HOW TO WRITE A GOOD REVIEW

As a *newcomer* in the field

OUTLINE

BASICS & MINDBLOCKS

First things first

Yes, *you* can do it!

THREE PILLARS OF REVIEWING

Critical Thinking, Constructive Input, and Empathy

AUDIENCE

Authors, Area Chairs and Fellow Reviewers

REVIEWING

Claims, Literature, Strengths & Weaknesses, Results

MAKING A DECISION

Ordering, Weighing, Strong Decisions & Borderline

GOOD PRACTICES

Practicals, Timing, Limits, Discussions

BASICS & MINDBLOCKS

First things first

- Domain conflicts
- Entering, updating your papers
- Bidding
- Review assignment
- Read the instructions
- Ethical guidelines
- Opting out, but early enough!

Yes, *you* can do it!

- Just put the effort and time.
- Shared responsibility
  other reviewers and the area chair
- What would your advisor say about it?
- Ask for help, learn by example
- Think of it as detailed, critical
  proof-reading for a *friend*
- It gets easier.

THREE PILLARS OF REVIEWING

Critical Thinking, Constructive Input, and Empathy

THREE PILLARS OF REVIEWING

Equally Important Principles

CRITICAL THINKING

Can it be done better?

- Questioning, doubting
- Judgemental
- Clear, precise
- For the community

CONSTRUCTIVE INPUT

How to do it?

- Corresponding suggestions
- What is missing?
- Like & dislike together
- Specific, detailed

EMPATHY

How to say it?

- What if it was you?
- Friend PoV
- Respect the work!
- Language
- Reasonable

THREE PILLARS OF REVIEWING

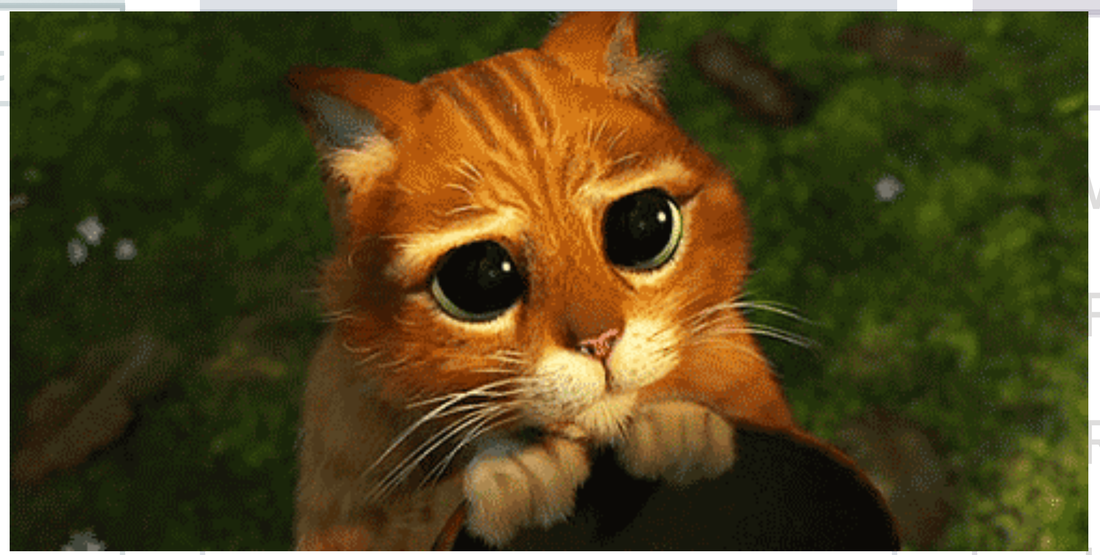
Equally Important Principles

CRITICAL THINKING

Can it be done bett

- Questioning, doubting
- Judgemental
- Clear, precise
- For the community

CONSTRUCTIVE INPUT



together

- Specific, detailed

EMPATHY

How to say it?

What if it was you?

Friend PoV

Respect the work!

- Language
- Reasonable

AUDIENCE

Authors, Area Chairs and Fellow Reviewers

AUDIENCE-I

Authors

- Suggestions or additional comments
  - to make the paper better
  - not serious enough to be a weakness
    not a reason to reject
  - additional experiments, references, pointers for future work
  - as detailed as possible
    e.g. format, typos, variable names, ordering
- Thoughtful, constructive feedback

AUDIENCE-II

Area Chairs (AC) and Fellow Reviewers

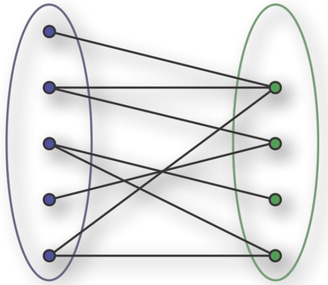
- Summary
  - problem, main contribution(s), approach, dataset(s) and experiments
  - objective
- Final recommendation
  - summary of the main points in the review
  - subjective
- Concise, organized feedback items, weighting of the items

REVIEWING

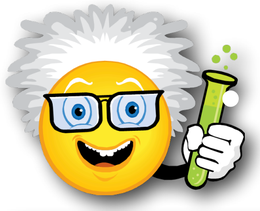
Claims, Literature, Strengths & Weaknesses, Results

CLAIM CHECK

Matching Claims to the Results



Actively look for claims!



Find the corresponding results.

- Usually can be found at the end of Intro:

with more diverse layout and appearance structures. Thus, our goal is to formulate the entire view synthesis pipeline as the inference procedure of a convolutional neural network, so that by training the network on large-scale video data for the 'meta'-task of view synthesis the network is forced to learn about intermediate tasks of depth and camera pose estimation in order to come up with a consistent explanation of the visual world. Empirical evaluation on the KITTI [15] benchmark demonstrates the effectiveness of our approach on both single-view depth and camera pose estimation. Our code will be made available at `https:`

Method	Dataset	Supervision		Error metric				Accuracy metric		
		Depth	Pose	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	K	✓		0.403	5.530	8.709	0.403	0.593	0.776	0.878
Eigen *et al.* [7] Coarse	K	✓		0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al.* [7] Fine	K	✓		0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [32]	K	✓		0.202	1.614	6.523	0.275	0.678	0.895	0.965
Godard *et al.* [16]	K		✓	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard *et al.* [16]	CS + K		✓	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Ours (w/o explainability)	K			0.221	2.226	7.527	0.294	0.676	0.885	0.954
Ours	K			0.208	1.768	6.856	0.283	0.678	0.885	0.957
Ours	CS			0.267	2.686	7.580	0.334	0.577	0.840	0.937
Ours	CS + K			0.198	1.836	6.565	0.275	0.718	0.901	0.960
Garg *et al.* [14] cap 50m	K		✓	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Ours (w/o explainability) cap 50m	K			0.208	1.551	5.452	0.273	0.695	0.900	0.964
Ours cap 50m	K			0.201	1.391	5.181	0.264	0.696	0.900	0.966
Ours cap 50m	CS			0.260	2.232	6.148	0.321	0.590	0.852	0.945
Ours cap 50m	CS + K			0.190	1.436	4.975	0.258	0.735	0.915	0.968

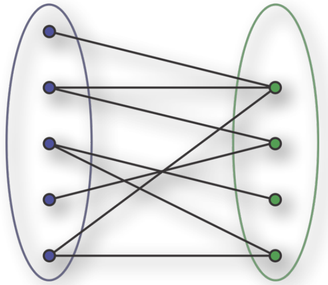
Table 1. Single-view depth results on the KITTI dataset [15] using the split of Eigen *et al.* [7] (Baseline numbers taken from [16]). For training, K = KITTI, and CS = Cityscapes [5]. All methods we compare with use some form of supervision (either ground-truth depth or calibrated camera pose) during training. Note: results from Garg et al. [14] are capped at 50m depth, so we break these out separately in the lower part of the table.

Method	**Seq.** 09	**Seq.** 10
ORB-SLAM (full)	$\mathbf{0.014 \pm 0.008}$	$\mathbf{0.012 \pm 0.011}$
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Mean Odom.	0.032 ± 0.026	0.028 ± 0.023
Ours	$\mathbf{0.021 \pm 0.017}$	$\mathbf{0.020 \pm 0.015}$

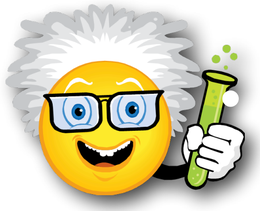
Table 3. Absolute Trajectory Error (ATE) on the KITTI odometry split averaged over all 5-frame snippets (lower is better). Our method outperforms baselines with the same input setting, but falls short of `ORB-SLAM (full)` that uses strictly more data.

CLAIM CHECK

Matching Claims to the Results



Actively look for claims!



Find the corresponding results.

- Usually can be found at the end of Intro
- Sometimes in Related Work:

tion [2, 25, 63, 38, 40]. Our approach is most related to the methods of Wang et al. [73, 74] and Pathak et al. [47], which use off-the-shelf tools for tracking and optical flow respectively, to provide supervisory signal for training. However, representations learned in this way are inherently limited by the power of these off-the-shelf tools as well as their failure modes. We address this issue by learning the representation and the tracker jointly, and find the two learning problems

model	Supervised	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Identity		22.1	23.6
Random Weights (ResNet-50)		12.4	12.5
Optical Flow (FlowNet2) [22]		26.7	25.2
SIFT Flow [39]		33.0	35.0
Transitive Inv. [74]		32.0	26.8
DeepCluster [8]		37.5	33.2
Video Colorization [69]		34.6	32.7
Ours (ResNet-18)		40.1	38.3
Ours (ResNet-50)		**41.9**	**39.4**
ImageNet (ResNet-50) [18]	✓	50.3	49.0
Fully Supervised [81, 7]	✓	55.1	62.1

Table 1: Evaluation on instance mask propagation on DAVIS-2017 [48]. We follow the standard metric on region similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$.

model	Supervised	PCK@.1	PCK@.2
Identity		43.1	64.5
Optical Flow (FlowNet2) [22]		45.2	62.9
SIFT Flow [39]		49.0	68.6
Transitive Inv. [74]		43.9	67.0
DeepCluster [8]		43.2	66.9
Video Colorization [69]		45.2	69.6
Ours (ResNet-18)		57.3	78.1
Ours (ResNet-50)		**57.7**	**78.5**
ImageNet (ResNet-50) [18]	✓	58.4	78.4
Fully Supervised [59]	✓	68.7	92.1

Table 2: Evaluation on pose propagation on JHMDB [26]. We report the PCK in different thresholds.

LITERATURE

How deep into the rabbit hole?

"What list gets longer and longer the more you read it?"

- ☑ All the relevant papers cited
- ☑ Cited *properly*
  both the method **and** the results
- ☑ *In comparison to*
  the proposed method
- ☑ Structured

Tracking as a Graph Problem. Data association can be done on a frame-by-frame basis for online applications [10, 22, 58] or track-by-track [7]. For video analysis tasks that can be done offline, batch methods are preferred since they are more robust to occlusions. The standard way to model data association is by using a graph, where each detection is a node, and edges indicates possible link among them. The data association can then be formulated as maximum flow [8] or, equivalently, minimum cost problem with either fixed costs based on distance [32, 59, 86], including motion models [48], or learned costs [45]. Both formulations can be solved optimally and efficiently. Alternative formulations typically lead to more involved optimization problems, including minimum cliques [85], general-purpose solvers, e.g., multi-cuts [75]. A recent trend is to design ever more complex models which include other vision input such as reconstruction for multi-camera sequences [49, 79], activity recognition [17], segmentation [55], keypoint trajectories [15] or joint detection [75].

Learning in Tracking. It is no secret that neural networks are now dominating the state-of-the-art in many vision tasks since [41] showed their potential for image classification. The trend has also arrived in the tracking community, where learning has been used primarily to learn a mapping from image to optimal costs for the aforementioned graph algorithms. The authors of [42] use a siamese network to directly learn the costs between a pair of detections, while a mixture of CNNs and recurrent neural networks (RNN) is used for the same purpose in [64]. More evolved quadruplet networks [70] or attention networks [89] have lead to improved results. In [63], authors showed the importance of learned reID features for multi-object tracking. All aforementioned methods learn the costs independently from the

Deep Learning on Graphs. Graph Neural Networks (GNNs) were first introduced in [66] as a generalization of neural networks that can operate on graph-structured domains. Since then, several works have focused on further developing and extending them by developing convolutional variants [11, 20, 40]. More recently, most methods were encompassed within a more general framework termed neural message passing [26] and further extended in [5] as graph networks. Given a graph with some initial features for nodes and optionally edges, the main idea behind these models is to embed nodes (and edges) into representations that take into account not only the node's own features but also those of its neighbors in the graph, as well as the graph overall topology. These methods have shown remarkable performance at a wide variety of areas, ranging from chemistry [26] to combinatorial optimization [51]. Within vision, they have been successfully applied to problems such as human action recognition [27], visual question answering [56] or single object tracking [25].

LITERATURE

How deep into the rabbit hole?

"What list gets longer and longer the more you read it?"

- ☑ All the relevant papers cited
- ☑ Cited *properly*
  both the method **and** the results
- ☑ *In comparison to*
  the proposed method
- ☑ Structured

- Start from the most relevant ones
  usually at the end of Related Work
- Going backwards
- *"What did the others say?"*
- Benchmarks
  only the published methods
- Last years' conferences

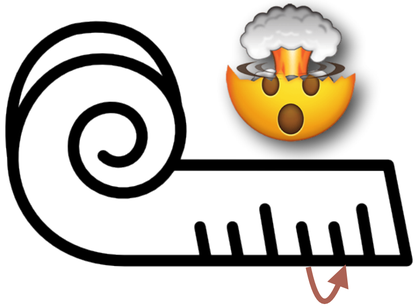
STRENGTHS & WEAKNESSES

A List of Questions

- Is the problem clearly stated and motivated?
- Does the related work cover the relevant approaches?
  Is the proposed approach clearly situated in the literature?
- Is the methodology clearly explained and well-organized?
  sub-sections for different parts, notation, equations, figures
- Is the evaluation sound and comprehensive?
  datasets, details, baselines, ablation studies, comparisons, runtime analysis
- Is the paper written clearly?
  figures, tables, grammar, organization
- Are ***all*** the claims validated? *What can we learn from this work?*

RESULTS

How important is SOTA?

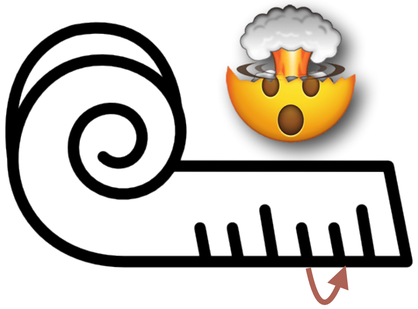


ABLATION

- Justification of design choices
- Contribution of each part
- Self-made baselines

RESULTS

How important is SOTA?



ABLATION

- Ju
- Co
- Se

net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs.* Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [9]			max	26.9	48.8	26.4
RoIWarp [7]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by ~3 points and AP_{75} by ~5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+7.3	*+ 5.3*	*+10.5*	*+5.8*	*+2.6*	*+9.5*

(d) **RoIAlign** (ResNet-50-**C5**, *stride 32*): Mask-level and box-level AP using *large-stride* features. Misalignments are more severe than with stride-16 features (Table 2c), resulting in massive accuracy gaps.

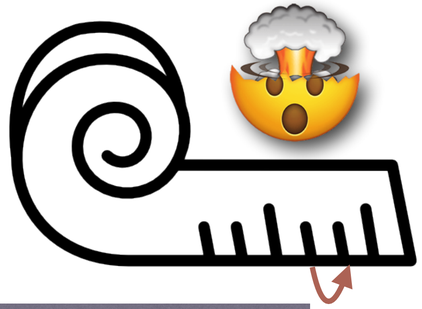
	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: 1024→1024→80·28^2	31.5	53.7	32.8
MLP	fc: 1024→1024→1024→80·28^2	31.5	54.0	32.6
FCN	conv: 256→256→256→256→256→80	**33.6**	**55.2**	**35.3**

(e) **Mask Branch** (ResNet-50-FPN): Fully convolutional networks (FCN) *vs.* multi-layer perceptrons (MLP, fully-connected) for mask prediction. FCNs improve results as they take advantage of explicitly encoding spatial layout.

Table 2. **Ablations** for Mask R-CNN. We train on `trainval35k`, test on `minival`, and report *mask* AP unless otherwise noted.

RESULTS

How important is SOTA?



ABLATION

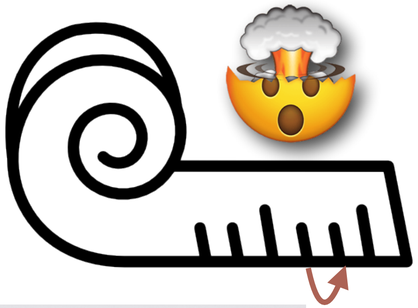
- Justification of design choices
- Contribution of each part
- Self-made baselines

COMPARISON

- The most promising setting(s)
- Previous work under the same cond.
- Submitted to a benchmark

RESULTS

How important is SOTA?



ABLATION

- Jus
- Co
- Sel

COMPARISON

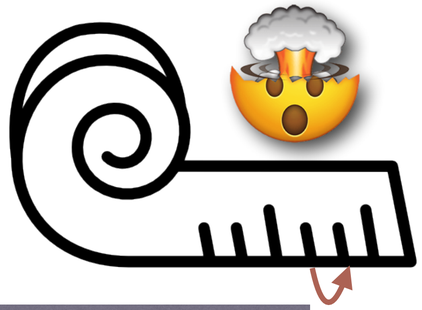
cond.

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**	**22.1**	**43.2**	51.2

Table 3. **Object detection** *single-model* results (bounding box AP), *vs.* state-of-the-art on `test-dev`. Mask R-CNN using ResNet-101-FPN outperforms the base variants of all previous state-of-the-art models (the mask output is ignored in these experiments). The gains of Mask R-CNN over [22] come from using RoIAlign (+1.1 AP^{bb}), multitask training (+0.9 AP^{bb}), and ResNeXt-101 (+1.6 AP^{bb}).

RESULTS

How important is SOTA?



ABLATION	COMPARISON
• Justification of design choices	• The most promising setting(s)
• Contribution of each part	• Previous work under the same cond.
• Self-made baselines	• Submitted to a benchmark

- ☑ Recommend missing baselines, datasets
- ☑ Verify the numbers and the evaluation setting from the cited papers
- ☑ Verify the numbers on the benchmark
- ☑ Comment on results

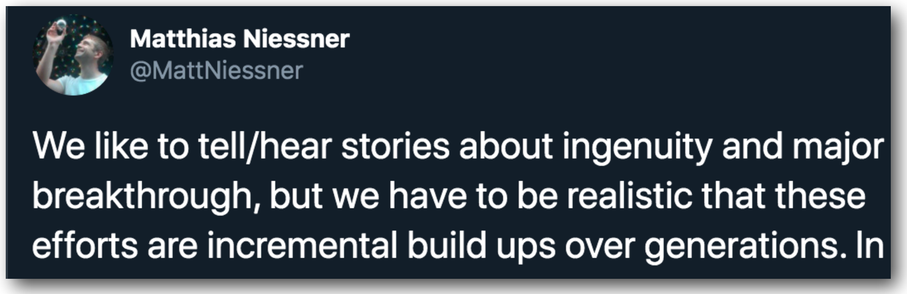
NOT-TO-DO LIST

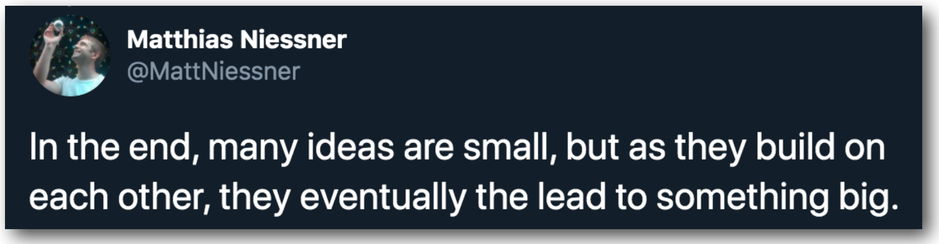
no-no's

- ✖ violating ethical rules
- ✖ skipping strengths or weaknesses
- ✖ failing to answer the questions listed

"use responsibly"

- ❗ "not novel"
- ❗ "fails to outperform the SOTA"
- ❗ "*incremental* contributions"
  "*incremental* improvements"





MAKING A DECISION

Ordering, Weighing, Strong Decisions & Borderline

ORDERING

The Process of Decision Making

1. Read the paper
2. List the strengths & weaknesses
3. Weigh the importance of points
4. Make a decision

1. Read the paper
2. Make a decision
3. Find reasons to justify it

WEIGHING

What is more important?

- A long list of weaknesses
  - Move to *Suggestions to Authors*?
  - "Is it a reason to reject?"
- Order the weaknesses by importance
- Can it be handled in
  - the rebuttal?
  - the final submission?

Problems in

1. Claim check
2. Methodology
3. Evaluation
4. Writing

FREQUENCIES

Roughly

STRONG DECISIONS

Rarely

- Obvious wrongdoings

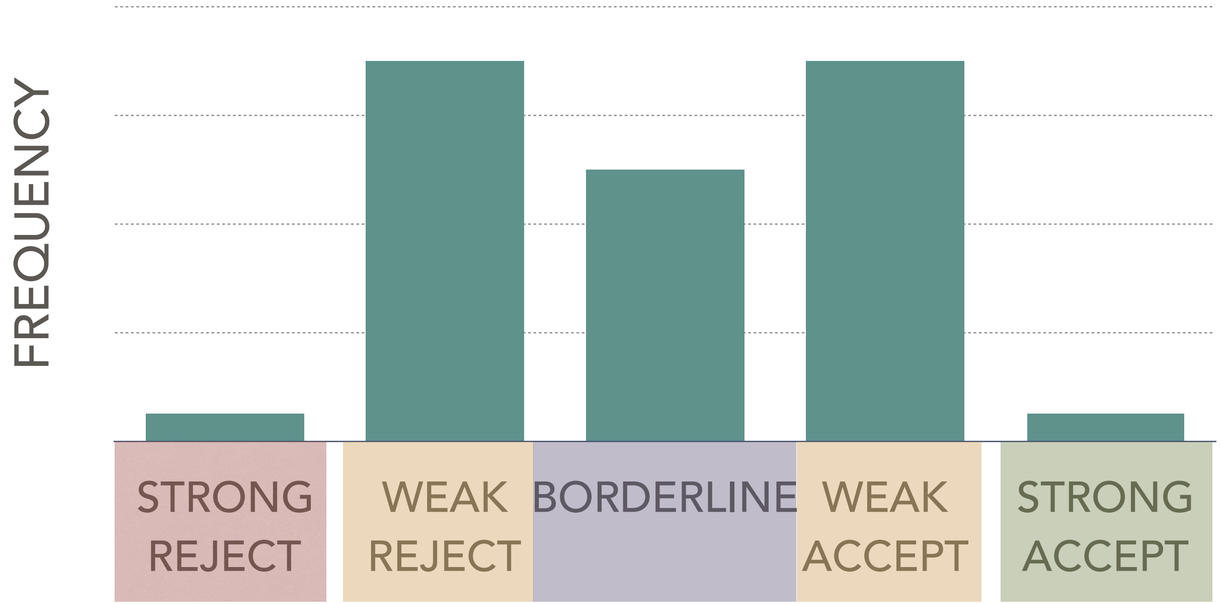
  e.g. plagiarism, no references, training on test data

- No weaknesses ***at all***

  nothing to criticize or correct already the final version

NORMAL DECISIONS

Ideally & Borderline



- List and contrast the main points
- Be clear about your reasoning
- Say how you'd change your mind

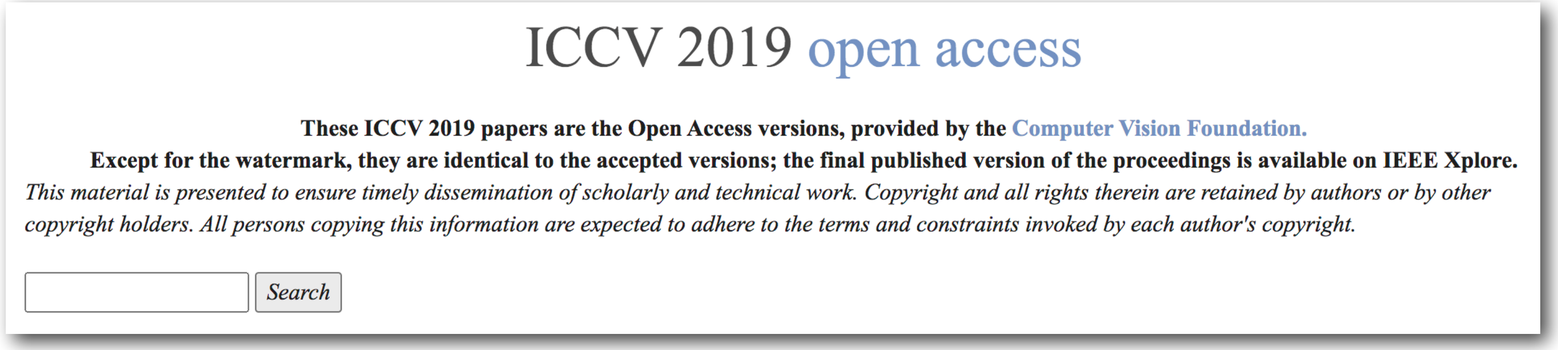
GOOD PRACTICES

Practicals, Timing, Limits, Discussions

PRACTICALS

to make your life easier

- Starting early
- Reading the paper and writing the review on different days
- Taking notes while reading
- Checking references and last years' conferences



TIMING

Bare Minimums, Ideals, and Exceptions

BARE MINIMUM

- 1 hour for reading
- Half an hour for the review
- 2 reviews a day at most

IDEALS

- 2 hours for reading
- 1 hour research
- 1 hour for the review
- 1 review a day

EXCEPTIONS

- Obvious reject/accept
- Unknown/well-known domain

TIMING

Bare Minimums, Ideals, and Exceptions

BARE MINIMUM

- 1 hour for reading
- Half an hour for the [illegible]
- 2 reviews a day

IDEALS

- 2 hours for reading
- 1 hour research
- 1 hour for the review
- 1 review a day

EXCEPTIONS

- Obvious reject/accept
- Unknown/well-known domain

LIMITS

Out of your Expertise

Let the AC know.

Early enough!

What if you have to do it?

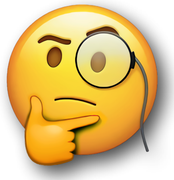
- Do the minimum!
  summary, claim check,
  writing, experiments
- More constructive than critical

DISCUSSIONS

not quite done yet!

Read the other reviews and the rebuttal.

Actively contribute to discussions.



Clearly explain and defend your point.

Be open to change your mind!