JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

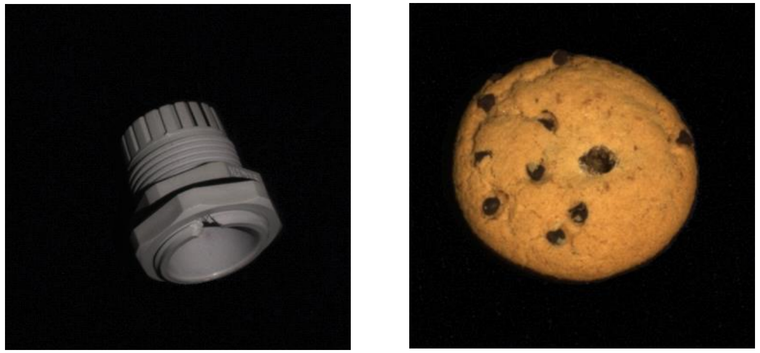
Multimodal Industrial Anomaly Detection via Hybrid Fusion

Yue Wang[1], Jinlong Peng[2], Jiangning Zhang[2], Ran Yi[1], Yabiao Wang[2], Chengjie Wang[21]

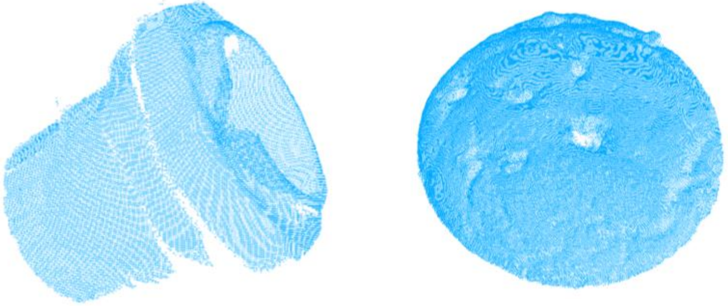
[1]Shanghai Jiao Tong University, Shanghai, China; [2]Youtu Lab, Tencent

TUE-PM-374

From 2D to 3D Anomaly Detection



2D RGB data

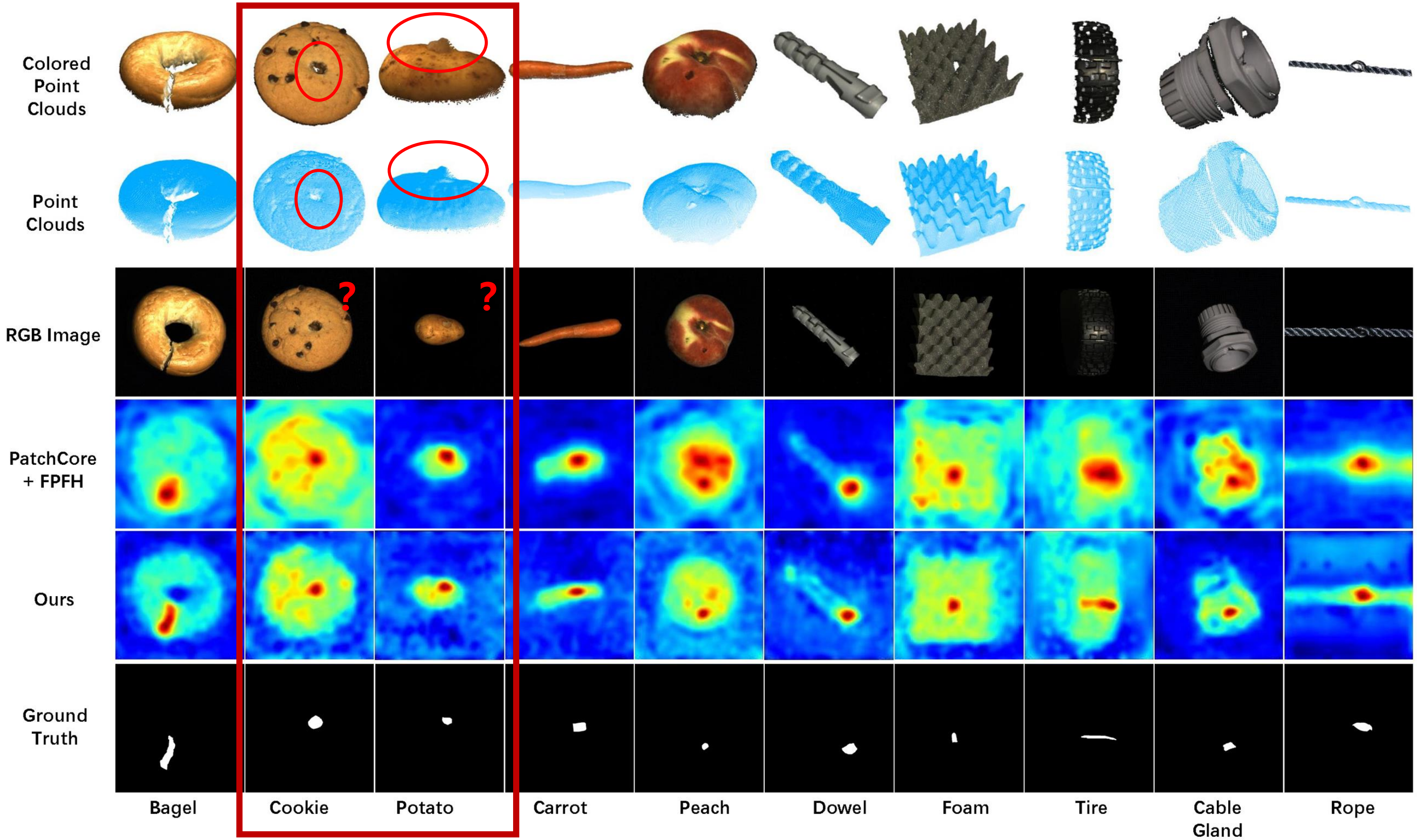


Point Clouds data

Colored Point Clouds data

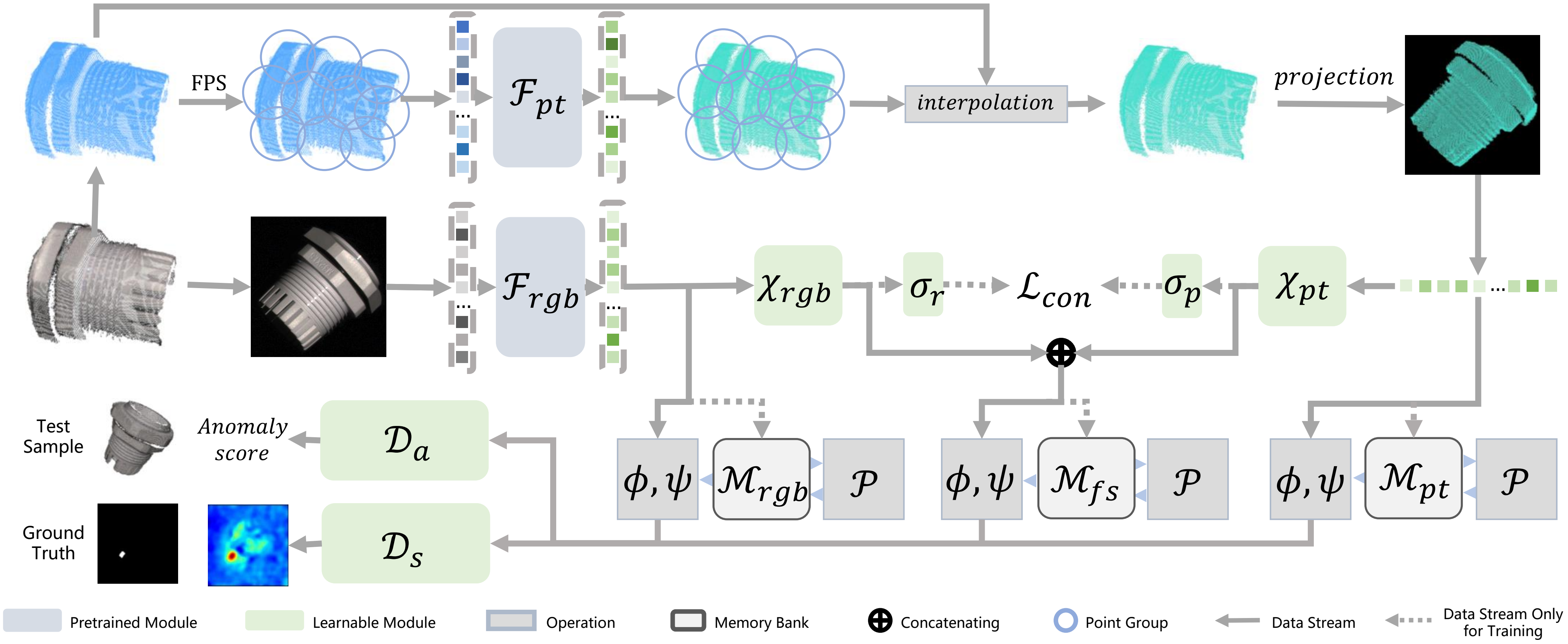
From 2D to 3D Anomaly Detection

- Results on MVTec-3D AD



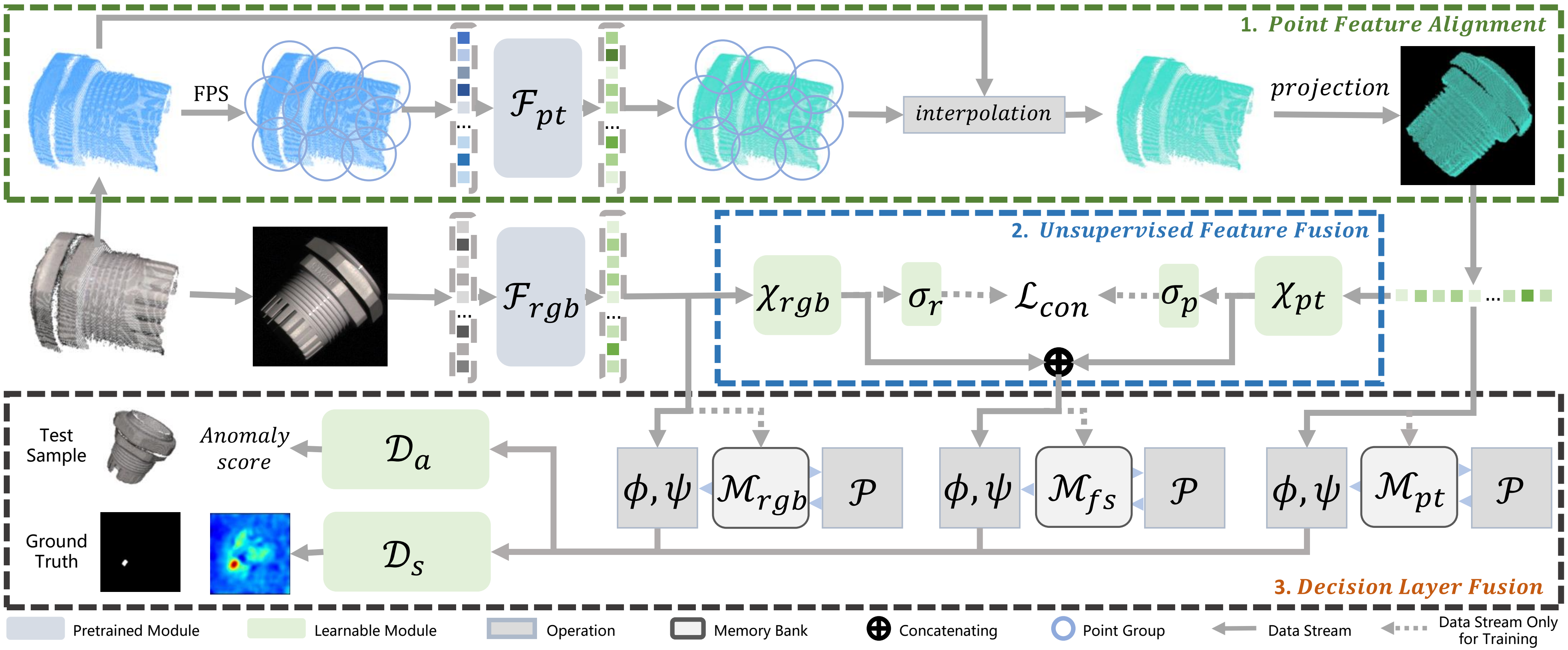
Overview

- We propose Multi-3D-Memory (M3DM), a novel multimodal anomaly detection method with hybrid fusion scheme.



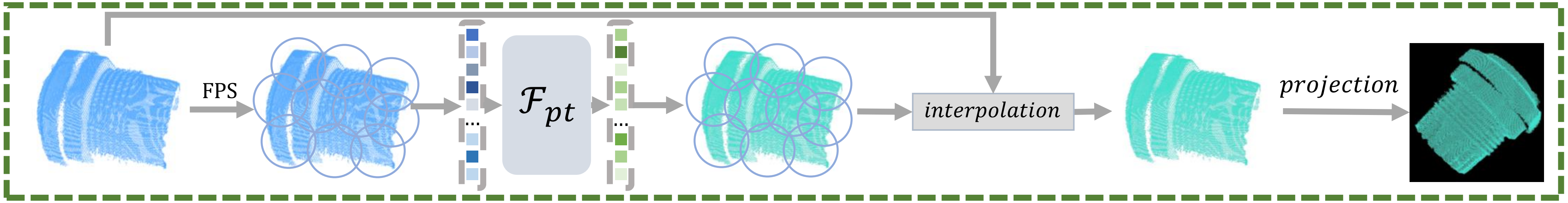
Overview

- We propose Multi-3D-Memory (M3DM), a novel multimodal anomaly detection method with hybrid fusion scheme.



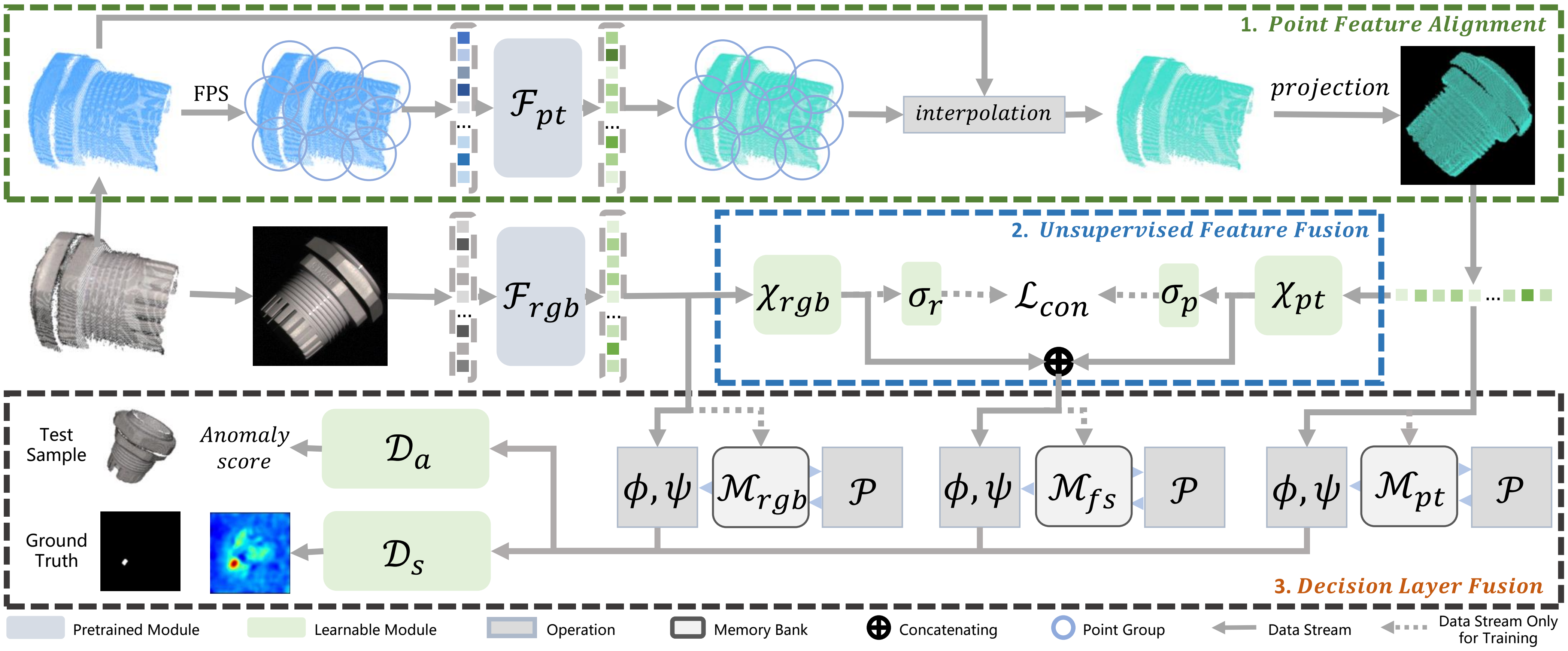
Point Feature Alignment

- **Point Feature Extraction.** We utilize a Point Transformer ($\mathcal{F}_{pt}$) to extract the point clouds feature.
- **Point Feature Interpolation.** We propose to interpolate the feature back to the original point cloud.
- **Point Feature Projection.** After interpolation, we project point feature into the 2D plane using the point coordinate and camera parameters.



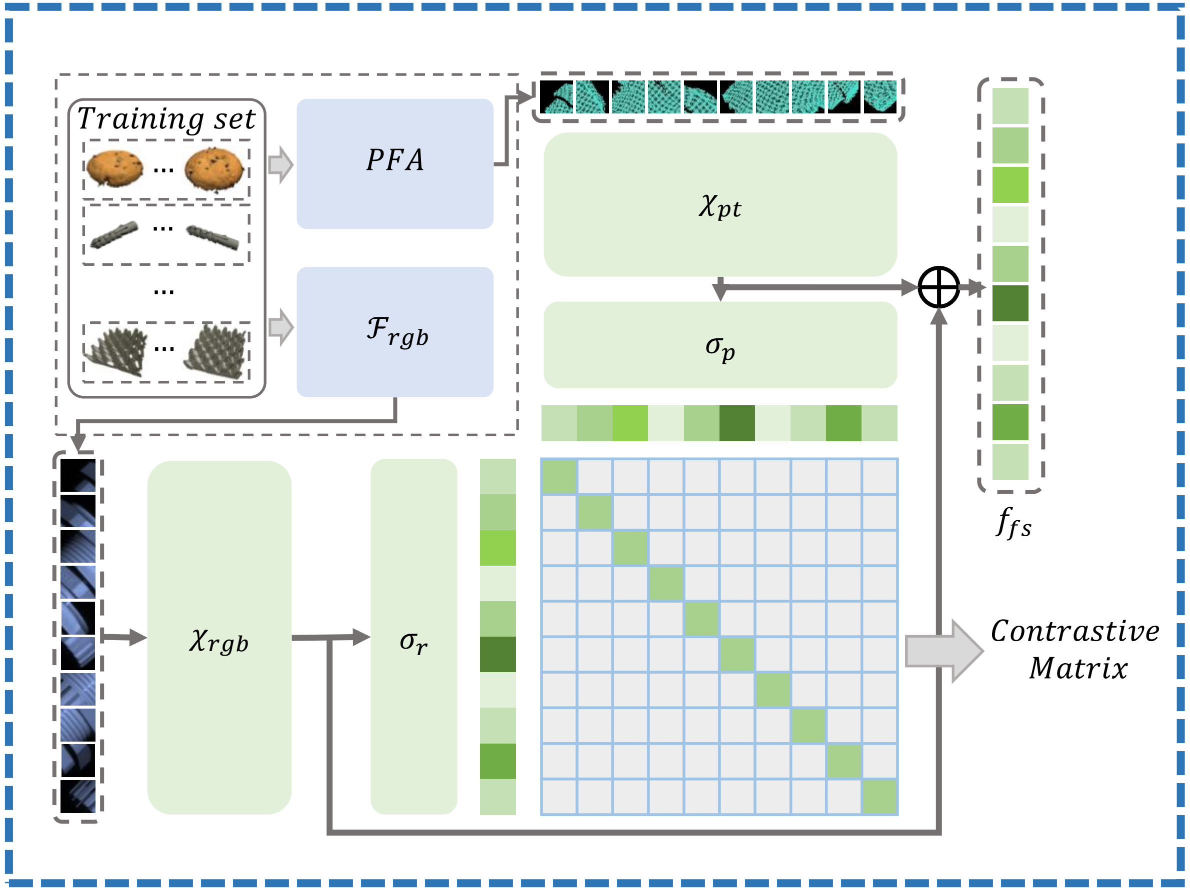
Overview

- We propose Multi-3D-Memory (M3DM), a novel multimodal anomaly detection method with hybrid fusion scheme.



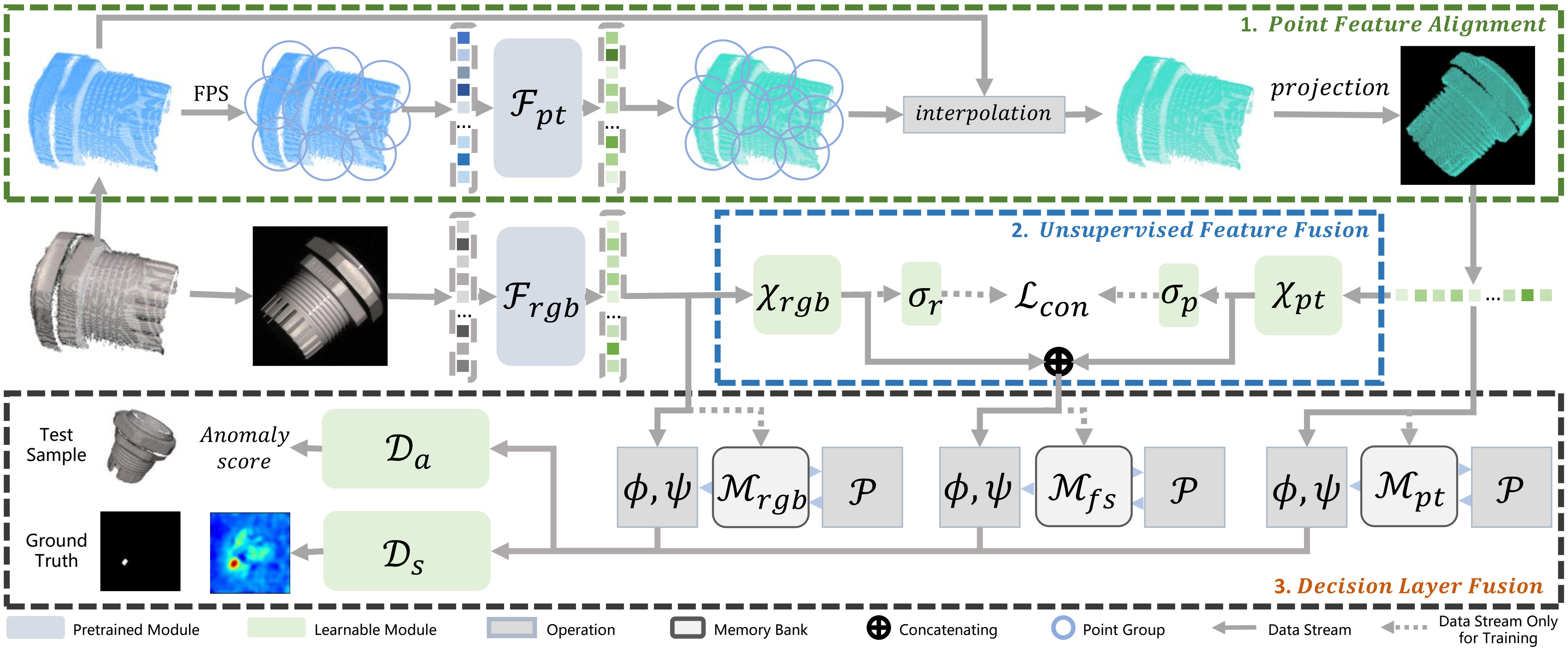
Unsupervised Feature Fusion

- The interaction between multimodal features can create new information that is helpful for industrial AD.
- UFF is a unified module trained with all training data of MVTec-3D AD.
- The patch-wise contrastive loss encourages the multimodal patch features in the same position to have the most mutual information.



Overview

- We propose Multi-3D-Memory (M3DM), a novel multimodal anomaly detection method with hybrid fusion scheme.

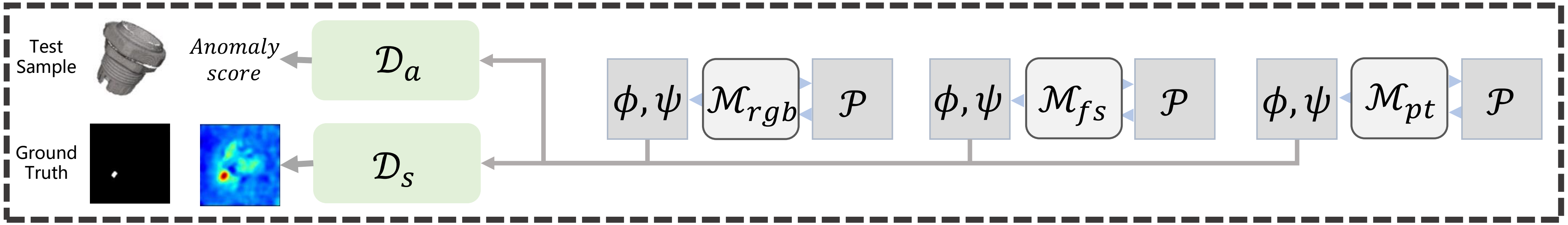


Decision Layer Fusion

- We propose to utilize multiple memory banks to store the original color feature, position feature and fusion feature.
- 2 learnable One-Class Support Vector Machines $\mathcal{D}_a$ and $\mathcal{D}_s$ to make the final decision for both anomaly score a and segmentation map S.
- Decision Layer Fusion can be described as:

$$a = D_a\left(\phi(\mathcal{M}_{rgb}, f_{rgb}), \phi(\mathcal{M}_{pt}, f_{pt}), \phi(\mathcal{M}_{fs}, f_{fs})\right),$$

$$S = D_S\left(\psi(\mathcal{M}_{rgb}, f_{rgb}), \psi(\mathcal{M}_{pt}, f_{pt}), \psi(\mathcal{M}_{fs}, f_{fs})\right).$$



Comparison on MVTec-3D AD

- I-AUROC score

Modality	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	Depth GAN [3]	0.530	0.376	0.607	0.603	0.497	0.484	0.595	0.489	0.536	0.521	0.523
	Depth AE [3]	0.468	0.731	0.497	0.673	0.534	0.417	0.485	0.549	0.564	0.546	0.546
	Depth VM [3]	0.510	0.542	0.469	0.576	0.609	0.699	0.450	0.419	0.668	0.520	0.546
	Voxel GAN [3]	0.383	0.623	0.474	0.639	0.564	0.409	0.617	0.427	0.663	0.577	0.537
	Voxel AE [3]	0.693	0.425	0.515	0.790	0.494	0.558	0.537	0.484	0.639	0.583	0.571
	Voxel VM [3]	0.750	**0.747**	0.613	0.738	0.823	0.693	0.679	0.652	0.609	0.690	0.699
	3D-ST [4]	0.862	0.484	0.832	0.894	0.848	0.663	0.763	0.687	**0.958**	0.486	0.748
	FPFH [16]	0.825	0.551	0.952	0.797	0.883	0.582	0.758	0.889	0.929	0.653	0.782
	AST [27]	0.881	0.576	**0.965**	0.957	0.679	**0.797**	**0.990**	**0.915**	0.956	0.611	0.833
	Ours	**0.941**	0.651	**0.965**	**0.969**	**0.905**	0.760	0.880	**0.974**	0.926	**0.765**	**0.874**
RGB	DifferNet [26]	0.859	0.703	0.643	0.435	0.797	0.790	0.787	0.643	0.715	0.590	0.696
	PADiM [8]	**0.975**	0.775	0.698	0.582	0.959	0.663	0.858	0.535	0.832	0.760	0.764
	PatchCore [25]	0.876	0.880	0.791	0.682	0.912	0.701	0.695	0.618	0.841	0.702	0.770
	STFPM [32]	0.930	0.847	0.890	0.575	0.947	0.766	0.710	0.598	0.965	0.701	0.793
	CS-Flow [29]	0.941	**0.930**	0.827	0.795	**0.990**	0.886	0.731	0.471	0.986	0.745	0.830
	AST [27]	0.947	0.928	0.851	**0.825**	0.981	**0.951**	0.895	0.613	**0.992**	**0.821**	**0.880**
	Ours	0.944	0.918	**0.896**	0.749	0.959	0.767	**0.919**	**0.648**	0.938	0.767	0.850
RGB + 3D	Depth GAN [3]	0.538	0.372	0.580	0.603	0.430	0.534	0.642	0.601	0.443	0.577	0.532
	Depth AE [3]	0.648	0.502	0.650	0.488	0.805	0.522	0.712	0.529	0.540	0.552	0.595
	Depth VM [3]	0.513	0.551	0.477	0.581	0.617	0.716	0.450	0.421	0.598	0.623	0.555
	Voxel GAN [3]	0.680	0.324	0.565	0.399	0.497	0.482	0.566	0.579	0.601	0.482	0.517
	Voxel AE [3]	0.510	0.540	0.384	0.693	0.446	0.632	0.550	0.494	0.721	0.413	0.538
	Voxel VM [3]	0.553	0.772	0.484	0.701	0.751	0.578	0.480	0.466	0.689	0.611	0.609
	3D-ST [4]	0.950	0.483	**0.986**	0.921	0.905	0.632	0.945	**0.988**	0.976	0.542	0.833
	PatchCore + FPFH [16]	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	**0.886**	0.865
	AST [27]	0.983	0.873	0.976	0.971	0.932	0.885	**0.974**	0.981	**1.000**	0.797	0.937
	Ours	**0.994**	**0.909**	0.972	**0.976**	**0.960**	**0.942**	0.973	0.899	0.972	0.850	**0.945**

Comparison on MVTec-3D AD

- AUPRO score

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	Depth GAN [3]	0.111	0.072	0.212	0.174	0.160	0.128	0.003	0.042	0.446	0.075	0.143
	Depth AE [3]	0.147	0.069	0.293	0.217	0.207	0.181	0.164	0.066	0.545	0.142	0.203
	Depth VM [3]	0.280	0.374	0.243	0.526	0.485	0.314	0.199	0.388	0.543	0.385	0.374
	Voxel GAN [3]	0.440	0.453	0.875	0.755	0.782	0.378	0.392	0.639	0.775	0.389	0.583
	Voxel AE [3]	0.260	0.341	0.581	0.351	0.502	0.234	0.351	0.658	0.015	0.185	0.348
	Voxel VM [3]	0.453	0.343	0.521	0.697	0.680	0.284	0.349	0.634	0.616	0.346	0.492
	FPFH [16]	**0.973**	**0.879**	**0.982**	**0.906**	**0.892**	0.735	**0.977**	**0.982**	**0.956**	**0.961**	**0.924**
	Ours	0.943	0.818	0.977	0.882	0.881	**0.743**	0.958	0.974	0.95	0.929	0.906
RGB	CFlow [15]	0.855	0.919	0.958	0.867	**0.969**	0.500	0.889	0.935	0.904	0.919	0.871
	PatchCore [25]	0.901	0.949	0.928	0.877	0.892	0.563	0.904	0.932	0.908	0.906	0.876
	PADiM [8]	**0.980**	0.944	0.945	0.925	0.961	0.792	0.966	0.940	0.937	0.912	0.930
	Ours	0.952	**0.972**	**0.973**	**0.891**	0.932	**0.843**	**0.97**	**0.956**	**0.968**	**0.966**	**0.942**
RGB + 3D	Depth GAN [3]	0.421	0.422	0.778	0.696	0.494	0.252	0.285	0.362	0.402	0.631	0.474
	Depth AE [3]	0.432	0.158	0.808	0.491	0.841	0.406	0.262	0.216	0.716	0.478	0.481
	Depth VM [3]	0.388	0.321	0.194	0.570	0.408	0.282	0.244	0.349	0.268	0.331	0.335
	Voxel GAN [3]	0.664	0.620	0.766	0.740	0.783	0.332	0.582	0.790	0.633	0.483	0.639
	Voxel AE [3]	0.467	0.750	0.808	0.550	0.765	0.473	0.721	0.918	0.019	0.170	0.564
	Voxel VM [3]	0.510	0.331	0.413	0.715	0.680	0.279	0.300	0.507	0.611	0.366	0.471
	3D-ST [4]	0.950	0.483	**0.986**	0.921	0.905	0.632	0.945	**0.988**	**0.976**	0.542	0.833
	PatchCore + FPFH [16]	**0.976**	0.969	0.979	**0.973**	0.933	0.888	0.975	0.981	0.950	0.971	0.959
	Ours	0.970	**0.971**	0.979	0.950	**0.941**	**0.932**	**0.977**	0.971	0.971	**0.975**	**0.964**

Comparison on MVTec-3D AD

- Our method performs well on both anomaly detection and anomaly segmentation.

Method	I-AUROC	P-AUROC	AUPRO
Depth-AE	0.595	-	0.481
Voxel-VM	0.609	-	0.471
3D-ST	0.865	-	0.833
PatchCore + FPFH	0.865	**0.992**	0.959
AST	0.937	0.976	-
Ours	**0.945**	**0.992**	**0.964**

Ablation Study

- Compared with directly concatenating feature with UFF, the single memory bank method get better performance.
- With DLF, the anomaly detection and segmentation performance gets great improvement.

Method	Memory bank	I-AUROC	AUPRO	P-AUROC
Only PC	$\mathcal{M}_{pt}$	0.874	0.906	0.970
Only RGB	$\mathcal{M}_{rgb}$	0.850	0.942	0.987
w/o UFF	$\mathcal{M}_{fs}$	0.857	0.944	0.987
w/ UFF	$\mathcal{M}_{fs}$	0.898	0.956	0.990
w/o DLF	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}$	0.929	0.953	0.987
w/ DLF	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}$	0.932	0.959	0.990
Ours	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}, \mathcal{M}_{fs}$	**0.945**	**0.964**	**0.992**

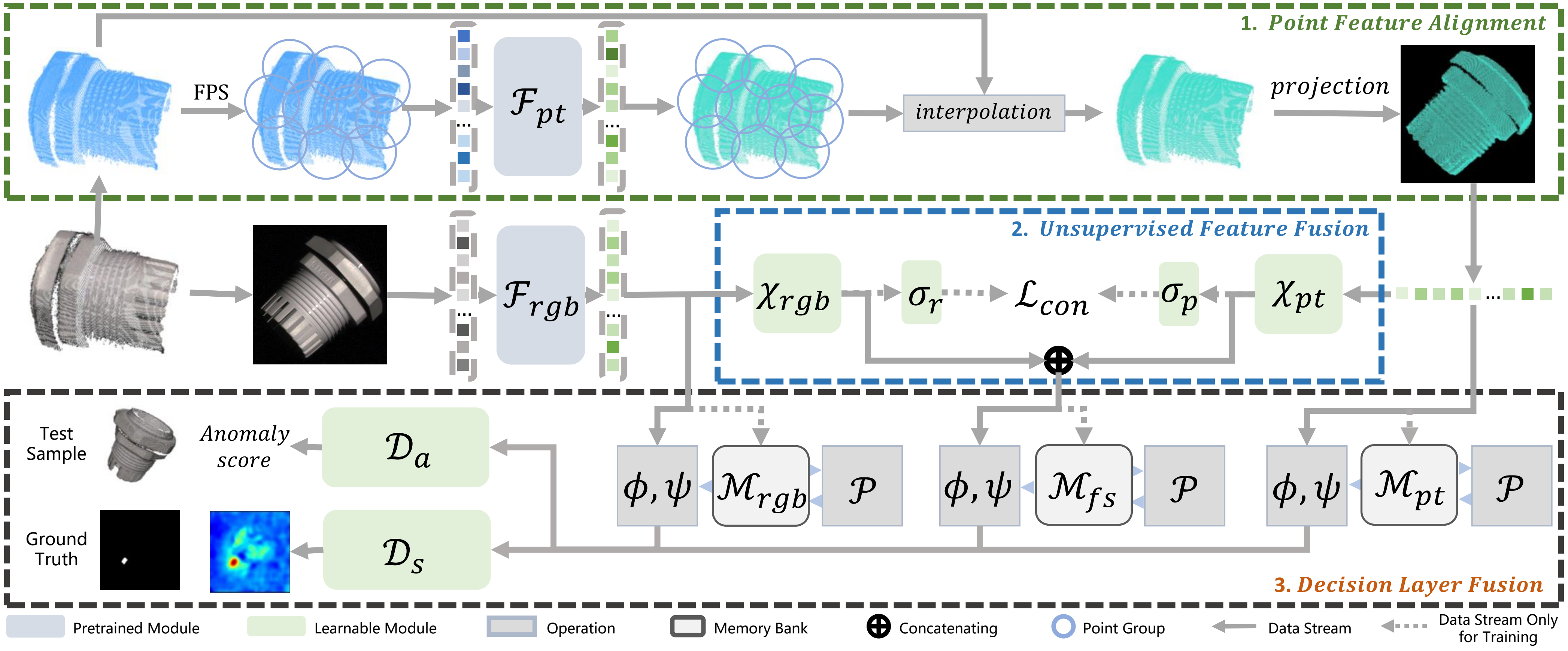
Exploring Point Transformer setting

- We get the best performance with 1024 point groups per sample and each point group contains 128 Points.
- Compared with directly calculating anomaly and segmentation scores on point groups, the method based on a 2D plane patch needs a small patch size towards high performance.

S.G	N.G	Sampling	I-AUROC	AUPRO	P-AUROC
64	784	point group	0.793	0.813	0.922
128	1024	point group	0.841	0.896	0.960
64	784	28×28 patches	0.805	0.879	0.963
128	1024	28×28 patches	0.819	0.896	0.967
128	1024	56×56 patches	**0.874**	**0.906**	**0.970**

Conclusion

- Our method is based on multiple memory banks and we propose a hybrid feature fusion scheme to process the multimodal data.



THANKS!