Supplemental material

In this supplemental material, we provide additional qualitative results and insights on the design choices dealing with the creation of our dataset. **We would like to highlight that, in case of acceptance of the paper, SiM3D will be publicly released along with the codebase created to implement the baselines and compute the performance evaluation metrics, so as to stimulate further research concerning the open challenges.**

S.1. Threshold employed for background removal

As anticipated in Sec. 4.2 of the main paper, in Tab. A we report the values of the threshold distance τ employed to determine the inliers of the consensus set and the offset α used to shift the fitted plane. The classes *Wooden Stool* and *Sink Cabinet* have not been filtered since the slight amount of background points does not provide enough support points to fit a plane with the employed algorithm.

Class	τ	α
Plastic Stool	30	-2
Rubbish Bin	30	-2
Wicker Vase	60	10
Bathroom Furniture	20	10
Container	30	2
Plastic Vase	60	2
Wooden Stool	–	–
Sink Cabinet	–	–

Table A. **Background removal τ and α values.**

S.2. Details on the defect distribution of the test set

In Tab. B we report several statistics on the distribution of the defects on the test set, such as the number of test instances, the number of anomalies and the mean number of anomalies per defective instance. Even though each defect is inherently multimodal, they can be distinguished as: *(i)* 3D, such as dents and bumps, characterised by significant structural variation but minimal changes in the visual appearance of the defective area; *(ii)* 2D, such as scratches and marker strokes, in which we have a negligible structural variation and a significant deviation in the appearance of the defective area; *(iii)* multimodal, such as cracks and contaminations, in which both structural and appearance variations can be appreciated.

Furthermore, we report in Fig. A the distribution of the size (in voxels) of the defects present in the test set of SiM3D. The distribution highlights the predominance of smaller anomalies, which renders the benchmark particularly challenging.

Class	Test Instances		No. Anomalies				Mean Anomalies
	Nominal	Anomalous	Tot.	3D	2D	Multimodal	per Instance
Plastic Stool	10	10	32	11	18	3	3.2
Rubbish Bin	20	20	48	8	30	10	2.4
Wicker Vase	10	10	14	1	3	10	1.4
Bathroom Furniture	8	10	32	0	24	8	3.2
Container	46	46	104	4	67	33	2.2
Plastic Vase	48	49	62	20	27	15	1.2
Wooden Stool	6	7	34	9	21	4	4.8
Sink Cabinet	9	8	27	4	15	8	3.3

Table B. **Test set defects statistics.**

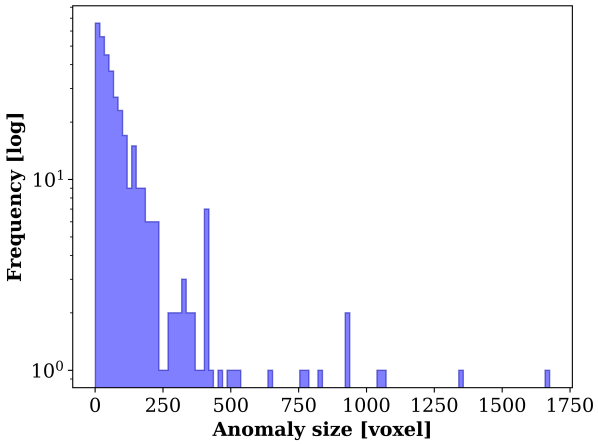

Figure A. **Anomaly size distribution.** The proposed dataset contains predominantly smaller anomalies, making the overall benchmark particularly challenging.

S.3. Additional details on the calibration procedure

The proposed pipeline deploys a custom procedure to estimate the transformation between the reference frame in which the Atos Q sensor provides the point cloud associated with a scan, from now on Sensor Reference Frame (SRF), and the Camera Reference Frame of the left camera (CRF). More in detail, we acquire several (i.e., 20) views, i.e., images and associated point clouds, of a high-precision dot pattern provided by ZEISS as part of the calibration toolkit of the Atos Q sensor (see Fig. 3 (2) in the main paper). Then, for each pair of images and point clouds, we apply the following steps:

(i) Detect and refine the centres of the dots from the image by a classical circle detection algorithm;

(ii) apply the Perspective-n-Points (PnP) algorithm between these 2D centres in the CRF and the known 3D coordinates of the centres expressed in the 3D reference frame attached to the calibration pattern to find the roto-translation between the CRF and the 3D reference frame attached to the calibration pattern;

(iii) roto-translate the 3D centres of the dots – expressed in the calibration pattern reference frame – into the CRF, exploiting the transformation between the calibration pattern and the camera previously estimated by PnP;

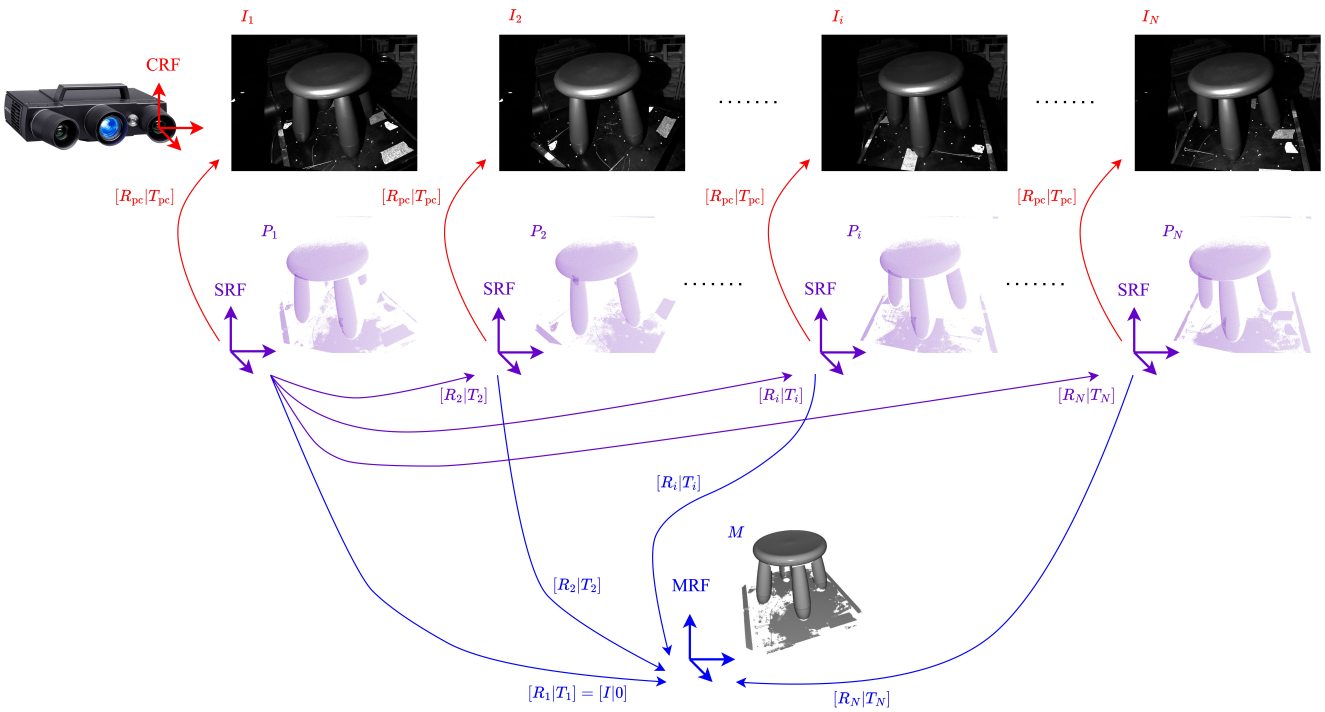

Figure B. **Reference frames.**

(iv) use the software tools provided with the Atos Q to get the 3D coordinates of the centres of the dots into the SRF;

(v) finally, given all the 3D-3D correspondences between the centres of the dots in the SRF and the CRF obtained from the different views of the pattern, we apply an absolute orientation algorithm (i.e., Kabsch-Umeyama) to estimate the roto-translation from the SRF to the CRF, i.e. $[R_{\text{pc}}|T_{\text{pc}}]$.

The accuracy of the estimation process can be assessed by computing the ℓ_2 norm of the difference between the coordinates of centres brought from the SRF into the CRF by $[R_{\text{pc}}|T_{\text{pc}}]$ and those brought from the pattern reference frame into the CRF via the transformation obtained by PnP This assessment is performed on an additional set of views with respect to those used to estimate $[R_{\text{pc}}|T_{\text{pc}}]$, resulting in a precision under 1 mm.

In Fig. B we depict the different reference frames and the relations between them. We highlight that, with the Atos Q sensor, the Mesh Reference Frame (MRF) is aligned in the 3D space to the SRF associated with the first point cloud acquired while performing a 360-degree scanning of an object, hence $[R_1|T_1] = [I|0]$.

S.4. Additional details on the labelling procedure

As anticipated in Sec. 3, the labelling process involves transferring 2D annotations – manually created on images – to the 3D integrated mesh. This is achieved by projecting the 3D mesh vertices into the 2D image space and associating each vertex with the ID colour (label information) carried by the corresponding pixel. Thus, the mesh vertices are first transformed by the inverse of the roto-translations $[R_i|T_i]$ between the mesh and the considered views. Then, they are brought into the CRF by employing $[R_{\text{pc}}|T_{\text{pc}}]$. Finally, the vertices' coordinates expressed in the CRF are projected onto the image plane of the 2D annotations by the intrinsic parameters of the camera, A. The corresponding pixel coordinates in the image are identified for each vertex, their ID colour (i.e. label) is extracted, and the information is assigned to the associated vertices of the 3D mesh, thereby lifting in 3D the annotations created on the original 2D inputs. Finally, the 3D mesh with vertex colours representing the annotations is manually refined.

S.5. Additional details on pre-processing of 3D data

In Fig. C we show the procedure to obtain the depth maps $\{D_i\}_{i=1}^n$, or organized point clouds $\{P_i\}_{i=1}^n$, employed in the experiments described in Sec. 5. In particular, after the raw singleview point clouds are acquired through a whole scan of the object, the ZEISS software shipped with the Atos Q integrates them to obtain a comprehensive mesh. Subsequently, we remove the background from this mesh with the procedure described in Sec. 4, obtaining a filtered mesh. Afterwards, by the knowledge of the roto-translation $[R_{\text{pc}}|T_{\text{pc}}]$ between the CRF and the SRF as well as those be-

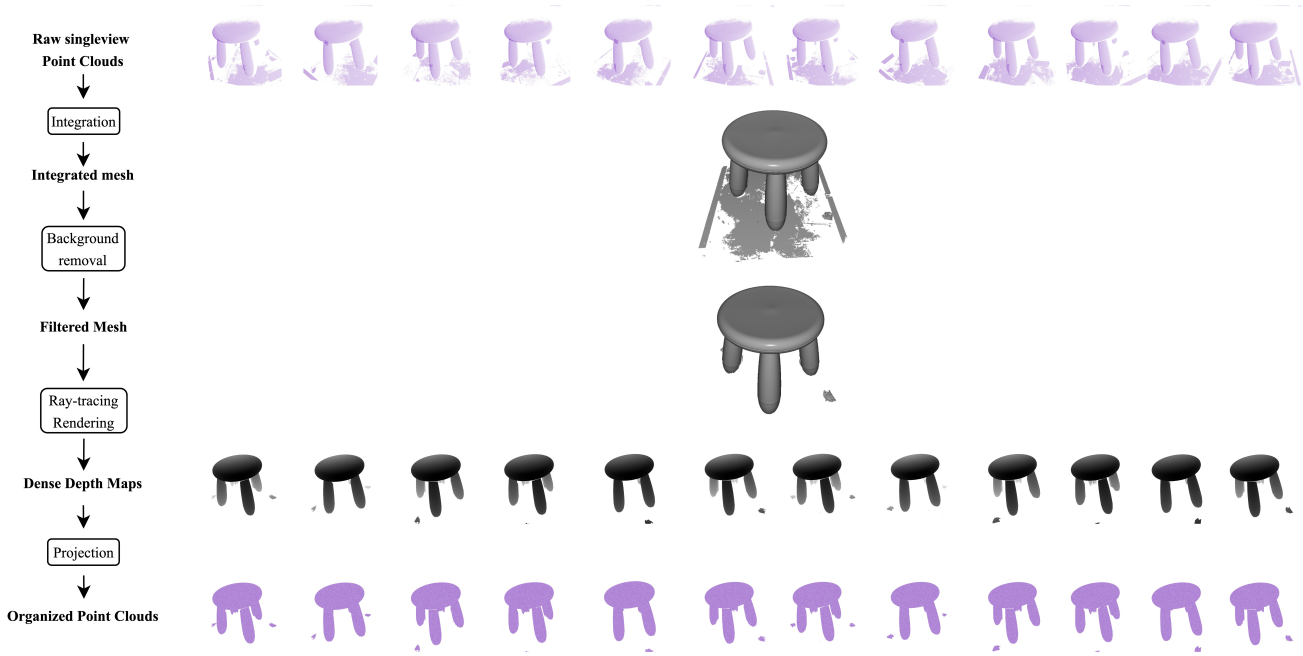

Figure C. **3D Data pre-processing.**

tween the the roto-translations and the mesh, $[R_i|T_i]$ (see Fig. B), we can render the depth maps $\{D_i\}_{i=1}^n$ associated with each view by ray-tracing the cleaned mesh. Finally, to obtain clean and organized point clouds $\{P_i\}_{i=1}^n$, which are pixel-aligned to the corresponding greyscale images $\{I_i\}_{i=1}^n$, we can simply reproject the pixel coordinates in 3D via corresponding rendered depths and the inverse of the camera matrix, A.

S.6. Additional details on the baselines

For the following baselines, we process greyscale images $\{I_i\}_{i=1}^n$, depth maps $\{D_i\}_{i=1}^n$ or organized point clouds $\{P_i\}_{i=1}^n$ downsampled to the common size of $1540 \times 1540 \times 3$, the highest achievable with the available hardware.

Input data processed with WideResNet101, either greyscale images or depth maps, returns features from the second and third layers, yielding feature maps with dimensions equal to $28 \times 28 \times 1536$. Input data processed with DINO-v2, either greyscale images or depth maps, return features from the last layer, yielding feature maps with dimensions equal to $110 \times 110 \times 768$. Input data processed with FPFH features, namely point clouds, are processed considering a voxel size equal to 2 mm in order to match the resolution attainable from the 3D voxel ground-truths and yield feature maps with dimensions equal to $1540 \times 1540 \times 33$. The spatial resolution of these maps is then downsampled to match either the spatial resolution of the WideResNet101 or DINO-v2 features, hence 28×28 or 110×110.

Unlike the common practice, the 2D Anomaly Maps obtained by the methods are not Gaussian-blurred since the subsequent 3D aggregation and discretisation introduce smoothing.

PatchCore. PatchCore [30] is a singleview, image-based ADS method that employs WideResNet101 to extract features from training data, which are subsequently stored in a memory bank. During inference, features from test samples are queried against this memory bank to compute an anomaly score.

We adapted this method on SiM3D, by either processing greyscale images or depth maps, using either its original feature extractor, WideResNet101, or the DINO-v2 backbone. We created the coreset with a 10% coverage and 0.9 projection radius and selected 3 as a reweight parameter for the anomaly map computation.

EfficientAD. Efficient [1] is a singleview, image-based ADS method that employs a Teacher–Student paradigm based on patch description networks paired with an autoencoder pre-trained on WideResNet101.

We implement EfficientAD by disabling the Teacher normalisation since there is no validation set available, and we upsample the intermediate encoder outputs to match the size of the Teacher and the Students fed with $1540 \times 1540 \times 3$ images. Moreover, we train the Students for 1000 epochs, unlike the 70000 expected from the adopted implementation, since the loss tends to stall earlier due to the limited number of training images that characterise our single-instance setup.

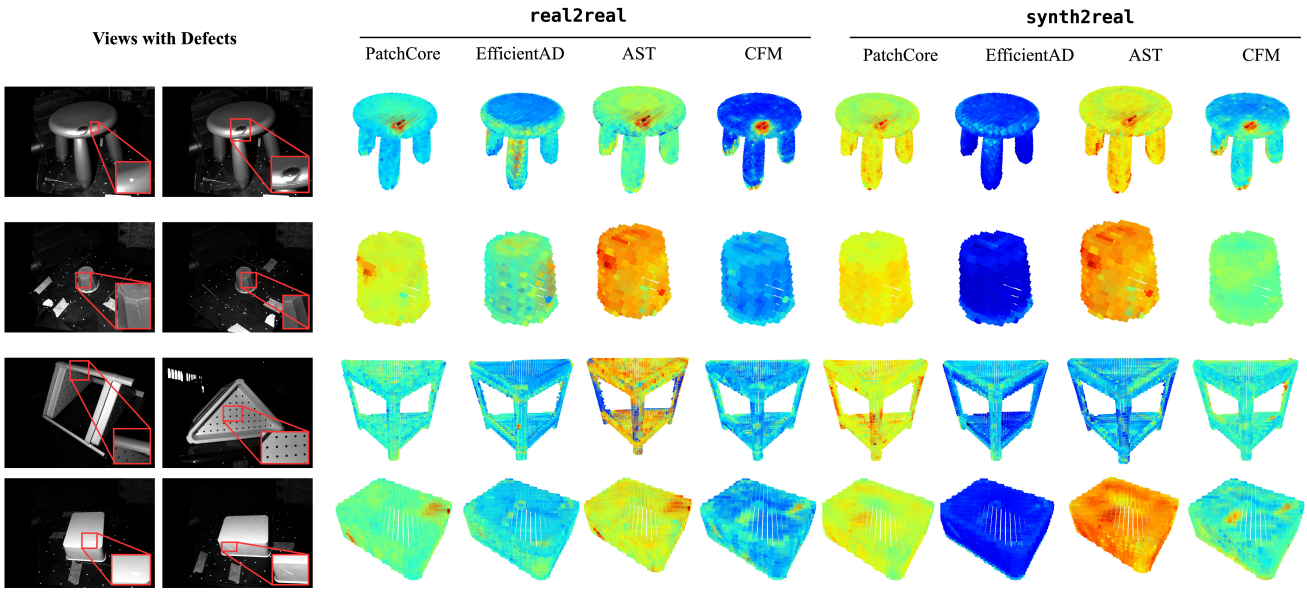

Figure D. **Qualitatives.** Views with defects (left) and Anomaly Volumes for several methods (right) downsampled for visualisation.

Back to the Features. BTF [18] is a singleview, multimodal ADS method that, similarly to PatchCore, employs WideResNet101 to extract 2D features from RGB images, and relies on FPFH to extract 3D features from point clouds. The 2D and 3D features are subsequently concatenated and stored in a single memory bank. During inference, 2D and 3D features from test samples are also concatenated and queried against this memory bank to compute an anomaly score.

We also implement BTF by using WideResNet101 to process both grayscale images and depth maps. We created the coreset with a 10% coverage and 0.9 projection radius and selected 3 as a reweight parameter for the anomaly map computation.

Crossmodal Feature Mapping. CFM [14] is a singleview, multimodal ADS method that exploits MLPs to map features from one modality to the other on nominal samples and then detect anomalies by pinpointing inconsistencies between observed and mapped features. This solution leverages DINO-v1 and Point-MAE to extract features from RGB images and point clouds, respectively.

Since SiM3D contains high-resolution images and point clouds, we adopt DINO-v2 and FPFH as feature extractors. Alternatively, to reduce computational demands, we also implement CFM by using DINO-v2 to process both greyscale images and depth maps. We trained the crossmodal feature mappings for 50 epochs, following the original implementation, with a unitary batch size to limit memory consumption.

Multi-3D-Memory. M3DM [36] is a singleview, multimodal ADS method that employs DINO-v1 and Point-MAE to extract features from RGB images and point clouds, which are subsequently stored in memory banks. Moreover, it also learns a function to fuse 2D and 3D features into multimodal features, which are then stored in memory banks alongside those computed from the individual modalities. During inference, features from test samples are queried against the memory banks to compute anomaly scores, which are then aggregated with One-Class SVMs.

Given that SiM3D contains high-resolution images and point clouds, we adopt DINO-v2 and FPFH as feature extractors. To reduce computational demands, we also implement M3DM by using DINO-v2 to process both greyscale images and the rendered depth maps. Furthermore, we disabled the feature fusion module due to computational limitations introduced by the high-resolution features. Following the original implementation, we created both coresets with a 10% coverage and 0.9 projection radius and selected 1 as a reweight parameter for the image-based anomaly map and 0.1 as a reweight parameter for the point cloud-based anomaly map. Moreover, both One-Class SVMs are trained with a ν parameter equal to 0.5 and a maximum number of SGD iterations fixed to 1000.

Asymmetric Student-Teacher. AST [32] is a singleview, multimodal ADS method that employs EfficientNet-B5 to extract features from RGB images and depth maps. Such features are subsequently employed to optimise a Normalizing Flow as Teacher network and a feed-forward network as a Student network. The idea is that, after optimisation, both networks are able to reconstruct nominal samples, begetting low discrepancies, while failing to reconstruct anomalous samples. Since these two networks present different archi-

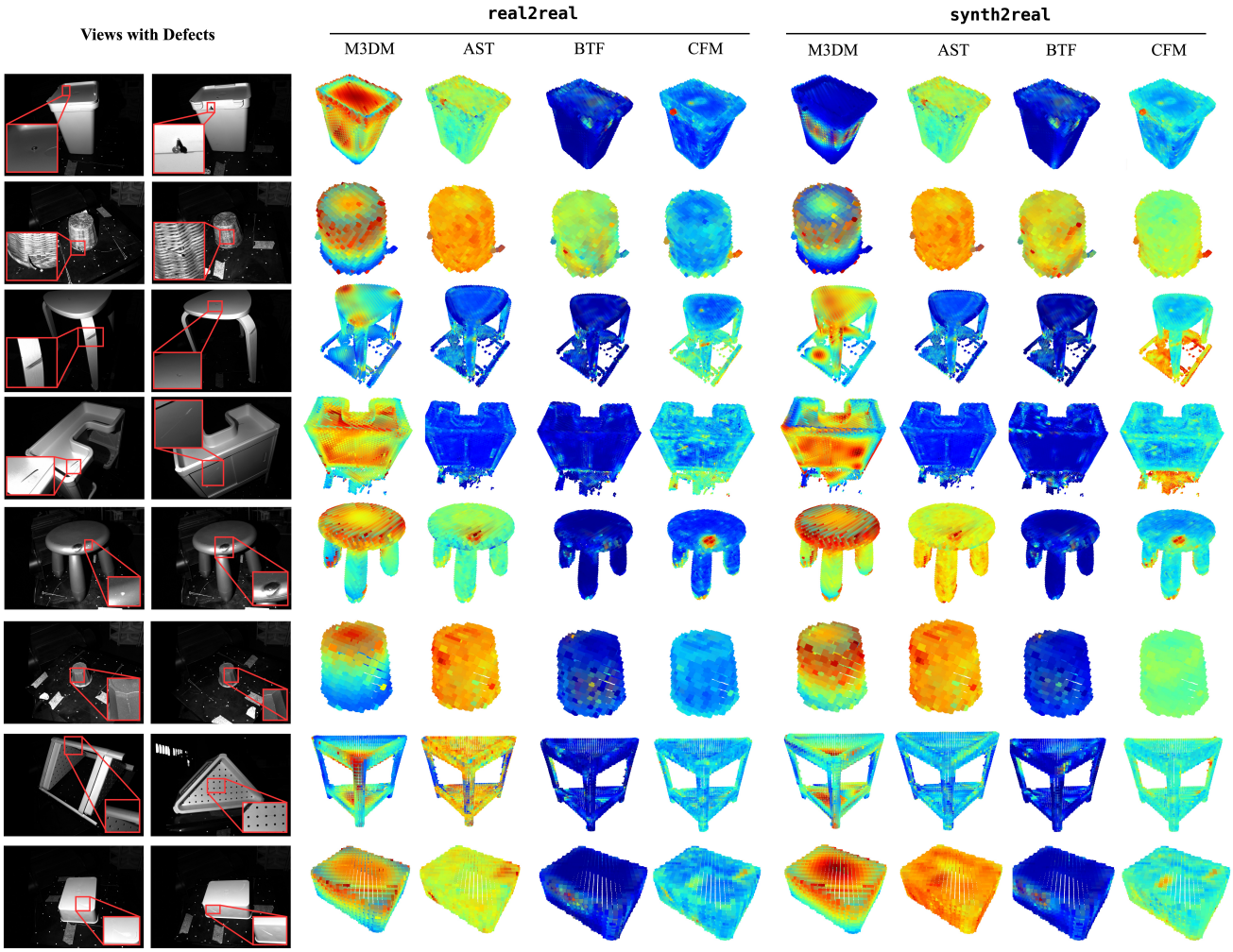

Figure E. **Multimodal methods qualitatives.** Views with defects (left) and Anomaly Volumes for all multimodal methods (right) downsampled for visualisation.

tectures, the way in which they fail to reconstruct anomalous samples is different, hence discrepancies will be exacerbated, highlighting anomalies. Due to the fact that this method work on the features, it is triviality extendable to multiple modalities.

Given that SiM3D contains high-resolution images and point clouds, we deployed AST by passing to the framework $1540 \times 1540 \times 3$ images and depth maps.

S.7. Assessment of data generation quality and insight of failure cases of the `synth2real` setup

We report in Tab. C the Fréchet Inception Distance (FID) between the train set, either from the `real2real` or `synth2real` setup, and the test set, along with the performance gaps (Δ I-AUROC, Δ V-AUPRO) for PatchCore. We do not observe any correlation between Δ FID and Δ I-AUROC or Δ V-AUPRO. Instead, in the `synth2real` scenario, we noticed strong outliers in anomaly maps of **nominal samples** (see Fig. F), for some objects. These outliers have a strong impact on detection since they increase the false detection rate, while affecting the segmentation performance much less.

S.8. Ancillary experiments to assess the SiM3D challenges

As suggested during the peer-reviewing process, we ran the following experiments to gain more insight into the unique challenges set forth by SiM3D:

- We treated the task as single-view ADS by deploying PatchCore, obtaining a mean I-AUROC of 0.488 vs. 0.754 of our multiview approach, **highlighting the necessity of addressing the task in a multiview fashion**. Notice that we cannot compare the segmentation performance since the ground-truths are voxel grids;

	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	**Mean**
FID `real2test`	10.088	8.844	7.345	15.784	12.590	9.560	8.910	23.750	12.108
FID `synth2test`	48.744	44.565	57.302	40.900	51.900	55.194	24.040	34.060	44.588
Δ FID	38.656	35.721	49.957	25.116	39.310	45.634	15.130	10.310	32.480
Δ I-AUROC (PatchCore)	0.458	0.061	0.223	0.158	0.371	0.121	0.393	0.476	0.282
Δ V-AUPRO (PatchCore)	0.044	0.012	0.003	0.229	0.061	0.019	0.114	0.115	0.075

Table C. **Data quality assessment and impact on performance.** The average Fréchet Inception Distance between the train samples and the test samples for both `real2real` and `synth2real` setups is reported.

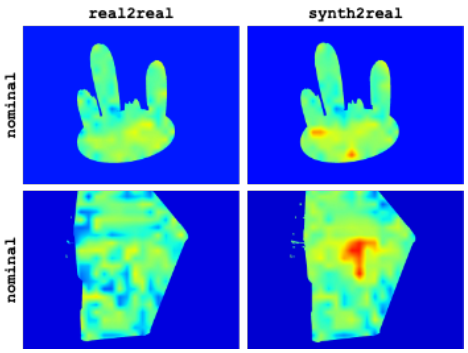

Figure F. **Outliers.** The same test views inferred with the same model, trained either with `real2real` or `synth2real` train set, highlights the presence of outliers in the `synth2real` scenario.

- We ran the official implementation of PatchCore on grayscale MVTec AD, obtaining a mean I-AUROC of 0.988 vs. 0.991, and a mean AUPRO@30% of 0.929 vs. 0.935, both official results from [30], **highlighting that benchmark complexity does not stem from grayscale images**.

As for metrics, I-AUROC is directly comparable to other benchmarks; hence, the **general lower performance confirms that SiM3D is more challenging**. Segmentation performance cannot be compared since the ground-truths of SiM3D are voxel grids.

S.9. Assessment of baselines on 2D and multimodal anomalies

As suggested during the peer-reviewing process, we performed a comparative analysis of some baselines while disentangling the different kinds of anomalies. To this end, starting from Tab. B, we split the anomalies into 2D and 3D+multimodal anomalies, hence, unifying 3D defects with the multimodal ones. Indeed, we argue that, while it is possible to unambiguously consider some anomalies, such as marker strokes and white-outs, as 2D-only, it is more difficult to pinpoint 3D-only anomalies, as structural deviations in the geometry, more often than not, tend to manifest themselves also in the image space.

After that, we selected the best-performing algorithm working on images (PatchCore w/ DINO-v2) and multimodal data (AST), The experiments, reported in Tab. D, show that PatchCore, the image-based method, tend to perform best on 2D anomalies and worse on multimodal anomalies, obtaining an overall weaker performance when all the anomalies are considered. On the other hand, AST, the multimodal method, tend to perform well on both 2D and multimodal anomalies, obtaining an overall stronger performance when all the anomalies are considered. These findings further confirm the necessity of a multimodal analysis when working in the SiM3D setup.

S.10. Attribution of existing assets

We adapted the baselines described in the main paper starting from the following codebases:

- PatchCore: `https://github.com/eliahuhorwitz/3D-ADS` released under MIT License;
- EfficientAD: `https://github.com/nelson1425/EfficientAD` released under Apache License 2.0;
- BTF: `https://github.com/eliahuhorwitz/3D-ADS` released under MIT License;
- M3DM: `https://github.com/nomewang/M3DM` released under MIT License;
- CFM: `https://github.com/CVLAB-Unibo/crossmodal-feature-mapping` released under non-commercial use only license;
- AST: `https://github.com/marco-rudolph/AST` released under MIT License.

S.11. Ethical statement

This research, which was carried out to produce this dataset, adheres to ethical principles and practices in computer vision research. The dataset introduced in this study does not contain any personally identifiable information or sensitive data. It has been collected and processed in a manner that respects individual privacy and avoids potential biases. The dataset and its intended use align with the ethical guidelines outlined by the CVPR community. We have taken care to ensure that the dataset and its potential applications do not

Method	Modality	Anomalies	Detection									Segmentation								
			Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	**Mean**	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	**Mean**
PatchCore w/ DINO-v2	RGB	All	0.500	0.958	0.636	0.622	0.578	0.563	1.000	0.563	0.678	0.745	0.469	0.775	0.792	0.709	0.753	0.435	0.690	0.671
PatchCore w/ DINO-v2	RGB	2D	0.646	0.902	0.602	0.723	0.678	0.498	0.980	0.612	0.705	0.832	0.498	0.674	0.801	0.718	0.687	0.455	0.678	0.667
PatchCore w/ DINO-v2	RGB	Multimodal	0.425	0.573	0.534	0.439	0.592	0.632	0.754	0.587	0.567	0.276	0.354	0.456	0.548	0.698	0.603	0.404	0.596	0.491
AST w/ EffNet-B5	RGB + Depth	All	0.950	0.927	0.785	0.474	0.542	0.470	0.428	0.925	0.687	0.750	0.503	0.792	0.807	0.716	0.764	0.467	0.798	0.699
AST w/ EffNet-B5	RGB + Depth	2D	0.910	0.905	0.773	0.427	0.539	0.320	0.423	0.892	0.648	0.654	0.443	0.730	0.687	0.708	0.797	0.303	0.679	0.625
AST w/ EffNet-B5	RGB + Depth	Multimodal	0.945	0.895	0.698	0.410	0.498	0.543	0.413	0.904	0.663	0.738	0.475	0.765	0.789	0.893	0.683	0.564	0.723	0.703

Table D. **Anomaly detection and segmentation results considering different kinds of anomalies.**

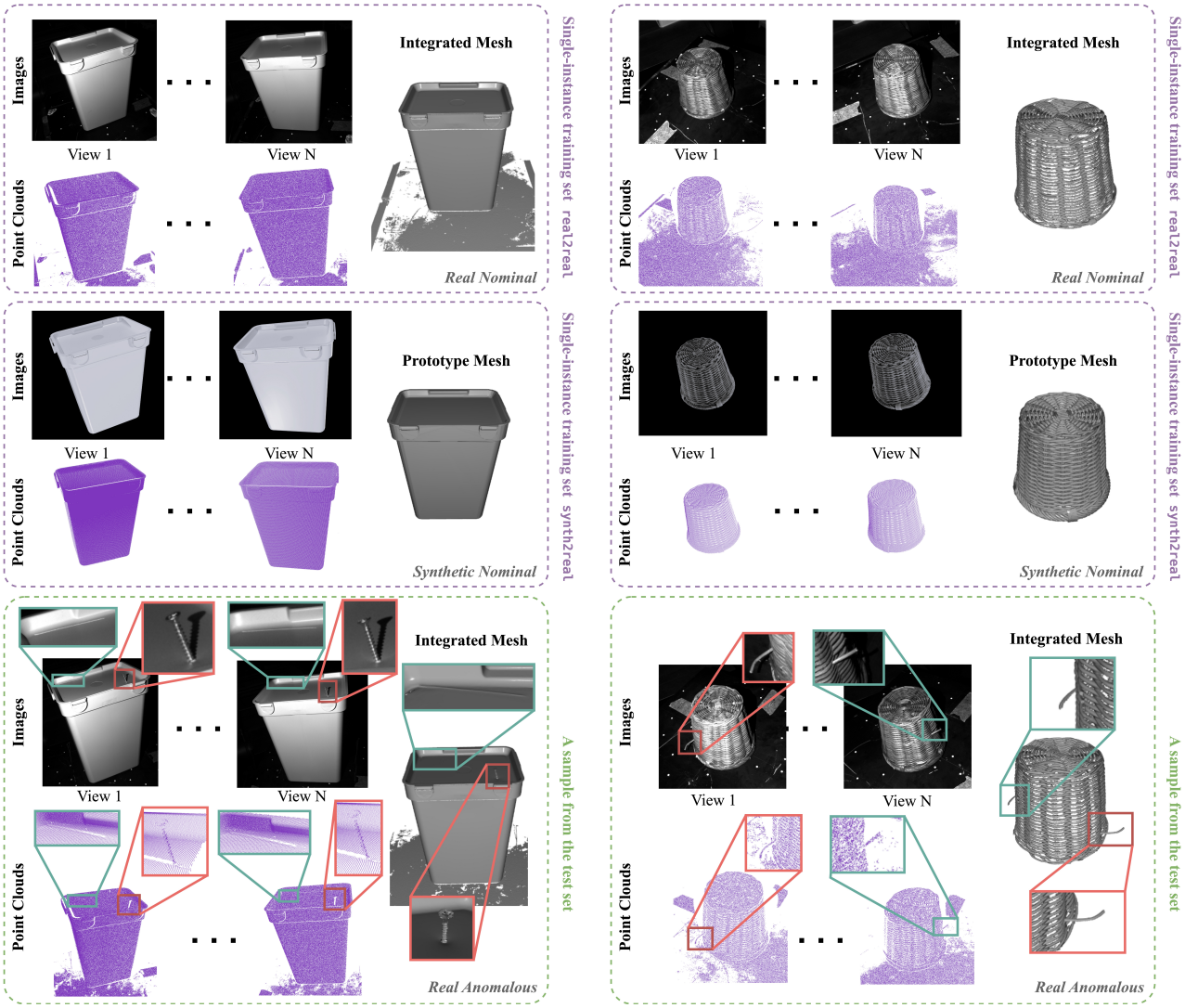

Figure G. **SiM3D dataset overview.** From top to bottom: the single-instance real and synthetic training samples for objects *Rubbish Bin* and *Wicker Vase*, one of the anomalous samples from the test set.

pose significant risks to individuals or society.

S.12. Additional qualitative results

We report in Fig. D additional qualitative results concerning the classes of SiM3D which have not been displayed in in Fig. 5. We additionally report in Fig. E qualitative results for top-performer multimodal methods reported in Tab. 5.

S.13. Extended dataset visualizations

Akin to Fig. 1 of the main paper, in Fig. G, Fig. H, and Fig. I, we show training samples, both real and synthetic, as well as defective test samples for other object types present in the SiM3D dataset.

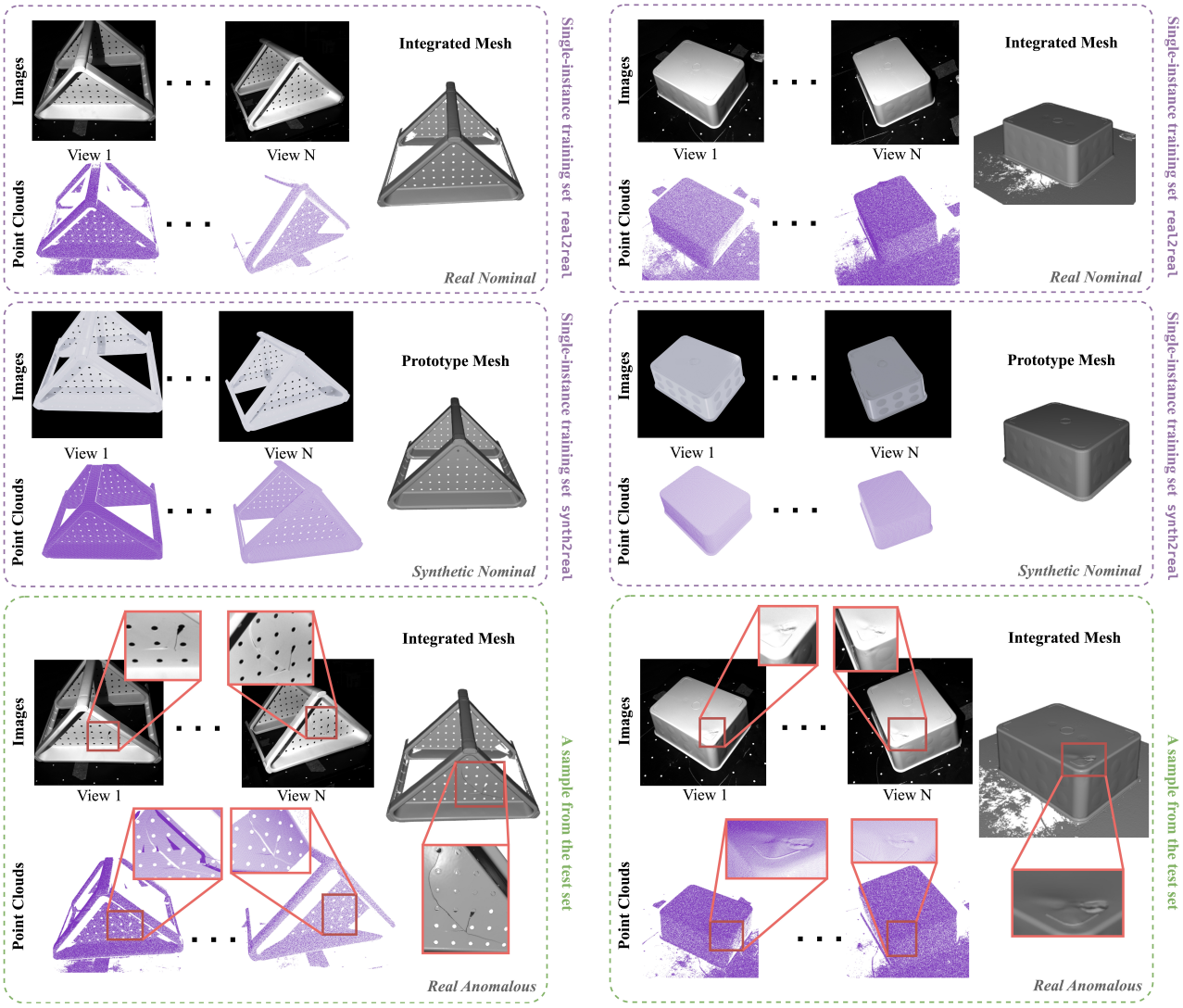

Figure H. **SiM3D dataset overview.** From top to bottom: the single-instance real and synthetic training samples for objects *Bathroom Furniture* and *Container*, one of the anomalous samples from the test set.

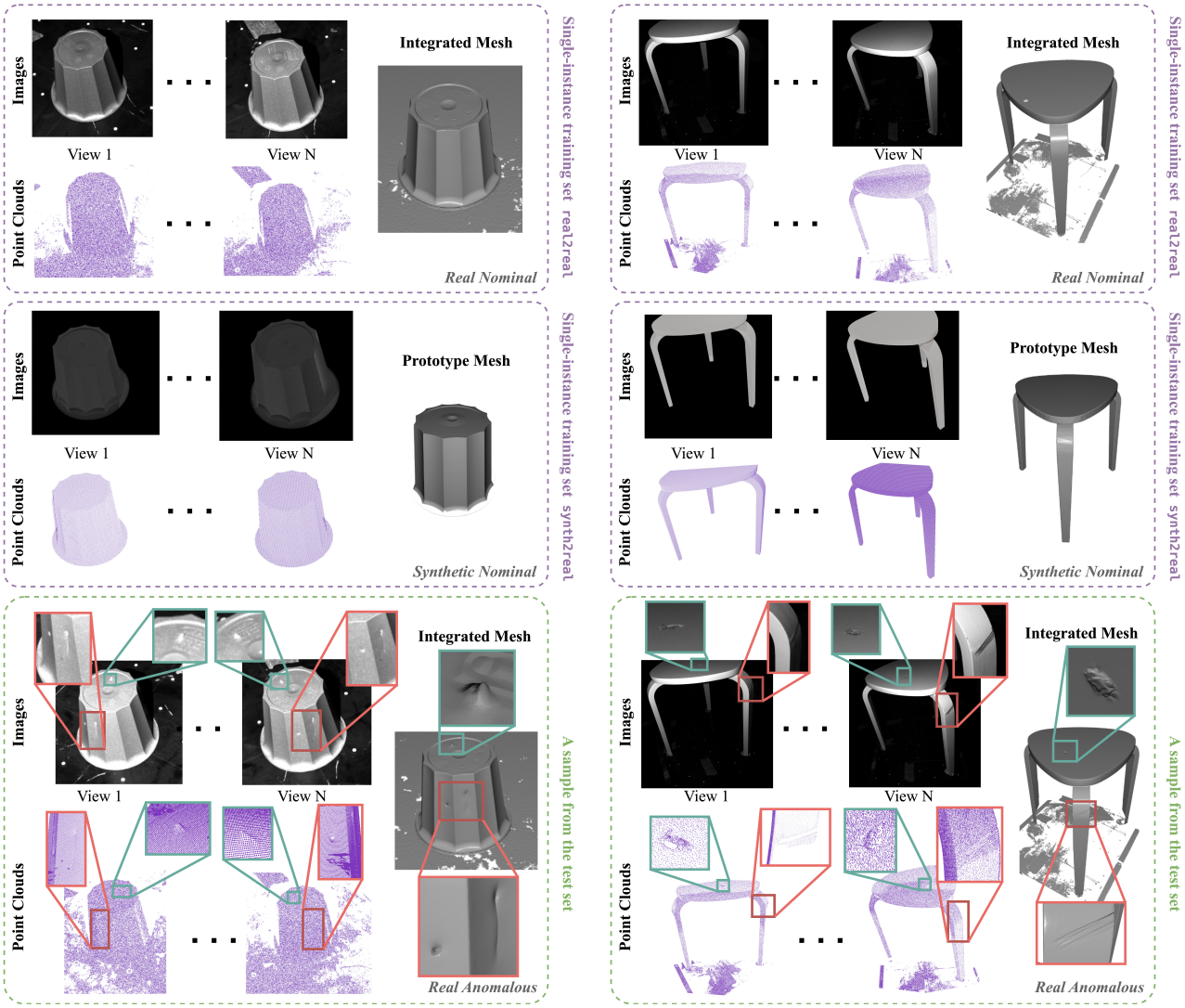

Figure I. **SiM3D dataset overview.** From top to bottom: the single-instance real and synthetic training samples for objects *Plastic Vase* and *Wooden Stool*, one of the anomalous samples from the test set.