Comics Datasets Framework: Mix of Comics datasets for detection benchmarking

Emanuele Vivoli, Irene Campaioli, Mariateresa Nardoni, Niccolò Biondi, Marco Bertini, and Dimosthenis Karatzas

Computer Vision Center, Barcelona, Spain

University of Florence, Florence, Italy

MANPU 2024

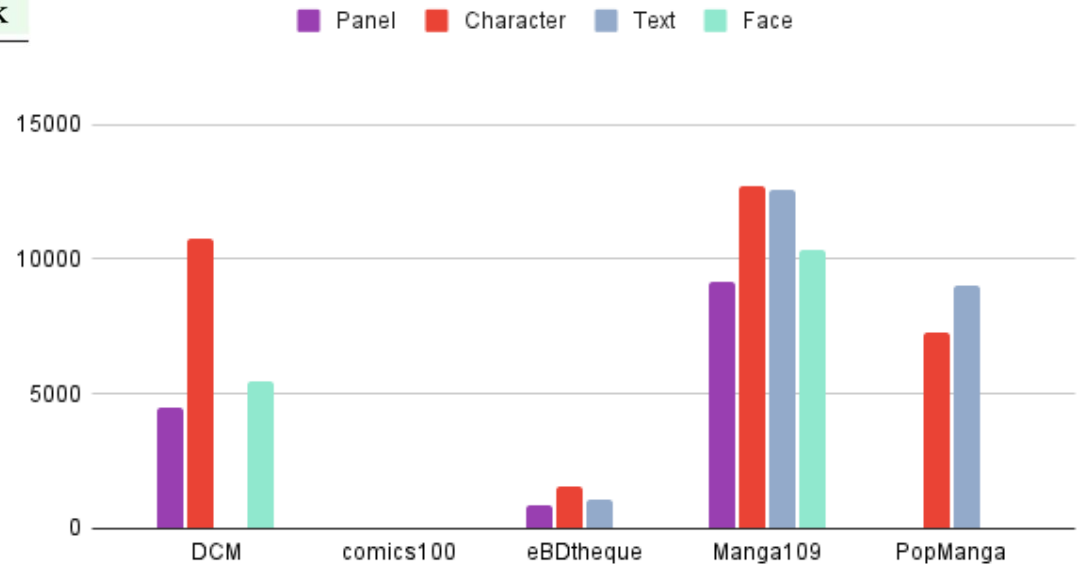
ICDAR Athens Greece 2024
International Conference on Document Analysis and Recognition

Outline

1. Landscape of Comics (Datasets-Benchmarks)
2. Challenges
3. Comics Datasets Framework
4. … Let’s annotate a bit …
5. Benchmarks
6. Conclusions
7. Future works

1. Landscape of **Comics Datasets**

Dataset	Release	Avail	Tasks	Years	Style	Books	Pages
eBDtheque [11]	2013	✓	d,t2c	1905-2012	mix	28	100
COMICS [15]	2017	✓	c	1938-1954	comics	3948	198k
GCN [5]	2017	✗	d,t2c	1978-2013	comics	*253	*38k
DCM772 [22]	2018	✓	d	1938-1954	comics	27	772
Manga109 [9, 24]	2018	✓	d,t2c,c2c	1970-2010	manga	109	10k
BCBId [6]	2022	✓	-	-	bangla	64	3k
VLRC	2023	✗	-	1940-now	-	*376	*7k
PopManga [27]	2024	✓	d,t2c,c2c	2010-2023	manga	25	1.8k



1. Landscape of **Comics Benchmarks**

DATASETS INFO			BENCHMARKS			
name	format	task	model type	work	metric	perf (%)
Fahad18 [13]	-	det [C]	-	[13]	mAP	41,7
Ho42 [7]	-	gm	**graph**	[7]	P	91,5
					R	71,5
		det [C]	**graph**	[7]	Acc	71,4
eBDtheque [9]	SVG/XML	seg [B]	**custom- CNN**	[6]	P	93,5
					R	96,2
					F1	94,8
		det [F]	**Faster R-CNN**	[23]	P	75,2
					R	49,8
					F1	60
sun70 [27]	-	det [C]	**SIFT**	[27]	P	97,8
					R	47
SSGCI [14]	XML	subg-s	**graph**	[14]	P	75,4
					R	9,8
					ScoreP	82,18
					ScoreR	80,71
COMICS [12]	TXT	T-c [easy]	**ComicVT5**	[29]	Acc	79,1
		T-c [hard]	**ComicVT5**	[29]	Acc	71,3
		V-c [easy]	**CNN + LSTM**	[12]	Acc	85,7
		V-c [hard]	**CNN + LSTM**	[12]	Acc	63,2
		C-c	**CNN + LSTM**	[12]	Acc	70,9
Comics3w [10]	-	det [P]	custom- Faster R-CNN	[11]	P	99,24
					R	99,16
					F1	99,2

DATASETS INFO			BENCHMARKS			
name	format	task	model type	work	metric	perf (%)
JC2463 [23]	-	det [F]	**Faster R-CNN**	[23]	P	95
					R	93,2
					F1	94,1
AEC912 [23]	-	det [F]	**Faster R-CNN**	[23]	P	82,4
					R	73,1
					F1	77,5
GNC [4]	CSV	det [B]	**U-Net (VGG-16)**	[3]	P	95,58
					R	94,04
					F1	94,48
DCM772 [18]	TXT	det [T]	custom- CNN	[6]	P	92,7
					R	96,9
					F1	94,7
Manga109-anns [8,21]	XML	seg [B]	custom- CNN	[6]	P	93,56
					R	95,49
					F1	94,51
		det [P]	**SSD300**	[21]	Acc	97,1
		det [T]	custom- SSD300	[21]	Acc	84.1
		det [F]	custom- SSD300	[21]	Acc	76.2
		det [C]	custom- SSD300	[21]	Acc	79.6
Sequencity4k [19]	-	seg [B]	**U-Net (VGG-16)**	[19]	P	91,01
					R	91,23
					F1	91.12 (+- 5.44)
EmoRecCom [20]	CSV	cls [E]	**CNN + BERT**	[20]	AUC	68,49
BCBId (Bangla) [5]	TXT/XML	seg [B]	custom- CNN	[6]	P	97,05
					R	98,81
					F1	97,92
		det [T]	custom- CNN	[6]	P	95,63
					R	98,52
					F1	97,05

1. Landscape of **Comics Benchmarks**

	PopManga (Test-S)		PopManga (Test-U)		Manga109	
method	Char	Text	Char	Text	Body	Panel
DASS [46]	0.8410	-	0.8580	-	**0.9251**	-
Grounding-DINO [24]	0.7250	0.7922	0.7420	0.8301	0.7985	0.5131
Magi [Ours]	**0.8485**	**0.9227**	**0.8615**	**0.9208**	0.9015	**0.9357**

Table 2. Detection Results. We report the average precision results, which have an upper bound of 1.0.

R. Sachdeva, A. Zisserman, "The Manga Whisperer: Automatically Generating Transcriptions for Comics", CVPR 2024

2. Challenges

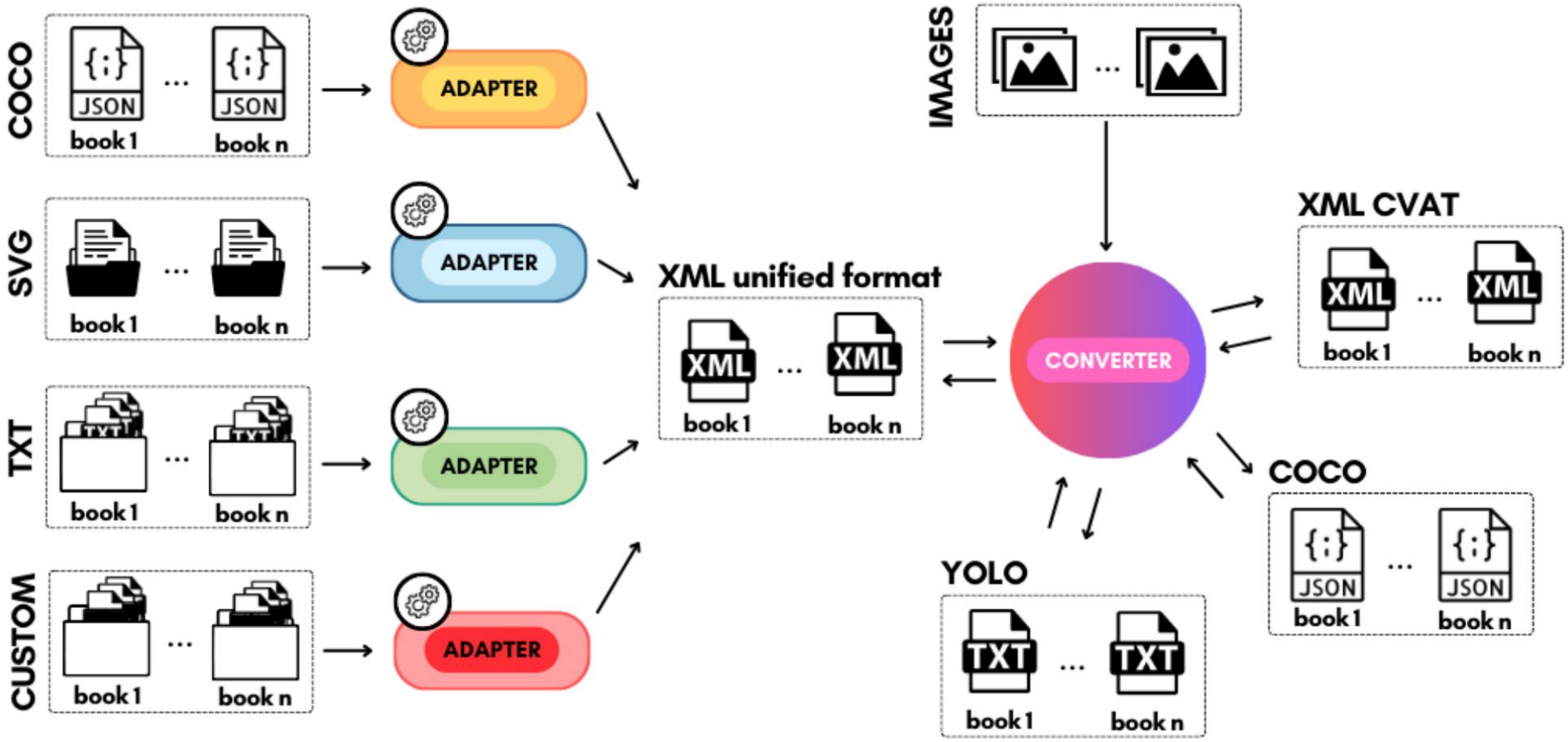
Dataset	Release	Avail	Tasks	Years	Style	Books	Pages
eBDtheque [11]	2013	✓	d,t2c	1905-2012	mix	28	100
COMICS [15]	2017	✓	c	1938-1954	comics	3948	198k
GCN [5]	2017	✗	d,t2c	1978-2013	comics	*253	*38k
DCM772 [22]	2018	✓	d	1938-1954	comics	27	772
Manga109 [9, 24]	2018	✓	d,t2c,c2c	1970-2010	manga	109	10k
BCBId [6]	2022	✓	-	-	bangla	64	3k
VLRC	2023	✗	-	1940-now	-	*376	*7k
PopManga [27]	2024	✓	d,t2c,c2c	2010-2023	manga	25	1.8k
comics100	2024	✓	d,t2c,c2c,N,D	1938-1954	comics	100	5.5k

2. Challenges

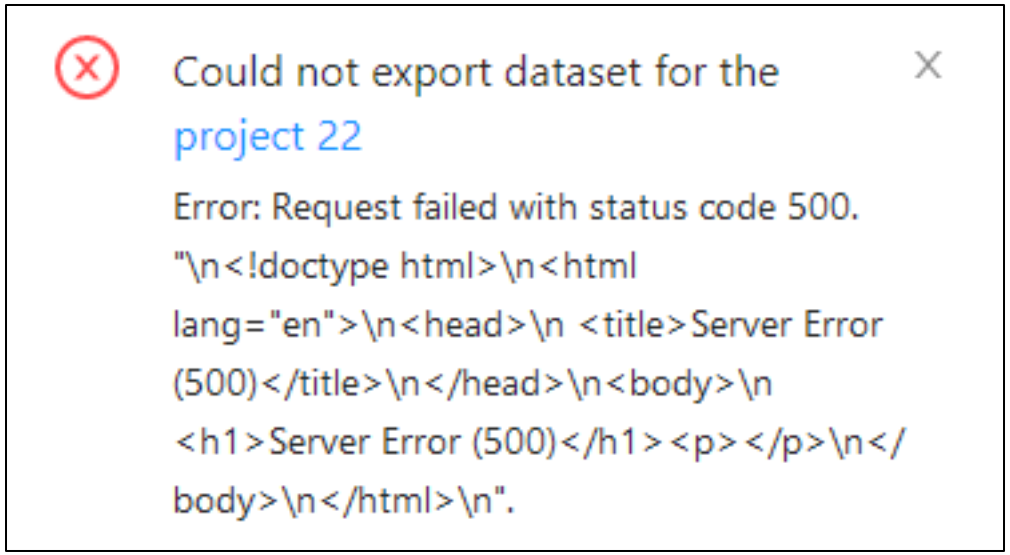
Dataset	Release	Avail	Tasks	Years	Style	Books	Pages
eBDtheque [11]	2013	✓	d,t2c	1905-2012	mix	28	100
COMICS [15]	2017	✓	c	1938-1954	comics	3948	198k
GCN [5]	2017	✗	d,t2c	1978-2013	comics	*253	*38k
DCM772 [22]	2018	✓	d	1938-1954	comics	27	772
Manga109 [9, 24]	2018	✓	d,t2c,c2c	1970-2010	manga	109	10k
BCBId [6]	2022	✓	-	-	bangla	64	3k
VLRC	2023	✗	-	1940-now	-	*376	*7k
PopManga [27]	2024	✓	d,t2c,c2c	2010-2023	manga	25	1.8k
comics100	2024	✓	d,t2c,c2c,N,D	1938-1954	comics	100	5.5k

Every dataset has it’s own annotations

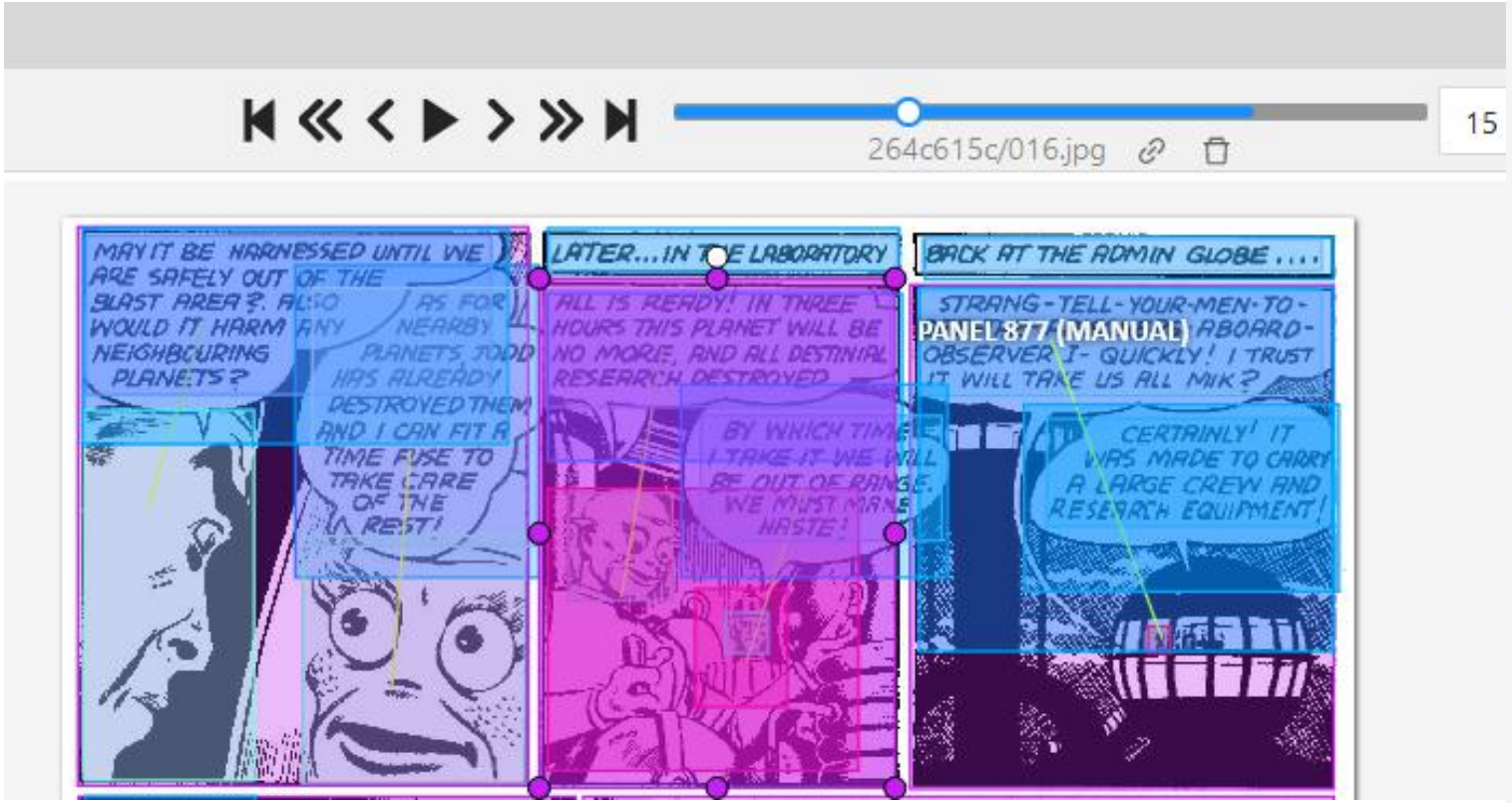
3. Comics Datasets Framework

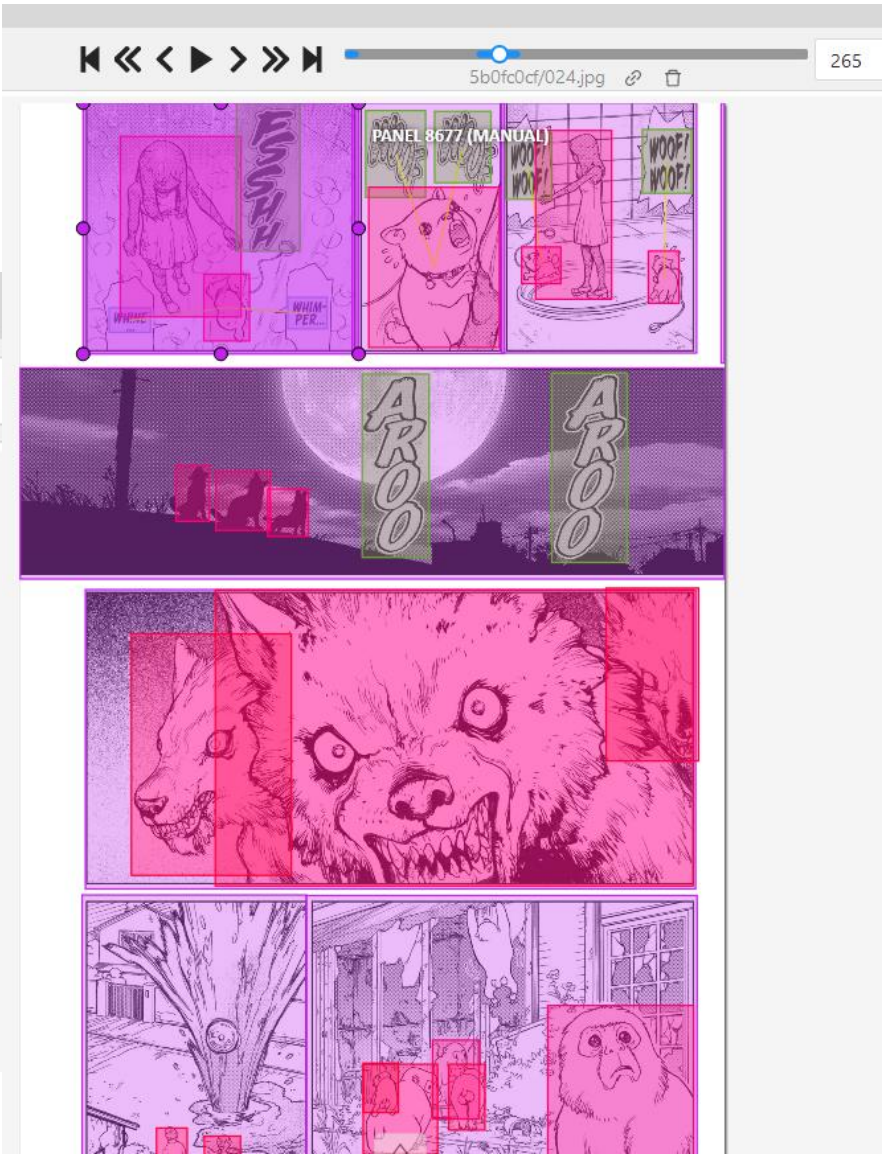


4. … Let’s annotate a bit …

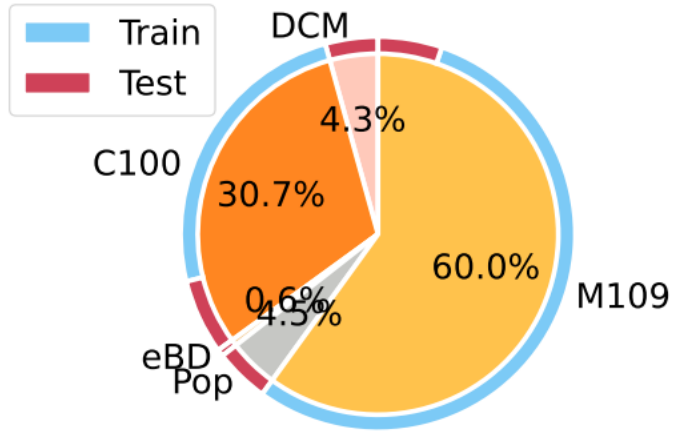


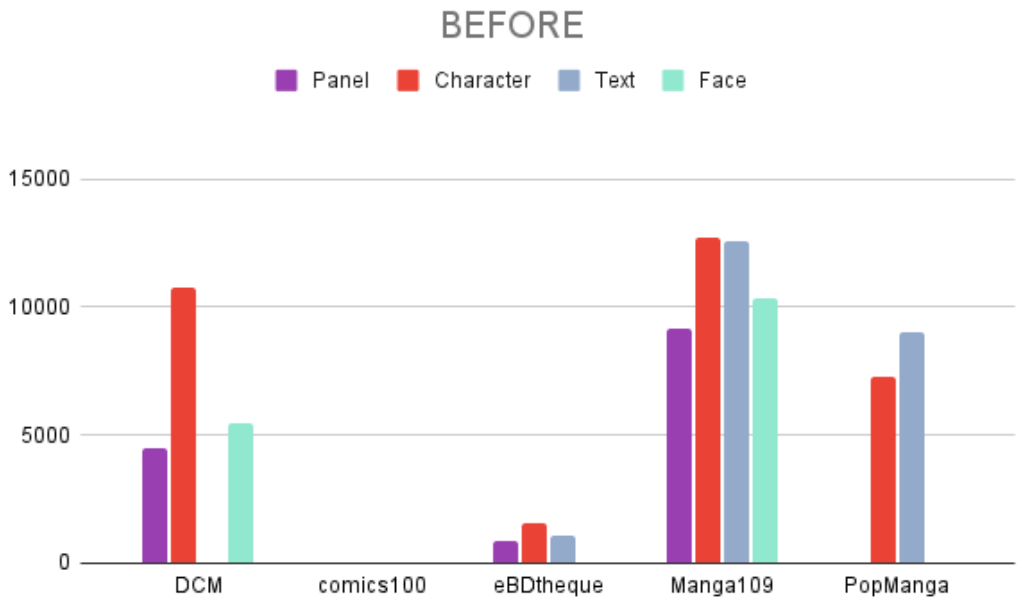
4. … Let's annotate a bit …

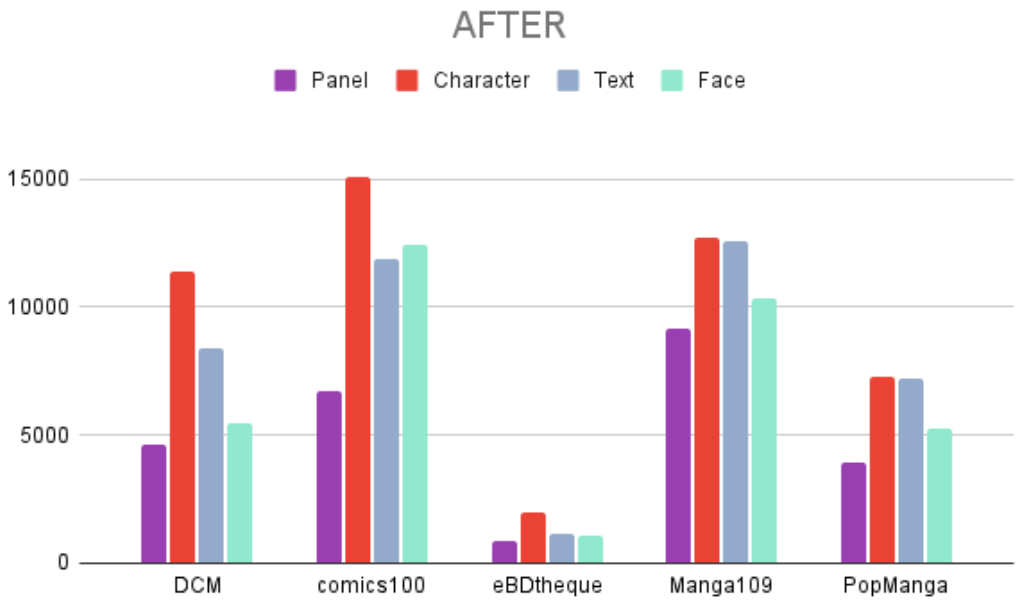




4. … Let’s annotate a bit …







5. Benchmarks

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	63,4	62,5	56,9	61,8	73,7	64,7
R-CNN	86,3	**88,9**	65,4	64,9	77,6	**79,5**
SSD	12,1	9,1	28,4	34,6	4,6	15,6
YOLO	81,4	75,0	**67,0**	**76,8**	64,5	74,3
DASS	-	-	-	-	-	-
Magi	**89,0**	73,9	62,1	65,3	**92,8**	78,5

Panel detection.

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	57,7	62,1	40,1	25,3	46,0	48,2
R-CNN	50,6	61,0	34,7	4,7	50,9	42,2
SSD	52,4	54,1	39,5	55,8	32,6	49,3
YOLO	45,6	55,4	30,1	9,4	42,0	38,6
DASS	**75,1**	76,0	**60,9**	**84,4**	70,5	**76,3**
Magi	71,8	**76,7**	56,6	50,4	**79,7**	69,3

Characters detection.

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	66,5	58,9	37,3	38,1	62,0	55,4
R-CNN	43,0	38,9	20,7	8,7	43,0	32,7
SSD	60,1	60,0	30,9	76,4	75,4	66,5
YOLO	43,1	48,8	20,6	16,2	42,1	37,5
DASS	**78,8**	**62,7**	**61,1**	**87,8**	**78,0**	**75,3**
Magi	-	-	-	-	-	-

Face detection.

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	20,7	23,0	17,8	9,9	27,6	20,1
R-CNN	64,2	**83,1**	41,9	14,4	48,5	54,0
SSD	58,5	70,2	38,5	**70,8**	31,7	59,1
YOLO	68,3	73,0	38,7	42,2	12,7	50,9
DASS	-	-	-	-	-	-
Magi	**84,0**	77,9	**73,6**	49,2	**93,4**	**75,2**

Text detection.

5. Benchmarks

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	34,7	36,9	25,3	27,0	36,8	33,7
R-CNN	40,7	**58,2**	27,1	18,5	44,2	41,1
SSD	30,5	33,3	22,9	**47,5**	17,2	32,6
YOLO	39,7	50,8	26,1	28,9	29,8	38,1
DASS	25,7	19,0	20,3	34,4	17,6	23,9
Magi	**40,8**	57,1	**32,1**	33,0	**66,5**	**49,2**

Average mAP across ALL CLASSES

	DCM	c100	eBD	M109	Pop	*avg*
G.Dino	52,1	51,1	38,0	33,8	52,0	46,9
R-CNN	61,0	66,0	40,7	23,1	59,0	52,3
SSD	45,8	44,5	34,3	59,4	23,0	43,7
YOLO	59,6	63,0	39,1	36,1	54,1	53,2
DASS	77,0	67,6	61,0	**86,1**	73,3	**75,1**
Magi	**81,6**	**75,6**	**64,1**	54,9	**88,0**	74,0

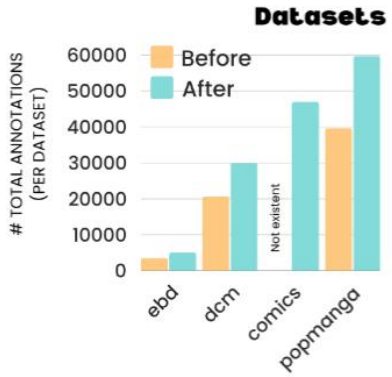
Average mAP across ALL **DETECTABLE** CLASSES

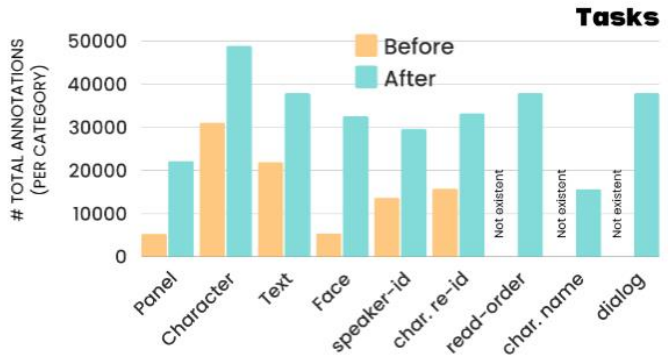
6. Conclusions

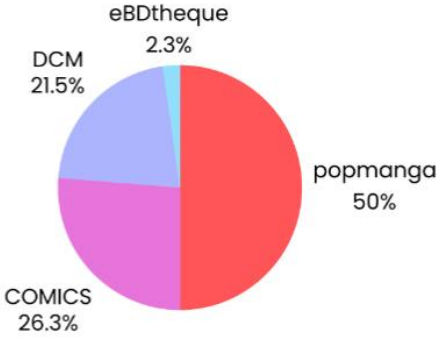
1. Landscape of Comics Datasets: **messy**
2. Challenges: **consistency and replicability**
3. Comics Datasets Framework: **consistency**
4. … Let’s annotate a bit …
5. Benchmarks: **replicability**
6. Conclusions: **hope for adoption**

7. What next?

CoMix: A Comprehensive Benchmark for Multi-Task Comic Understanding







7. What next?

CoMix: A Comprehensive Benchmark for Multi-Task Comic Understanding

Object Detection

Panels Characters Textboxes Faces

Speaker id.

Character Re-Id

Reading order

Character Naming

Narrator, Sailor 1, McWhustle, Captain
Matey, Sailor 2, Sailor 3, Sailor 4

Dialog generation

① **Narrator:** "As the tramp steamer SS. Clementine crosses the Equator [...]"
② Sailor 1: "THAT DISGUISE DOESN'T FOOL ME! YOU'RE CHIEF ENGINEER [...]"
③ McWhustle: "HOOT, LADDIE! BEFORE YE CAN BE A SON OF NEPTUNE, YE [...]"
④ McWhustle: "HOOT, MON! WE'RE AGROUND ON A REEF!"
Captain: "THERE ARE NO REEFS IN THESE PARTS! BACK TO YOUR ENGINE [...]"
Captain: "WHAT'S WRONG WITH YOUR BLASTED ENGINES, MCWHUSTLE ?"
McWhustle: "WHY DON'T YE LEARN TO NAVIGATE YER SHIP, LASSITER! [...]"
McWhustle: "NO HARD FEELINGS, CAP'N! BUT THE ENGINES ARE DOING FINE!"
Captain: "I KNOW THAT, MAC... WE'VE BEEN SHIPMATES TOO LONG TO [...]"
Captain: "WE SHOULD BE MAKING HEADWAY... BUT SOMETHING'S [...]"
McWhustle: "MON, IT'S NO CANNY!"
Narrator: "Meanwhile, below..."
Matey: "THE SKIPPER' SAYS WE CAN'T HAVE RUN AGROUND!"
Sailor 2: "HEY, MATEY! LOOK!"
Sailor 3: "AHOY, TOPSIDES! STAND BY TO REPEL BOARDERS IN THE [...]"
Sailor 4: "BOARDERS IN THE STOKEHOLD? I NEVER HEARD OF SUCH A THING."

7. What next?

Paper will be on arXiv
the next days
(**BEFORE END OF ICDAR**)

A **Survey** on **Comics Understanding**

Layer	Category	Task	Input	Output
5 (Sec.10)	Synthesis (Sec. 10.2)	Narrative-Based Complex Scene Generation (NCSG)	Detailed Narrative Text	Series of Images
		Video Generation from Text (VGT)	Complex Long Text	Video
		3D Model Generation from Images (3DGI)	Collection of Images	3D Model
	Generation (Sec. 10.1)	Sound Generation from Single Panel	Single Comic Panel	Sound/Audio
		Scene Graph Generation for Captioning	Comic Panel	Scene Graph
		Image Generation [text-2-img] (IG)	Text	Image
		Grounded Image Captioning (GIC)	Image	Text + Bbox
		Image Captioning [img-2-text] (IC)	Image	Text
4 (Sec.9)	Understanding (Sec. 9.1)	Visual Reasoning (VR)	Image + Text	Text
		Visual Dialog (VisDial)	Image + Dialog + Text	Text
		Visual Question Answering (VQA)	Image + Text	Text
		Visual Entailment (VE)	Image + Text	Tag
3 (Sec.8)	Modification (Sec. 8.2)	Image Editing via Text (IET)	Text + Image	Image
		Image Inpainting (II)	Text + [prompt] + Image	Image
	Retrieval (Sec. 8.1)	Composed Image Retrieval (CIR)	Text + Image	Image
		Text-Image Retrieval (TR)	Image	Text
		Image-Text Retrieval (IR)	Text	Image
2 (Sec.7)	Segmentation (Sec. 7.3)	Instance Segmentation (IS)	[Prompt] + Image	Segments
	Analysis (Sec. 7.2)	Translation	Image	Text
		Dialog transcription	Image	Text
		Character-Balloon Association (Speaker ID)	Character + Balloons	Tag
	Grounding (Sec. 7.1)	Grounding (IG)	[Prompt] + Image	Bounding Boxes
		Character Re-identification (Character ID)	Multiple Panels	Tag
		Object Detection	Tag/s + Image	Bounding Boxes
1 (Sec.6)	Augmentation (Sec. 6.2)	Depth Estimation	Comic Panels/Images	Depth Map
		Vectorization	Comic Panels/Images	Vector Image
		Style Transfer (ST)	Image	Image
		Image Super-Resolution (ISR)	Image	Image
	Tagging (Sec. 6.1)	Page Stream Segmentation (PSS)	Images	Tags sequence
		Action Detection	Multiple Panels	Tag
		Emotion Classification	Comic Panels/Images	Emotion Labels
		Image Classification (I-CLS)	Image	Tag
0 (Sec.5)	View (Sec. 5.1)	Basic Image Viewing (BIV)	Text Command	Image Display

Computer Vision Center

Thanks for your Attention

Comics Datasets Framework: Mix of Comics datasets for detection benchmarking