732**	**0.695**	**0.840**	**0.502**	**0.491**

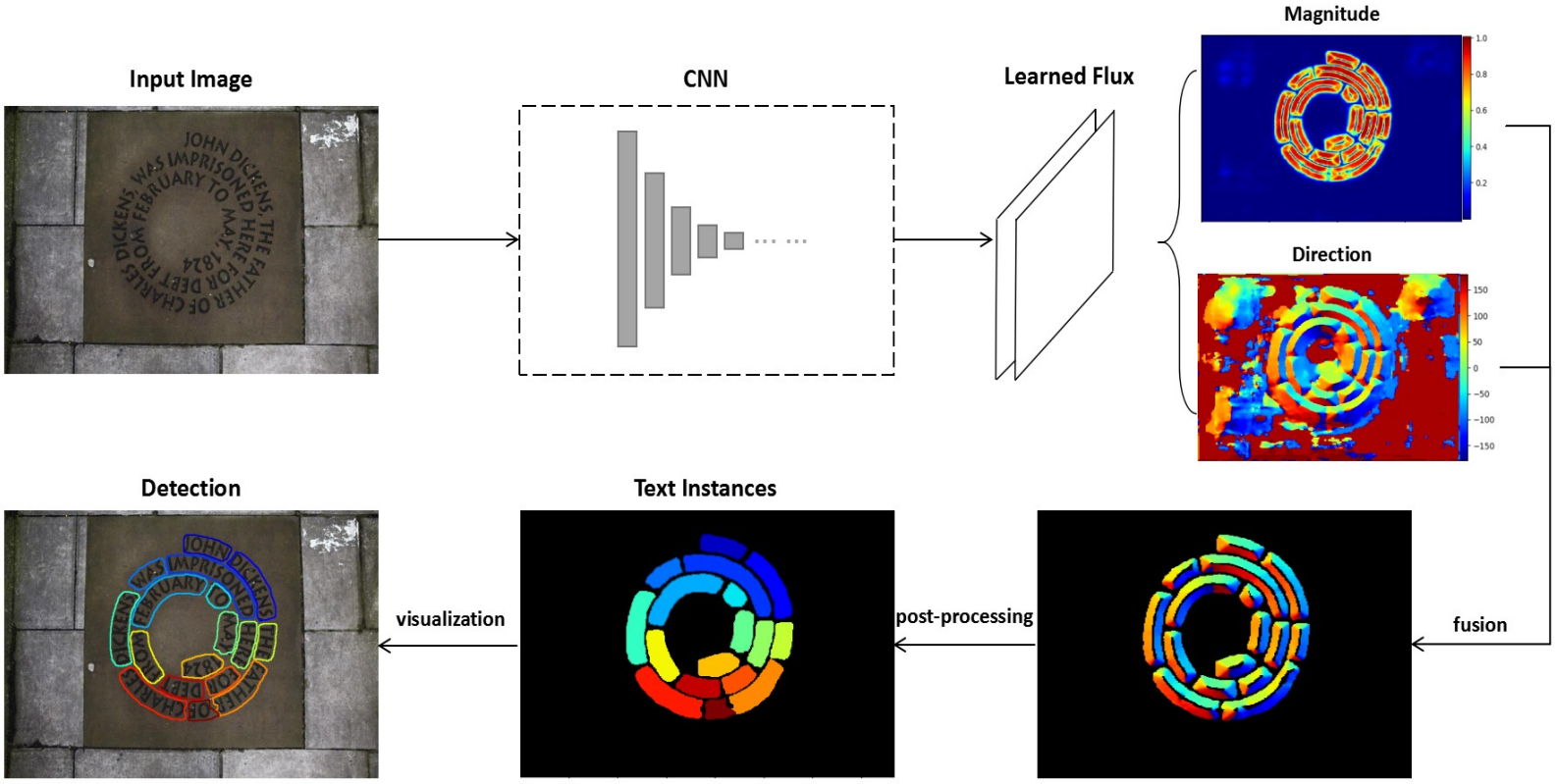
Table 1. F-measure comparison. * indicates scale supervision was also used. Results for competing methods are from the respective papers.

Method	F-measure	Runtime (in *sec*)
HED [47]	0.497	0.014
FSDS [35]	0.633	0.017
LMSDS [34]	0.649	0.017
LSN [22]	0.668	0.021
SRN [17]	0.678	0.016
Hi-Fi [51]	0.724	0.030
DeepFlux (ours)	0.732	0.019

Table 2. Runtime and performance on SK-LARGE. For DeeFlux we list the total inference (GPU) + post-processing (CPU) time.

TextField for Irregular Scene Text Detection

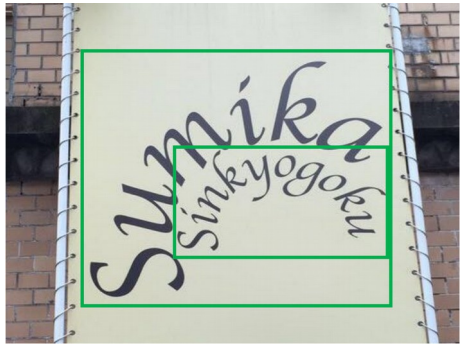
How to use Flux Representation to detect text ?



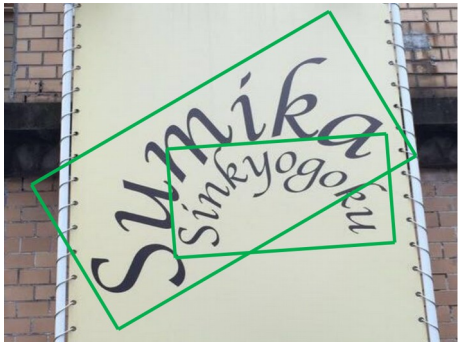
Pixel Relatic Flux → Regions → Trees —cluster→ Text Instances

TextField for Irregular Scene Text Detection

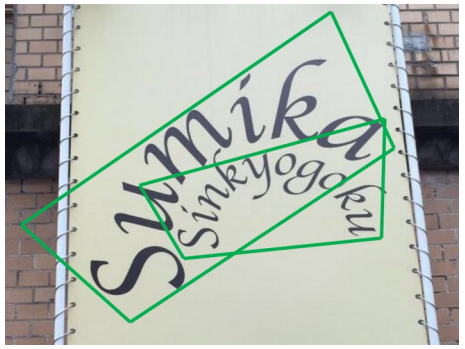
Bounding Box vs. Flux Representation



Horizontal box



Rotated rectangle



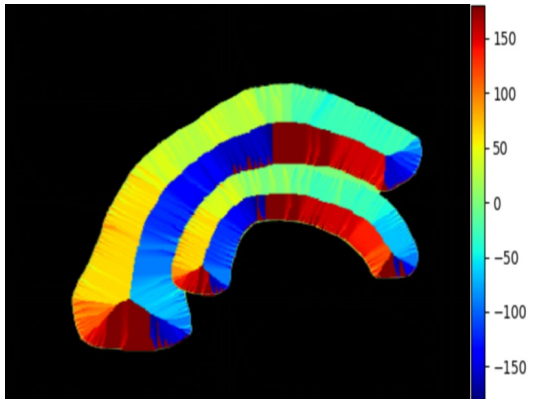
Quadrilateral

Bounding Box:

- proposal-based methods
- fail to accurately delimit irregular texts

Flux Representation:

- flexible representation
- precisely describe irregular texts

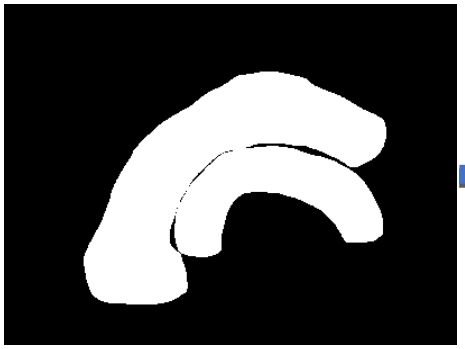


Flux direction visualization

TextField for Irregular Scene Text Detection

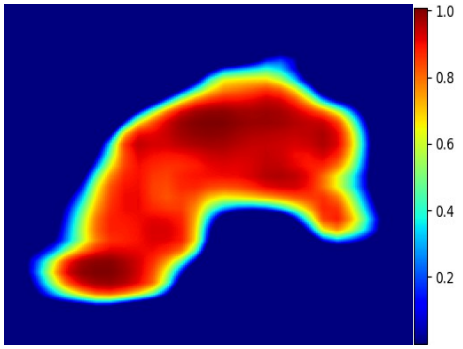
Text Mask vs. Flux Representation

Image



Text Mask

CNN



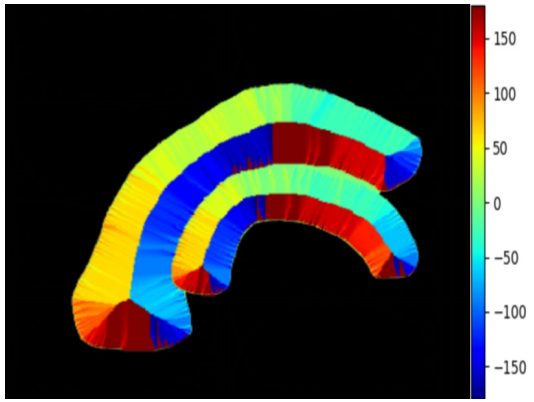
Predicted Text Areas

Text Mask:

- segmentation-based methods
- hard to extract text instances from the predicted text areas

Flux Representation:

- instance-level representation
- easy to separate adjacent text instances



Flux direction visualization

TextField for Irregular Scene Text Detection

Qualitative Results

Fig. 7: Some qualitative detection results on SCUT-CTW500 in (a), Total-Text in (b), IC15 in (c), and MSRA-TD500 in (d). The arbitrary-shaped texts are correctly detected with accurate text instance boundaries.

TextField for Irregular Scene Text Detection

Quantitative Results

TABLE I: Quantitative results of different methods evaluated on SCUT-CTW1500. * indicates the result obtained from [47].

Methods	recall	precision	f-measure
SegLink * [34]	0.400	0.423	0.408
CTPN * [33]	0.538	0.604	0.569
EAST * [25]	0.491	0.787	0.604
DMPNet * [23]	0.560	0.699	0.622
CTD [47]	0.652	0.743	0.695
CTD+TLOC [47]	0.698	0.774	0.734
TextField (Ours)	**0.798**	**0.830**	**0.814**

TABLE II: Quantitative results of different methods evaluated on Total-Text.

Methods	recall	precision	f-measure
Ch'ng *et al.* [41]	0.400	0.330	0.360
Liao *et al.* [24]	0.455	0.621	0.525
TextField (Ours)	**0.799**	**0.812**	**0.806**

TABLE III: Comparison of methods on ICDAR2015 Incidental Scene Text. † means that the base net of the model is not VGG16. Note that only single scale test result is depicted.

Methods	recall	precision	f-measure	FPS
Zhang *et al.* [37]	0.430	0.708	0.536	0.48
CTPN [33]	0.516	0.742	0.609	7.1
Yao *et al.* [38]	0.587	0.723	0.648	1.61
DMPNet [23]	0.682	0.732	0.706	-
SegLink [34]	0.768	0.731	0.750	-
MCN [36]	0.800	0.720	0.760	-
EAST [25]	0.728	0.805	0.764	6.52
SSTD [26]	0.730	0.800	0.770	7.7
RRPN [29]	0.730	0.820	0.770	-
WordSup [28]	0.770	0.793	0.782	2
ITN [32]	0.741	0.857	0.795	-
EAST † [25]	0.735	0.836	0.782	13.2
Lyu *et al.* [35]	0.707	**0.941**	0.807	3.6
He *et al.* † [27]	0.800	0.820	0.810	1.1
TextBoxes++ [30]	0.767	0.872	0.817	11.6
RRD [31]	0.790	0.856	0.822	6.5
TextField (Ours)	**0.805**	0.843	**0.824**	5.2

TABLE IV: Comparison of methods on MSRA-TD500. † stands for the base net of the model is not VGG16.

Methods	recall	precision	f-measure
He *et al.* [61]	0.610	0.760	0.690
EAST [25]	0.616	0.817	0.702
ITN [32]	0.656	0.803	0.722
Zhang *et al.* [37]	0.670	0.830	0.740
RRPN [29]	0.680	0.820	0.740
He *et al.* † [27]	0.700	0.770	0.740
Yao *et al.* [38]	0.753	0.765	0.759
EAST † [25]	0.674	0.873	0.761
Wu *et al.* [39]	0.780	0.770	0.770
SegLink [34]	0.700	0.860	0.770
RRD [31]	0.730	0.870	0.790
Lyu *et al.* [35]	0.762	0.876	0.815
MCN [36]	**0.790**	**0.880**	**0.830**
TextField (Ours)	0.759	0.874	0.813

Conclusion

- **Flux representation encodes pixel relationship**
- **Accurately detect skeleton in the wild**
- **Efficiently detect irregular texts in natural images**
- **More applications and better use of direction information**

URLs for DeepFlux and TextField:
DeepFlux: https://arxiv.org/abs/1811.12608
TextField: https://arxiv.org/abs/1812.01393