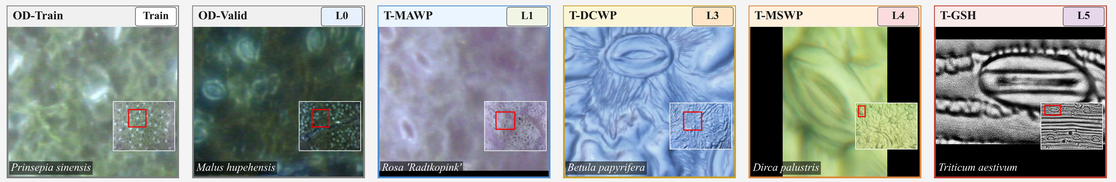
NeurIPS 2026 · Evaluations & Datasets Track · Spotlight

StomataBench

Measuring Taxonomic Generalization in Stomatal Detection

Sainath Reddy Gummi[1] · Hankui K. Zhang[1] · Chulwoo Pack[1] · Shyam Solanki[1] · Young Chang[1,2]
[1]South Dakota State University [2]Manitoba Agriculture

Increasing biological distance from OD-Train



1 Detectors are scored on boxes. Biology uses density.

Stomata are the microscopic pores plants use to exchange gas and water. Their density, stomata per mm² of plant surface, links cellular anatomy to photosynthesis, transpiration and water-use efficiency. Object detectors now count them at scale.

What detection benchmarks report
COCO AP
Do the predicted boxes land on the stomata?

What biologists use
Stomatal density
One count per image, from every box above a confidence threshold.

Does a good detector give a good measurement when the species and the microscope change?

2 One training set, four held-out test sets

Test species sit on a six-level OOD ladder, set by the first taxonomic rank where they differ from the training species.

L0	L1	L2	L3	L4	L5
Familiar species, new accession	New species, familiar genus	New genus, familiar family	New family, familiar order	New order, same clade	New clade: monocots

OD-Train training
3,154 images · 134,802 stomata · 21 taxa
Woody plants: 20 eudicots and one magnoliid. No monocots, gymnosperms or ferns.

OD-Valid validation L0
138 images · 8,792 stomata · 5 taxa
OD-Train species from new accessions. Used to select thresholds.

T-MAWP T L1 L2 L3 L4
456 images · 17,164 stomata · 17 taxa
New species. Same microscope and annotation protocol as OD-Train.

T-DCWP T+I L0 L1 L2 L3 L4
119 images · 2,858 stomata · 39 taxa
Adds a different microscope, the Olympus BX-63.

T-MSWP T+I+A L1 L3 L4
10,715 images · 514,568 stomata · 18 taxa
Adds another lab's imaging and its own annotators.

T-GSH T+I+A L4 L5
8,229 images · 319,161 stomata · 9 taxa
Multi-source, from soybean and poplar to wheat, barley, rice and maize.

Shift from OD-Train: T taxonomic, I imaging system, A annotator. Images and stomata as released.

3 Every detector is scored twice

As a detector
COCO AP
$AP_{50:95}$ over the ranked boxes

As a measurement tool
Count and density
Count MAE, density MAE and within-species density R^2 at one threshold τ

$$R^2_s = 1 - \frac{\Sigma_i (\rho_i - \hat{\rho}_i)^2}{\Sigma_i (\rho_i - \bar{\rho}_s)^2}$$

ρ_i is the density of image i of species s. R^2 below 0 means the predicted densities are worse than predicting the species mean $\bar{\rho}_s$ for every image.

Two-stage CNN	FRCNN-R50-1×, FRCNN-R50-3×, FRCNN-R101-1×, FRCNN-X101
One-stage CNN	Retina-R50-1×, Retina-R50-3×, Retina-R101-3×
YOLO	YOLOv8, YOLOv11, YOLOv12, YOLOv26
Transformer	RT-DETRv2, RT-DETRv4
Foundation-model-derived	GDINO-SB, SAM3-OD

All 15 detectors are fine-tuned once on OD-Train and evaluated unchanged. Results use τ = 0.25 for every detector. Thresholds selected on OD-Valid are reported too.

4 No detector leads every shift

Test set	Shifted from OD-Train: Species	Imaging	Annotators	AP leader	$AP_{50:95}$ (%)	Second
T-MAWP	●	○	○	**RT-DETRv4**	39.08	YOLOv11 37.76
T-DCWP	●	●	○	**SAM3-OD**	46.46	GDINO-SB 43.26
T-MSWP	●	●	●	**GDINO-SB**	45.28	YOLOv11 44.57
T-GSH	●	●	●	**SAM3-OD**	24.93	GDINO-SB 20.96

The best detector depends on the kind of shift, so one held-out test set is not enough to judge transfer.

5 Same AP, opposite measurement

The same 45 apple images (*Malus domestica*, T-MAWP, L1) at τ = 0.25. Brackets are 95% bootstrap intervals.

YOLOv12

$AP_{50:95}$ (%)	Density R^2	Density MAE (mm^{-2})
44.59	**0.83** [0.70, 0.90]	**10.2** [7.7, 12.6]

RT-DETRv4

$AP_{50:95}$ (%)	Density R^2	Density MAE (mm^{-2})
43.57	**−0.67** [−1.71, −0.14]	**37.6** [32.6, 43.2]

A species' AP leader is also its density R^2 leader for only 1 of the 17 T-MAWP species.

6 Density recovery is strongest near the training species

■ R^2 below 0: worse than the species-mean reference ■ R^2 of 0 or above

Each bar is the 15 detectors at τ = 0.25. Grey number: calibrated images. Right: detectors with R^2 below 0.

OD-Valid 5 species

L0	*Actinidia arguta*	37	0
L0	*Malus hupehensis*	41	4
L0	*Malus spontanea*	40	5
L0	*Malus toringoides*	9	5
L0	*Vitis vinifera*	11	9

T-MAWP 17 species

L1	*Actinidia coriacea*	37	3
L1	*Malus domestica*	45	2
L1	*Prunus Tai-Haku*	47	4
L1	*Prunus maritima*	12	15
L1	*Prunus sargentii*	18	7
L1	*Prunus subhirtella*	20	8
L1	*Rosa Radtkopink*	12	1
L1	*Rosa centifolia*	10	13
L1	*Rosa chuckles*	29	2
L1	*Vitis empire*	9	3
L1	*Vitis riparia*	33	8
L2	*Fagus crenata*	17	13
L2	*Amelanchier sanguinea*	32	9
L3	*Rhododendron Roseum Superbum*	32	5
L4	*Populus tremula*	40	13
L4	*Populus tremuloides*	23	12
L4	*Populus trichocarpa*	40	7

T-MSWP 18 species

L1	*Quercus michauxii Nutt.*	129	5
L1	*Quercus nigra L.*	500	10
L1	*Quercus pagoda Raf.*	286	12
L1	*Quercus phellos L.*	568	14
L1	*Quercus shumardii Buckley*	1348	7
L1	*Quercus stellata Wangenh.*	96	14
L1	*Quercus texana Buckley*	2434	7
L3	*Acer rubrum L.*	100	15
L3	*Carya ovata (Mill.) K. Koch*	91	14
L3	*Nyssa sylvatica Marshall*	50	14
L3	*Ulmus alata Michx.*	205	15
L3	*Ulmus americana Planch*	185	14
L3	*Vaccinium stamineum L.*	75	15
L4	*Dirca palustris L.*	65	13
L4	*Fraxinus caroliniana Miller*	100	13
L4	*Ilex opaca Aiton*	100	14
L4	*Populus L.*	3065	10
L4	*Salix L.*	82	13

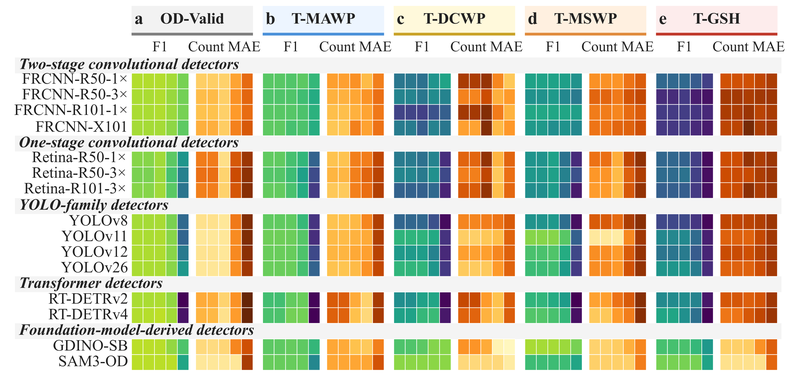
T-GSH 4 species

L4	*Populus balsamifera*	113	11
L5	*Triticum aestivum*	348	15
L5	*Hordeum vulgare*	542	14
L5	*Zea mays*	927	13

Wheat (L5): on the 348 calibrated wheat images every detector is below 0. OD-Train has no monocots, so wheat is a stress test.

The 37 T-DCWP taxa, most with three images, are in Figure 3 of the paper.

7 The confidence threshold changes the winner



15 of 17 T-MAWP species whose density R^2 leader changes between τ = 0.25 and 0.75. The count MAE leader changes for 14.

On T-MAWP, RT-DETRv4 has the highest AP but one of the highest count MAEs at τ = 0.25, and the lowest count MAE at 0.50. Wheat R^2 stays below 0 at every threshold.

8 Takeaways

1 Test across several kinds of shift, not one held-out set.

2 Report count and density error next to AP.

3 State the confidence threshold with every result.

Benchmark scores are not a deployment guarantee. Calibrate in the target domain, choose thresholds on independent validation data, and keep human review.

9 Data, weights and code

Project page
gummi001.github.io/StomataBench

Dataset
huggingface.co/datasets/Sainath001/StomataBench

Weights
huggingface.co/Sainath001/stomatabench_neurips_weights

Code
github.com/gummi001/stomatabench_neurips_eval

Images and annotations are released under CC-BY-NC-4.0 with Croissant metadata.