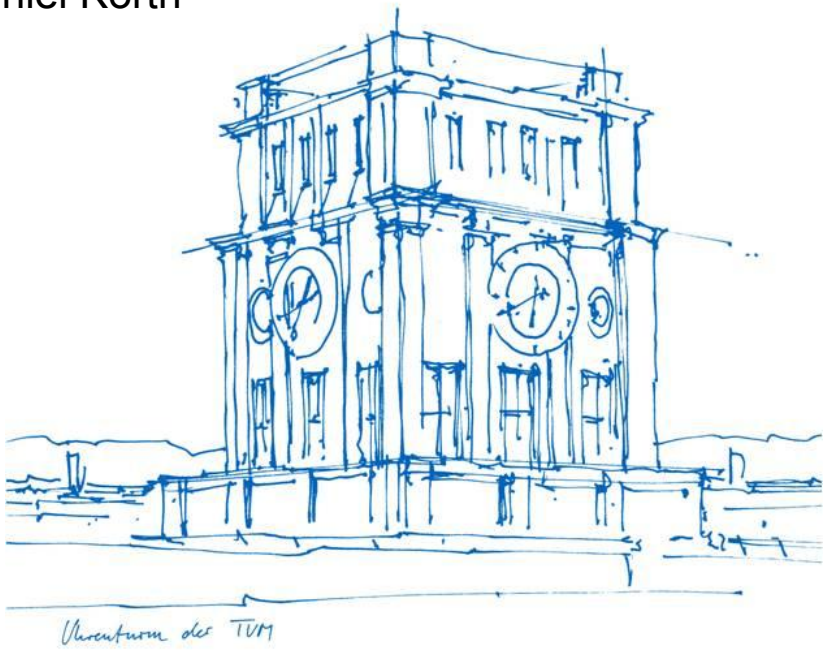
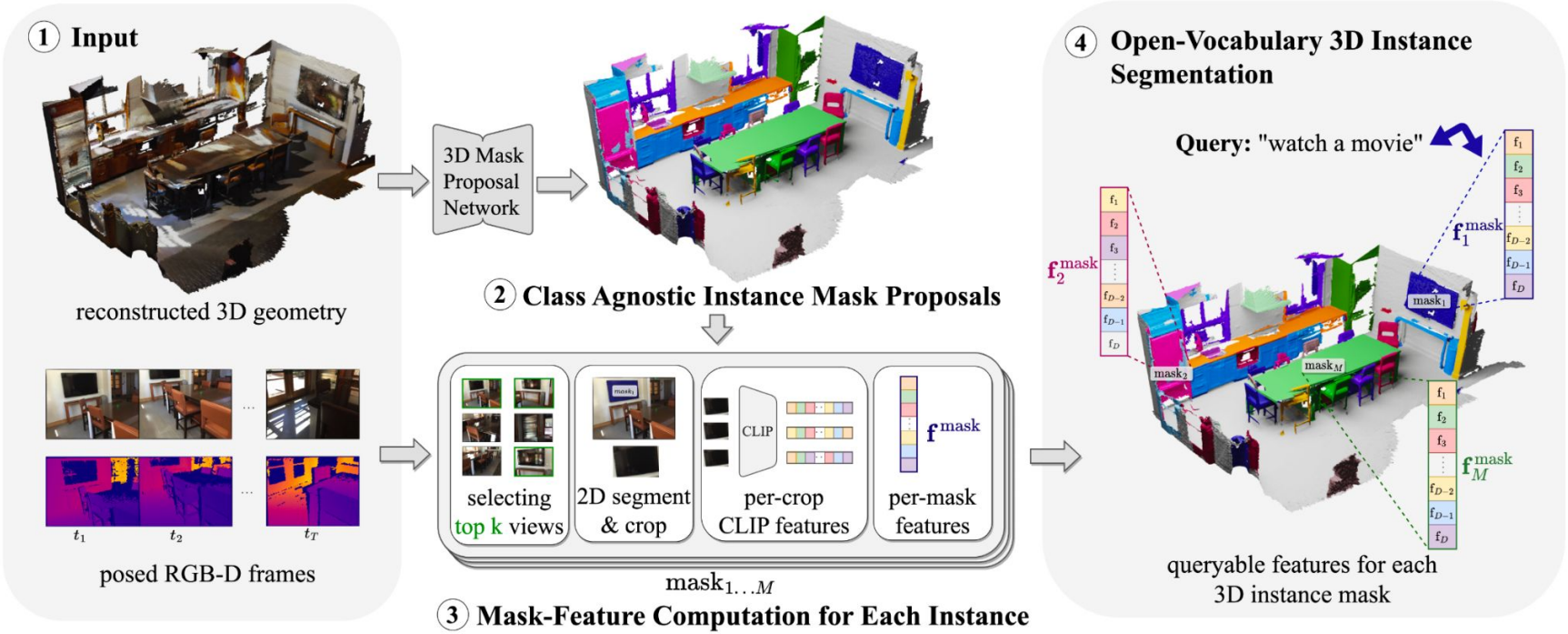
OpenMask3D++: Open-Vocabulary 3D Instance Segmentation

Machine Learning for 3D Geometry

Michael Neumayr, Michal Stary, Simon Pannek, Daniel Korth



OpenMask3D



[1] Takmaz, Ayça, et al. "OpenMask3D: Open-Vocabulary 3D Instance Segmentation." *arXiv preprint arXiv:2306.13631* (2023).

Initial Ideas / Challenges

- Use OpenMask3D on outdoor dataset
- Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]
- Focus on more diverse view selection to enhance semantic context of instances
- Experiment with different prompting strategies for SAM

Initial Ideas / Challenges

- ~~Use OpenMask3D on outdoor dataset~~ not enough interesting classes, availability
- ~~Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]~~ make little sense for indoor scenes
- Focus on more diverse view selection to enhance semantic context of instances
- ~~Experiment with different prompting strategies for SAM~~ bad proposition

Initial Ideas / Challenges

- ~~Use OpenMask3D on outdoor dataset~~ not enough interesting classes, availability
- ~~Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]~~ make little sense for indoor scenes
- Focus on more diverse view selection to enhance semantic context of instances
- ~~Experiment with different prompting strategies for SAM~~ bad proposition
- Replace the Mask3D backbone by a SAM3D [4]
- Retrain Mask3D backbone on Scannet++ [5]
- Replace mean pooling by max pooling to take surely valid vector from embedding space
- Replace the Mask3D backbone by Segment3D [6]
- Remove outlier embeddings caused by imperfect SAM masks due to noisy depth maps
- Replace CLIP embeddings by SigLIP [7, 8]
- Support image queries with CLIP/SigLIP/DINOv2 embeddings for finegrained querying [9]

Initial Ideas / Challenges

- ~~Use OpenMask3D on outdoor dataset~~ not enough interesting classes, availability
- ~~Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]~~ make little sense for indoor scenes
- Focus on more diverse view selection to enhance semantic context of instances
- ~~Experiment with different prompting strategies for SAM~~ bad proposition
- ~~Replace the Mask3D backbone by a SAM3D [4]~~ degrades performance
- Retrain Mask3D backbone on Scannet++ [5]
- Replace mean pooling by max pooling to take surely valid vector from embedding space
- Replace the Mask3D backbone by Segment3D [6]
- Remove outlier embeddings caused by imperfect SAM masks due to noisy depth maps
- Replace CLIP embeddings by SigLIP [7, 8]
- Support image queries with CLIP/SigLIP/DINOv2 embeddings for finegrained querying [9]

Initial Ideas / Challenges

- ~~Use OpenMask3D on outdoor dataset~~ not enough interesting classes, availability
- ~~Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]~~ make little sense for indoor scenes
- Focus on more diverse view selection to enhance semantic context of instances
- ~~Experiment with different prompting strategies for SAM~~ bad proposition
- ~~Replace the Mask3D backbone by a SAM3D [4]~~ degrades performance
- ~~Retrain Mask3D backbone on Scannet++ [5]~~ too compute heavy
- Replace mean pooling by max pooling to take surely valid vector from embedding space
- Replace the Mask3D backbone by Segment3D [6]
- Remove outlier embeddings caused by imperfect SAM masks due to noisy depth maps
- Replace CLIP embeddings by SigLIP [7, 8]
- Support image queries with CLIP/SigLIP/DINOv2 embeddings for finegrained querying [9]

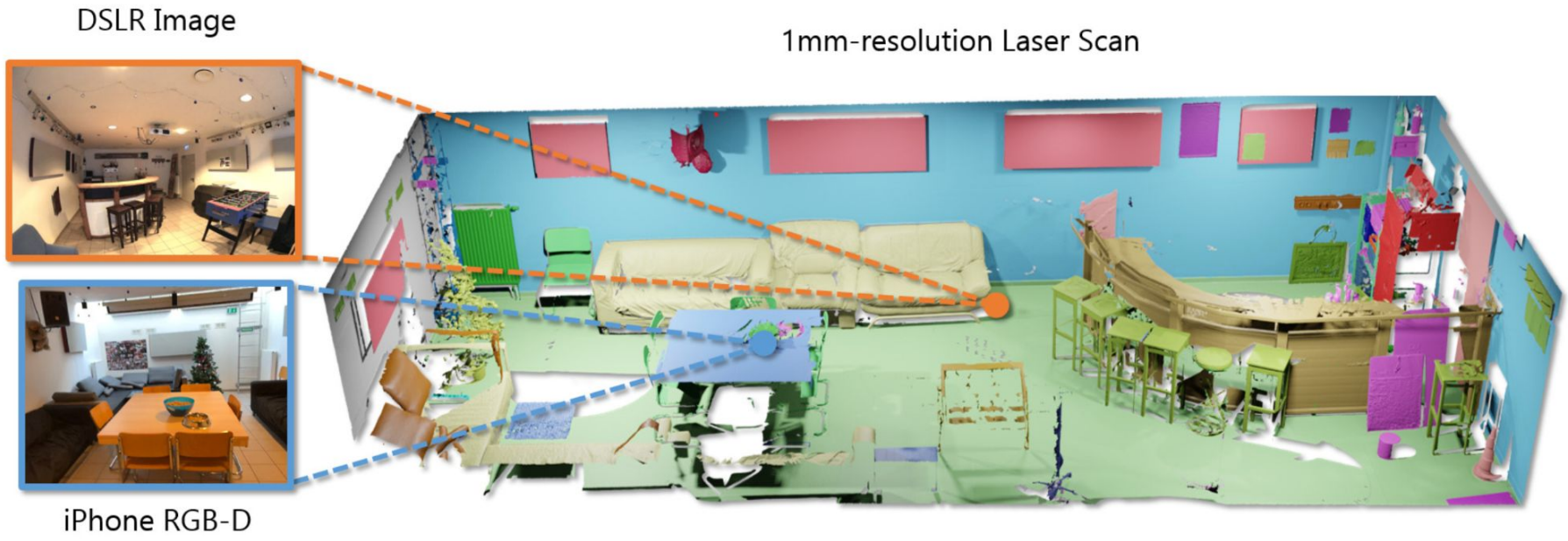
Initial Ideas / Challenges

- ~~Use OpenMask3D on outdoor dataset~~ not enough interesting classes, availability
- ~~Replace SAM by Semantic SAM to take in account outdoor objects granularity [2, 3]~~ make little sense for indoor scenes
- Focus on more diverse view selection to enhance semantic context of instances
- ~~Experiment with different prompting strategies for SAM~~ bad proposition
- ~~Replace the Mask3D backbone by a SAM3D [4]~~ degrades performance
- ~~Retrain Mask3D backbone on Scannet++ [5]~~ too compute heavy
- Replace mean pooling by max pooling to take surely valid vector from embedding space
- ~~Replace the Mask3D backbone by Segment3D [6]~~ code not open source
- Remove outlier embeddings caused by imperfect SAM masks due to noisy depth maps
- Replace CLIP embeddings by SigLIP [7, 8]
- Support image queries with CLIP/SigLIP/DINOv2 embeddings for finegrained querying [9]

Dataset

ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes

Chandan Yeshwanth*, Yueh-Cheng Liu*, Matthias Nießner, Angela Dai
Technical University of Munich
*equal contribution

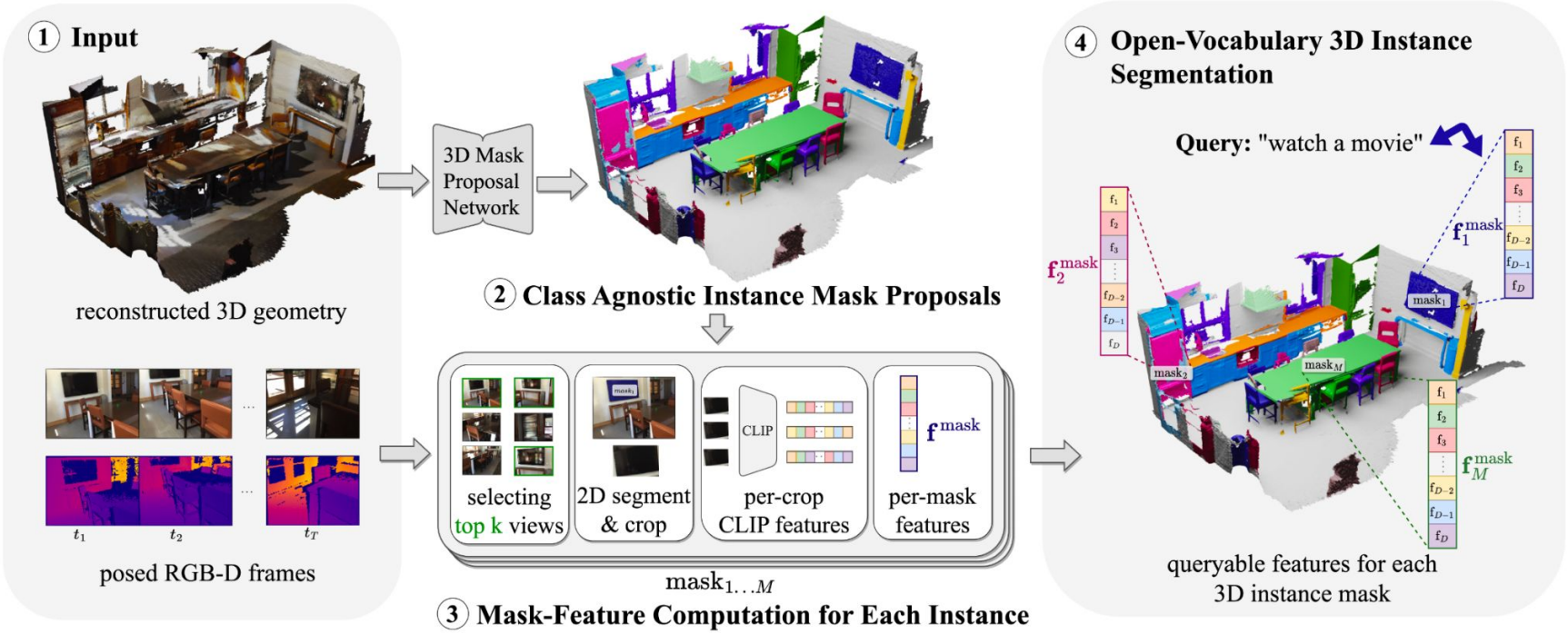


[5] Yeshwanth, Chandan, et al. "Scannet++: A high-fidelity dataset of 3d indoor scenes." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

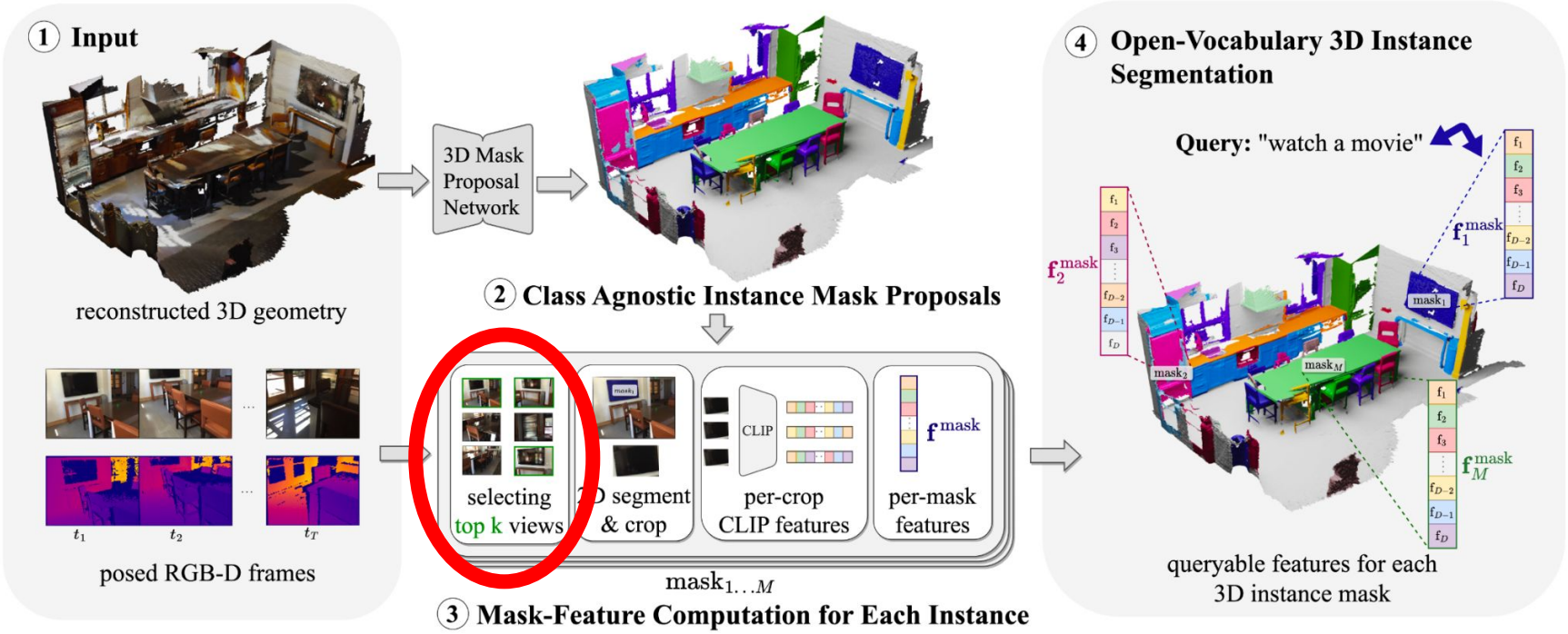
Improvement 1

More Diverse Views

View Selection



View Selection



View Selection

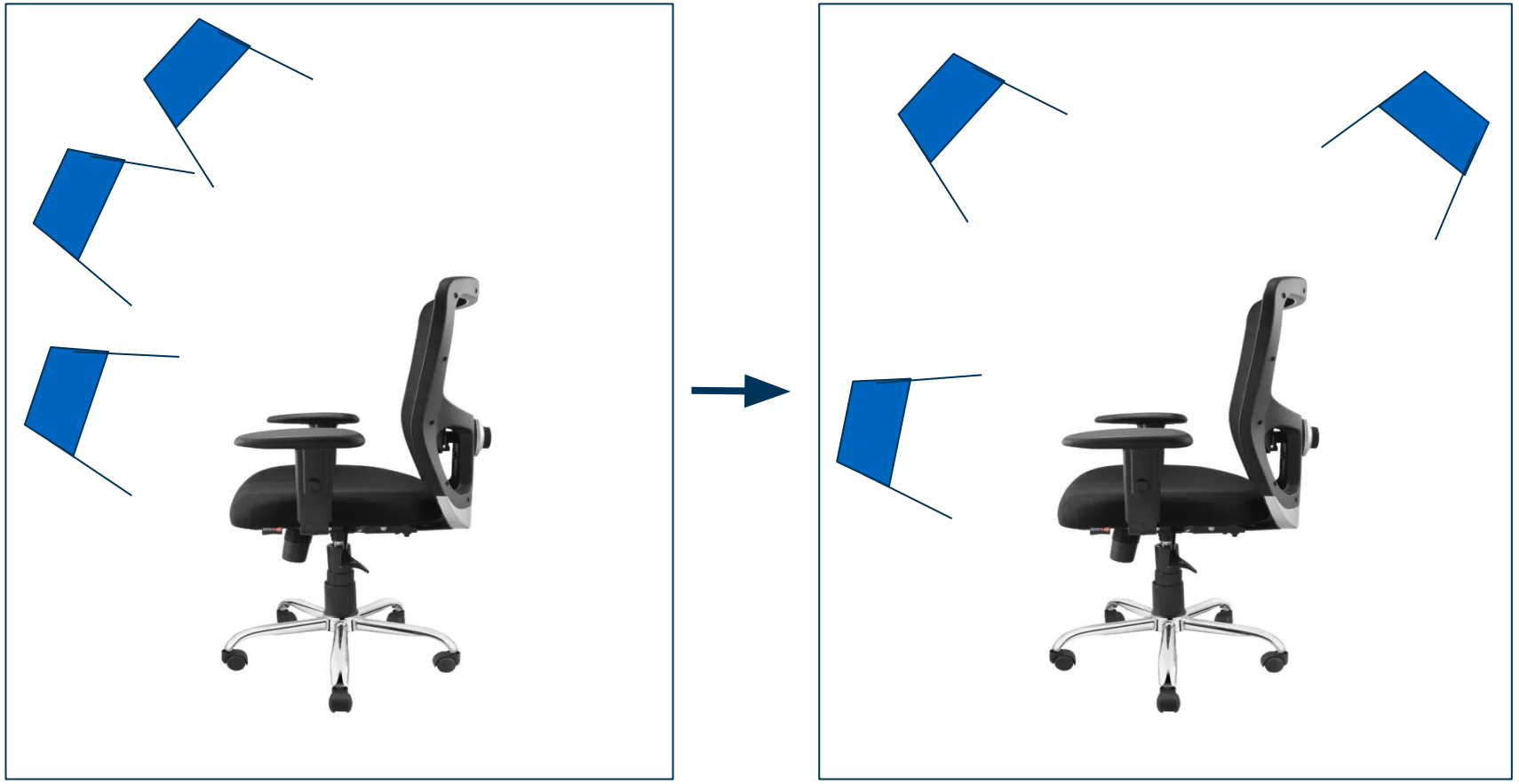


Image source:https://adikosystems.com/product/adiko-medium-back-ergonomic-office-chair-adpn-jz-mb-020/

View Selection

Base

View Selection

Base

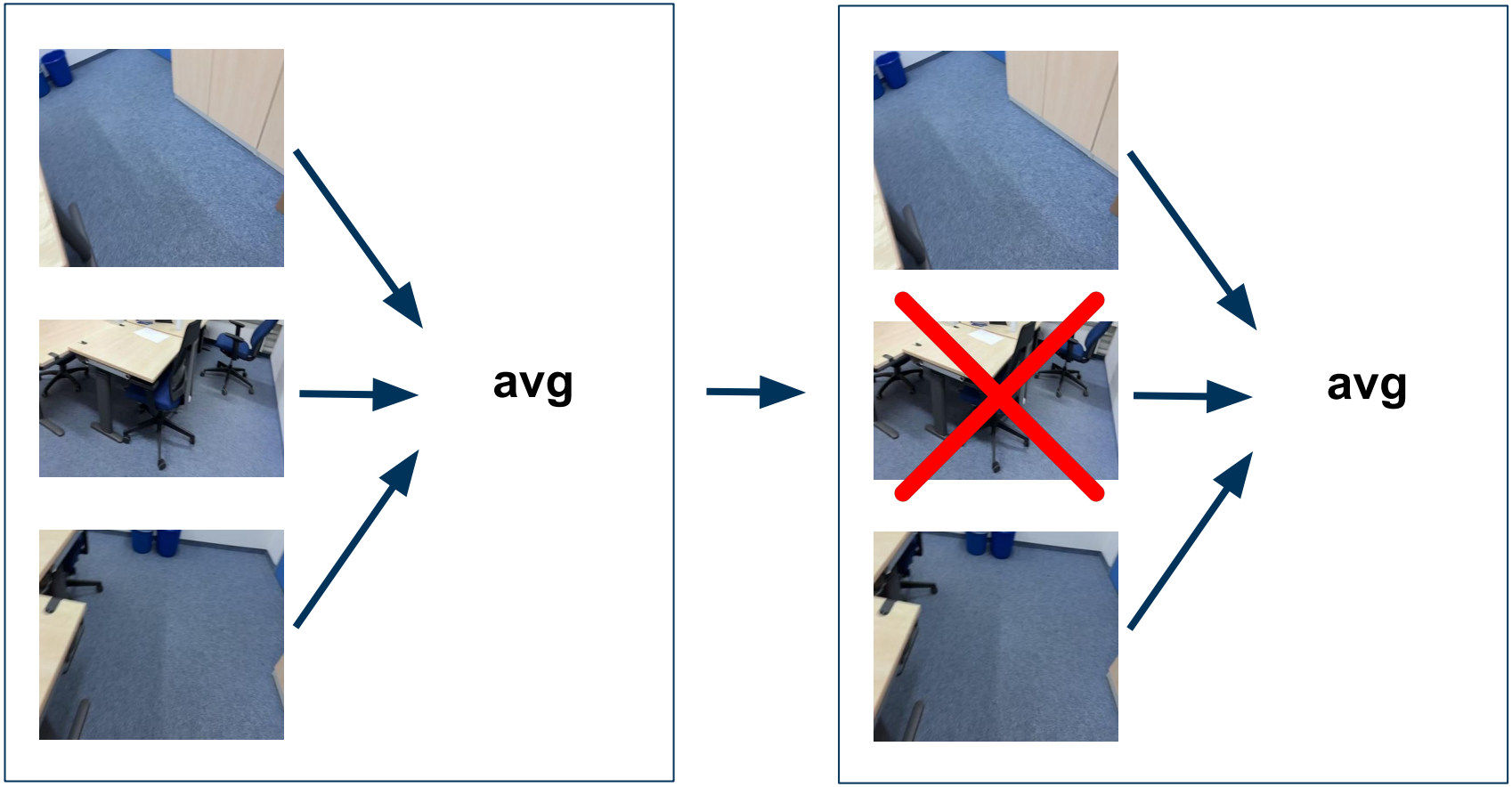
Diverse Views

Improvement 2

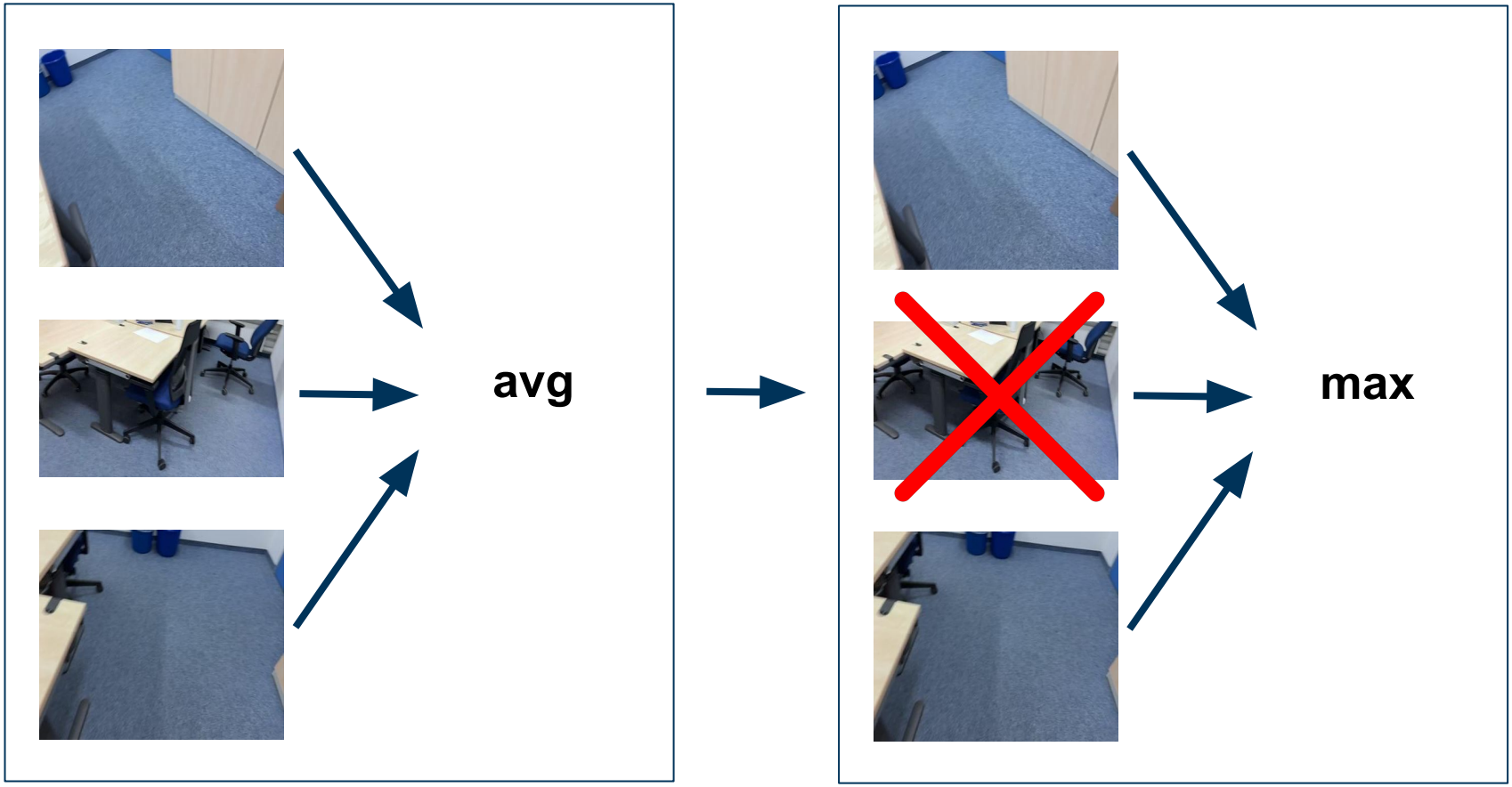
Outlier Removal

Outlier Removal

Outlier Removal



Outlier Removal

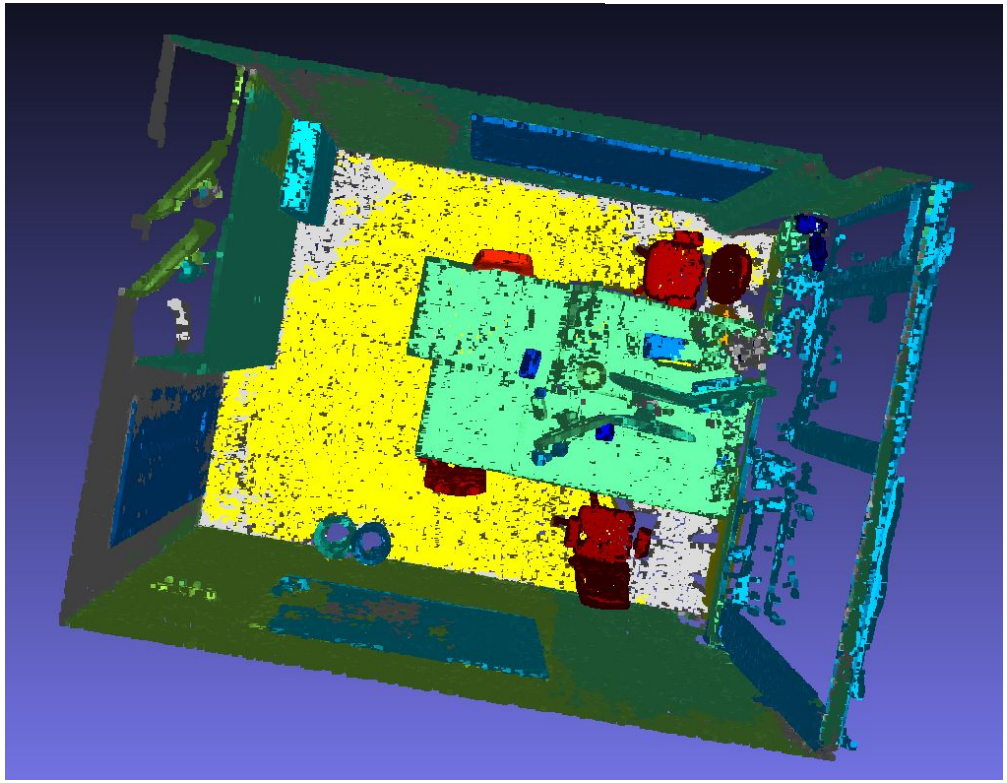


Outlier Removal

Text Query: “a chair”

Similarity

low high

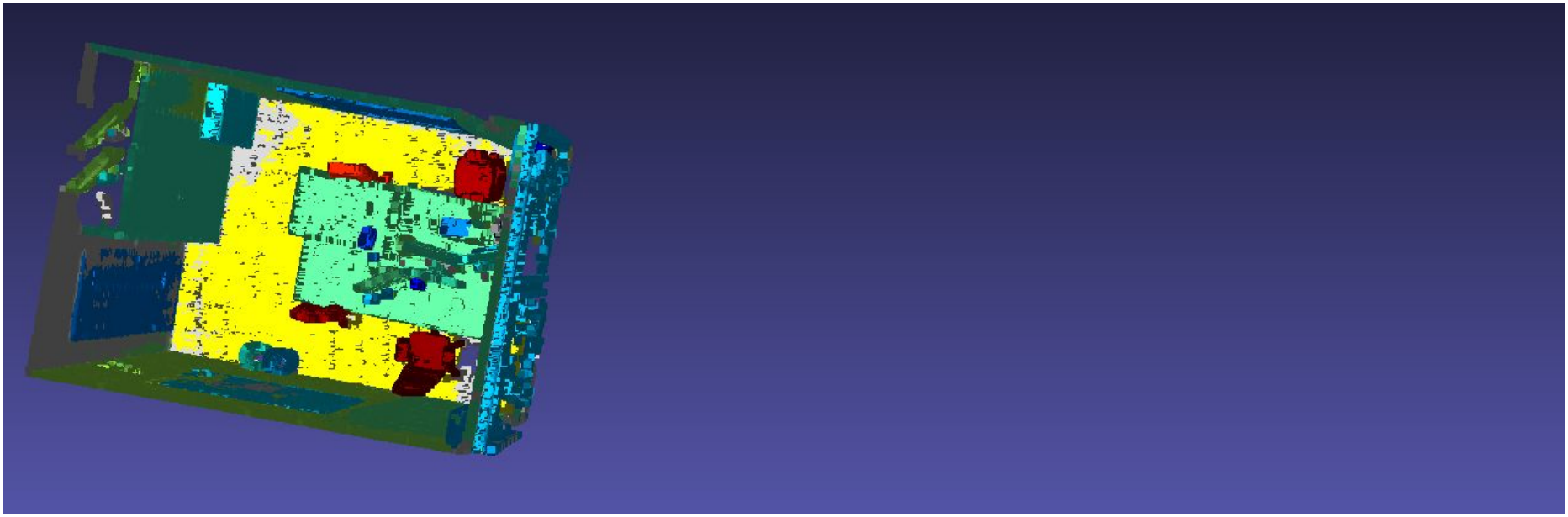


Outlier Removal

Text Query: “a chair”

Similarity

low high



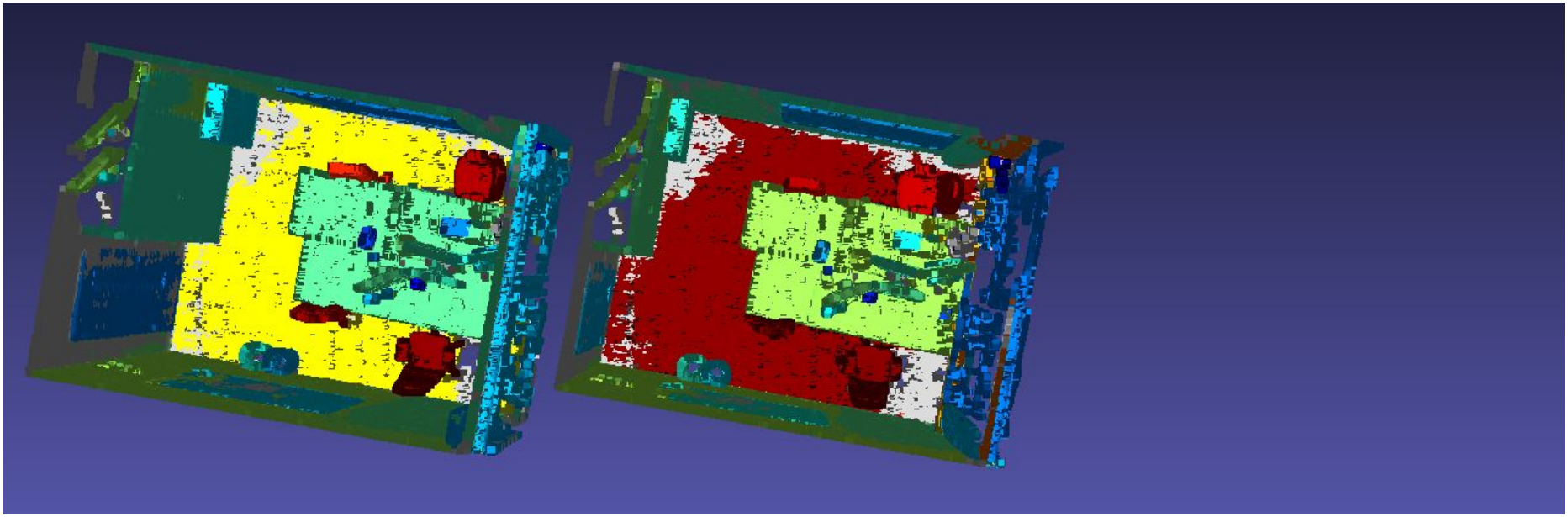
Baseline **Max Pooling** **Outlier Removal**

Outlier Removal

Text Query: “a chair”

Similarity

low high



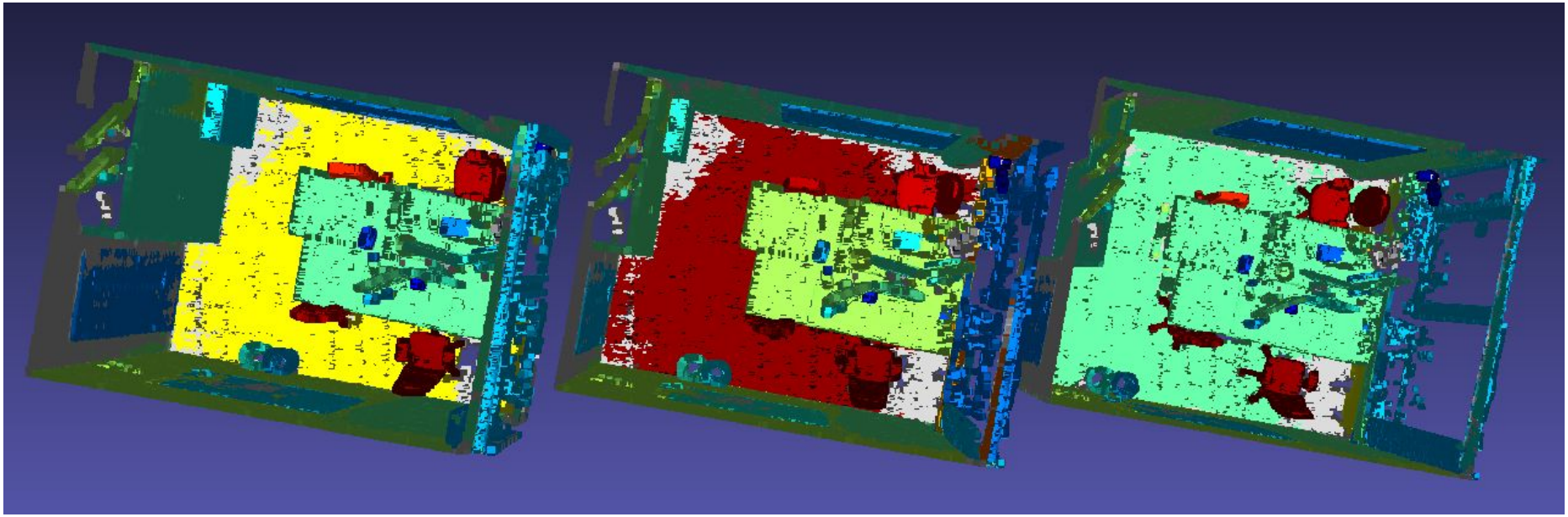
Baseline **Max Pooling** **Outlier Removal**

Outlier Removal

Text Query: “a chair”

Similarity

low high



Baseline **Max Pooling** **Outlier Removal**

Improvement 3

Replacing CLIP by SigLIP

CLIP: SoftMax

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

SigLIP: Sigmoid

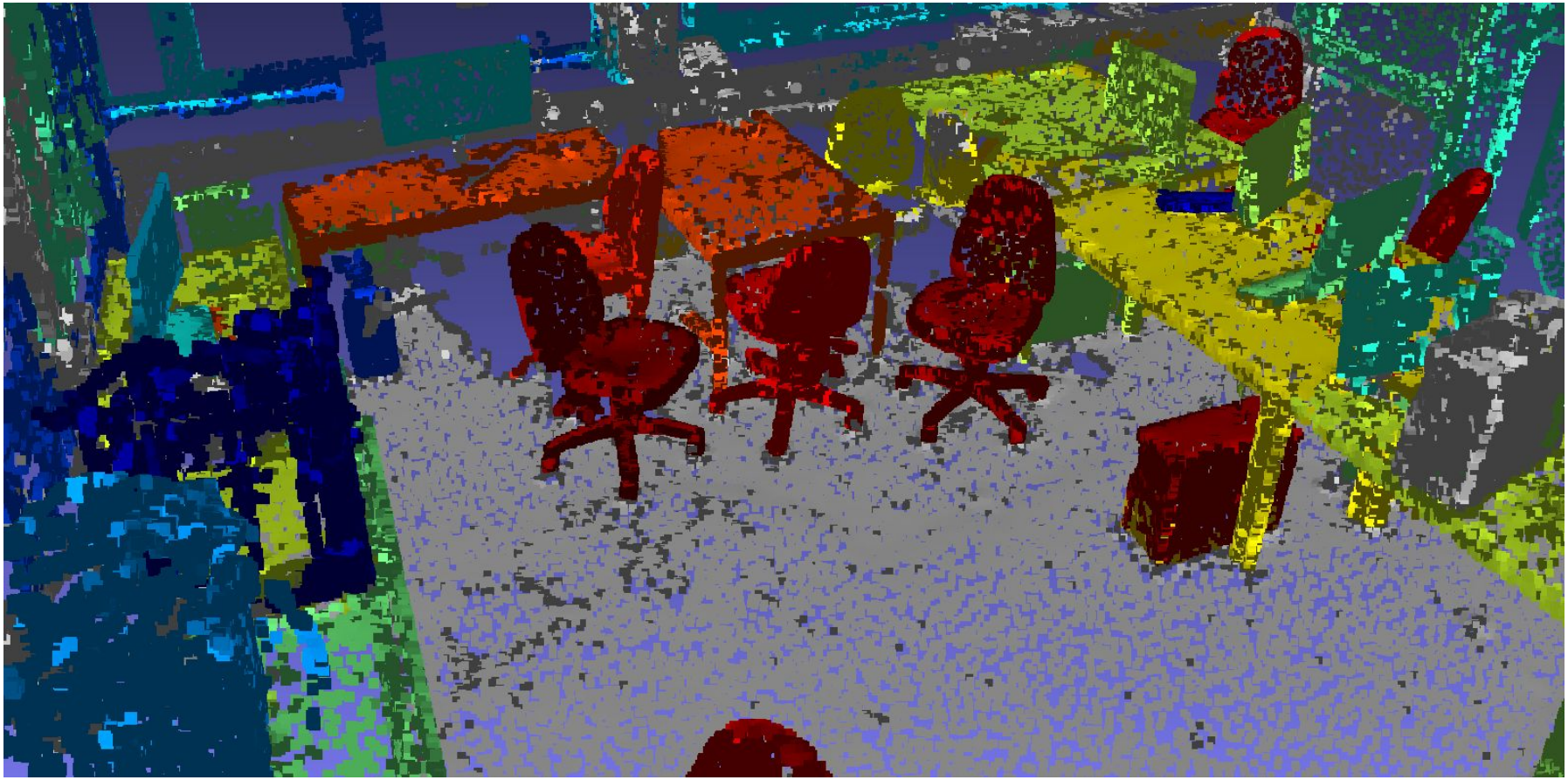
$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

[7] Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023)

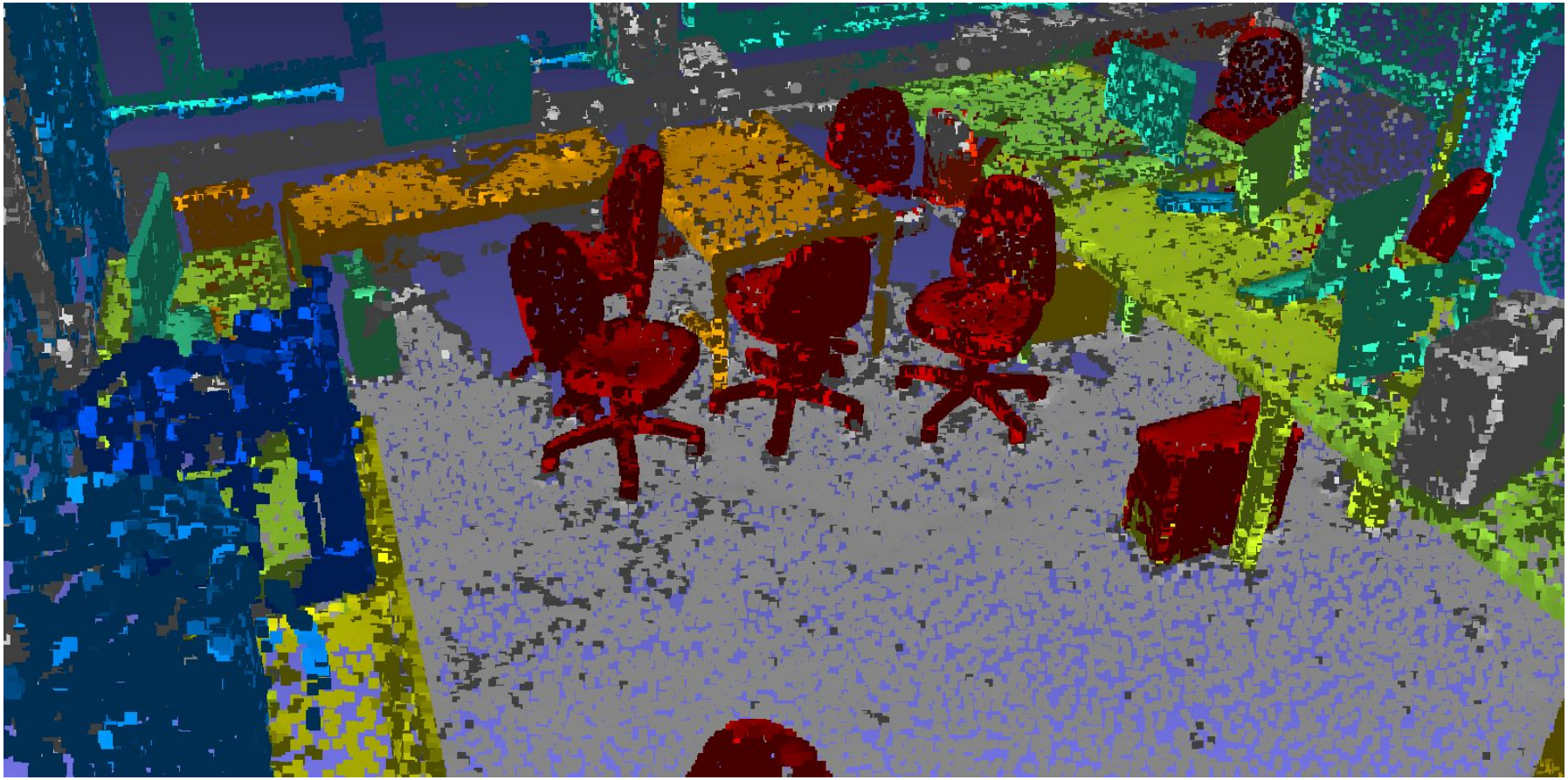
Replacing CLIP by SigLIP

Method	Image Encoder		ImageNet-1k				COCO	
	Size	Patch #	Validation	v2	Real	ObjectNet	I→T	T→I
CLIP	B	196	68.3	61.9	-	55.3	52.4	33.1
OpenCLIP	B	196	70.2	62.3	-	56.0	59.4	42.3
EVA-CLIP	B	196	74.7	67.0	-	62.3	58.7	42.2
SigLIP	B	196	76.3	69.6	82.8	70.7	64.4	47.2
SigLIP	B	256	**76.6**	**70.0**	**83.1**	**71.3**	**65.1**	**47.4**
CLIP	L	256	75.5	69.0	-	69.9	56.3	36.5
OpenCLIP	L	256	74.0	61.1	-	66.4	62.1	46.1
EVA-CLIP	L	256	79.8	75.3	-	72.9	63.7	47.5
SigLIP	L	256	**80.6**	**74.2**	**85.9**	77.9	**69.5**	**51.1**

CLIP



SigLIP



Results

Table 1: Top 5 Views Max Aggregation

	AP	**AP_50%**	**AP_25%**
Base	9.6	13.7	16.8
w/ max aggregation	9.9	14.5	16.9

❓

Table 2: Diverse View Point Selection

	AP	**AP_50%**	**AP_25%**
Base	9.6	13.7	16.8
w/ diverse view selection	9.1	12.8	15.9

❌

Table 3: Top 5 Views Outlier Removal

	AP	**AP_50%**	**AP_25%**
Base	9.6	13.7	16.8
w/ outlier removal	10.5	15.4	18.6

✅

Table 4: SigLip Embeddings

	AP	**AP_50%**	**AP_25%**
Base	9.6	13.7	16.8
w/ SigLip embeddings	**11.1**	**16.6**	**22.4**

✅✅

Quantitative instance segmentation results in [%] on top 5 semantic class matches evaluated on 6 diverse scenes

Improvement 4

Image Queries

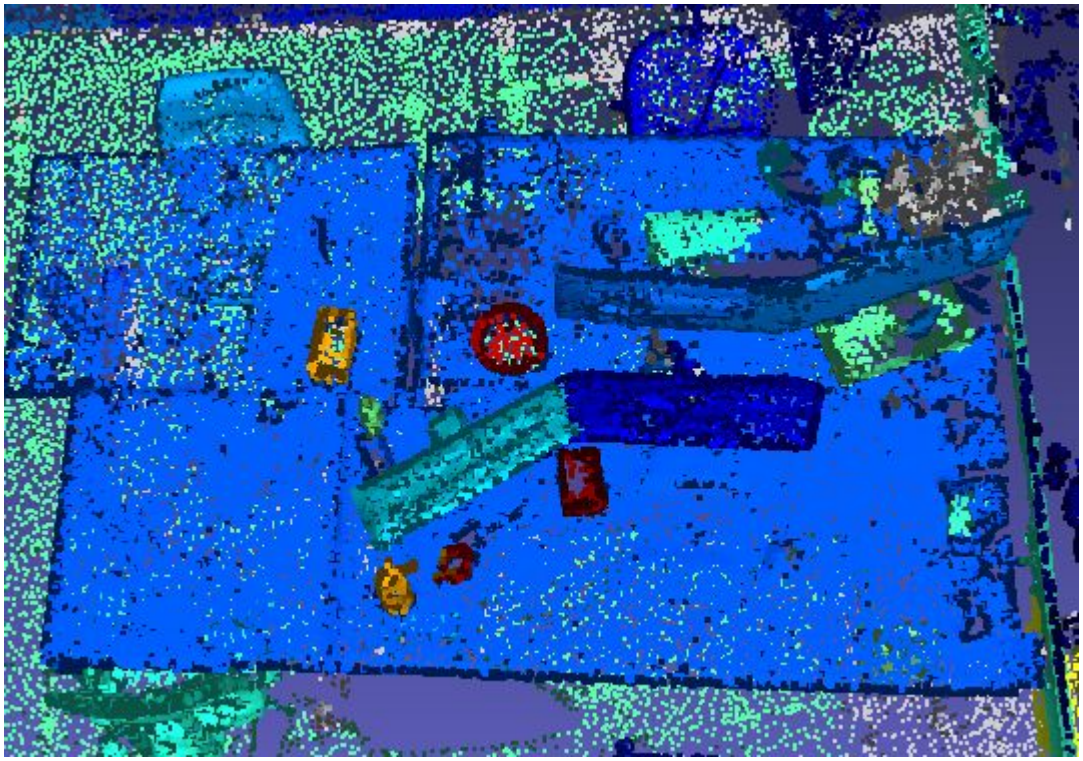
Sometimes objects are quite complex…

https://www.cielo.co.za/coffee-tables/kestara-coffee-table

Open Vocab

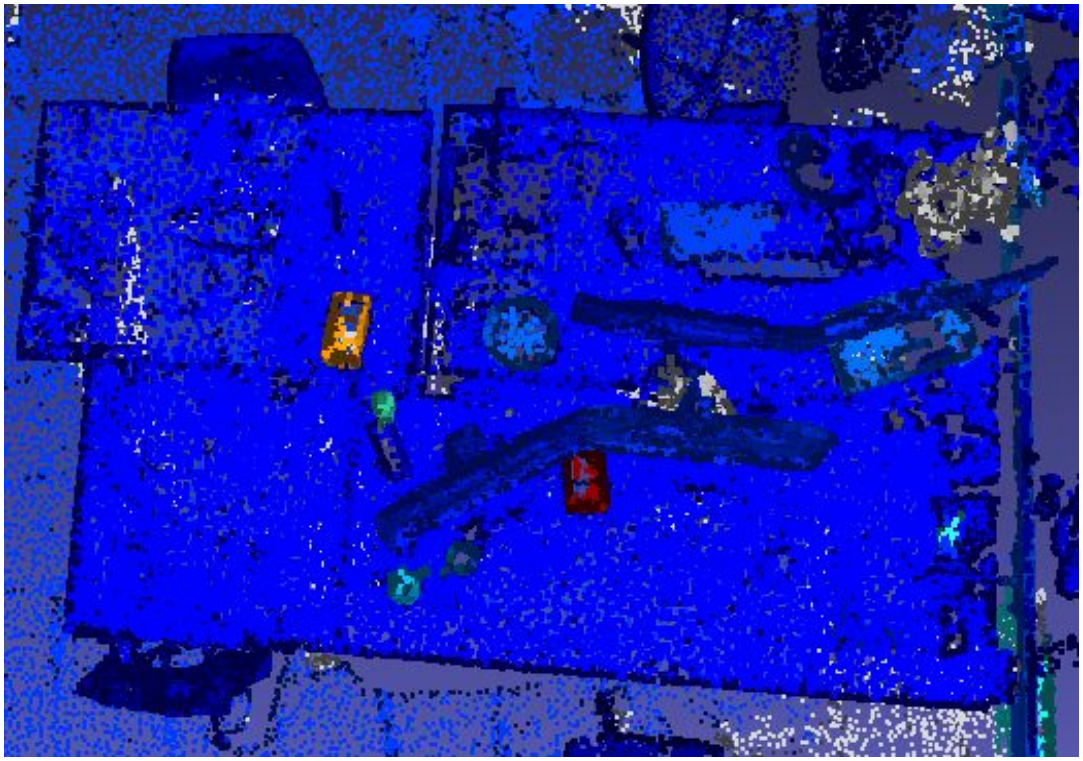
Target:

Text query:
“a tea package”



Open Vocab

DINOv2



[9] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

Thanks!

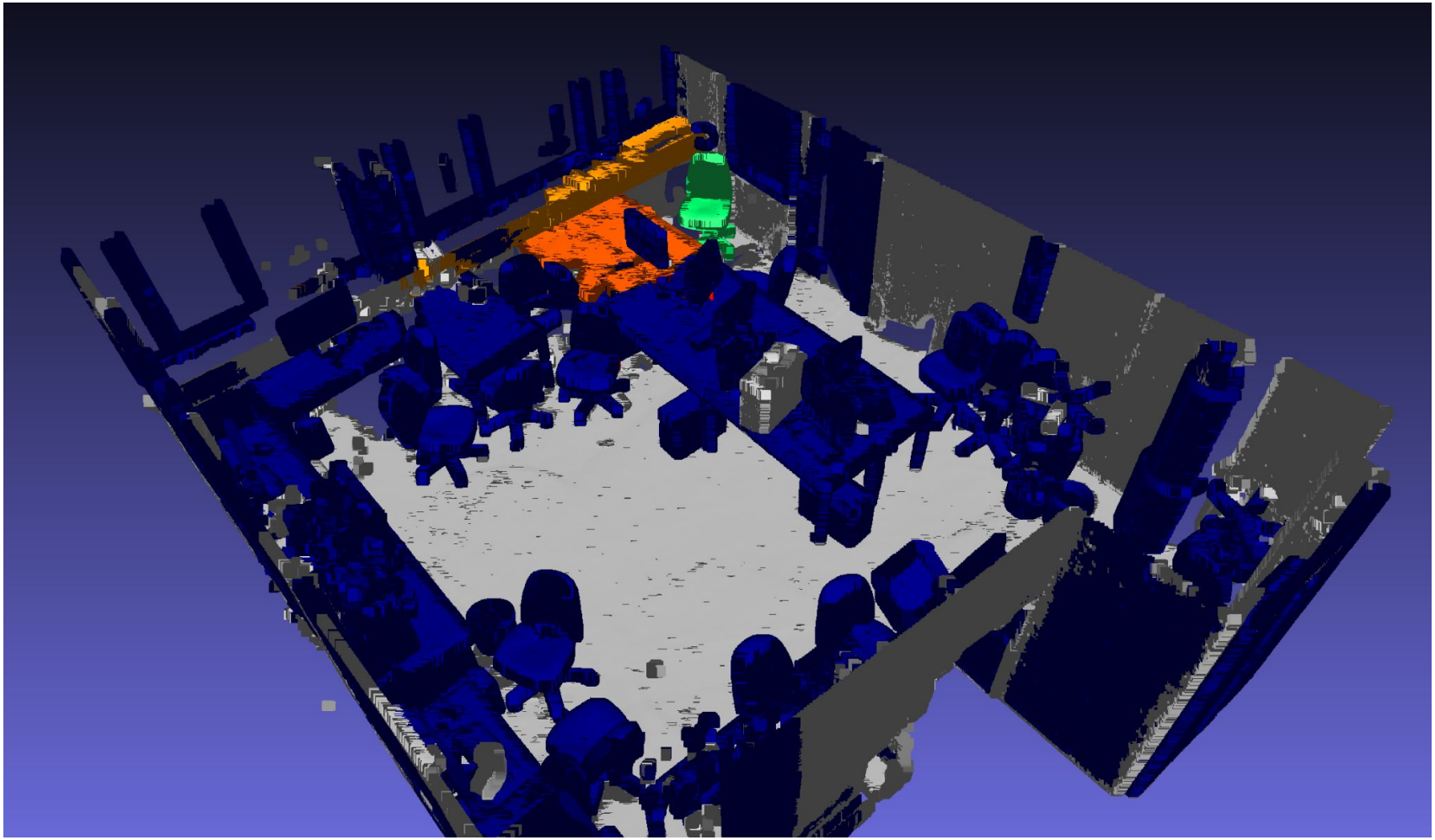
References

- [1] Takmaz, Ayça, et al. "OpenMask3D: Open-Vocabulary 3D Instance Segmentation." *arXiv preprint arXiv:2306.13631* (2023).
- [2] Kirillov, Alexander, et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).
- [3] Li, Feng, et al. "Semantic-sam: Segment and recognize anything at any granularity." *arXiv preprint arXiv:2307.04767* (2023).
- [4] Schult, Jonas, et al. "Mask3d: Mask transformer for 3d semantic instance segmentation." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.
- [5] Yeshwanth, Chandan, et al. "Scannet++: A high-fidelity dataset of 3d indoor scenes." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
- [6] Huang, Rui, et al. "Segment3D: Learning Fine-Grained Class-Agnostic 3D Segmentation without Manual Labels." *arXiv preprint arXiv:2312.17232* (2023).
- [7] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
- [8] Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023)
- [9] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

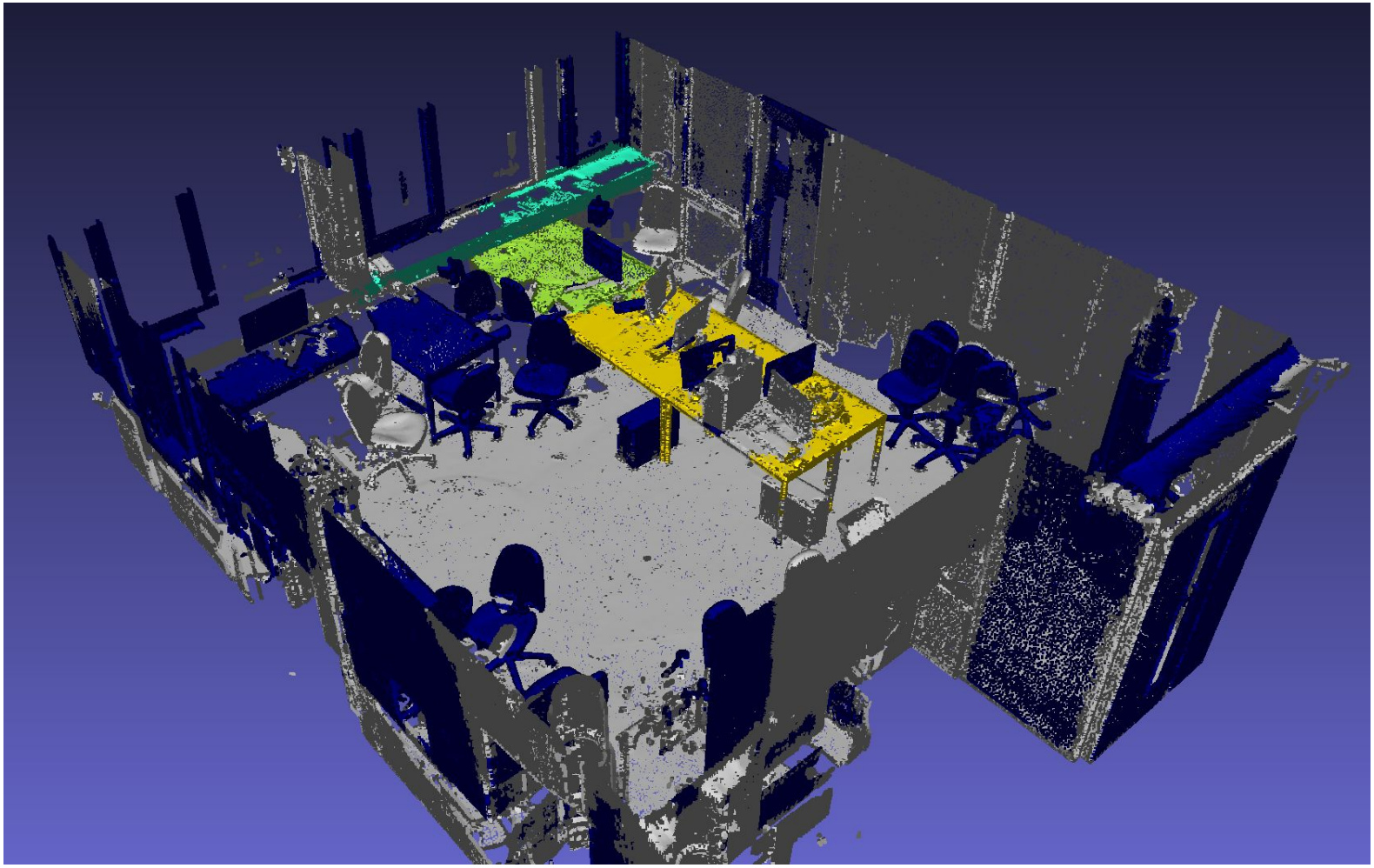
Issue: Box crops

Box crops cannot crop out chairs under tables
→ Tables get flagged as chairs as well

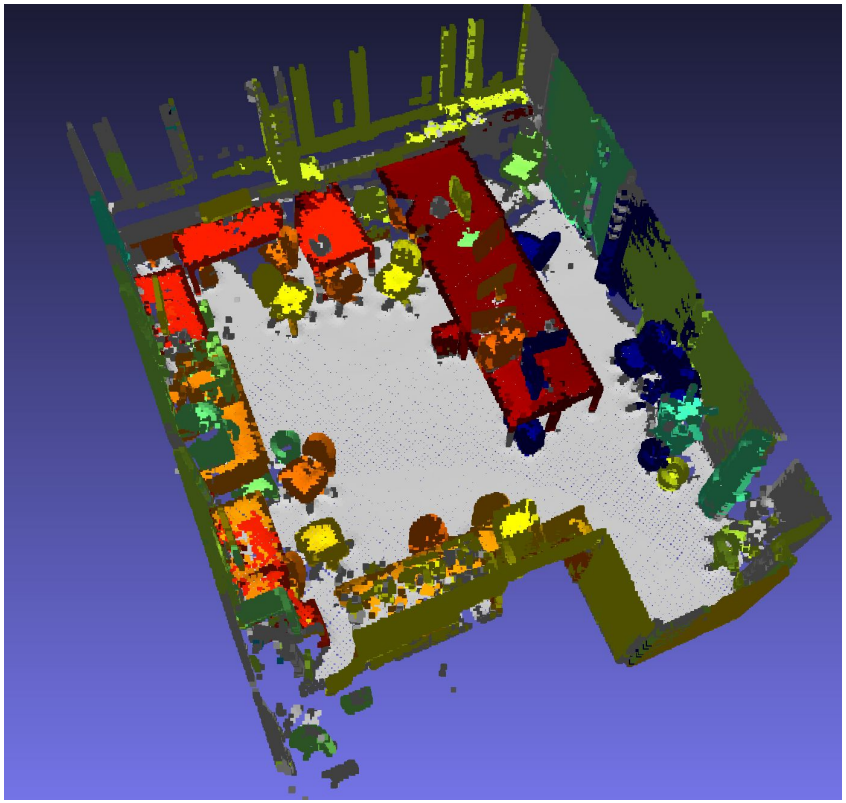
Scannet++ porting issues

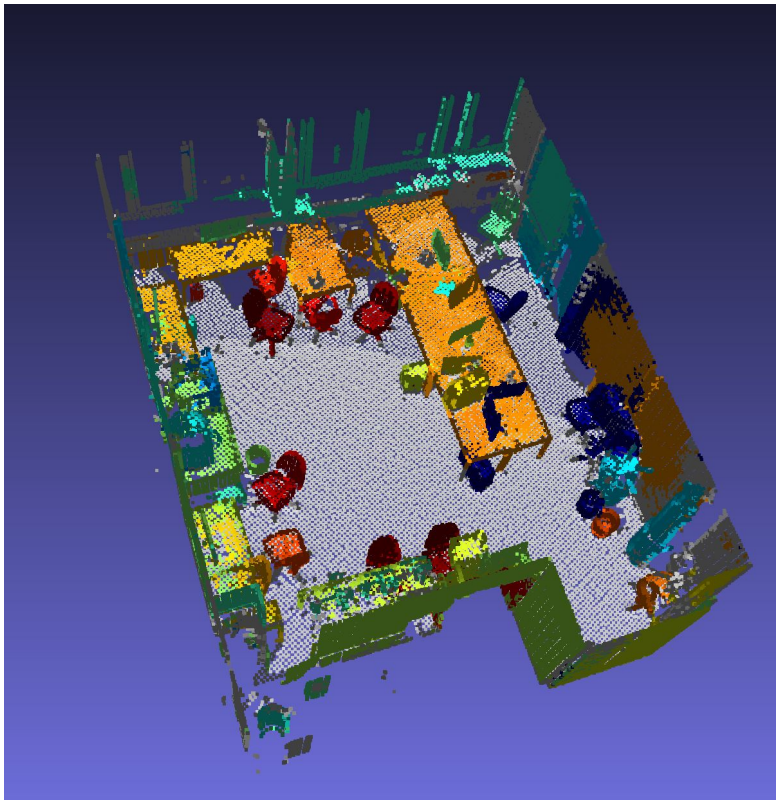


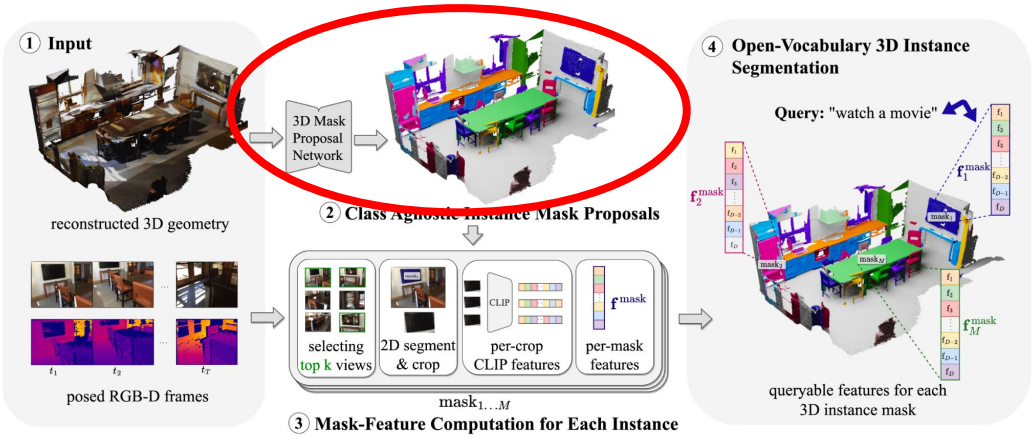
Scannet++ porting issues



JET colormaps





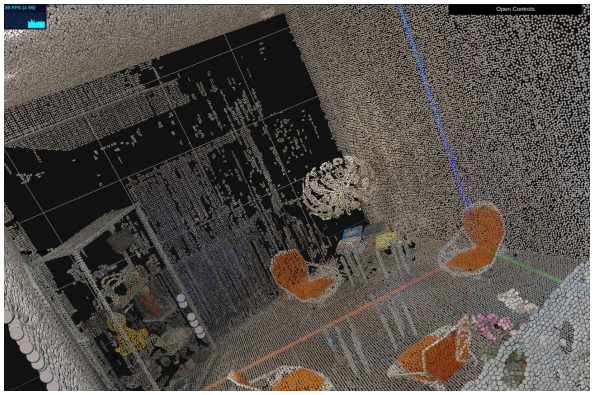


Model	Segmentor	ScanNet++
OpenMask3D [47]	Mask3D [43]	15.0
OpenMask3D [47]	Segment3D (Ours)	17.7 (+2.7)

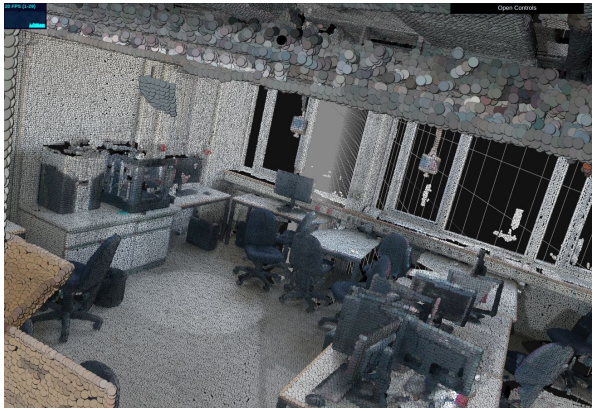
Our selected scenes

c50d2d1d42 (standard office)

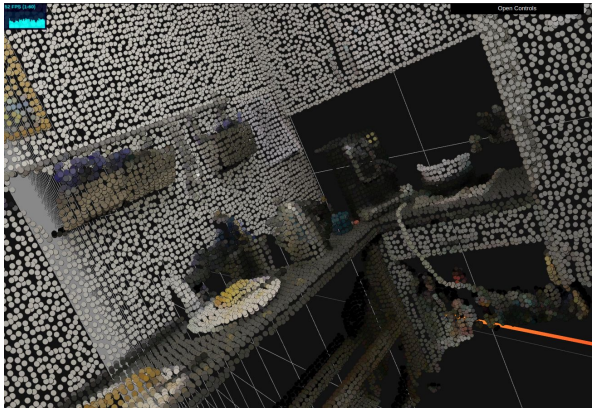
104acbf7d2 (studentenwerk)



45d2e33be1 (apartment)



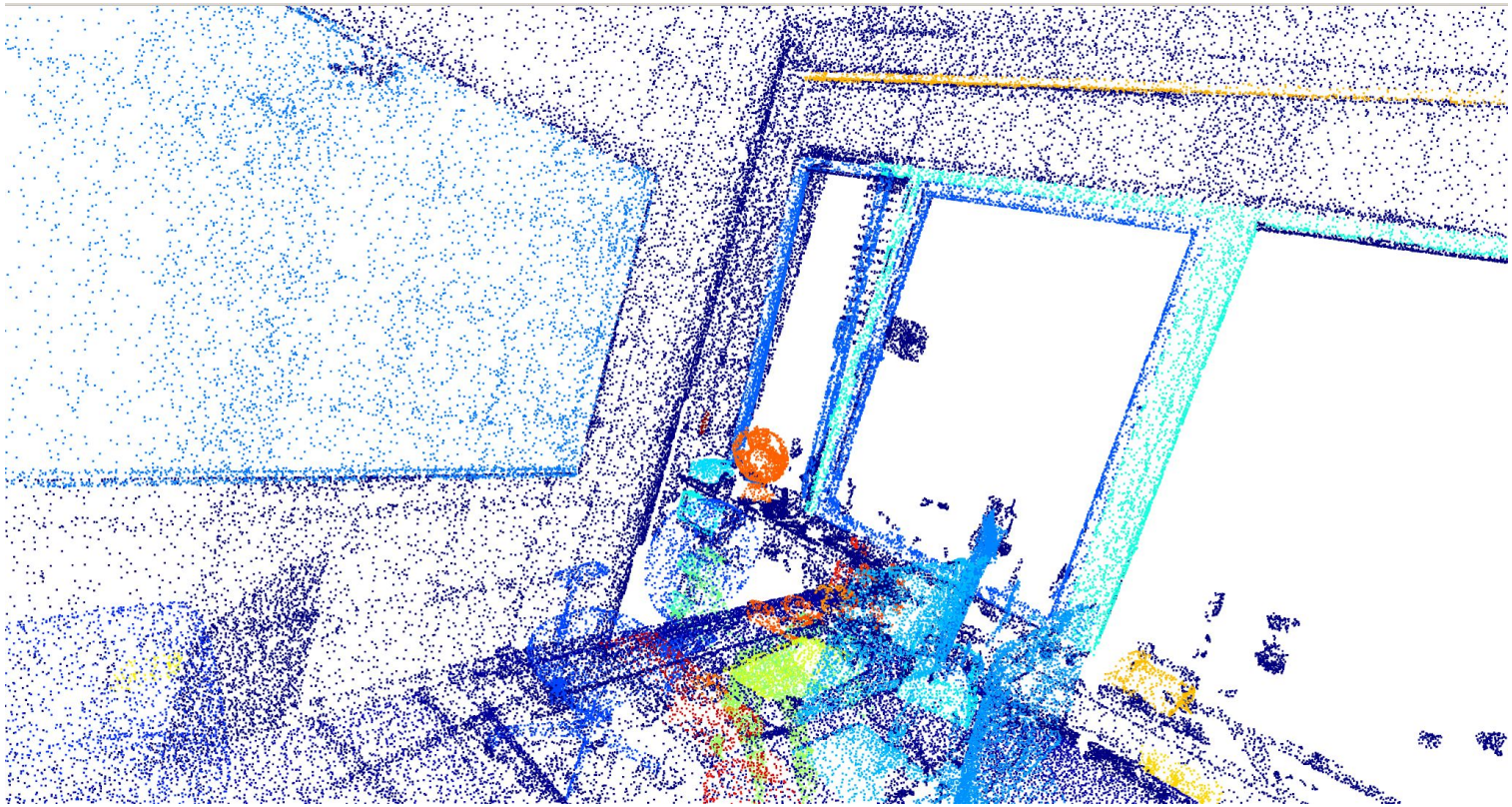
41b00feddb (messy office)



f9f95681fd (messy kitchen)

75d29d69b8 (storage dinosaurs)

downsampled pcl and downsampled gt instances match:



Schrift

Das Grundprinzip ist, Informationen bestmöglich zu transportieren. Dazu muss vor allem die Schrift einheitlich und für alle im Raum lesbar sein.

Schriftart: Arial

Schriftgrößen:30 | 22 | 16 | 12

Zeilenabstand: 1,15mm

Die Einstellungen sind in den Textfeldern und Textfeldvorlagen dieses ppt-Masters als Standard eingestellt. Bei Diagrammen und Tabellen muss die Schriftgröße ggf. angepasst werden. Für Auszeichnungen im Fließtext kann auch **fett** markiert werden.
Bei großer Distanz bzw. kleinem Präsentationsmedium kann der Schriftgrad notfalls proportional erhöht werden.

Farben

Als erstes soll mit schwarz und weiß gearbeitet werden.
Für Aufwändigere Darstellungen sind Farben mit Bedacht und in möglichst geringem Umfang einzusetzen.

In diesem Folienmaster ist die Farbpalette festgelegt.

Zuerst mit den Primärfarben arbeiten.

Für z.B. komplexe Diagramme stehen nochSekundärfarben zur Verfügung.

Gering im Einsatz sind die Akzentfarben.

Texte

Kurze und knappe Texte, Fließtexte linksbündig, kein Blocksatz

Beispiel:
Tem soluptam, nisi as verum ereprehendam at acculpa quidisq uissit volupta tusdant utem as etur, odi odis es doluptiae dem nimaion con nossinctenis pora quam voloria consenimus blabore everfer epeliquo maio etur.

Aufzählung

Bei kleinen Aufzählungen auf Aufzählungszeichen verzichten und ggf. zusätzliche Leerzeile
Nur die wesentlichen Punkte nennen und Themen auf verschiedene Seiten splitten.

Punkt 1

Punkt 2

Wenn Unterpunkte in einer Aufzählung nötig sind ist ein Einrücken mit – möglich

- Unterpunkt 1
  - Unterpunkt 1
  - Unterpunkt 2

Bei größeren Listen die Standardeinstellung • verwenden

- Unterpunkt 1
- Unterpunkt 2
- Unterpunkt 3

Bilder - Allgemein

schlichte Darstellung von Informationen

reduzierte Farben

Rahmen und Überlagerungen nach Möglichkeit vermeiden

Bilder

Bildbeschreibung
oberer Bildrand: Begrenzung durch Text

Bilder

Überschrift 2
Hier steht ein einleitender oder
beschreibender Fließtext und nach Wunsch
eine Aufzählung

Punkt 1

Punkt 2

Punkt 3

Punkt 4

Bilder

Bildbeschreibung
oberer Bildrand: Begrenzung durch Text

Bilder

Bildbeschreibung
oberer Bildrand: Begrenzung durch Text

Nicht formatfüllende Bilder

Weißer bzw. transparenter Hintergrund
mit genug Freiraum anordnen

Bilder Format füllend - maximale Bildgröße

Nicht Format füllende Bilder

Alternativ mit formatfüllendem Hintergrund: 5 % schwarz
Beschriftungen können zusätzlich neben den Bildern angebracht werden

Bilderklärung

Tabelle – Beispiel 1

Tabelle ohne Farbe und kein Rand
innerer Seitenrand links 0 cm, oben z.B. 0,5 cm (für genug Zeilenabstand innerhalb)

Ø - Strecke	39 km/Tag (14.360 km/Jahr)
Ø - Geschwindigkeit	25 km/h
Ø - Verfügbare Ladezeit	22 h/Tag
Kosten	Kleinwagen mit Verbrennungsmotor
Einsatzgebiet	Stadt und Umland

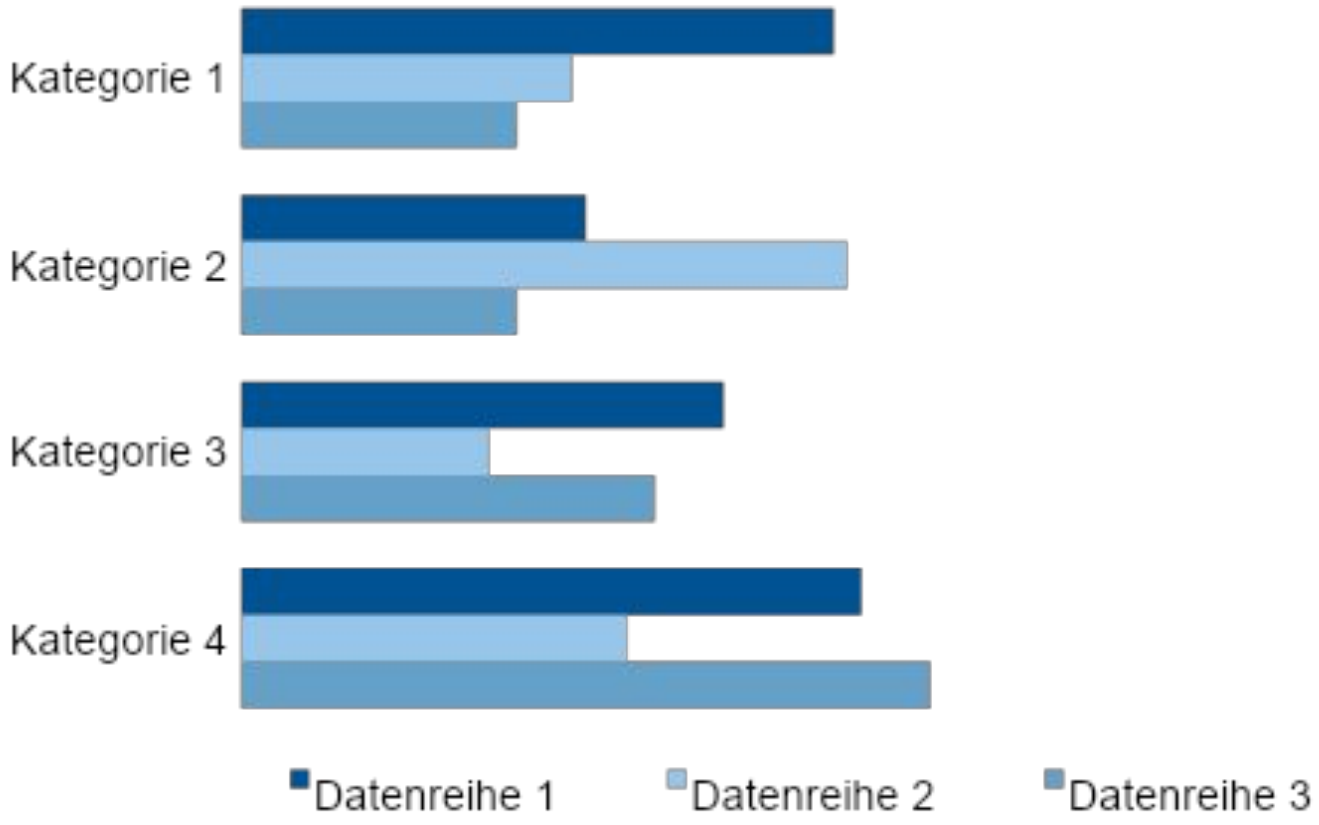
Tabelle – Beispiel 2

Tabelle mit schwarzem Rand
innerer Seitenrand links 0,15 cm, oben z.B. 0,5 cm (für genug Zeilenabstand innerhalb)

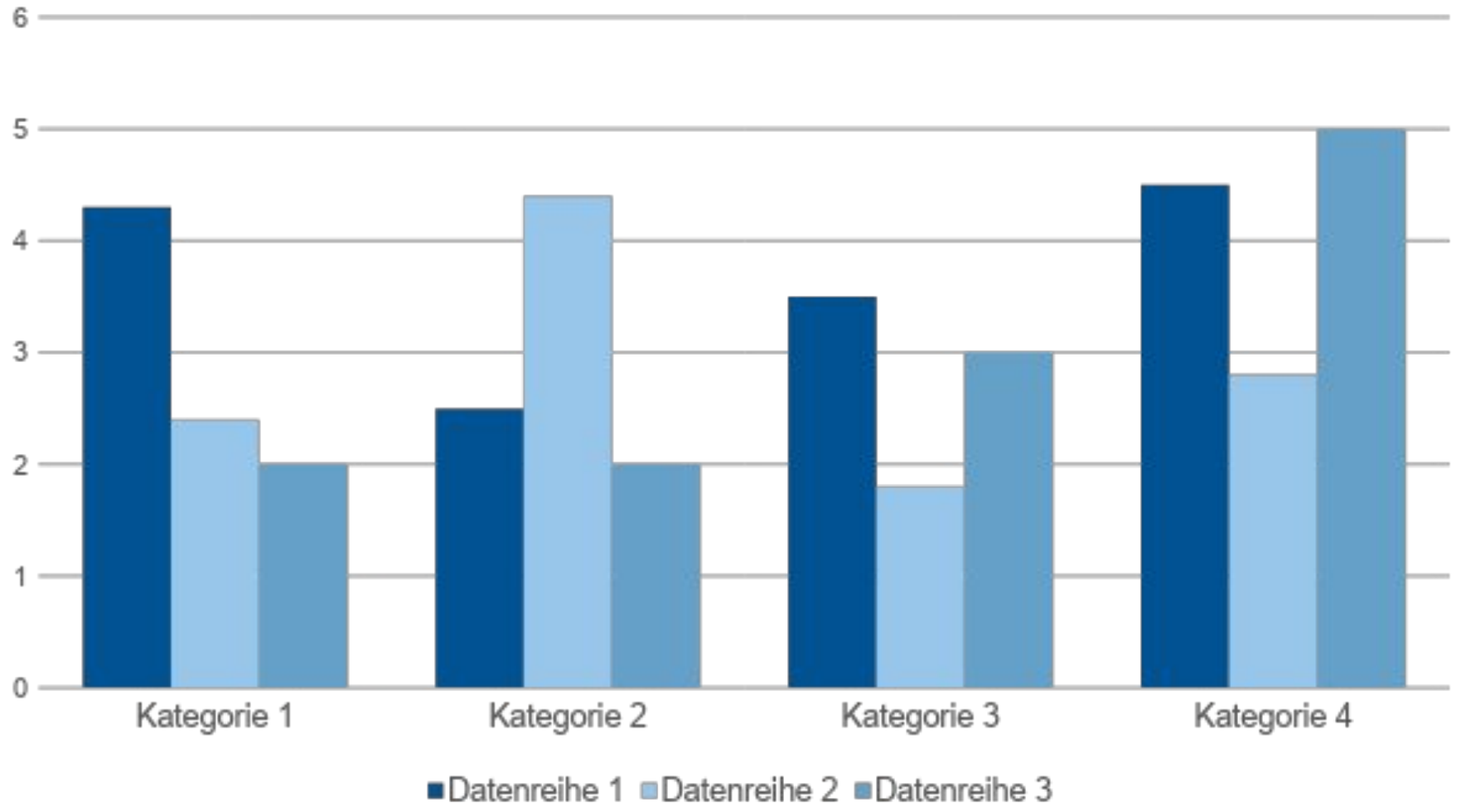
Ø - Strecke	39 km/Tag (14.360 km/Jahr)
Ø - Geschwindigkeit	25 km/h
Ø - Verfügbare Ladezeit	22 h/Tag
Kosten	Kleinwagen mit Verbrennungsmotor
Einsatzgebiet	Stadt und Umland

Diagramme – Beispiel 1

Nach Möglichkeit linksbündig bleiben
Unnötige Striche und Balken vermeiden



Diagramme


6
5
4
3
2
1
0
Kategorie 1
Kategorie 2
Kategorie 3
Kategorie 4
Datenreihe 1
Datenreihe 2
Datenreihe 3