Explaining Human Preferences via Metrics for Structured 3D Reconstruction

Jack Langerman[1*] Denys Rozumnyi[2,3†] Yuzhong Huang[4] Dmytro Mishkin[3,4]

[1]Independent Researcher, USA [2]ETH Zurich, Switzerland

[3]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic [4]Hover Inc., USA

jack@jackml.com rozumden@gmail.com yuzhong.huang@hover.to dmytro.mishkin@hover.to

ICCV 2025

*Now at Apple †Now at Meta

Explaining Human Preferences via Metrics for Structured 3D Reconstruction

Jack Langerman[1*] Denys Rozumnyi[2,3†] Yuzhong Huang[4] Dmytro Mishkin[3,4]
[1]Independent Researcher, USA [2]ETH Zurich, Switzerland
[3]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic [4]Hover Inc., USA
jack@jackml.com rozumden@gmail.com yuzhong.huang@hover.to dmytro.mishkin@hover.to

ICCV 2025

Highlight

*Now at Apple †Now at Meta

The Problem:
Measurements from Images

Areas	Siding	Other
Facades	2538 ft²	476 ft²
Openings	426 ft²	112 ft²
Unknown (no photos)	-	0 ft²
Total	2964 ft²	588 ft²

Corners	Siding	Other
Inside Qty	3	-
Inside Length	17' 4"	-
Outside Qty	8	-
Outside Length	64' 11"	-

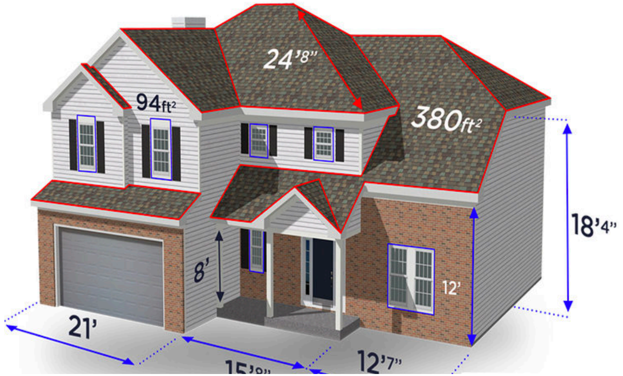
This key is not a representation of the building contained in this report. It is strictly for reference.

Trim	Siding	Other
Level Starter	232'	121' 5"
Sloped Trim	68' 11"	42' 9"
Vertical Trim	58' 2"	186' 9"

Roofline	Length	Avg. Depth	Soffit Area
Eaves Fascia (Gutters)	221' 3"	-	-
Level Frieze Board	180' 9"	1' 8"	643 ft²
Rakes Fascia	160' 6"	-	-
Sloped Frieze Board	137' 7"	1' 2"	266 ft²

Openings	Siding	Other
Quantity	37	2
Tops Length	87'	-
Sills Length	100' 4"	16'
Sides Length	265' 3"	27' 11"

Accessories	Siding	Other
Shutter Qty	32	0
Shutter Area	200 ft²	0 ft²
Vents Qty	3	0
Vents Area	6 ft²	0 ft²

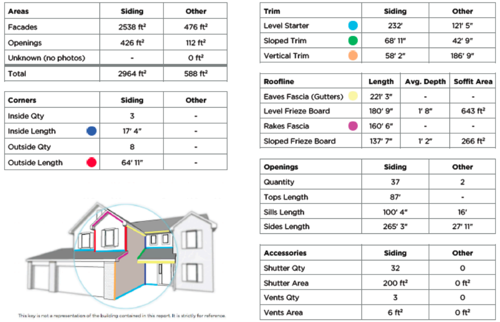


Produce semantically meaningful measurements

or a representation from which we can easily read off the measurements we care about

The Problem:
Structured 3D from Images

representation from which we can easily (automatically) read off the measurements we care about



Areas	Siding	Other
Facades	2538 ft²	476 ft²
Openings	426 ft²	112 ft²
Unknown (no photos)	-	0 ft²
Total	2964 ft²	588 ft²

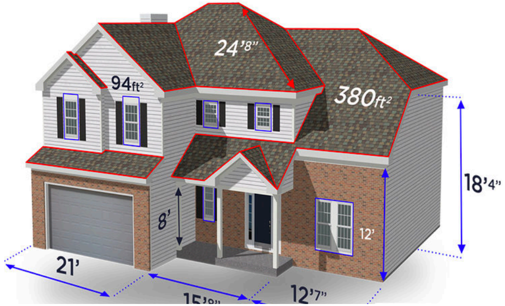
Corners	Siding	Other
Inside Qty	3	-
Inside Length	17' 4"	-
Outside Qty	8	-
Outside Length	64' 11"	-

Trim	Siding	Other
Level Starter	232'	121' 5"
Sloped Trim	68' 11"	42' 9"
Vertical Trim	58' 2"	186' 9"

Roofline	Length	Avg. Depth	Soffit Area
Eaves Fascia (Gutters)	221' 3"	-	-
Level Frieze Board	180' 9"	1' 8"	643 ft²
Rakes Fascia	160' 6"	-	-
Sloped Frieze Board	137' 7"	1' 2"	266 ft²

Openings	Siding	Other
Quantity	37	2
Tops Length	87'	-
Sills Length	100' 4"	16'
Sides Length	265' 3"	27' 11"

Accessories	Siding	Other
Shutter Qty	32	0
Shutter Area	200 ft²	0 ft²
Vents Qty	3	0
Vents Area	6 ft²	0 ft²



Produce semantically meaningful measurements
*or a **representation** from which we can easily read off the measurements we care about*

The Problem:
Structured 3D from Images

Structured 3D Representation

Areas	Siding	Other
Facades	2538 ft²	476 ft²
Openings	426 ft²	112 ft²
Unknown (no photos)	-	0 ft²
Total	2964 ft²	588 ft²

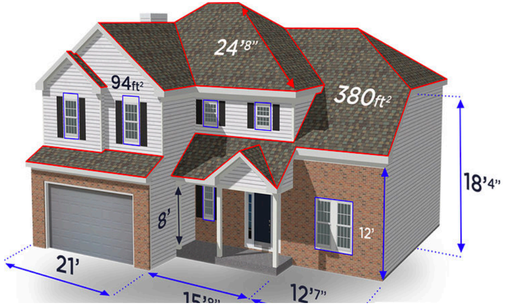
Corners	Siding	Other
Inside Qty	3	-
Inside Length	17' 4"	-
Outside Qty	8	-
Outside Length	64' 11"	-

Trim	Siding	Other
Level Starter	232'	121' 5"
Sloped Trim	68' 11"	42' 9"
Vertical Trim	58' 2"	186' 9"

Roofline	Length	Avg. Depth	Soffit Ar
Eaves Fascia (Gutters)	221' 3"	-	-
Level Frieze Board	180' 9"	1' 8"	643 ft
Rakes Fascia	160' 6"	-	-
Sloped Frieze Board	137' 7"	1' 2"	266 ft

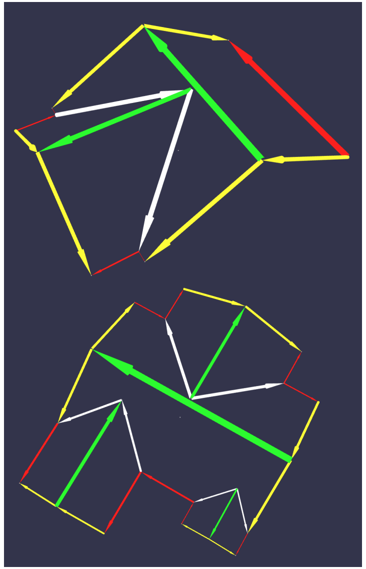
Openings	Siding	Other
Quantity	37	2
Tops Length	87'	-
Sills Length	100' 4"	16'
Sides Length	265' 3"	27' 11"

Accessories	Siding	Other
Shutter Qty	32	0
Shutter Area	200 ft²	0 ft²
Vents Qty	3	0
Vents Area	6 ft²	0 ft²



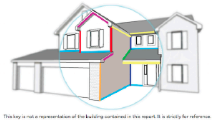
Produce semantically meaningful measurements
*or a **representation** from which we can easily read off the measurements we care about*

The Problem:
Structured 3D from Images



Areas	Siding	Other
Facades	2538 ft²	476 ft²
Openings	426 ft²	112 ft²
Unknown (no photos)	-	0 ft²
Total	2964 ft²	588 ft²

Corners	Siding	Other
Inside Qty	3	-
Inside Length	17' 4"	-
Outside Qty	8	-
Outside Length	64' 11"	-

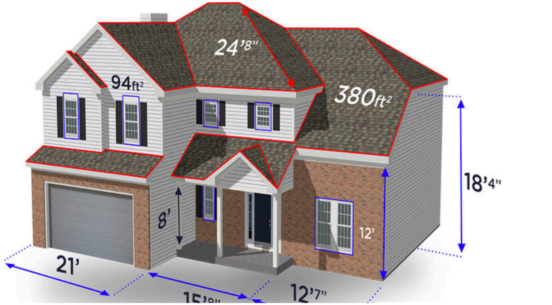


Trim	Siding	Other
Level Starter	232'	121' 5"
Sloped Trim	68' 11"	42' 9"
Vertical Trim	58' 2"	186' 9"

Roofline	Length	Avg. Depth	Soffit Ar
Eaves Fascia (Gutters)	221' 3"	-	-
Level Frieze Board	180' 9"	1' 8"	643 ft²
Rakes Fascia	160' 6"	-	-
Sloped Frieze Board	137' 7"	1' 2"	266 ft²

Openings	Siding	Other
Quantity	37	2
Tops Length	87'	-
Sills Length	100' 4"	16'
Sides Length	265' 3"	27' 11"

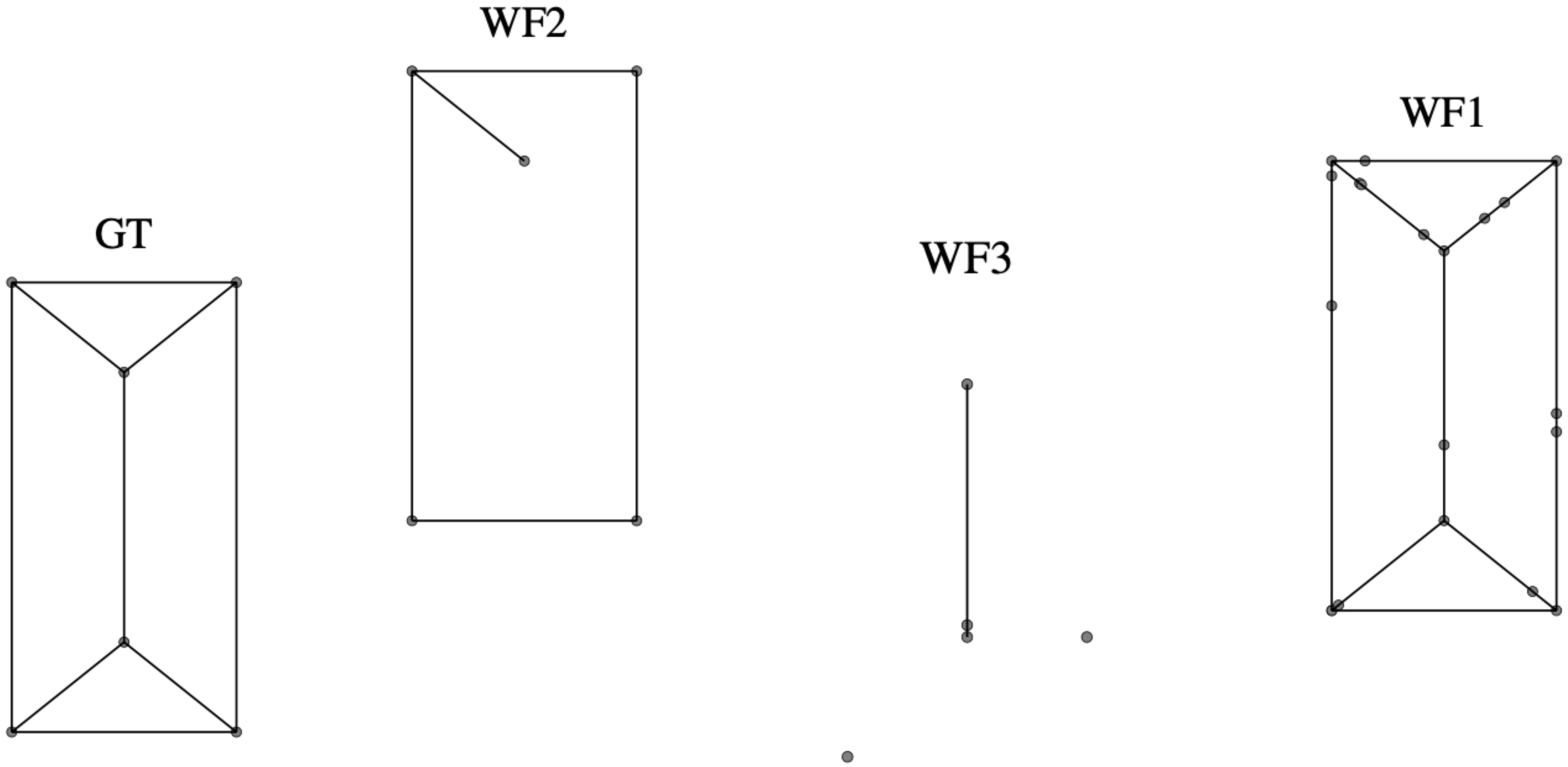
Accessories	Siding	Other
Shutter Qty	32	0
Shutter Area	200 ft²	0 ft²
Vents Qty	3	0
Vents Area	6 ft²	0 ft²



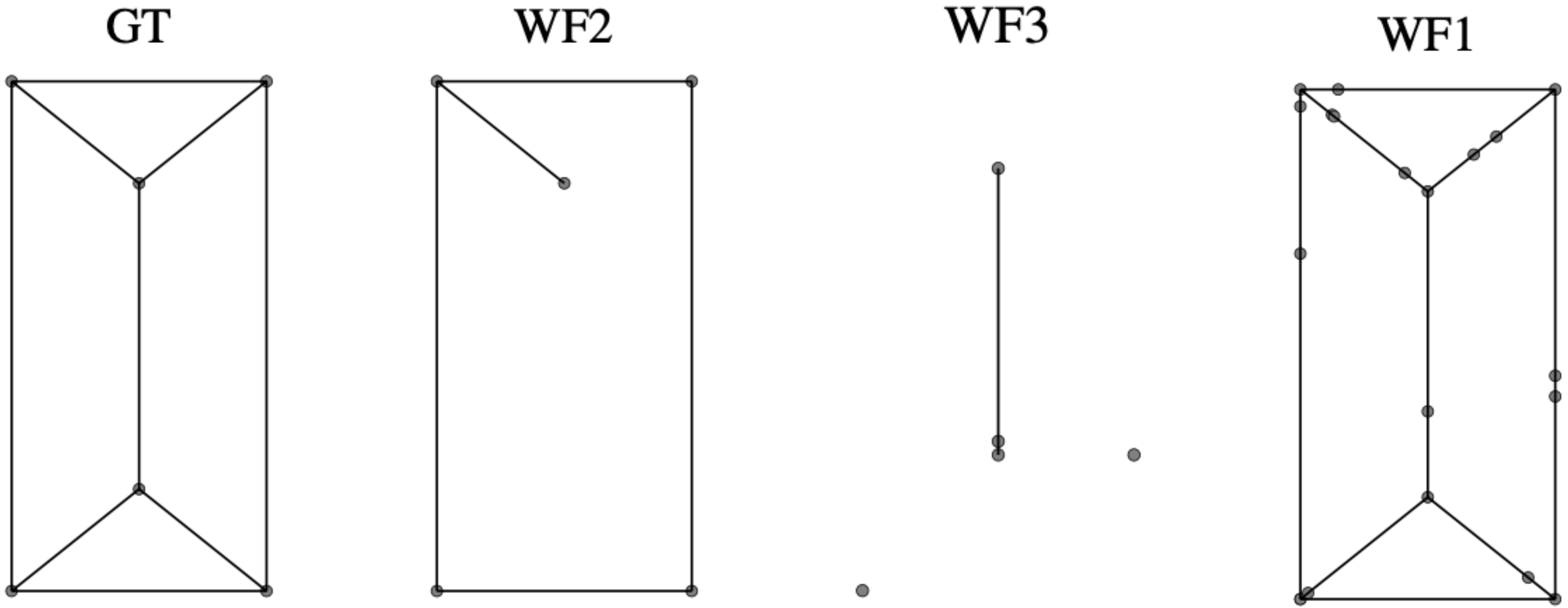
Extracting a structured 3D representation (Semantic Wireframes) from a set of multi view ground level images

”What cannot be measured cannot be improved”

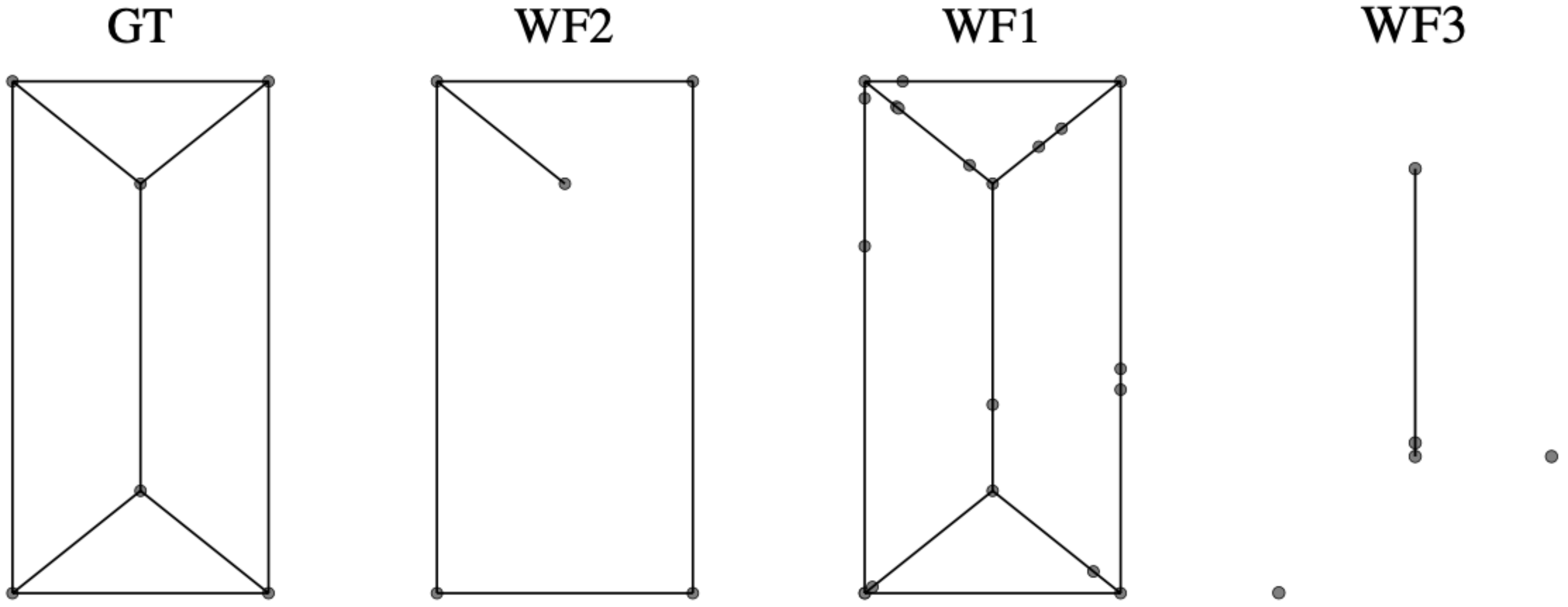
How would you order these from best to worst?



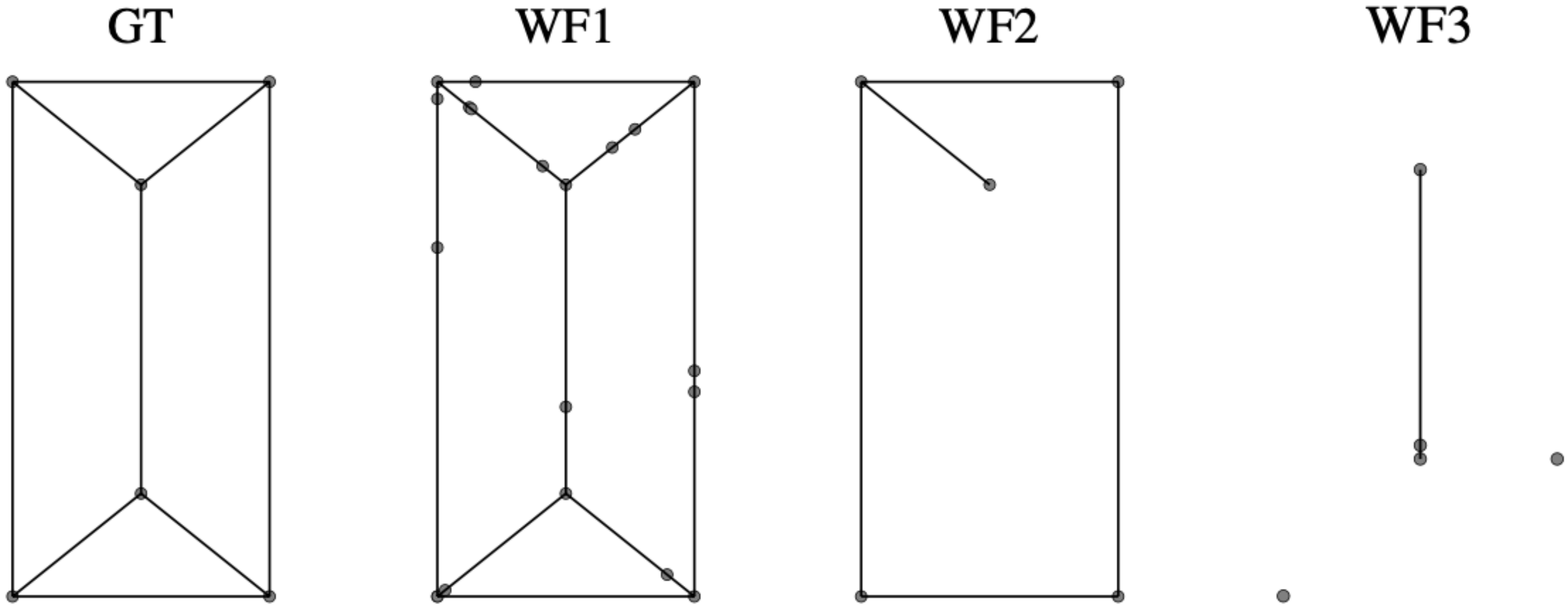
WED Ranks them like this



Edge and Vertex F1 rank them like this



Most people (including expert 3D modelers) choose this ranking



How can we choose (or design) metrics that:

1) Have “good” properties

2) are well aligned with (expert) human preferences?

Properties

Identity of Indiscernibles: This property ensures that identical inputs receive a dissimilarity score of zero, indicating perfect similarity. For any reconstruction x, a metric d satisfies this property if $d(x, x) = 0$.

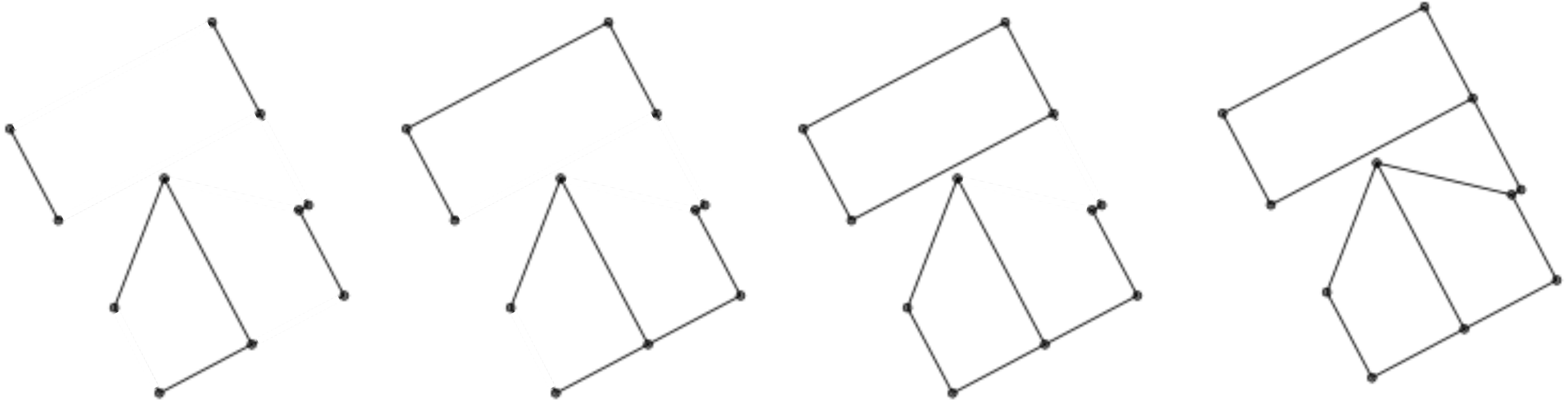
Symmetry: A symmetric metric produces the same dissimilarity score regardless of the order of the inputs. For reconstructions x and y, a metric satisfies symmetry if $d(x, y) = d(y, x)$.

Triangle Inequality: The triangle inequality ensures that for any three reconstructions x, y, and z, the dissimilarity between x and z is less than or equal to the sum of dissimilarities between x and y, and y and z. This relationship is expressed as $d(x, z) \leq d(x, y) + d(y, z)$.

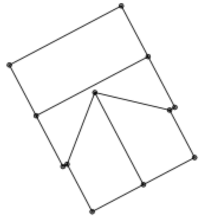
Monotonicity: This property describes how the dissimilarity score behaves when components (such as vertices or edges) are removed from a reconstruction. A metric satisfies monotonicity if the dissimilarity score does not increase when wrong vertices or edges are deleted. Similarly, the dissimilarity must not increase when correct vertices or edges are added.

Quasi-proportionality: This property holds when the metric changes smoothly under perturbations. This is evaluated by moving random vertices with small increments and checking the variance of the differences in the score. We use the following perturbations to simulate better or worse reconstructions: (i) remove correct edges from the ground truth wireframe; (ii) add wrong edges to the ground truth wireframe; (iii) disconnect ground truth edges; (iv) remove correct vertices; (v) move ground truth vertices to the wrong location. For every perturbation, we apply it 10 times and declare an example monotonic if it is strictly increasing (or decreasing as appropriate) for those continuous 10 perturbations.

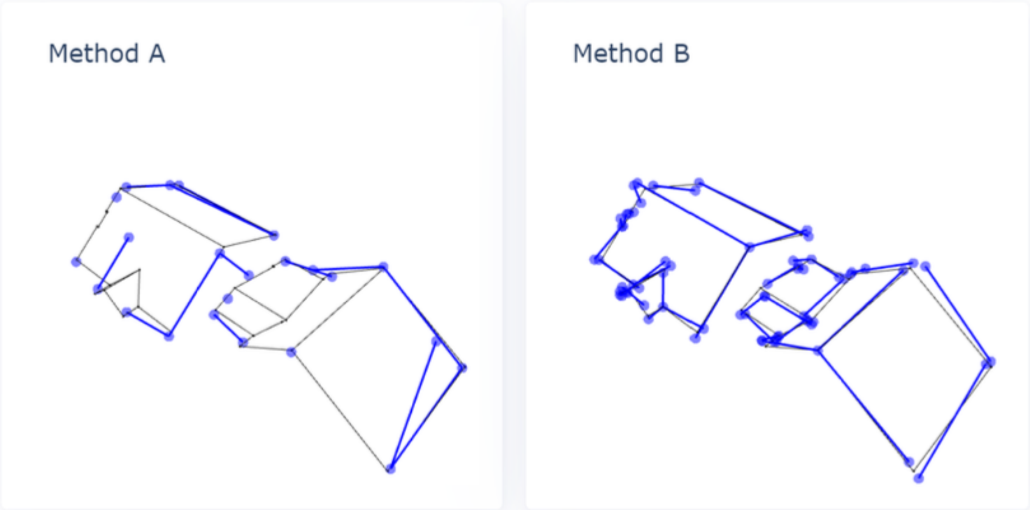
Example: Monotonicity



Metric Value Decreasing As Correct Edges Are Added



Expert Judgements



Which of the blue solutions is closest to the black solution?

A is Better (A/←) | Equal Quality (W/↑) | B is Better (D/→)

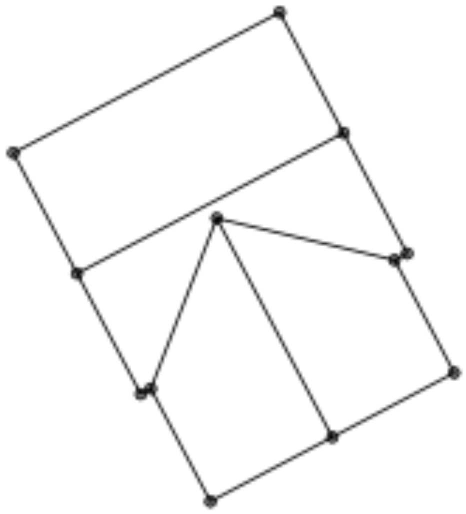
Where did we get the wireframes reconstructions?

“Real World” reconstructions from the 2024 S²3DR Challenge

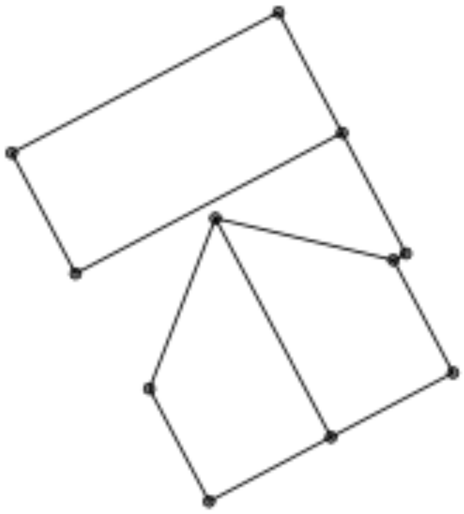
Pool1– $S^2 3DR$. We acquire a representative set of $S^2 3DR$ challenge entries [22] as well as a PC2WF [24] baseline. These wireframes were algorithmically (and with the help of deep-learning models) reconstructed from multiview inputs with the goal of minimizing a variant of WED. We include the top-10 entries with team names used as identifiers. The ground truth models were created by human experts and have undergone significant validation. The input data were captured by users on mobile phones in North America.

Synthetically Generated
(with Known Rankings)

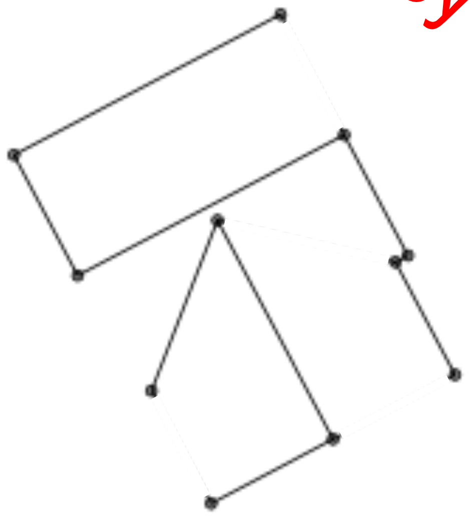
98% accuracy



GT



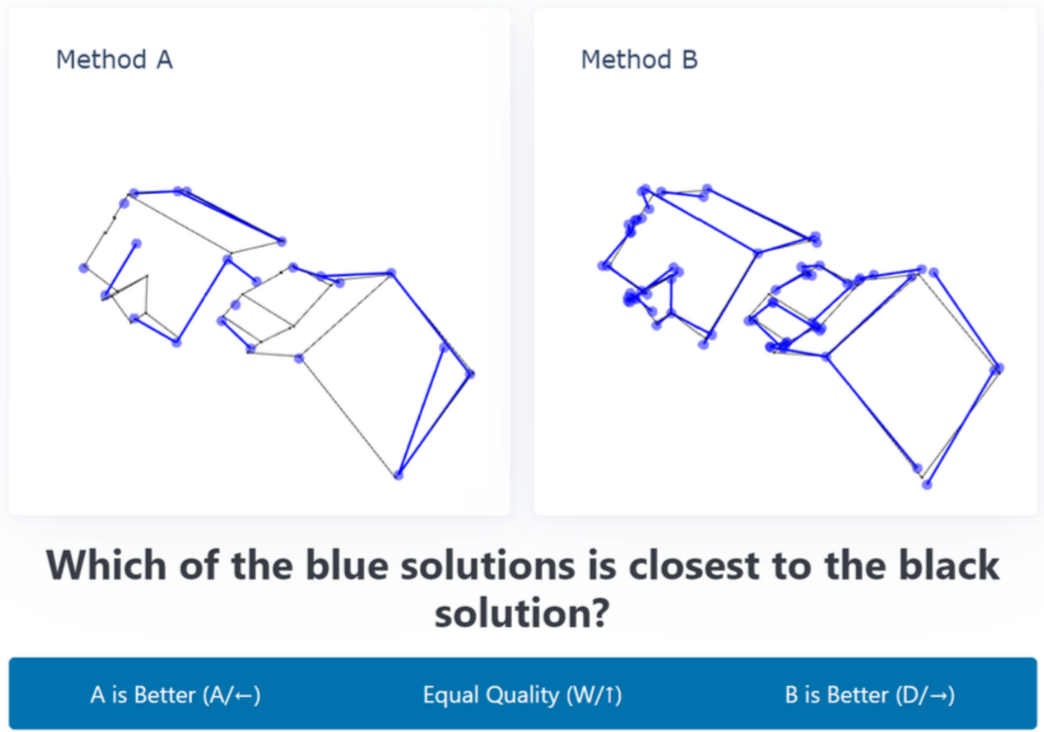
A



B

Recall: our goal is to find metrics that agree with human expert preferences

Preference Collection



Do Humans Agree with Themselves (are they self-consistent)?

Yes!

When showing the same pairs repeatedly humans pick the same winner **≈90%** of the time.

Do Humans Agree with Each Other?

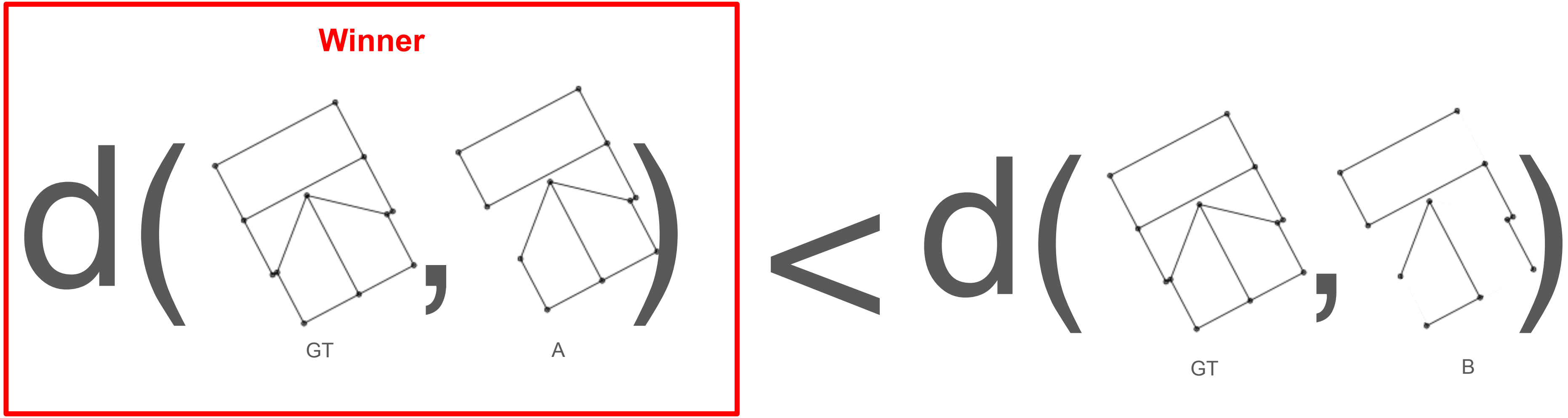
Yes!

Excluding ties (within clusters) raters agree **≈90%** of the time

*Two clusters emerge, one weighting **edges** (edge F1, Jaccard) and one weighting **vertices** (corner F1).
See paper for details

Let’s make clear what we mean by “metric” and how we use these metrics to rank pairs:

Metrics Picking Winners



We use several methods to quantify the degree of agreement (rank correlation).

Empirical Win Rate

1) For each pair of reconstructions give **1** point to the winning method, **0.5** for a tie (“equal”), and **0** for a loss.

2) Human rankings: For each pair, average over all raters to get human win rate for each method per pair

3) Global: Average scores over all pairs (to get global rankings for humans and each metric

4) Measure rank correlation against each metric

Method	Empirical Win Rate
add_low	0.89
add_med	0.86
perturb_med	0.85
add_high	0.82
perturb_low	0.79
remove_low	0.79
perturb_high	0.67
deform_med	0.67
deform_low	0.66
remove_high	0.65
remove_med	0.63
kc92	0.51
Siromanec	0.50
deform_high	0.50
maximivashechkin	0.39
rozumden	0.38
kcml	0.35
rozumden	0.34
Ana-Geneva	0.32
pc2wf_retrain	0.29
Yurii	0.29
snuggler	0.25
baseline	0.25
Hunter-X	0.22
TUM	0.22
Fudan EDLAB	0.21
pc2wf_pretrained	0.20

Bradley-Terry Abilities

1) Goal: Estimate latent ability scores θ_i for each method that explain pairwise preference probabilities.

2) Model: $P(i \succ j) = \sigma(\theta_i - \theta_j)$

3) Min_θ $\text{BCE}(\sigma(\theta_i - \theta_j), y_{ij})$ with $y_{ij}=1$ iff $i \succ j$ else 0

4) Measure rank correlation against each metric

Method	Empirical Win Rate	BT Ability
add_low	0.89	2.79
add_med	0.86	2.47
perturb_med	0.85	2.33
add_high	0.82	2.04
perturb_low	0.79	1.79
remove_low	0.79	1.71
perturb_high	0.67	0.83
deform_med	0.67	0.83
deform_low	0.66	0.78
remove_high	0.65	0.67
remove_med	0.63	0.60
kc92	0.51	-0.25
Siromanec	0.50	-0.34
deform_high	0.50	-0.34
maximivashechkin	0.39	-1.01
rozumden	0.38	-1.10
kcml	0.35	-1.28
rozumden	0.34	-1.35
Ana-Geneva	0.32	-1.47
pc2wf_retrain	0.29	-1.65
Yurii	0.29	-1.63
snuggler	0.25	-1.92
baseline	0.25	-1.91
Hunter-X	0.22	-2.10
TUM	0.22	-2.13
Fudan EDLAB	0.21	-2.18
pc2wf_pretrained	0.20	-2.28

Low Rank Factor Scoring (via SVD)

Goal: explain the Methods x Raters empirical log odds matrix with a single low-rank "Quality" factor

Method	Empirical Win Rate	BT Ability	Quality Factor
add_low	0.89	2.79	0.03
add_med	0.86	2.47	0.02
perturb_med	0.85	2.33	0.02
add_high	0.82	2.04	-0.02
perturb_low	0.79	1.79	-0.01
remove_low	0.79	1.71	-0.02
perturb_high	0.67	0.83	-0.09
deform_med	0.67	0.83	-0.09
deform_low	0.66	0.78	-0.10
remove_high	0.65	0.67	-0.10
remove_med	0.63	0.60	-0.11
kc92	0.51	-0.25	-0.18
Siromanec	0.50	-0.34	-0.19
deform_high	0.50	-0.34	-0.19
maximivashechkin	0.39	-1.01	-0.27
rozumden	0.38	-1.10	-0.28
kcml	0.35	-1.28	-0.26
rozumden	0.34	-1.35	-0.26
Ana-Geneva	0.32	-1.47	-0.25
pc2wf_retrain	0.29	-1.65	-0.23
Yurii	0.29	-1.63	-0.25
snuggler	0.25	-1.92	-0.25
baseline	0.25	-1.91	-0.25
Hunter-X	0.22	-2.10	-0.25
TUM	0.22	-2.13	-0.26
Fudan EDLAB	0.21	-2.18	-0.26
pc2wf_pretrained	0.20	-2.28	-0.25

Why so many ways to determine rankings?

- Using different and varied methods to extract the pseudo ground truth ranking of the methods from the expert pairwise judgements allows us to verify that their judgements are stable and meaningful.

- We check the correlation between the different scoring methods (Bradley-Terry, SVD)

- We find a Kendall correlation coefficient >0.7 (showing moderate to strong agreement) between the rankings implied by SVD and those implied by BT.

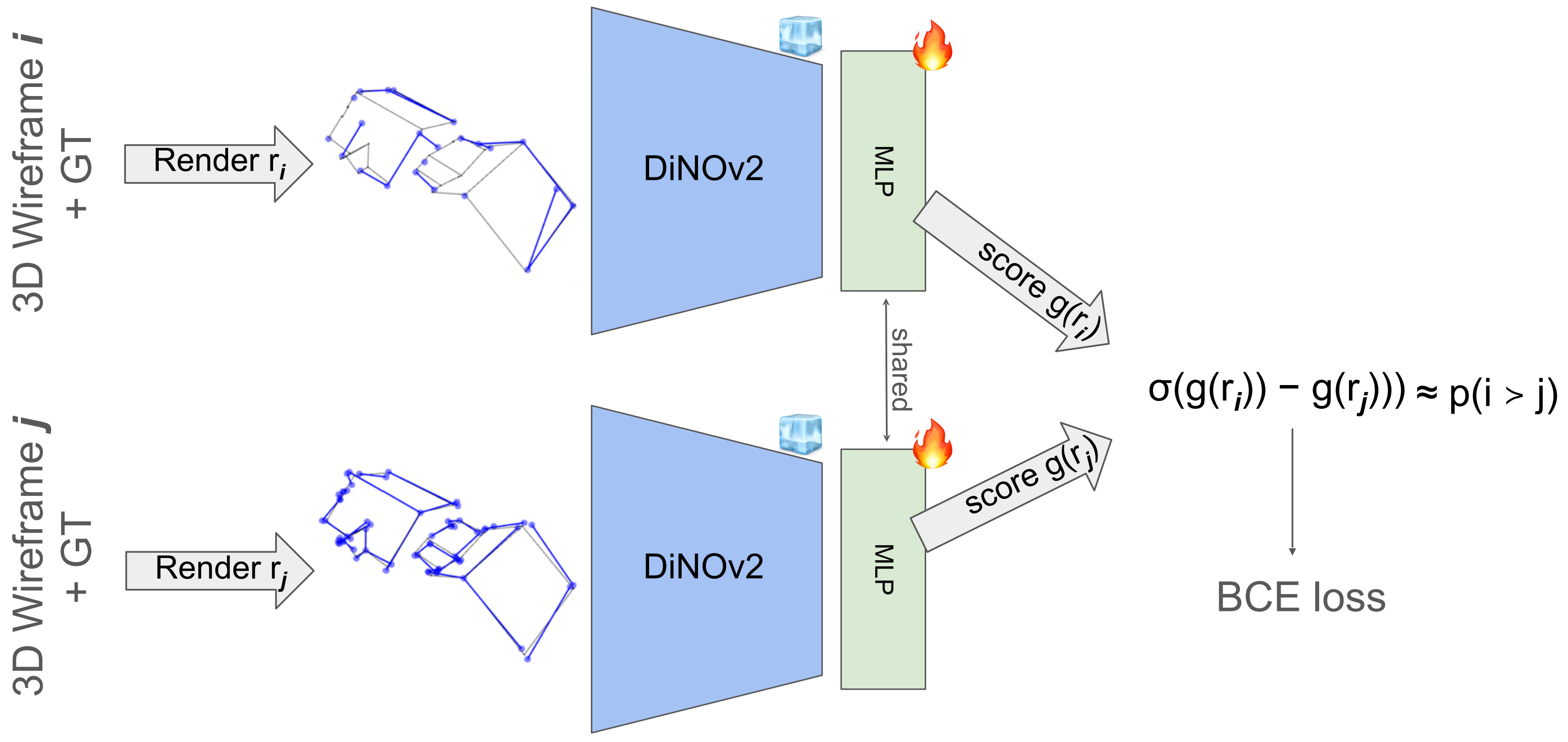
- This lends additional evidence to the hypothesis that there is a true ”quality” factor driving the raters’ views.

(Observation 4.6 in our paper)

Families of Hand Crafted Metrics Under Consideration

- **Wireframe Edit Distance (WED)**
- **Chamfer / Edge Chamfer Distance (ECD)**
- **Corner and Edge Detection** (precision, recall, F1)
- **Jaccard / IoU-based** (over cylinderized edges)
- **Hausdorff Distance**
- **Spectral Graph Distances**

Learned Metric



Selected Metrics: Average Agreement With Raters

Observation 4.2

Human annotators pay more attention to correct parts of the reconstruction than the incorrect parts. Regardless of whether edges or vertices are considered, recall metrics agree more with human preferences than precision ones.

Observation 4.3

The average agreement with human preferences of the top handcrafted metrics does not vary significantly. WED-based scores correlate with annotators the least.

Metric	Average Agreement With Raters
corner f1	80.94
learned metric_xval	80.33
edge f1	79.61
corner offset	74.0
jaccard dist	76.0
edge chamfer bi	75.28
spectral optimal_l1	73.56
hausdorff	68.5
WED mnn	69.94
WED prereg	63.06
WED ap	66.5
random	52.39

Recommendations

- Our learned metric shows strong agreement with expert judgements
- Humans care more about recall than precision (both on edges and vertices)
- Both recall only metrics and neural nets are hackable
- In environments subject to aggressive optimization (RL, Gradients, Cash Prizes), we recommend using the harmonic mean of the vertex f1 score and (cylinderized) edge IoU which we denote Hybrid Structure Score

More about our work

ArXiv: https://arxiv.org/abs/2503.08208

Code & Data: https://github.com/s23dr/wireframe-metrics-iccv2025

ICCV Poster Page: https://iccv.thecvf.com/virtual/2025/poster/1220

Jack Langerman (@JackLangerman) Denys Rozumnyi (@DRozumnyi) Yuzhong Huang (@YuzhongHuang) Dmytro Mishkin (@ducha_aiki)
jack@jackml.com rozumden@gmail.com yuzhong.huang@hover.to dmytro.mishkin@hover.to