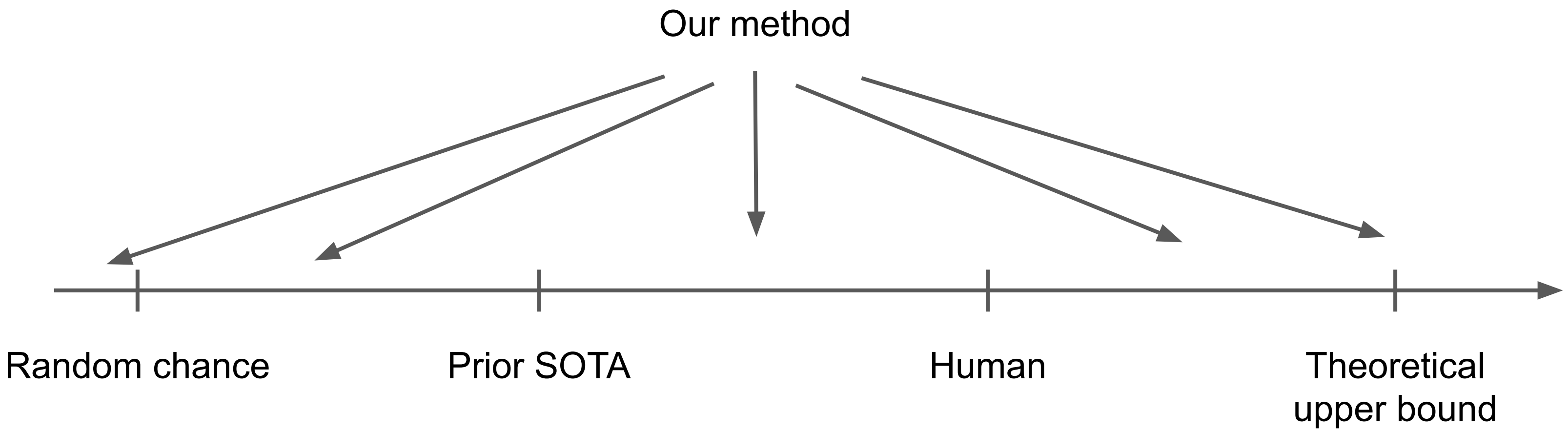
Evaluating weakly-supervised models

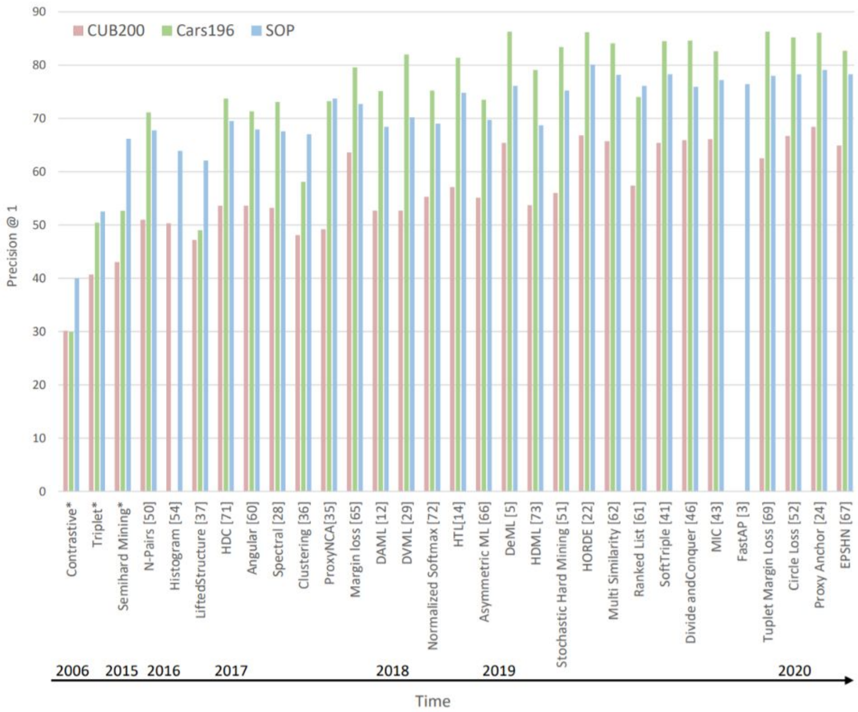
Junsuk Choe.

Why do we do evaluation?

It enables ranking.



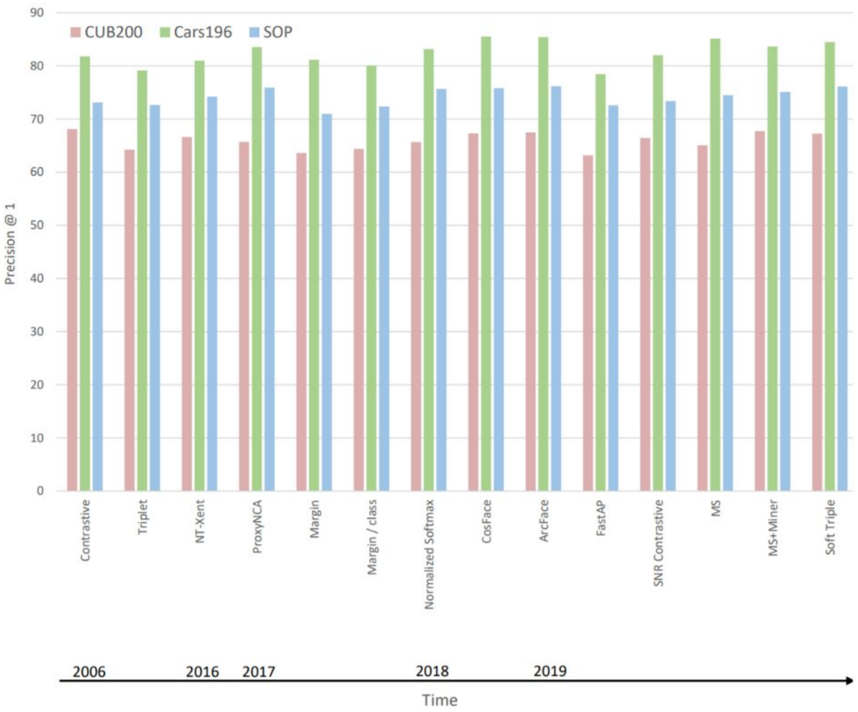
What are the costs of wrong evaluation?



(a) The trend according to papers

Musgrave et al. A Metric Learning Reality Check. ECCV'20.

What are the costs of wrong evaluation?



(b) The trend according to reality

Musgrave et al. A Metric Learning Reality Check. ECCV'20.

What are the costs of wrong evaluation?

Researchers

- 4+ years efforts put into pursuing the wrong metric.
- Opportunity cost: what if they have worked on other “real” challenges?

Practitioners

- Misinformed selection of methods based on the wrong ranking.
- Cost of neglecting a simple solution that works equally well.

Musgrave et al. A Metric Learning Reality Check. ECCV’20.

Similar “evaluation scandals”
in many CV & ML tasks.

Face detection: Mathias et al. Face Detection without Bells and Whistles. ECCV'14.

Zero-shot learning: Xian et al. Zero-Shot Learning-The Good, the Bad and the Ugly. CVPR'17.

Semi-supervised learning: Oliver et al. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. NeurIPS'18.

Unsupervised disentanglement: Locatello et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. ICML'19.

Image classification: Recht et al. Do ImageNet Classifiers Generalize to ImageNet? ICML'19.

Scene text recognition: Baek et al. What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. ICCV'19.

Weakly-supervised object localization: Choe et al. Evaluating Weakly-Supervised Object Localization Methods Right. CVPR'20.

Deep metric learning: A Metric Learning Reality Check. ECCV'20.

Natural language QA: Lewis et al. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. ArXiv'20.

Recipes for wrong evaluation.

1. Confound multiple factors when comparing methods.

k	1	10	100	1000	NMI
Histogram [34]	63.9	81.7	92.2	97.7	-
Binomial Deviance [34]	65.5	82.3	92.3	97.6	-
Triplet Semi-hard [25, 29]	66.7	82.4	91.9	-	89.5
LiftedStruct [22, 29]	62.5	80.8	91.9	-	88.7
StructClustering [29]	67.0	83.7	93.2	-	89.5
N-pairs [28]	67.7	83.8	93.0	97.8	88.1
HDC [41]	69.5	84.4	92.8	97.7	-
Margin	**72.7**	**86.2**	**93.8**	**98.0**	**90.7**

Improvements come from the loss function?

Musgrave et al. A Metric Learning Reality Check. ECCV'20.
Wu et al. Sampling Matters in Deep Embedding Learning. ICCV'17.

1. Confound multiple factors when comparing methods.

Architecture	k	1	10	100	1000	NMI
GoogleNet	Histogram [34]	63.9	81.7	92.2	97.7	-
GoogleNet	Binomial Deviance [34]	65.5	82.3	92.3	97.6	-
Inception-BN	Triplet Semi-hard [25, 29]	66.7	82.4	91.9	-	<u>89.5</u>
Inception-BN	LiftedStruct [22, 29]	62.5	80.8	91.9	-	88.7
Inception-BN	StructClustering [29]	67.0	83.7	<u>93.2</u>	-	<u>89.5</u>
Inception-BN	N-pairs [28]	67.7	83.8	93.0	<u>97.8</u>	88.1
GoogleNet	HDC [41]	<u>69.5</u>	<u>84.4</u>	92.8	97.7	-
ResNet50	**Margin**	**72.7**	**86.2**	**93.8**	**98.0**	**90.7**

Or from the architecture?

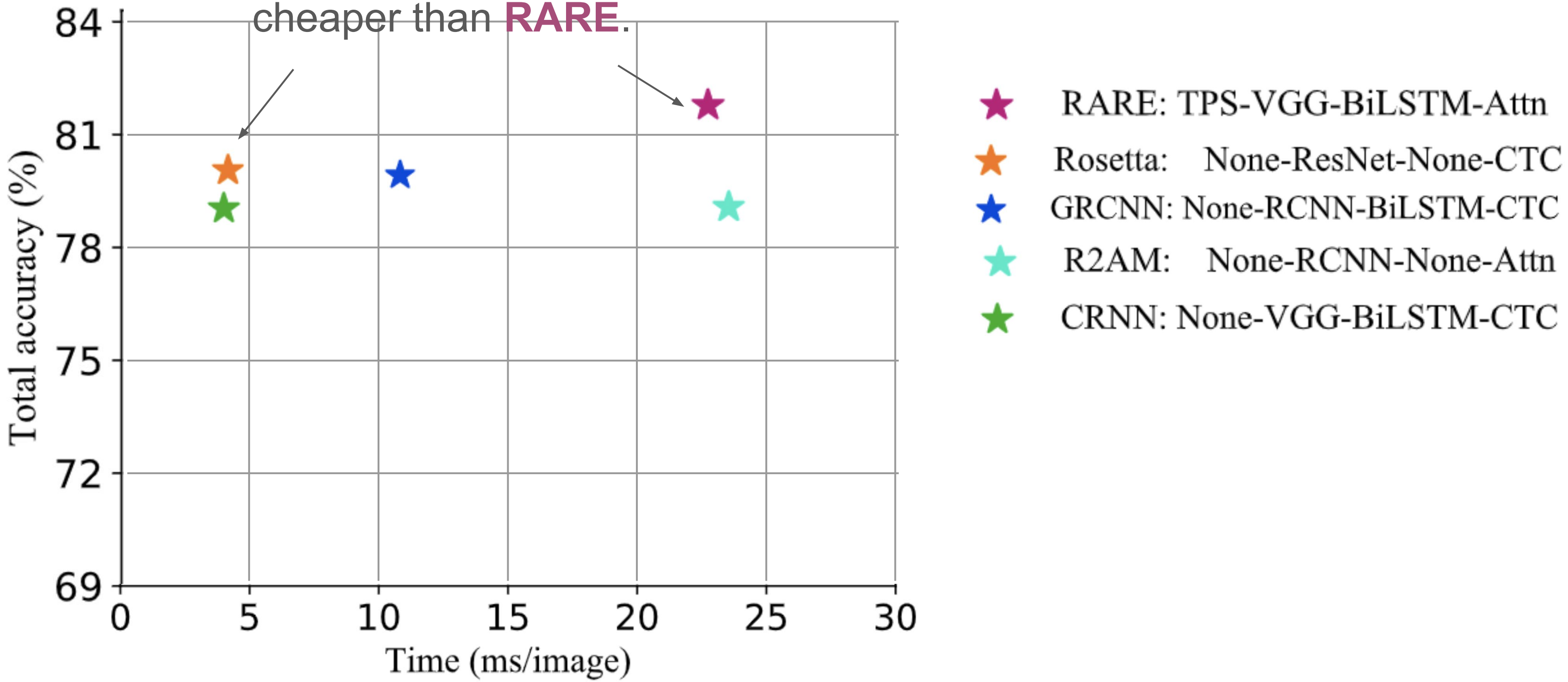
Musgrave et al. A Metric Learning Reality Check. ECCV’20.
Wu et al. Sampling Matters in Deep Embedding Learning. ICCV’17.

2. Hide extra resources needed to make improvements.

Baek et al. What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. ICCV'19.

2. Hide extra resources needed to make improvements.

But **Rosetta** is much cheaper than **RARE**.



Baek et al. What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. ICCV'19.

3. Train and test samples overlap.

Dataset	% Answer overlap	% Question overlap
Natural Questions	63.6	32.5
TriviaQA	71.7	33.6
WebQuestions	57.9	27.5

Fraction of test sets overlapping with the training set for natural language Q & A task.

Lewis et al. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. ArXiv'20.

3. Train and test samples overlap.

Model		Open Natural Questions			
		Total	Question Overlap	Answer Overlap Only	No Overlap
Closed book	T5-11B+SSM	36.6	77.2	22.2	9.4
	BART	26.5	67.6	10.2	0.8
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0
	TF-IDF	22.2	56.8	4.1	0.0

Model performances in different partitions of the test set.

Models have solved the task by **memorising**, rather than by **generalising**.

Lewis et al. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. ArXiv'20.

This talk:
What can go wrong with evaluation?

This talk:

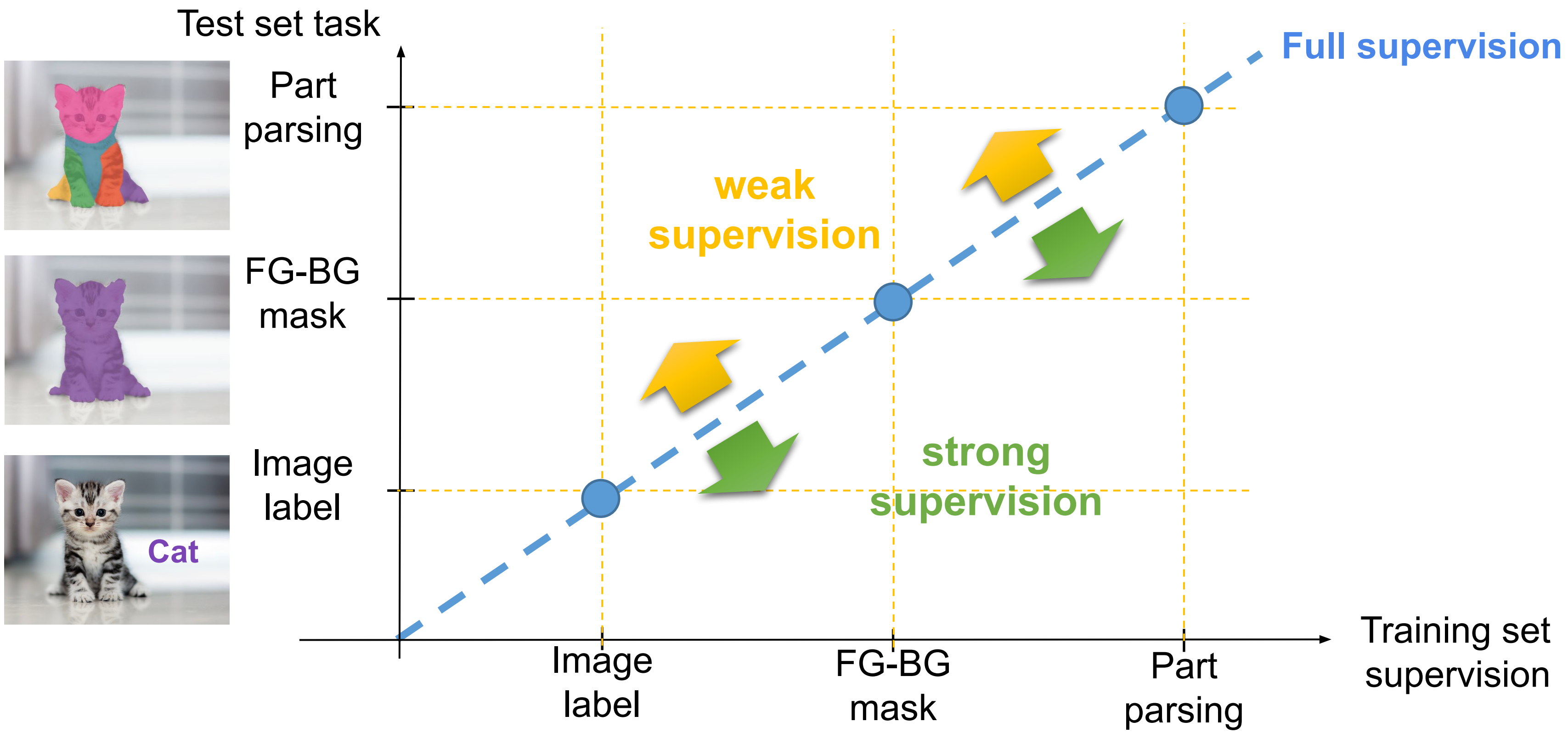
~~What can go wrong with evaluation?~~

What can go wrong with **weakly-supervised X** evaluation?

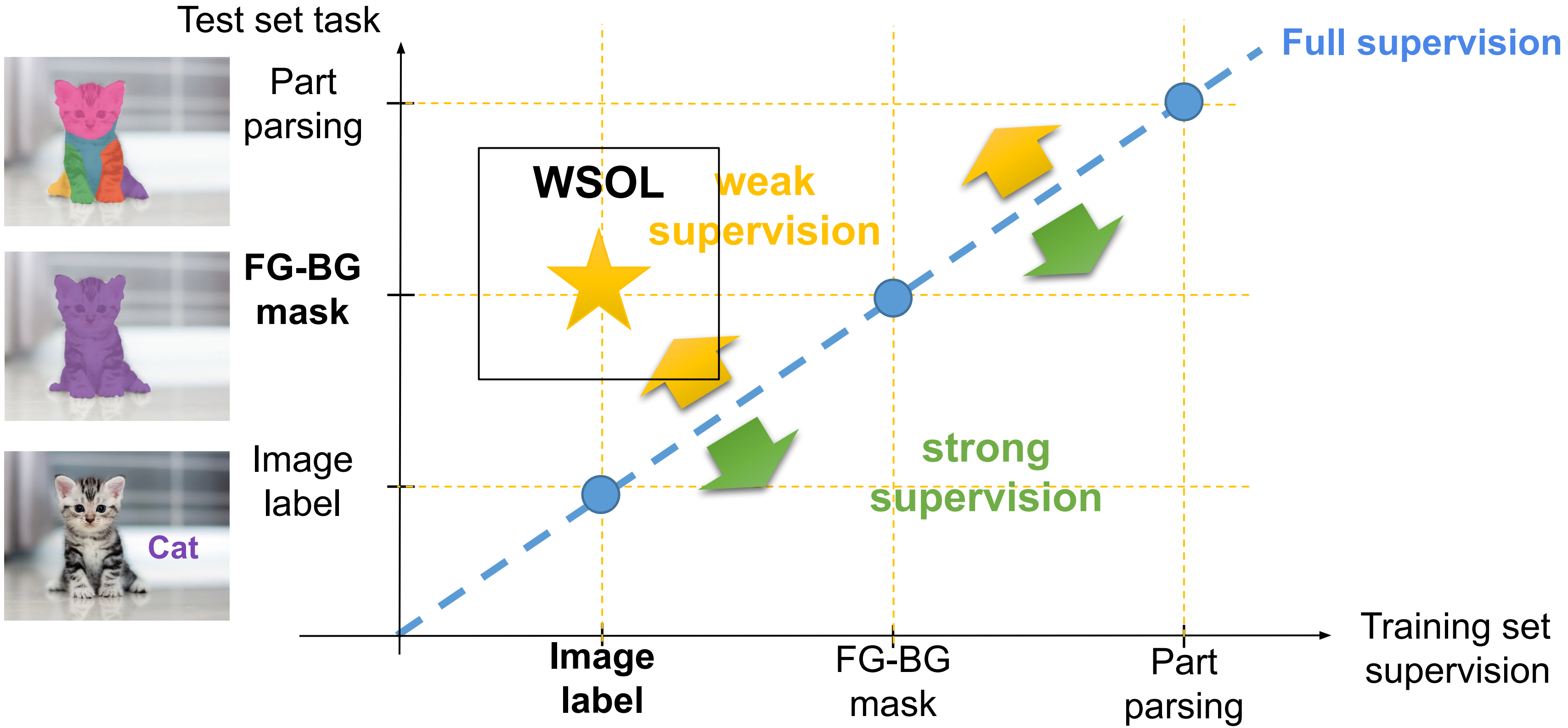
Weak supervision?

Appealing setup for ML - minimal annotation costs.

What do you mean by weak supervision?



What type of weak supervision?



Added complications for WSX evaluation.

Train/val/test splits for regular ML task.

Train set (Full sup)	Val set (Full sup)	Test set (Full sup)
Model fitting.	Model design choices. Tuning HPs.	Report final numbers. Comparison across methods.

Added complications for WSX evaluation.

Train/val/test splits for weakly-supervised X task.

Train set
(Weak sup)

Val set
(Full sup)

Test set
(Full sup)

“Weakly-supervised” method is supposed to use this set **ONLY**.

Added complications for WSX evaluation.

Train/val/test splits for weakly-supervised X task.

Train set (Weak sup)	Val set (Full sup)	Test set (Full sup)

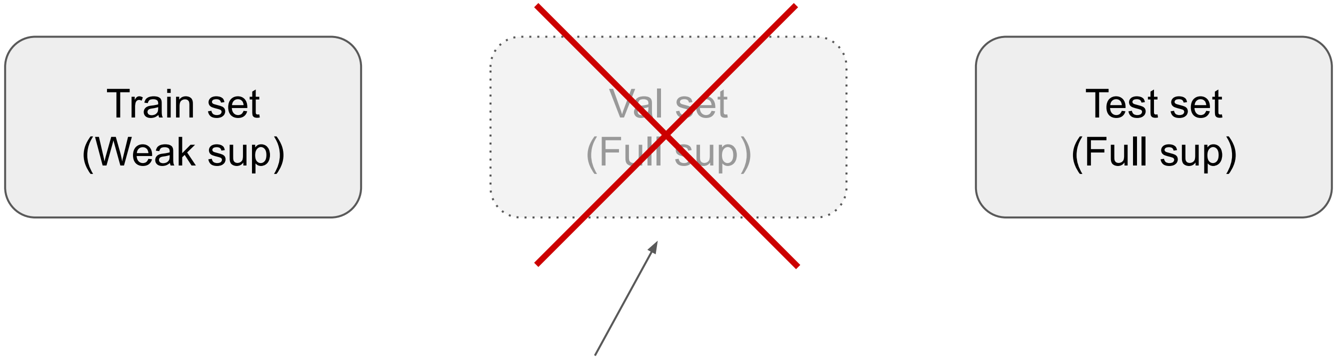
Usually used for tuning HPs.

Lack of unified agreement on “how to use”.

Some methods extensively make use of
val set for HP search (e.g. grid search)
→ Unfair !

Added complications for WSX evaluation.

Train/val/test splits for weakly-supervised X task.



Even worse, there is no val set
in many WSX benchmarks.

Added complications for WSX evaluation.

Train/val/test splits for weakly-supervised X task.

Train set
(Weak sup)

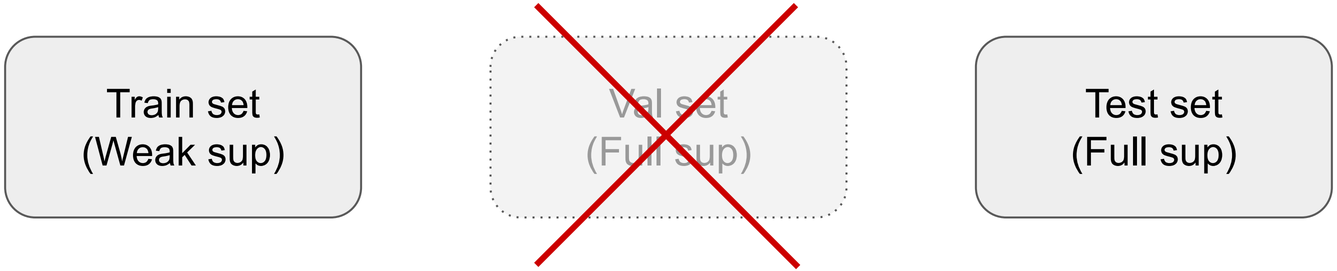
~~Val set~~
~~(Full sup)~~

Test set
(Full sup)

And people tune their HPs
over the test set !

Added complications for WSX evaluation.

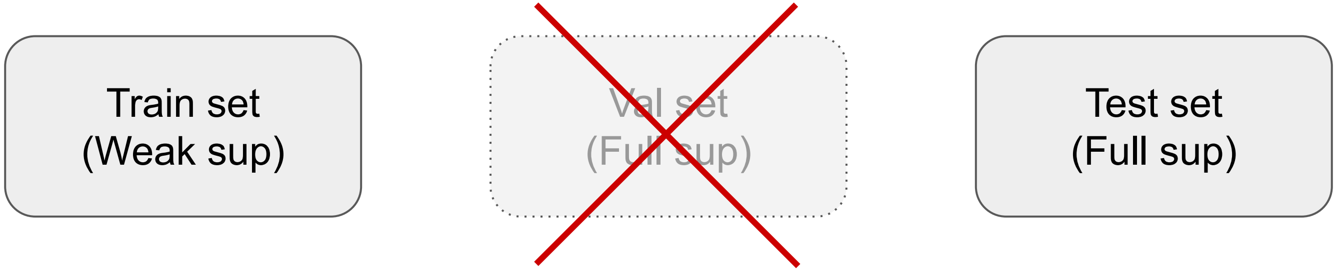
Correct evaluation is even more tricky for WSX.



1. Implicit tuning on the test set (problem shared by regular ML tasks).
2. **Implicit use of full supervision (specific to WSX tasks).**

Added complications for WSX evaluation.

These evaluation issues with WSX are not widely known yet.



Many researchers and practitioners are still misinformed by wrong evaluation results.

Case study: Weakly-supervised object localization.

WSOL is the "minimal working example" for the WSX evaluation issues.

Same problem in WSSS (semantic segmentation), WSOD (object detection), WSIS (instance segmentation), SSL (semi-supervised learning), UD (unsupervised disentanglement), ZSL (zero-shot learning), etc.

Other motivations

- Popularity: 100+ papers in the last 5+ years.
- Applicability: Ingredient for other WSX tasks.

CVPR 2020

Evaluating Weakly-Supervised Object Localization Right.






Junsuk Choe*
Sogang University

Seong Joon Oh*
NAVER

Seungho Lee
Yonsei Univ.

Sanghyuk Chun
NAVER

Zeynep Akata
University of Tübingen /
Max Plank Institute

Hyunjung Shim
Yonsei Univ.

* Equal contribution

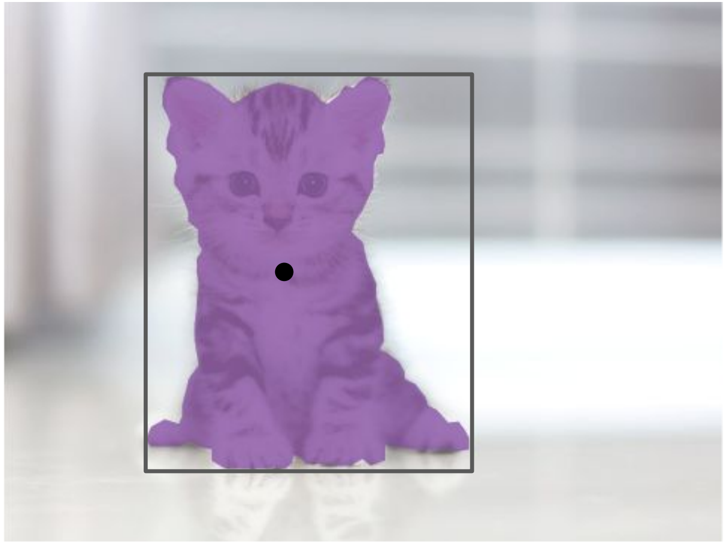
What is WSOL?

WSOL = Weak supervision + Object localization.

- What is object localization?
- What type of weak supervision?

What is object localization?

Where's the cat?



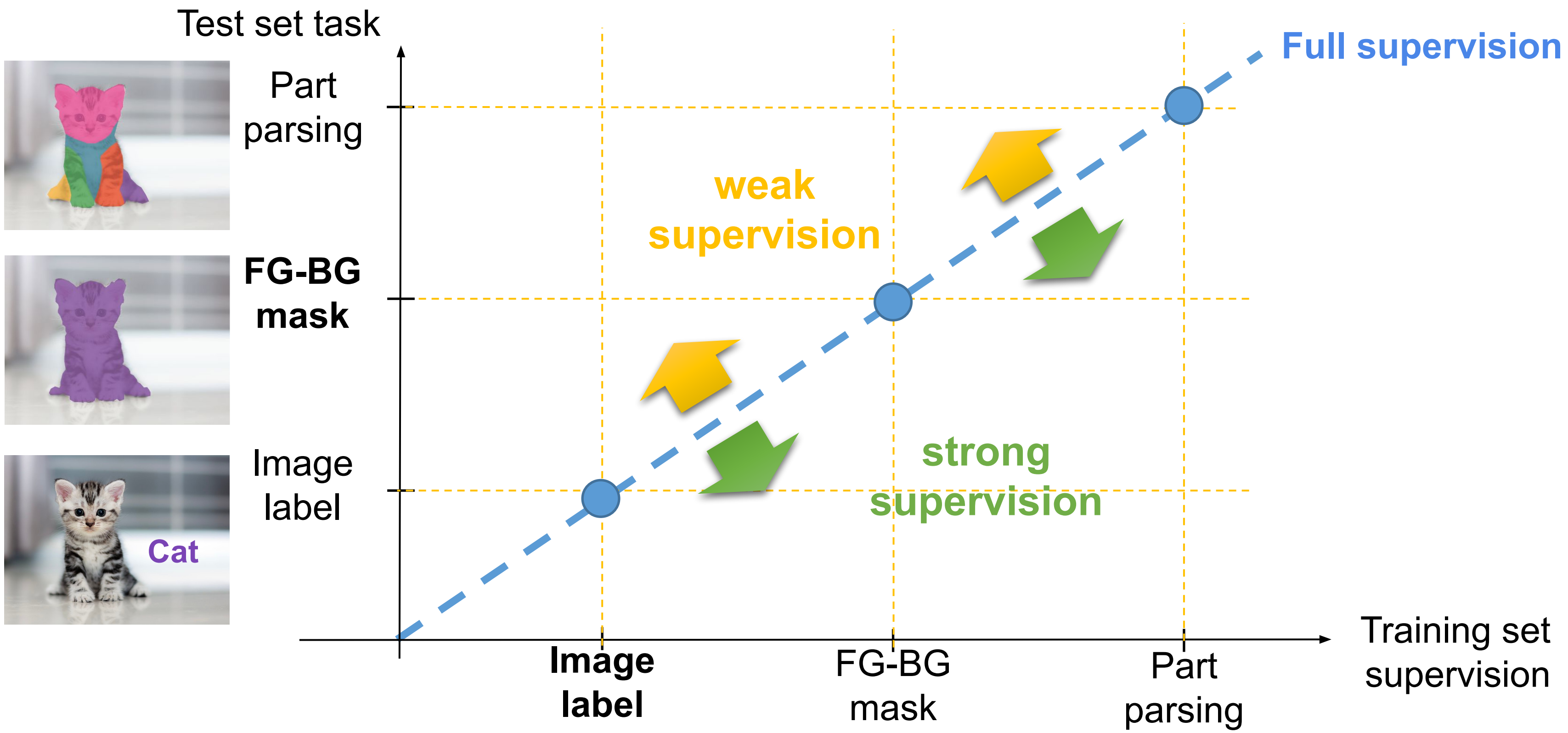
Object localization

- One class per image.
- Class is known (there's a cat).
- *Point me out where the cat is.*

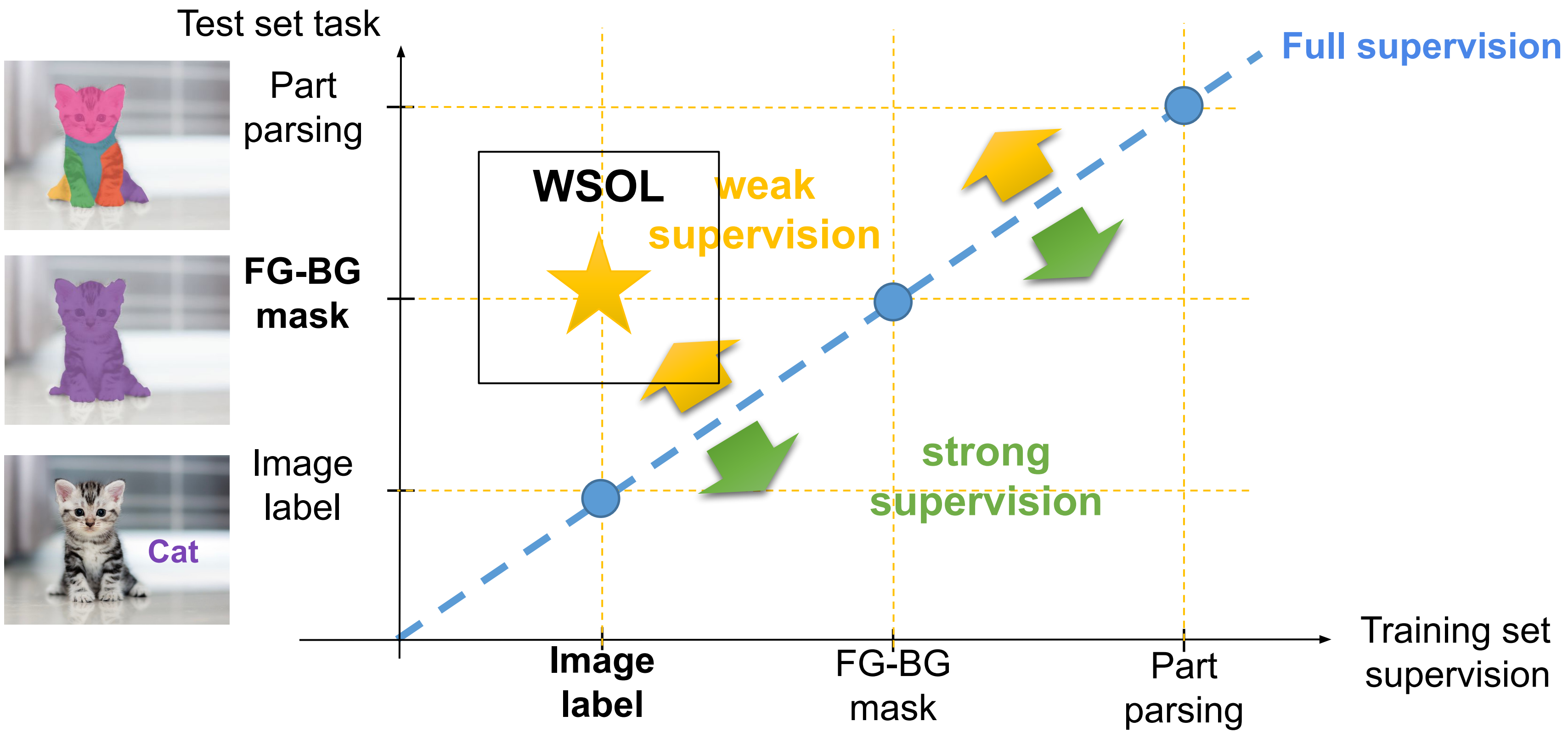
Output format:

- Point
- Box
- Mask.

What is weak supervision?



What type of weak supervision?



WSOL methods:
How are they trained & evaluated?

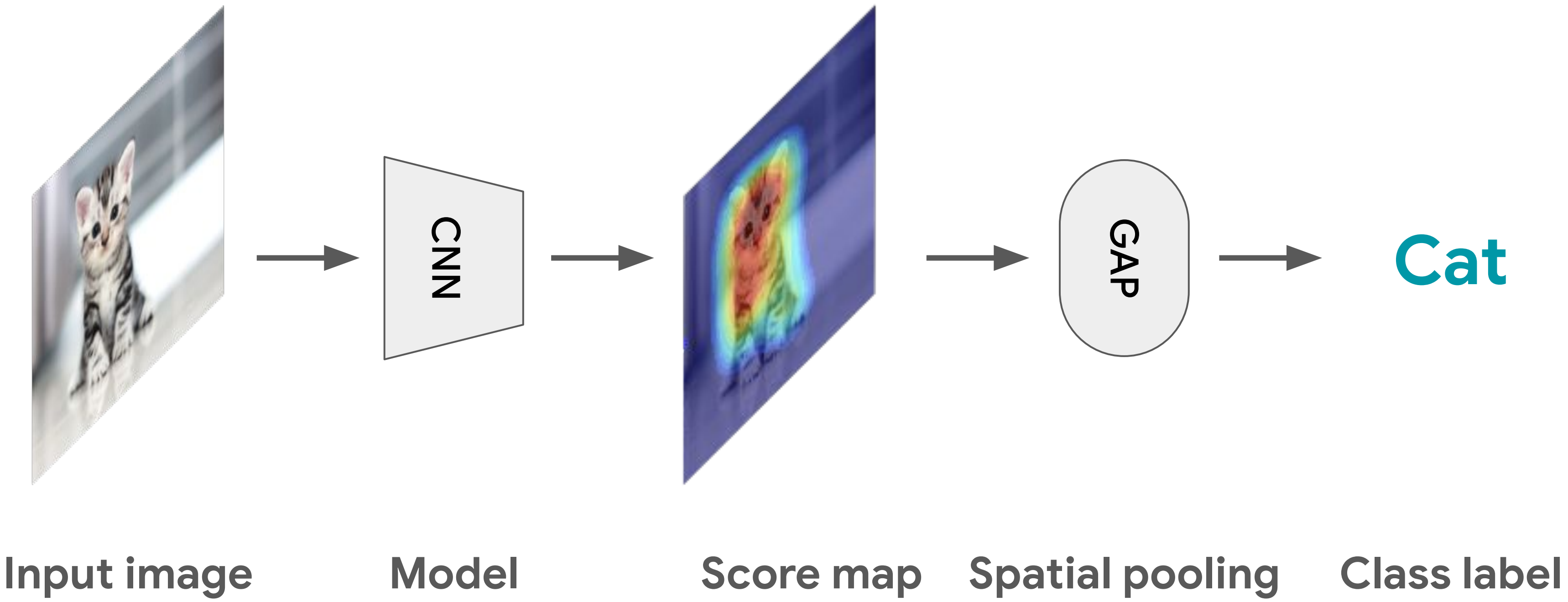
CAM as an “explanation tool” for visual models

CAM for “person” category.

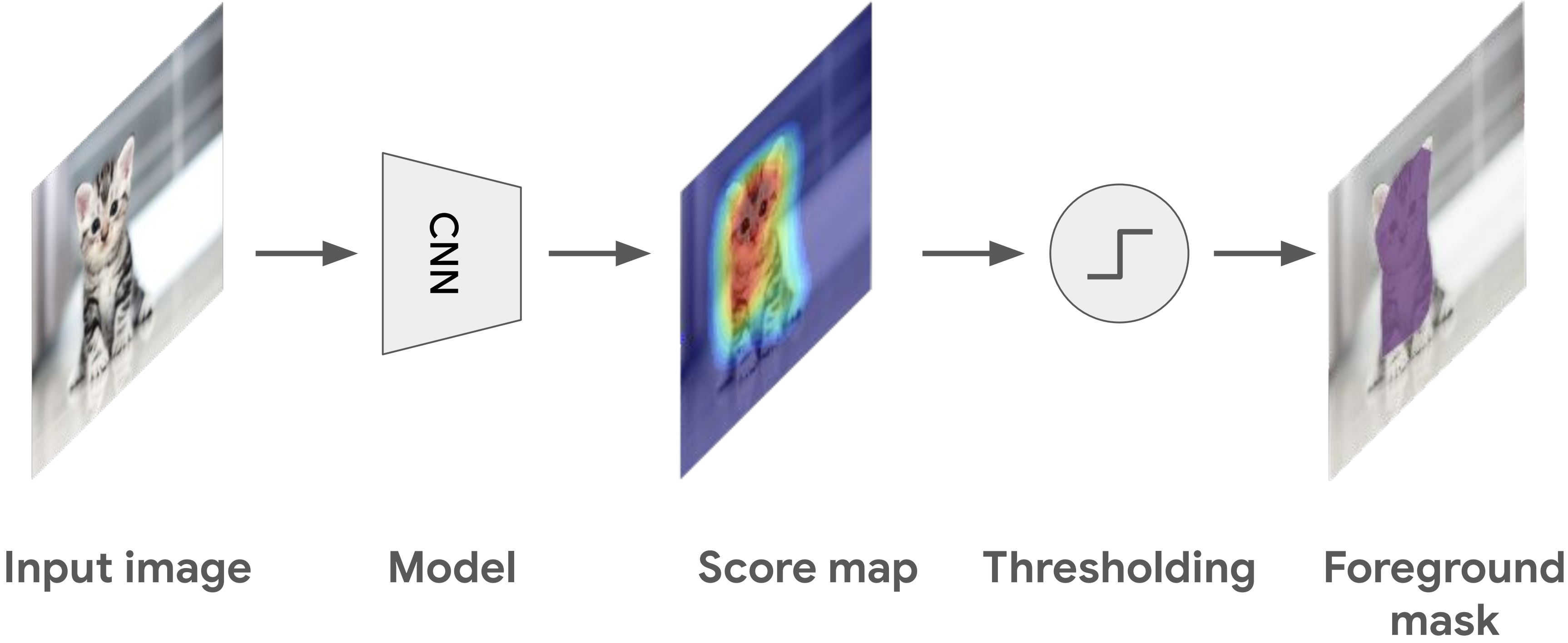
“It’s weird that the model is not attending to faces!”

→ Guidance for further regularization, data augmentation, etc.

How to train a WSOL model. CAM (CVPR'16) example

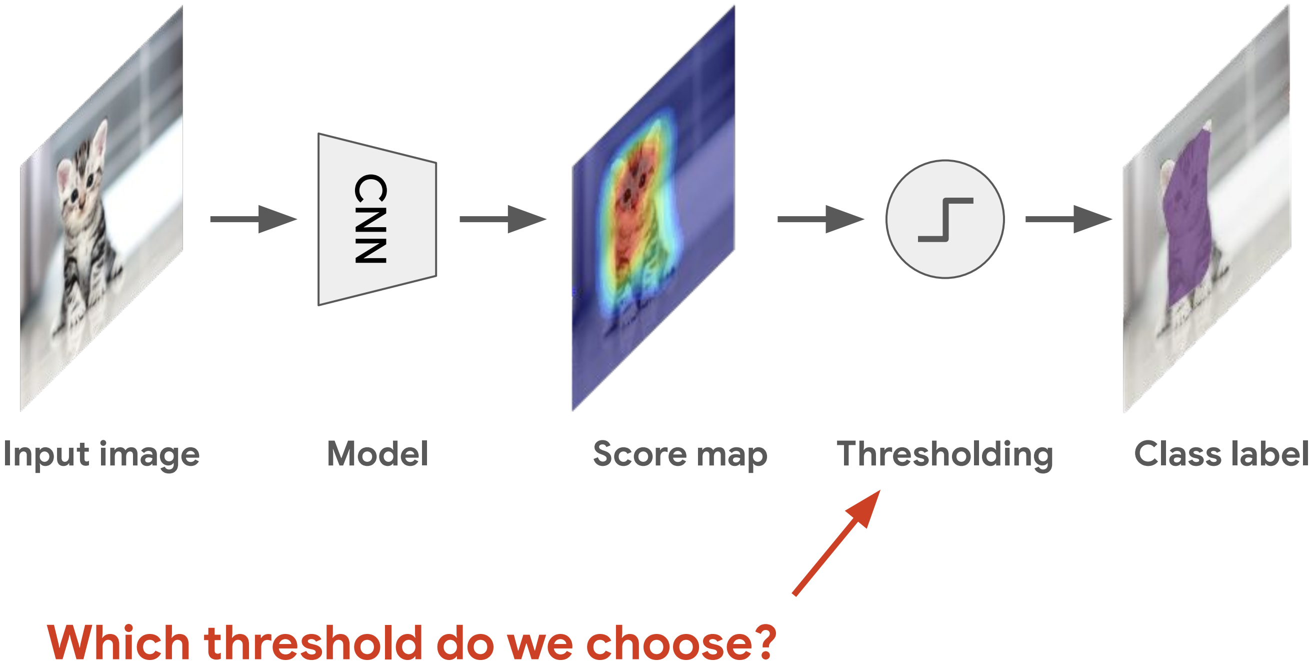


CAM at test time.

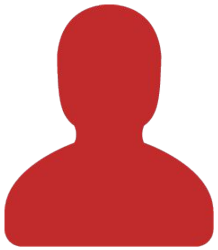


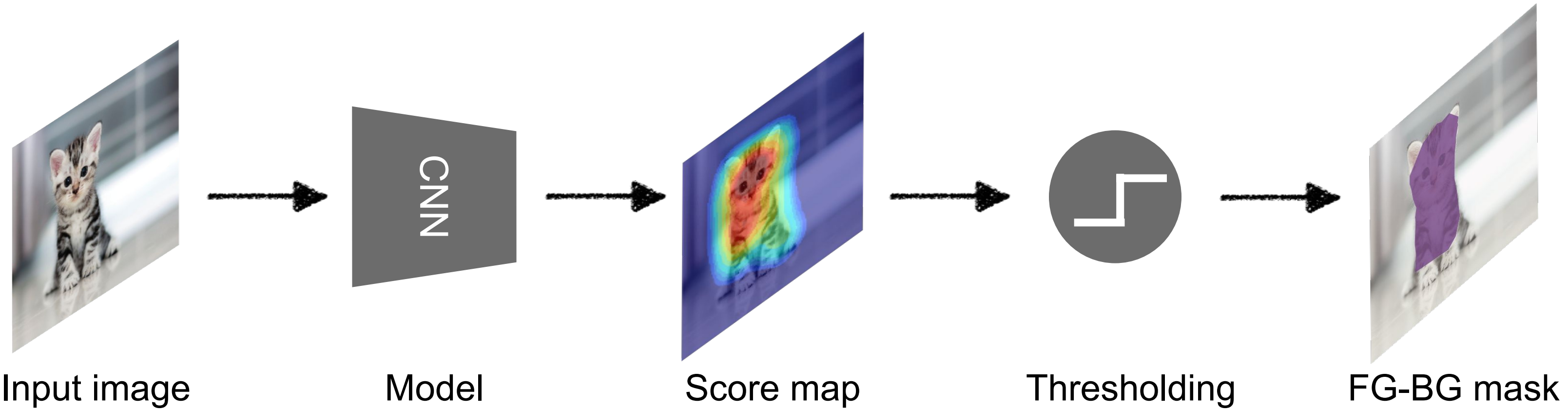
CAM does not use any full supervision, does it?

Implicit full supervision for WSOL.

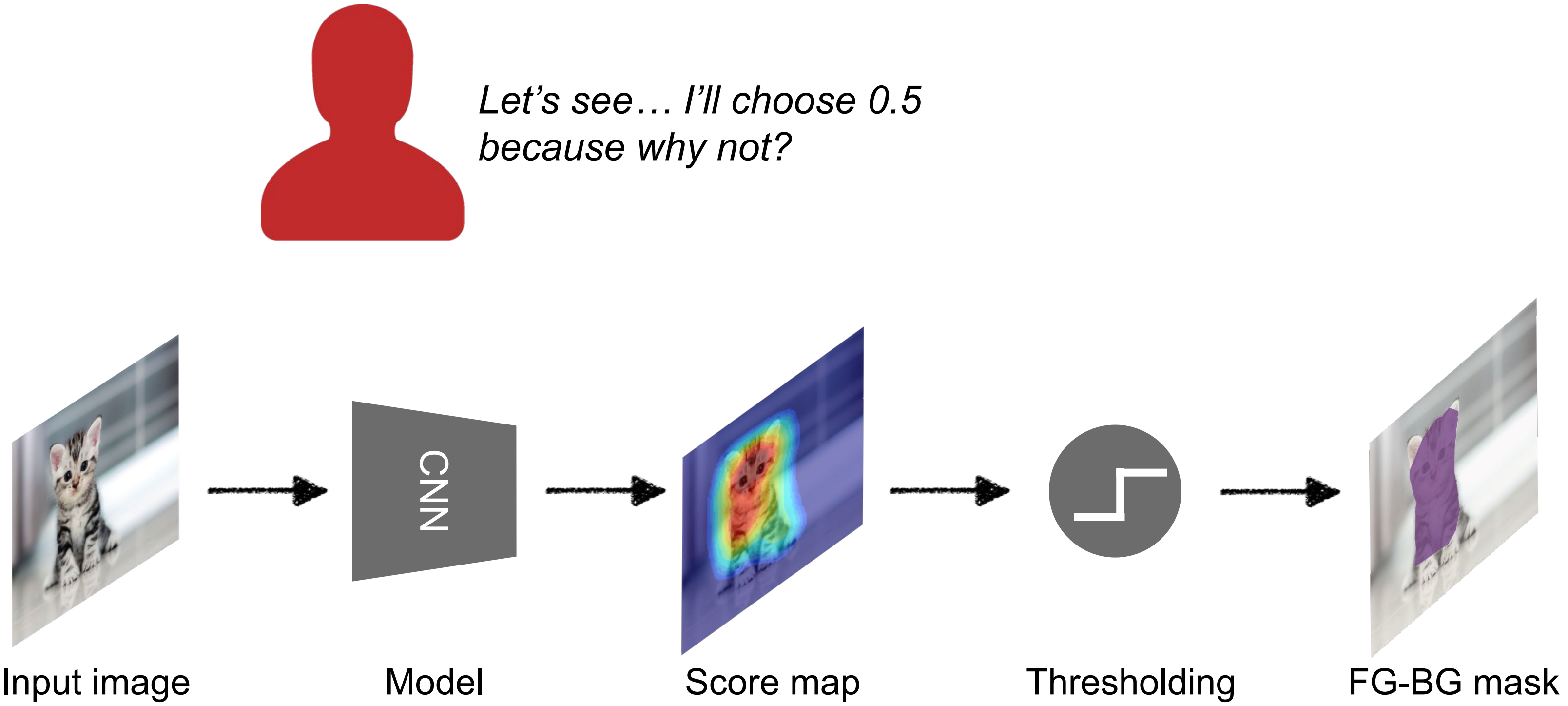


How things can go wrong (**Version 1**)

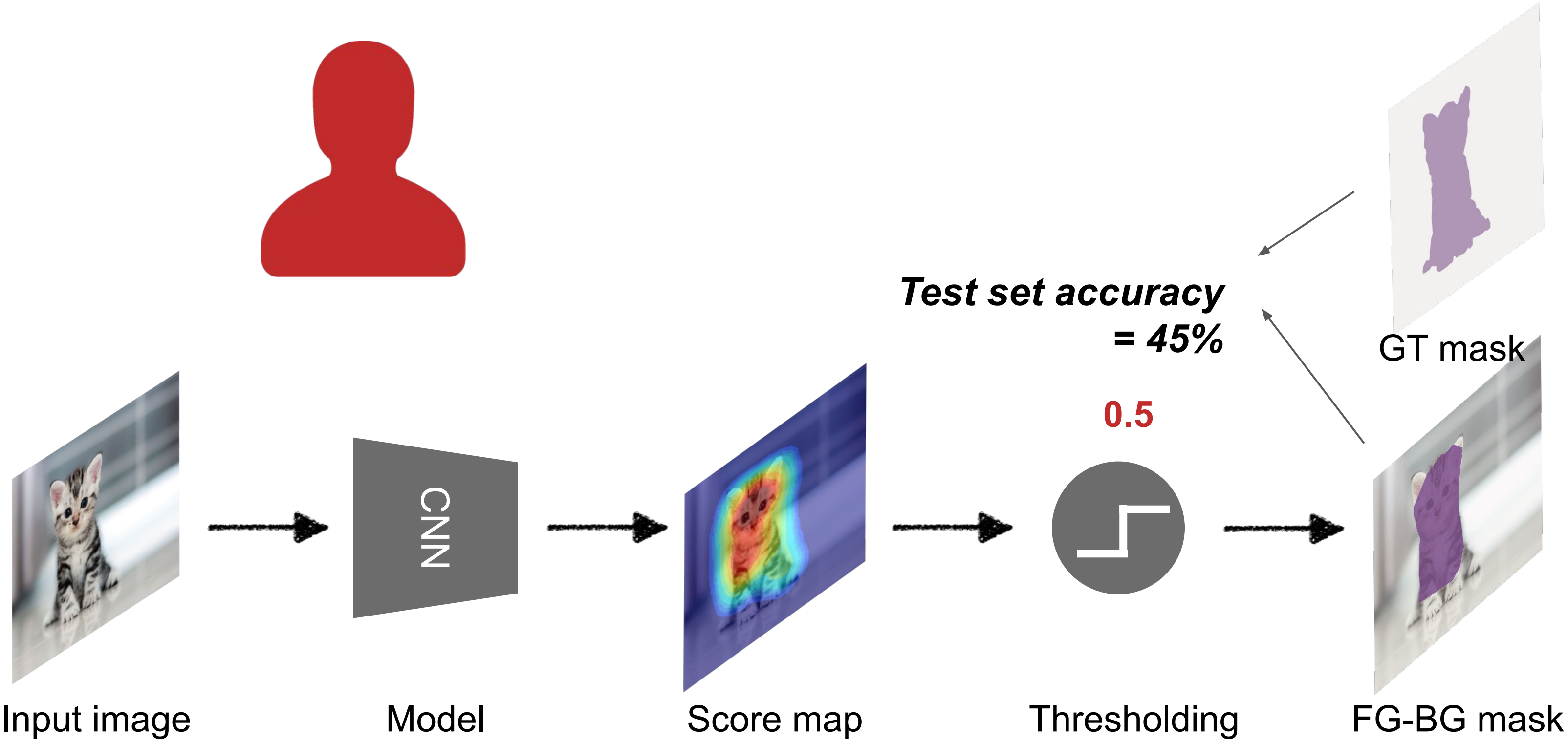




How things can go wrong (Version 1)

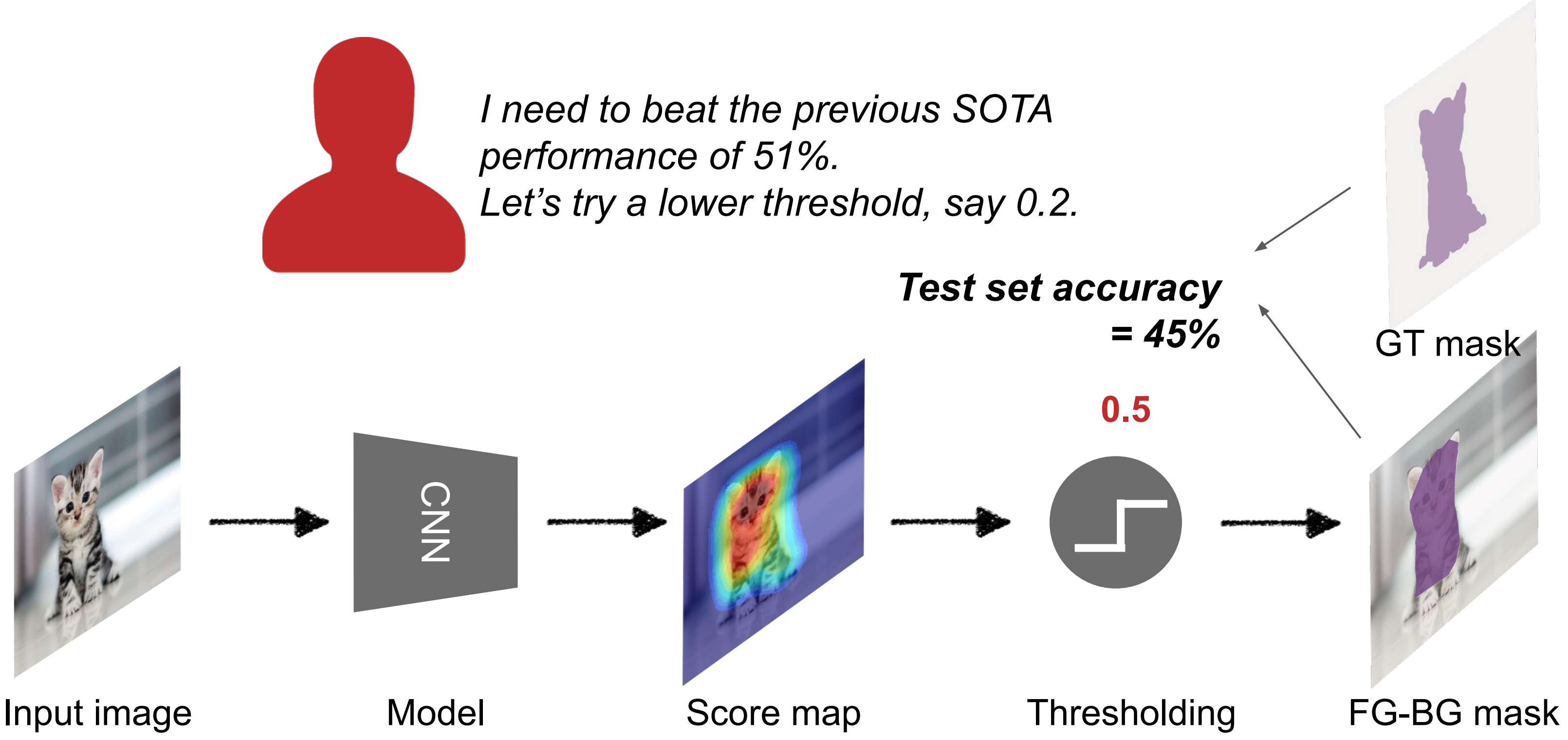


How things can go wrong (Version 1)

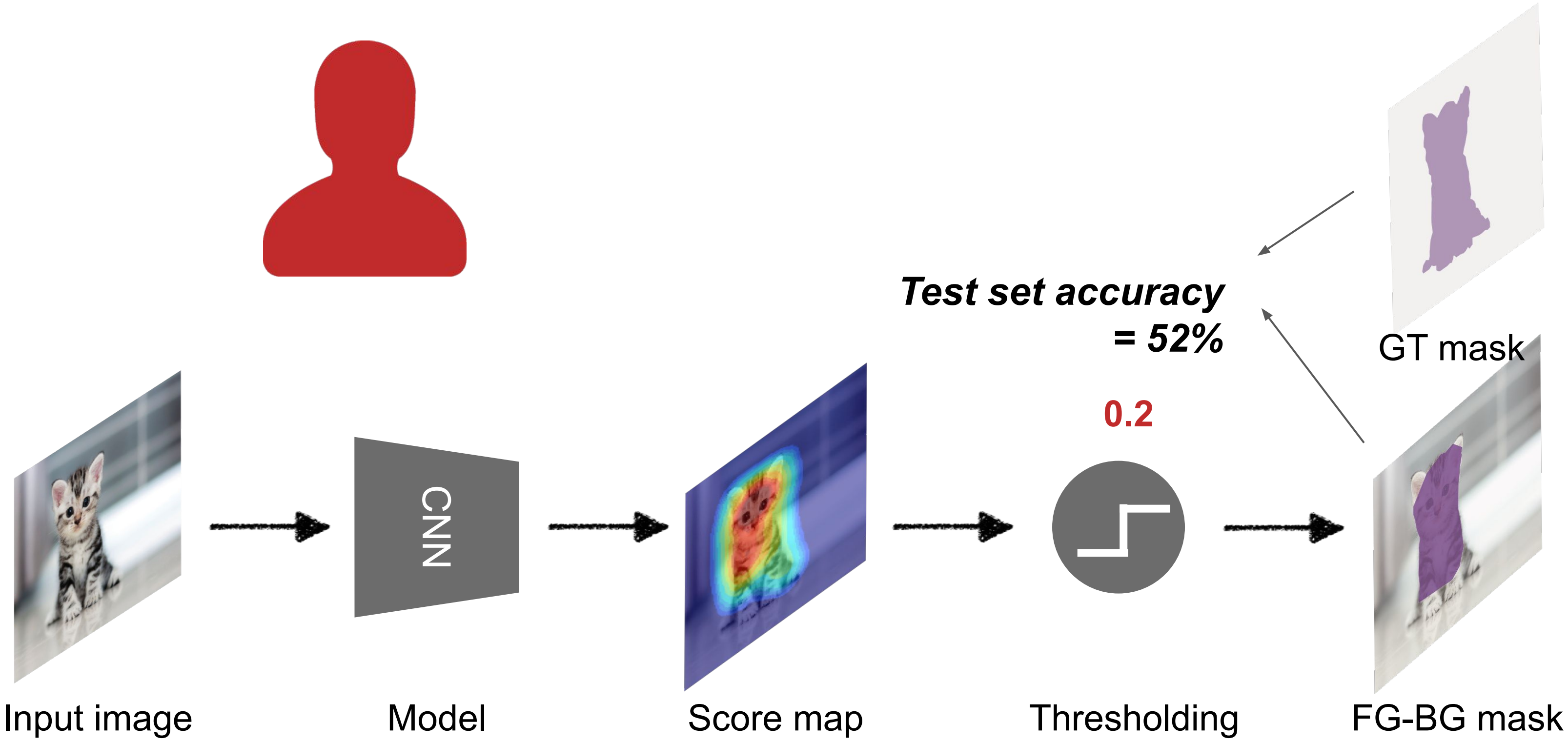


How things can go wrong (Version 1)

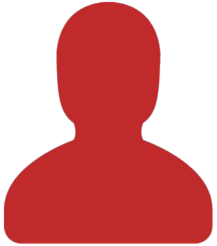
I need to beat the previous SOTA performance of 51%.
Let's try a lower threshold, say 0.2.



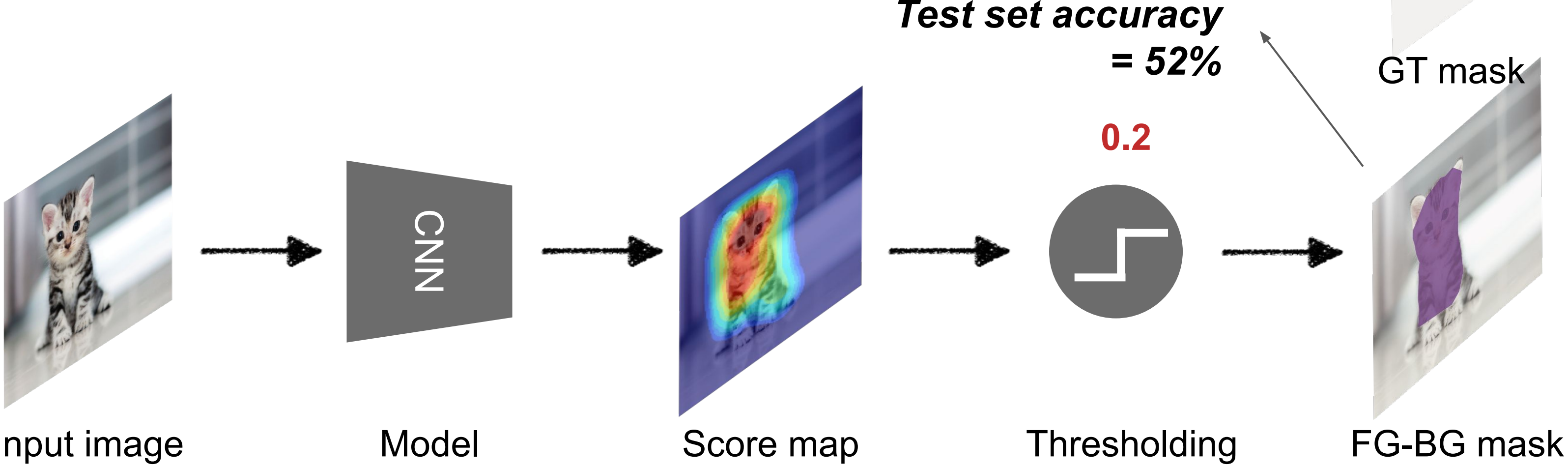
How things can go wrong (Version 1)



How things can go wrong (**Version 1**)

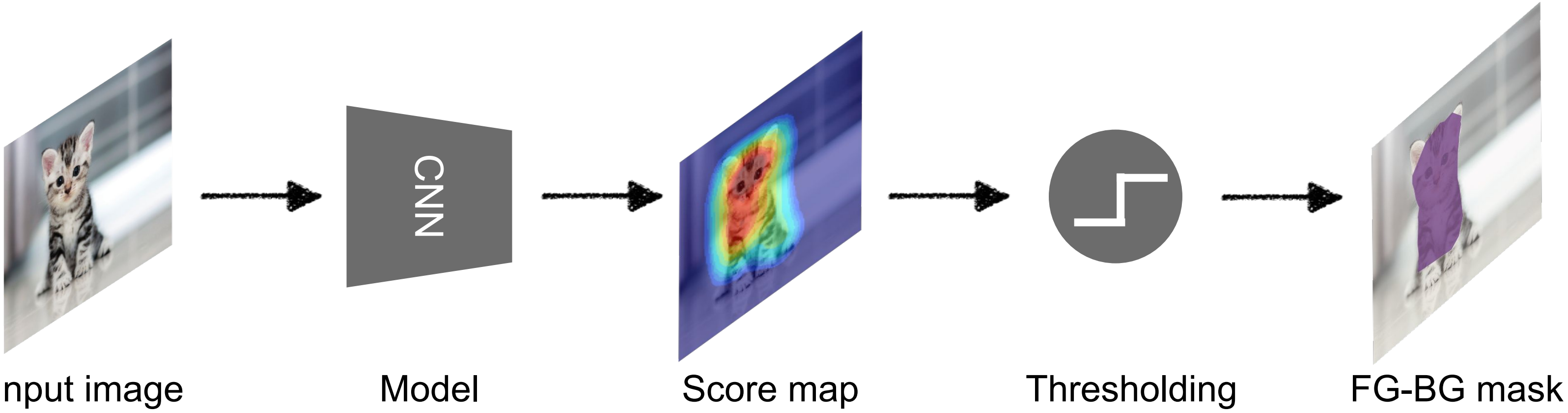


Yes! Our WSOL method is the SOTA now! Let's write a paper on this.



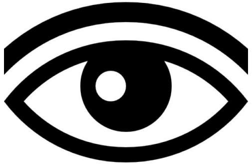
How things can go wrong (Version 2)

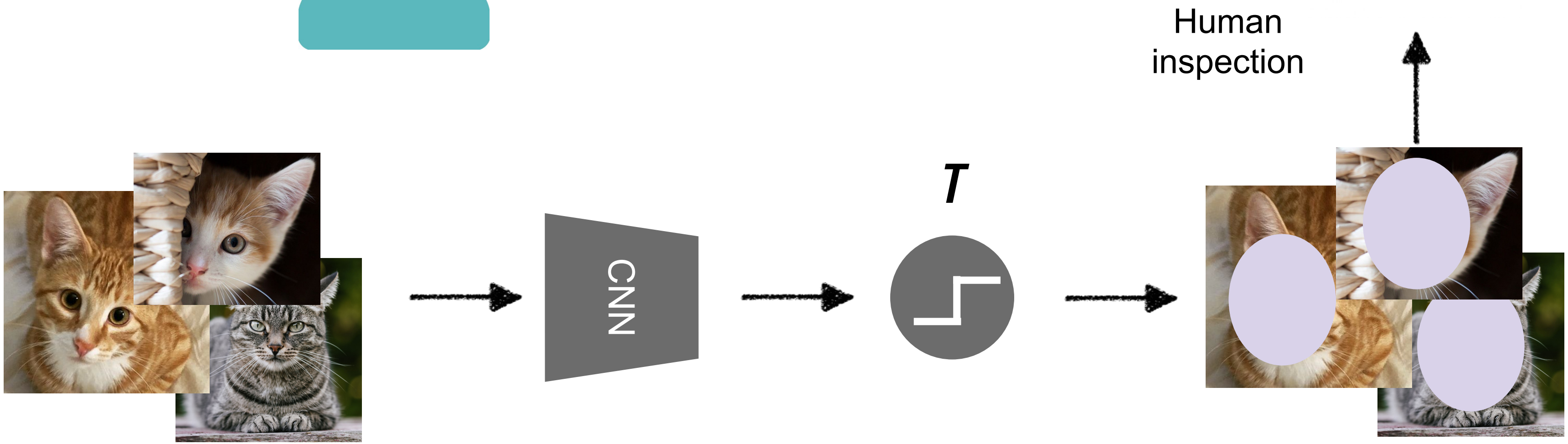
Let’s not tune the HPs on the test set.
Let’s never touch the full supervision.



How things can go wrong (Version 2)

Instead, let's inspect a few outputs in the training set.





A few training samples

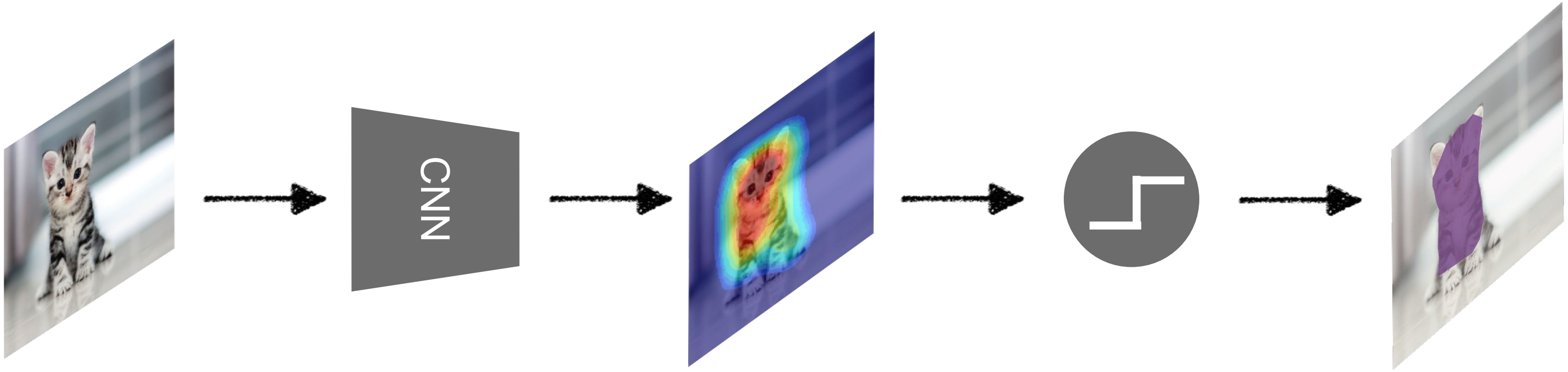
Model

Thresholding

FG-BG masks

How things can go wrong (**Version 3**)

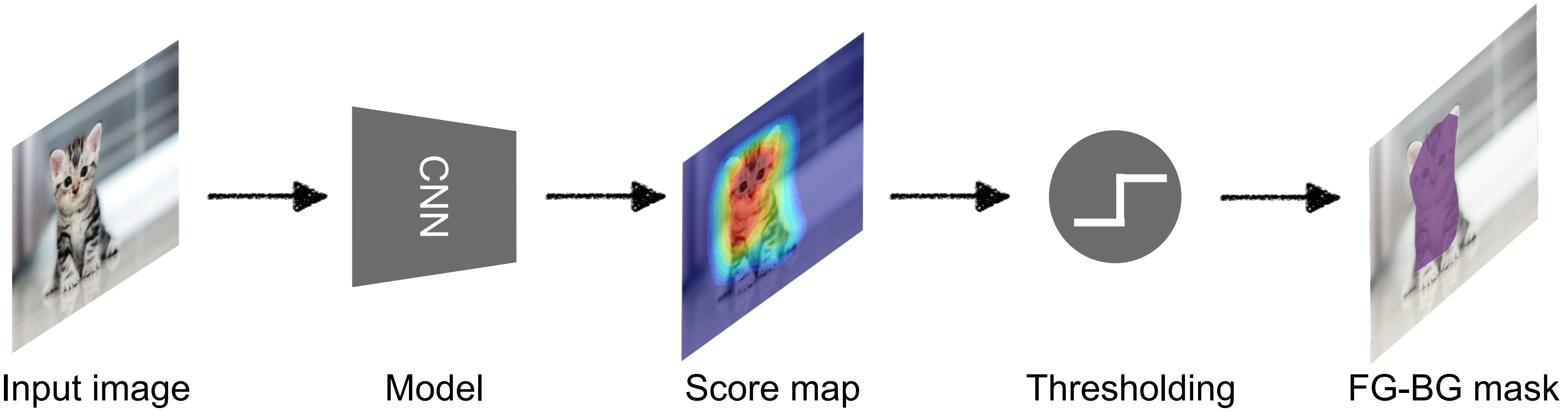
Human-in-the-loop is also violating the weak-supervision policy.



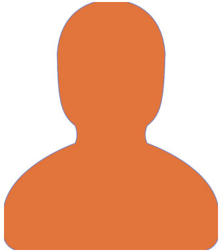
Input image Model Score map Thresholding FG-BG mask

How things can go wrong (Version 3)

We are going to adopt whatever HPs previous papers have been using.

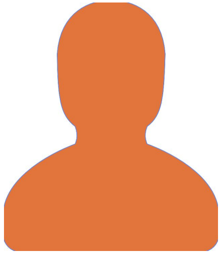


How things can go wrong (**Version 4**)



(Black magic happening)

How things can go wrong (**Version 4**)

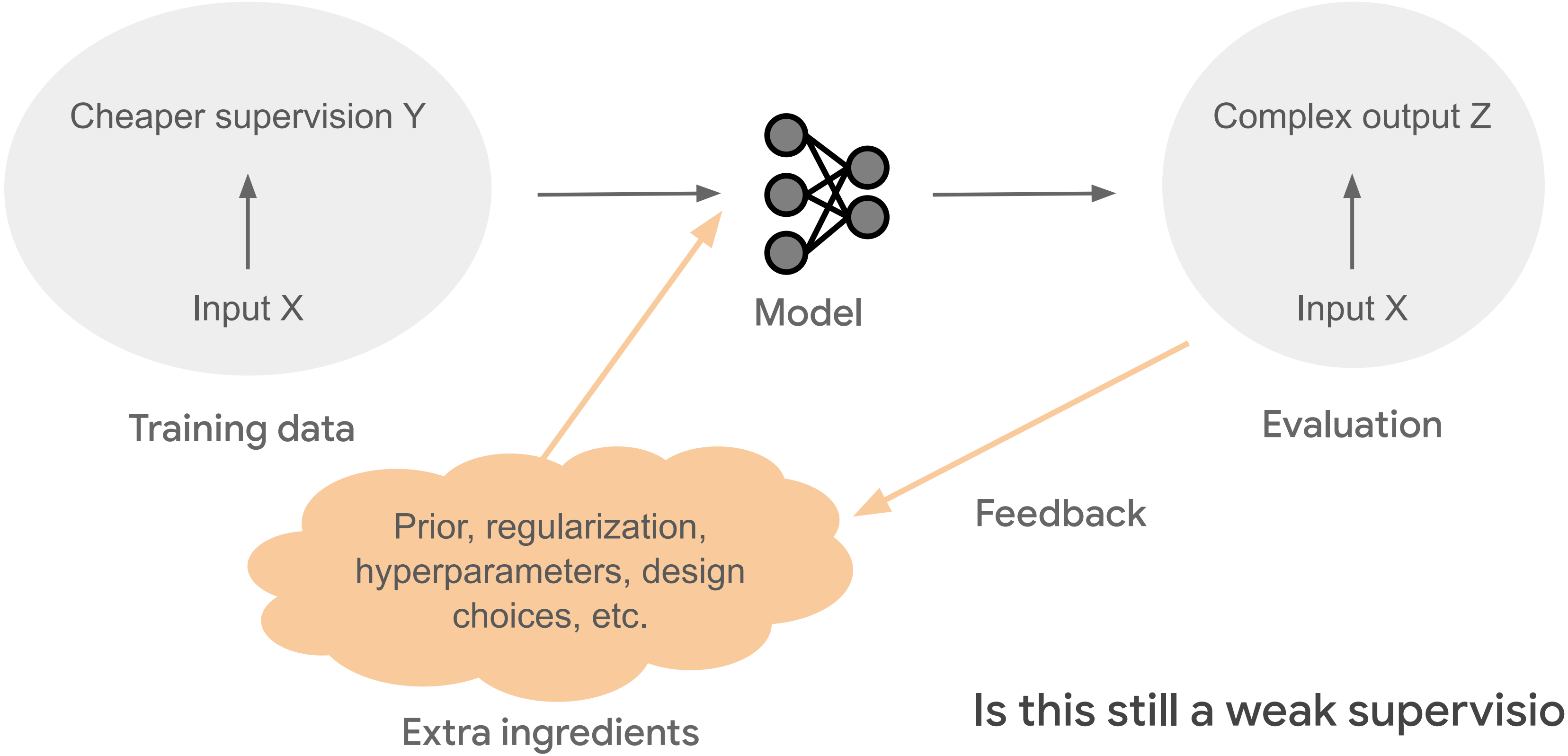


In paper:
"We use threshold 0.5 [full stop]"

How methods tune their HPs.

WSOL method	Hyperparameters	How to tune them
CAM, CVPR'16	Threshold / Learning rate / Feature map size	**(4) Black magic**
HaS, ICCV'17	Threshold / Learning rate / Feature map size / Drop rate / Drop area	**(2) Human in the loop, (3) “Not our fault”**
ACoL, CVPR'18	Threshold / Learning rate / Feature map size / Erasing threshold	**(1) Tune HP with full sup**
SPG, ECCV'18	Threshold / Learning rate / Feature map size / Threshold 1L / Threshold 1U / Threshold 2L / Threshold 2U / Threshold 3L / Threshold 3U	**(1) Tune HP with full sup**
ADL, CVPR'19	Threshold / Learning rate / Feature map size / Drop rate / Erasing threshold	**(1) Tune HP with full sup**
CutMix, ICCV'19	Threshold / Learning rate / Feature map size / Size prior / Mix rate	**(3) “Not our fault”**

Evaluating weakly-supervised models.



Is this still a weak supervision?

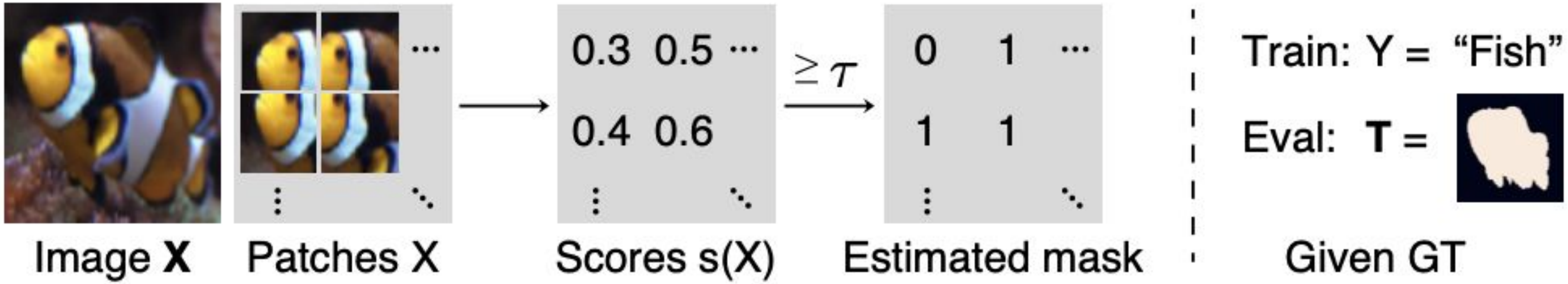
We are not trying to
blame the researchers.

We argue instead that
extra information is inevitable for WSX.

Problem formulation: **WSOL as MIL**

Interpreted as a patch classification trained with multiple instance learning (MIL).

The score map s(X) is thresholded at tau to estimate the mask **T**.



Junsuk Choe, et al. "Evaluating Weakly Supervised Object Localization Methods Right." CVPR. 2020.

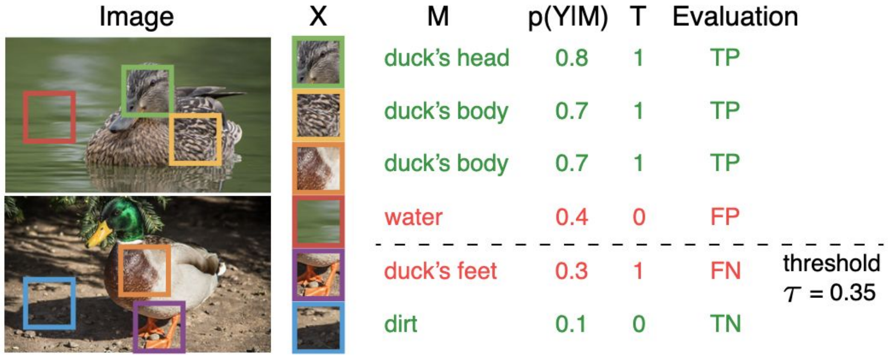
WSOL is an ill-posed problem.

Pathological case:

A class (e.g. duck) correlates better with a BG concept (e.g. water) than a FG concept (e.g. feet).

Then, WSOL is not solvable even with infinite supply of training data.

Solution: Let’s use full supervision!



The full supervision is **inevitable.**

So, let's use full supervision.

But in a controlled manner.

The four strategies then make sense!

- **Version 1: Validation on test set**
- **Version 2: Human-in-the-loop**
- **Version 3: “It’s not our fault”**
- **Version 4: Black magic**

For fair comparison, we need to let methods use

- Equal amount of extra information
- Identical HP search strategy with same amount of computational budget

Solution: Introduce the validation set!

Fair comparison with legalised full supervision.

Define the role of validation set for weakly-supervision tasks.

⟶ HP search with full supervision.

Introducing the validation set for WSOL.

Roles of train/val/test splits in textbook.

Train set	Val set	Test set
Model fitting.	Model design choices. Tuning HPs.	Report final numbers. Comparison across methods.

What about for WSOL?

Train set (Weak sup)	Val set (Full sup)	Test set (Full sup)
Model fitting, using weak supervision.	??? (no agreement on how to use it)	Report final numbers. Comparison across methods.

What about for WSOL?

Train set
(Weak sup)

“Weakly-supervised” method is supposed to use this set **ONLY**.

Val set
(Full sup)

Test set
(Full sup)

What about for WSOL?

Train set
(Weak sup)

Val set
(Full sup)

Test set
(Full sup)

Usually used for tuning HPs.

Lack of unified agreement on “how to use”.

Some methods extensively make use of val set
for HP search (e.g. grid search)

→ Unfair !

What about for WSOL?

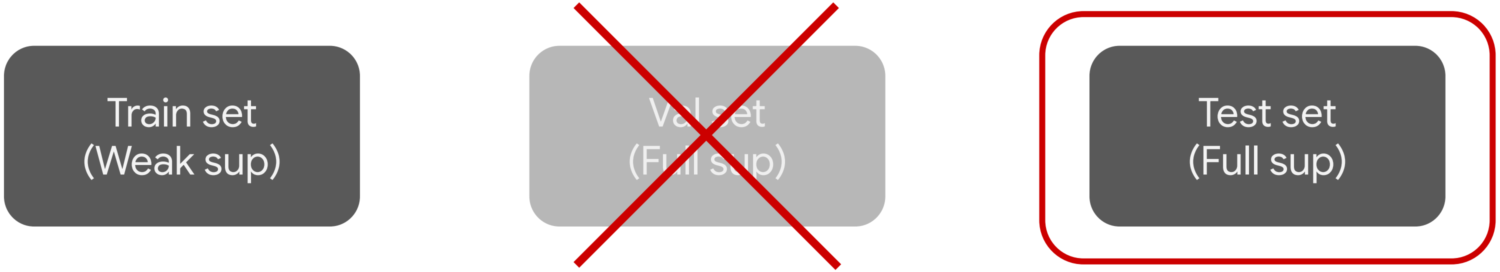
Train set
(Weak sup)

~~Val set~~
~~(Full sup)~~

Test set
(Full sup)

Even worse, there is no val set
in many WS-X benchmarks.

What about for WSOL?



And people tune their HPs
over the test set !

Existing benchmarks did not have the validation split.

Dataset	Training set (Weak sup)	Validation set (Full sup)	Test set (Full sup)
ImageNet	✔	✘ ImageNetV2[a] exists, but no full sup.	✔
CUB	✔	✘ No images, nothing.	✔

[a] Recht et al. Do ImageNet classifiers generalize to ImageNet? ICML 2019.

Our benchmark proposal.

Dataset	Training set (Weak sup)	Validation set (Full sup)	Test set (Full sup)
ImageNet (box annot.)	✓	✓ ImageNetV2 + Our annotations.	✓
CUB (box annot.)	✓	✓ Our image collections + Our annotations.	✓
OpenImages (mask annot.)	✓ Curation of OpenImages30k train set.	✓ Curation of OpenImages30k val set.	✓ Curation of OpenImages30k test set.

Fair algorithm, fair budget, fair resource.

WSOL method	Hyperparameters	How to tune them
CAM, CVPR'16	Threshold / Learning rate / Feature map size	**Version 4**
HaS, ICCV'17	Threshold / Learning rate / Feature map size / Drop rate / Drop area	**Version 2, Version 3**
ACoL, CVPR'18	Threshold / Learning rate / Feature map size / Erasing threshold	**Version 1**
SPG, ECCV'18	Threshold / Learning rate / Feature map size / Threshold 1L / Threshold 1U / Threshold 2L / Threshold 2U / Threshold 3L / Threshold 3U	**Version 1**
ADL, CVPR'19	Threshold / Learning rate / Feature map size / Drop rate / Erasing threshold	**Version 1**
CutMix, ICCV'19	Threshold / Learning rate / Feature map size / Size prior / Mix rate	**Version 3**

Previous search strategies

Fair algorithm, fair budget, fair resource.

WSOL method	Hyperparameters	How to tune them
CAM, CVPR'16	Threshold / Learning rate / Feature map size	**Random search on val set, 30 iterations**
HaS, ICCV'17	Threshold / Learning rate / Feature map size / Drop rate / Drop area	**Random search on val set, 30 iterations**
ACoL, CVPR'18	Threshold / Learning rate / Feature map size / Erasing threshold	**Random search on val set, 30 iterations**
SPG, ECCV'18	Threshold / Learning rate / Feature map size / Threshold 1L / Threshold 1U / Threshold 2L / Threshold 2U / Threshold 3L / Threshold 3U	**Random search on val set, 30 iterations**
ADL, CVPR'19	Threshold / Learning rate / Feature map size / Drop rate / Erasing threshold	**Random search on val set, 30 iterations**
CutMix, ICCV'19	Threshold / Learning rate / Feature map size / Size prior / Mix rate	**Random search on val set, 30 iterations**

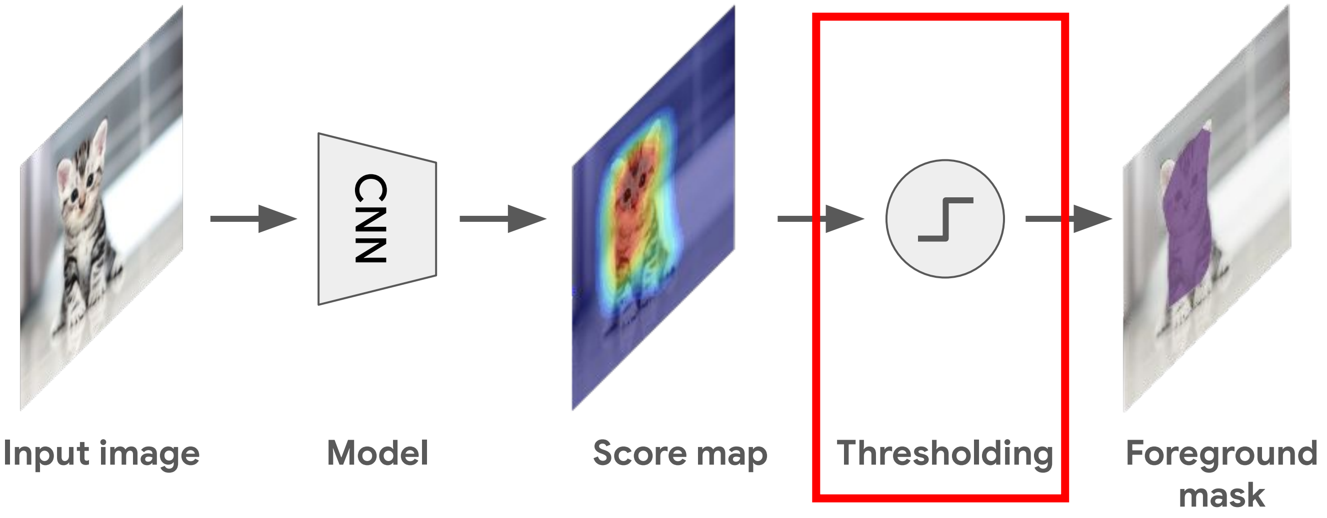
CVPR’20: Unified search algorithm

Do the validation explicitly, with the *same* algorithm.

- For each WSOL method, tune hyperparameters with
- Optimization algorithm: **Random search.**
- Search space: **Feasible** range (not "reasonable range").
- Search iteration: **30 tries.**

- **Select top-1 hyperparameter combination according to validation performance.**

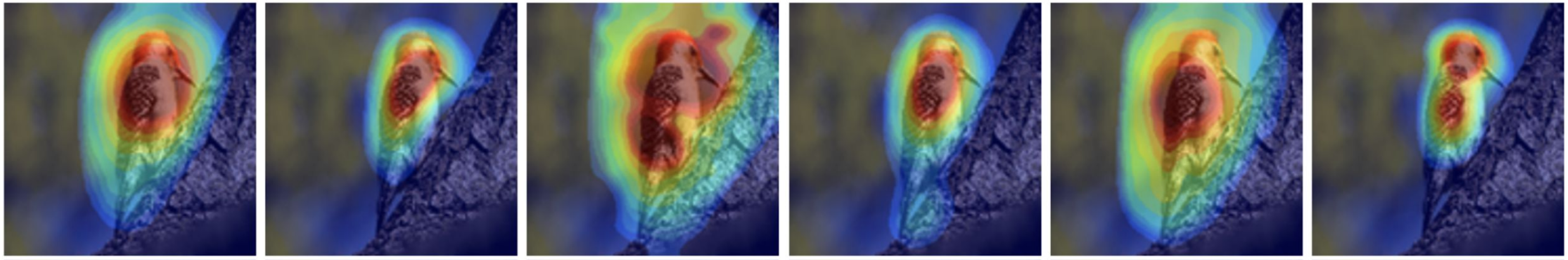
Previous treatment of the score map threshold.



- Score maps are natural outputs of WSOL methods.
- The binarizing threshold is sometimes tuned, sometimes set as a "common" value.

But setting the right threshold is critical.

CAM	HaS	ACoL	SPG	ADL	CutMix

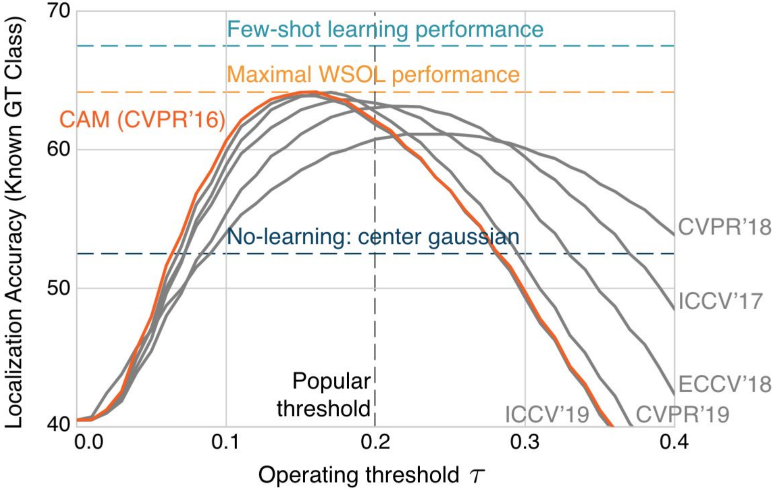


Best IoU: 0.543	Best IoU: 0.622	Best IoU: 0.612	**Best IoU: 0.797**	Best IoU: 0.623	Best IoU: 0.628
Thres: 0.50	Thres: 0.25	Thres: 0.75	**Thres: 0.20**	Thres: 0.45	Thres: 0.20

- Using a fixed threshold may be unfair!
- **We propose to use the oracle threshold for every method.**

Evaluation Crisis due to the Threshold

Metrics depend on threshold hyperparameter *tau*.



Solution: new evaluation metrics that are independent of the threshold *tau*.

We propose to remove the threshold dependence.

- **MaxBoxAcc:** For box GT, report accuracy at the best score map threshold.
    - ***Max*** performance over score map thresholds.

- **PxAP:** For mask GT, report the AUC for the pixel-wise precision-recall curve parametrized by the score map threshold.
    - ***Average*** performance over score map thresholds.

Unifying metrics, datasets, and architectures.

Reported results in existing papers

Metrics →	Top1-Loc														GT-known			
Datasets →	ImageNet								CUB						ImageNet			
Architectures →	V	I	R	A	G	N	S	M	V	I	R	G	S	M	V	I	A	G
CAM CVPR'16	42.8	-	46.3	36.3	43.6	34.5	-	41.7	37.1	43.7	49.4	41.0	42.7	43.7	-	62.7	55.0	58.7
HaS ICCV'17	-	-	-	37.7	45.5	-	-	41.9	-	-	-	-	-	44.7	-	-	58.7	60.6
ACoL CVPR'17	45.8	-	-	-	46.7	-	-	-	45.9	-	-	-	-	-	-	-	-	63.0
SPG ECCV'18	-	48.6	-	-	-	-	-	-	-	46.6	-	-	-	-	-	64.7	-	-
ADL CVPR'19	44.9	48.7	-	-	-	-	48.5	43.0	52.4	53.0	-	-	62.3	47.7	-	-	-	-
CutMix ICCV'19	43.5	-	47.3	-	-	-	-	-	-	52.5	54.8	-	-	-	-	-	-	-

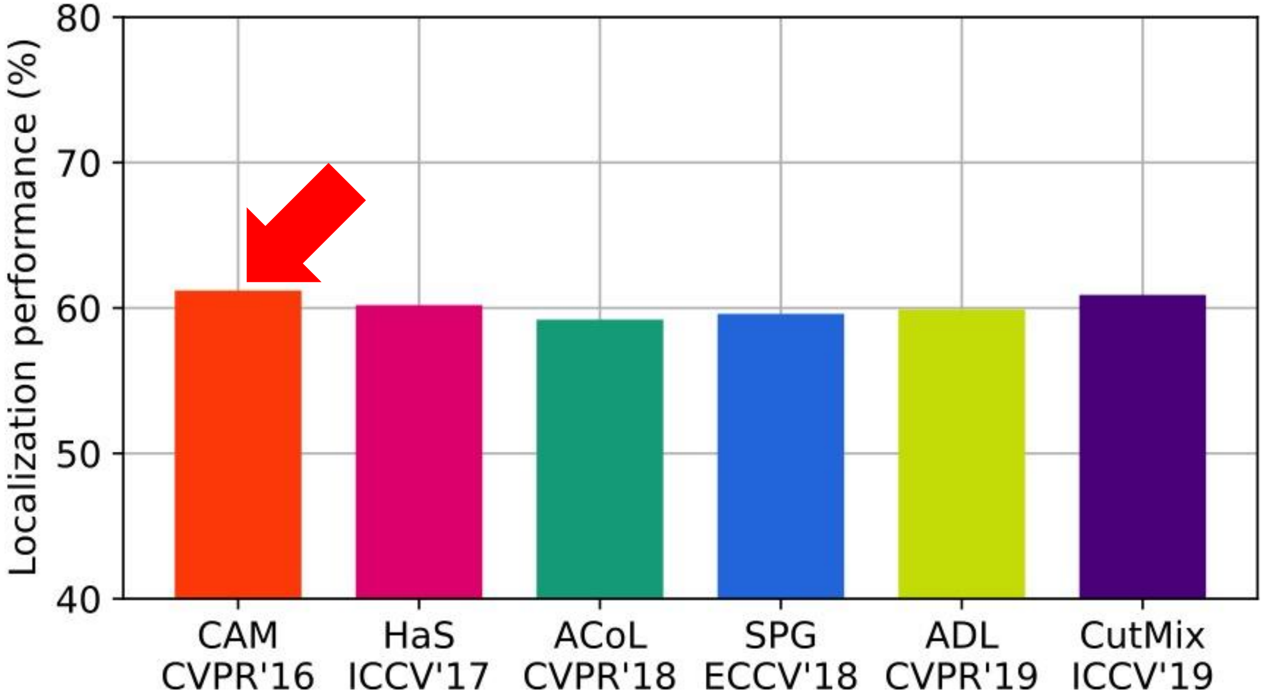
Unifying metrics, datasets, and architectures.

Coverage of our re-evaluation.

Dataset →	ImageNet (MaxBoxAccV2)				CUB (MaxBoxAccV2)				OpenImages (PxAP)				Total
Architecture →	V	I	R	Mean	V	I	R	Mean	V	I	R	Mean	Mean
CAM CVPR'16	60.0	63.4	63.7	62.4	63.7	56.7	63.0	61.1	58.3	63.2	58.5	60.0	61.2
HaS ICCV'17	60.6	63.7	63.5	62.6	63.7	53.4	64.7	60.6	58.1	58.1	55.9	57.4	60.2
ACoL CVPR'17	57.4	63.7	62.3	61.2	57.4	56.2	66.5	60.0	54.3	57.2	57.3	56.3	59.2
SPG ECCV'18	59.9	63.3	63.3	62.2	56.3	55.9	60.4	57.5	58.3	62.3	56.7	59.1	59.6
ADL CVPR'19	59.8	61.4	63.7	61.7	66.3	58.8	58.4	61.1	58.7	56.8	55.2	56.9	59.9
CutMix ICCV'19	59.4	63.9	63.3	62.2	62.3	57.5	62.8	60.8	58.1	62.5	57.7	59.4	60.9

Choe et al. Evaluation for Weakly-Supervised Object Localization: Protocol, Metrics, and Datasets. ArXiv 2020.

Simple is the best!



Choe et al. Evaluation for Weakly-Supervised Object Localization: Protocol, Metrics, and Datasets. ArXiv 2020.

ClovaAI/wsolevaluation: Open source library for end-to-end WSOL training and evaluation.

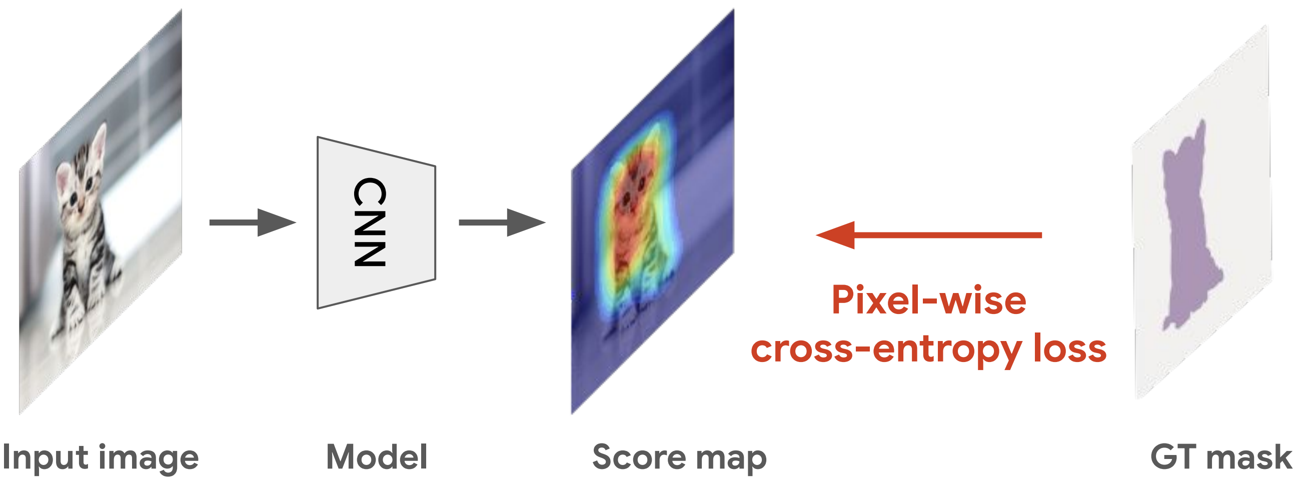
- Dataset download.
  - CUB v2, ImageNet v2, OpenImages 30k: images and annotations

- End-to-end train / evaluation for *six different WSOL methods* for *three different datasets* and *three different backbones*.
  - CAM / HaS / ACoL / SPG / ADL / CutMix
  - CUB / ImageNet / OpenImages
  - ResNet / Inception / VGG.

- https://github.com/clovaai/wsolevaluation

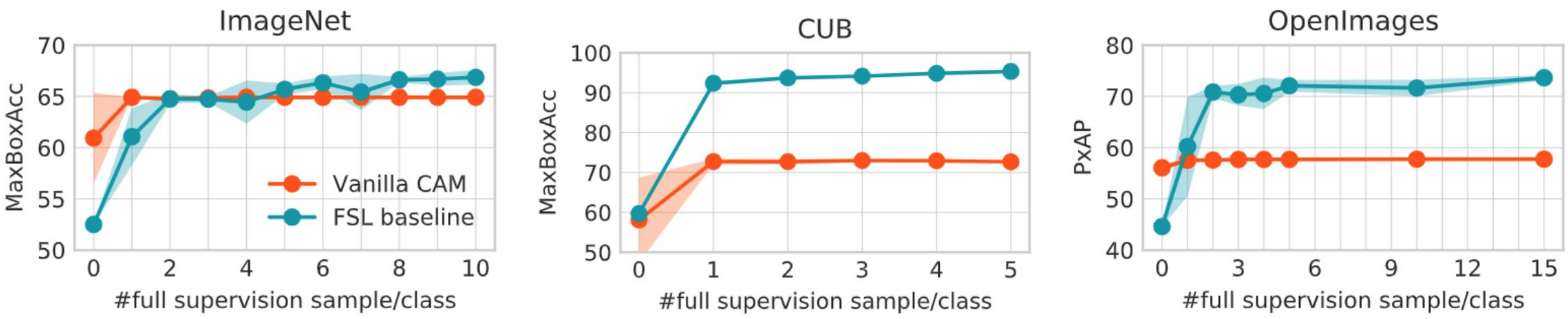
What if
the validation samples are used for model
training?

Few-shot learning baseline.



- # Validation samples: 1-5 samples/class.
- What if they are used for training the model itself?

Few-shot learning results.



- FSL > WSOL at only 2-3 full supervision / class.
- FSL is an important baseline to compare against.
- New research directions: semi-weak supervision

Conclusion and take-aways.

1. WSOL benchmarks are set up like this:

The common strategy for WSOL and other WSX methods is:

(1) introduce **many hyperparameters**.

(2) implicitly tune them with the **full-supervised samples**.

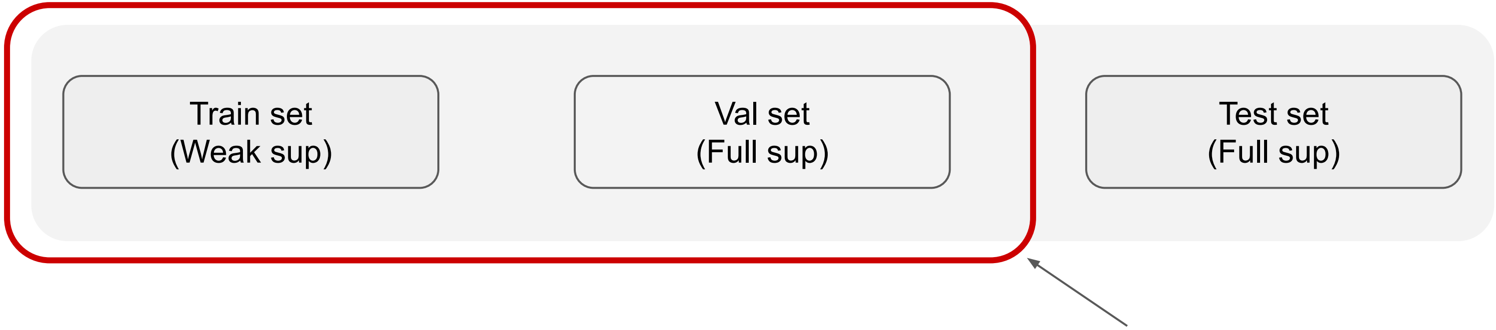
Conclusion and take-aways.

2. This is against the WSOL (and WSX) philosophy, but understandable.

WSOL and many other WSX tasks are ill-posed without extra sources of information or inductive bias.

Conclusion and take-aways.

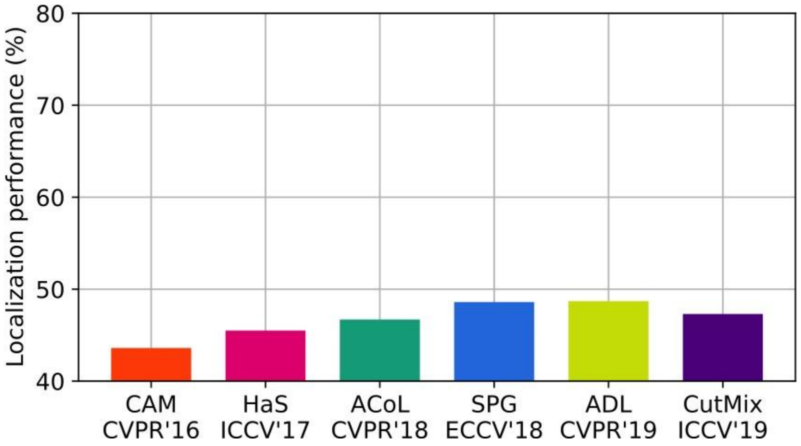
3. Let's legalise the use of full supervision (called "val set").



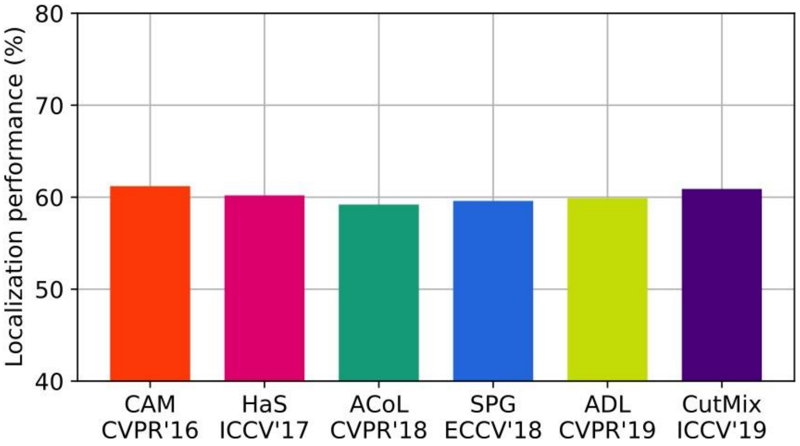
Same amount of full sup ensured for every method.

Conclusion and take-aways.

4. WSX methods can then be compared on the level ground.



Before evaluation clean-up.



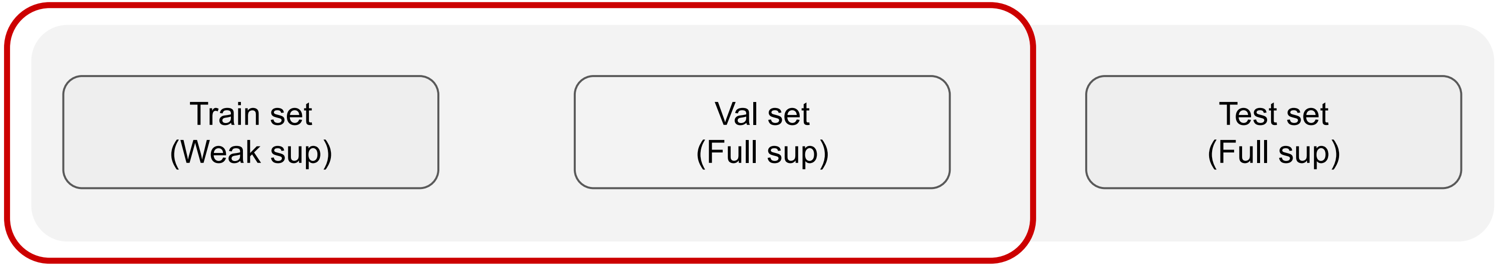
After evaluation clean-up.

Conclusion and take-aways.

5. “Val set” doesn’t need to be used for validation.

This opens up the new phase for WSX research:

- Hybrid weakly-supervised X.



Users can use val set for model fitting as well.

Implication: New phase of WSOL and WSX research.

Acknowledging the need for extra information opens up new research questions:

- **How to make best use of full supervision?**

  Validation? Model fitting? Or something else?

- **How to exploit existing datasets with diverse supervision types?**

  How to combine multi-modal supervision types?

  OpenImages, COCO, Pascal, ImageNet, Flickr, ...

- **Okay we need extra information - but can we minimise it?**

  Maybe under a constraint on the minimal required performance?

Future direction : Hybrid weakly-supervised X.

Hybrid-weakly-supervised X Hoffman et al. CVPR’15, Tang et al. CVPR’16

- Combination of different levels and amounts of supervision.

Why relevant?

- Abundance of well-curated and raw data on the web with different levels of supervision. OpenImages, COCO, Pascal, ImageNet, YFCC, Web crawl, …

Some non-trivial research questions:

- Setting up the benchmarks.
- Combining multiple supervision modalities.