Self-Supervised Cross-modality Feature Learning using 3D Gaussian Splatting

1. Introduction

Current robotic perception systems utilize a variety of sensors to estimate and understand a robot's surroundings. This paper focuses on a novel data representation technique that makes use of a recent scene reconstruction algorithm, known as 3D Gaussian Splatting (3DGS) [1], to explicitly represent and reason about an environment using only a sparse set of camera views as input. The point cloud provided by the 3DGS algorithm encodes a spatial representation of the environment, from which features can be learned.

The **research questions** that this study investigates are:

1. How can we learn cross-modal features using 3D Gaussians and the original views as a data representation in a self-supervised manner?
2. What is the impact of the sampling technique on the quality of the learned features, when evaluating for the supervision signals?
3. How does scaling up the point sub-network impact the performance on the pretext tasks?

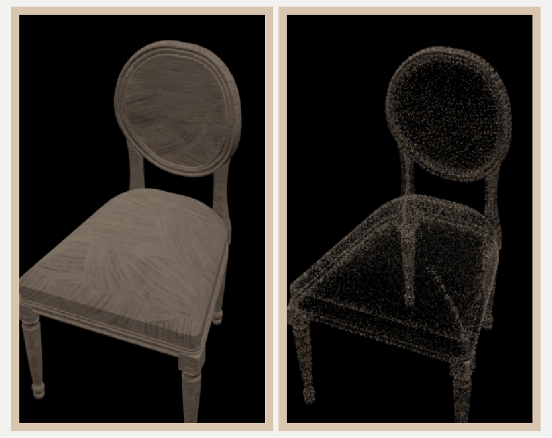

Figure 1: A rendered view (Left) and the optimized Gaussians (Right) for a 3D model of a chair

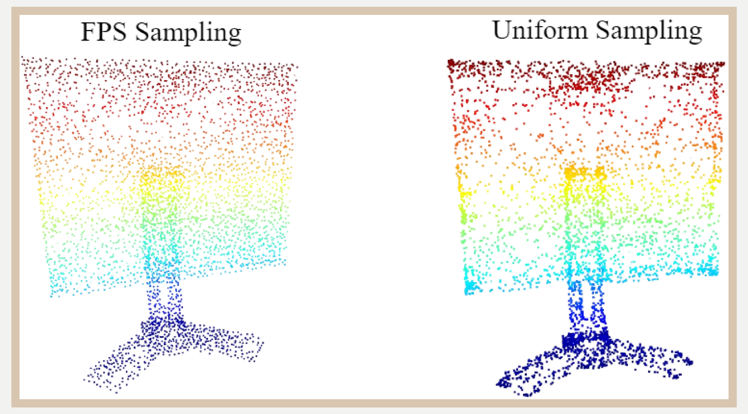

Figure 2: Vizualisation of the sampled gaussians on the model of a monitor using Furthest Point Sampling (Left) and uniform sampling (Right)

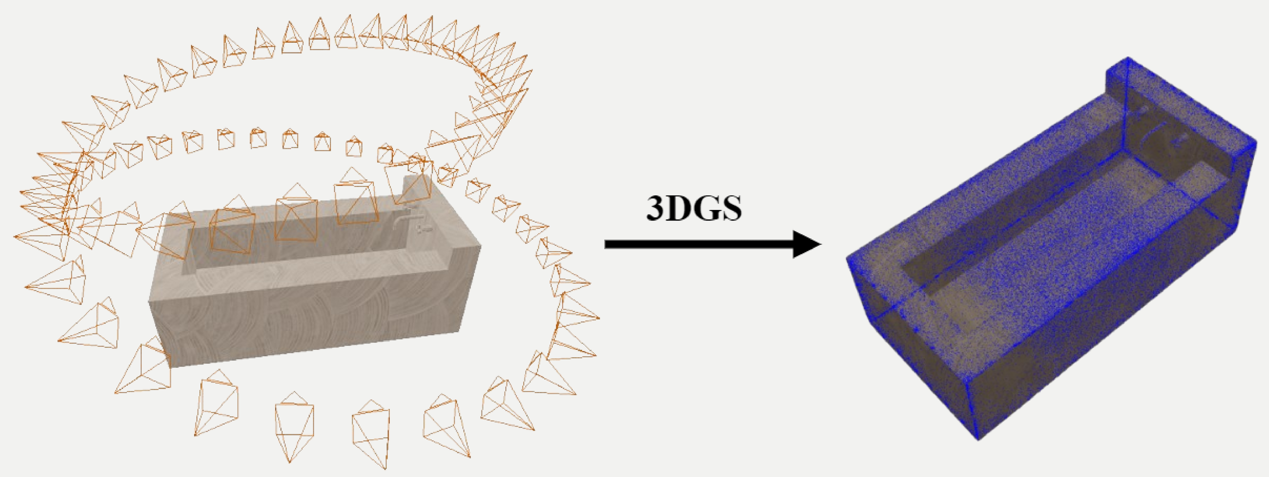

Figure 3: The camera trajectory (Left), and the resulting 3D Gaussian splat render and point cloud overlayed (Right). The bathtub mesh is up-scaled, and the point cloud (in blue) is down-scaled for visualization purposes.

2. Data Generation

The dataset is generated by taking 64 800x800 images around a texturized 3D model (Fig. 3), following a spiralling shaped motion around the object of interest. Each set of frames, alongside the camera trajectory is then turned into a point cloud using the 3DGS algorithm [1].

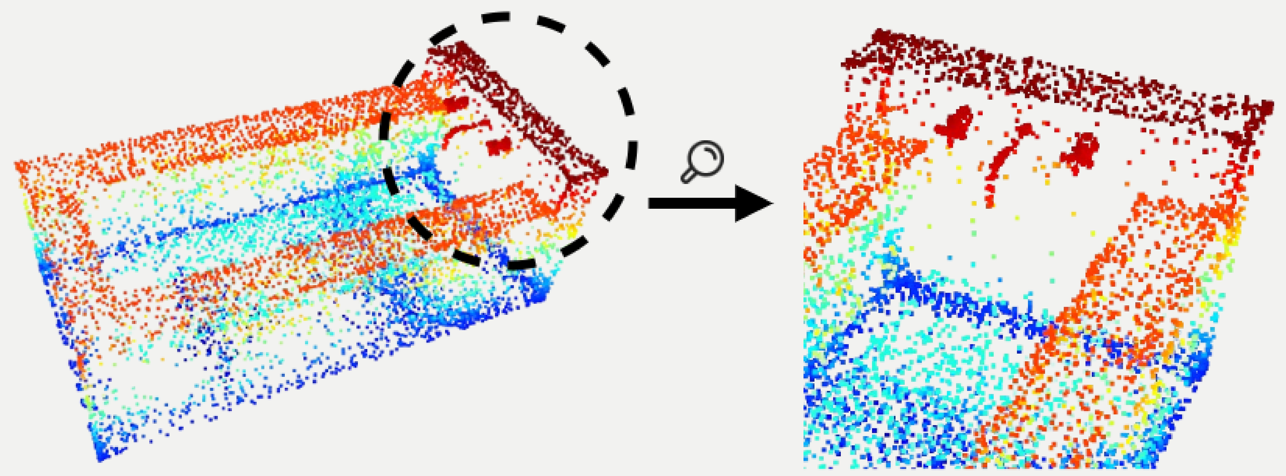

Figure 4: Visualization of the 3D Gaussian point cloud positions of a bathtub model (Left) uniformly sampled, and zoomed in on the reconstruction of its faucet and wall (Right). The faucet is an example where 3DGS uses more Gaussians to represent complex geometries.

3. Model Architecture

Self-supervised model, inspired by Jing et al. work [2], featuring two feature extractors for the image and point modalities.

The model is trained using two losses:

- *Cross-View* (Triplet Loss): Learns whether distinct 2D views belong to the same object
- *Cross-Modality* (Binary Cross-Entropy Loss): Learns whether the combined 2D and 3D features belong to the same object

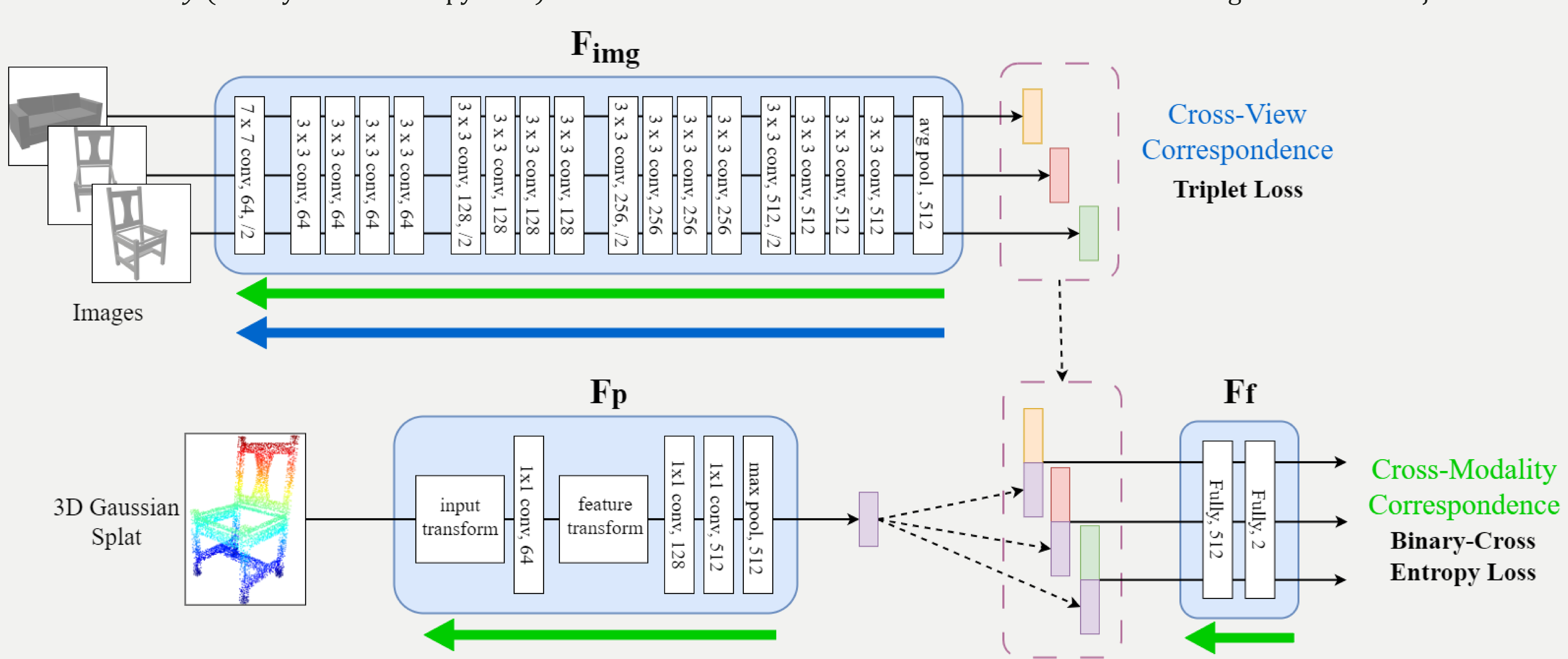

Figure 5: The proposed Model Architecture featuring a ResNet-18 for image processing, and the PointNet model as the backbone for gaussians processing

4. Results

The jointly-optimized model is evaluated on the two pretext tasks it was trained on:

- *Cross-View:* the mean paird distance (mPD) between **positive pairs is 5.34** and between **negative pairs 13.57**,
- *Cross-Modality:* obtains **95.2%** accuracy.

CM accuracy (%)	PointNet		PointNet++	
	FPS	Unif	FPS	Unif
Positions	0.941	0.924	**0.952**	0.931
+ Scale & Rotation	0.938	0.942	0.938	0.906
+ Spherical Harmonics	0.943	0.943	0.939	0.918

Table 1: Performance comparison for the cross-modality pretext task for the two point backbones, sampling techniques, and varying number of features. + indicates the accumulation of the features for each row.

CV mPD	FPS		Unif	
	Pos	Neg	Pos	Neg
Positions	**5.34**	**13.57**	4.94	11.65
+ Scale & Rotation	4.94	12.07	5.05	12.6
+ Spherical Harmonics	5.12	12.91	5.1	12.79

Table 2: Performance comparison for the cross-view pretext task for the two sampling techniques, and varying number of features. + indicates the accumulation of the features for each row.

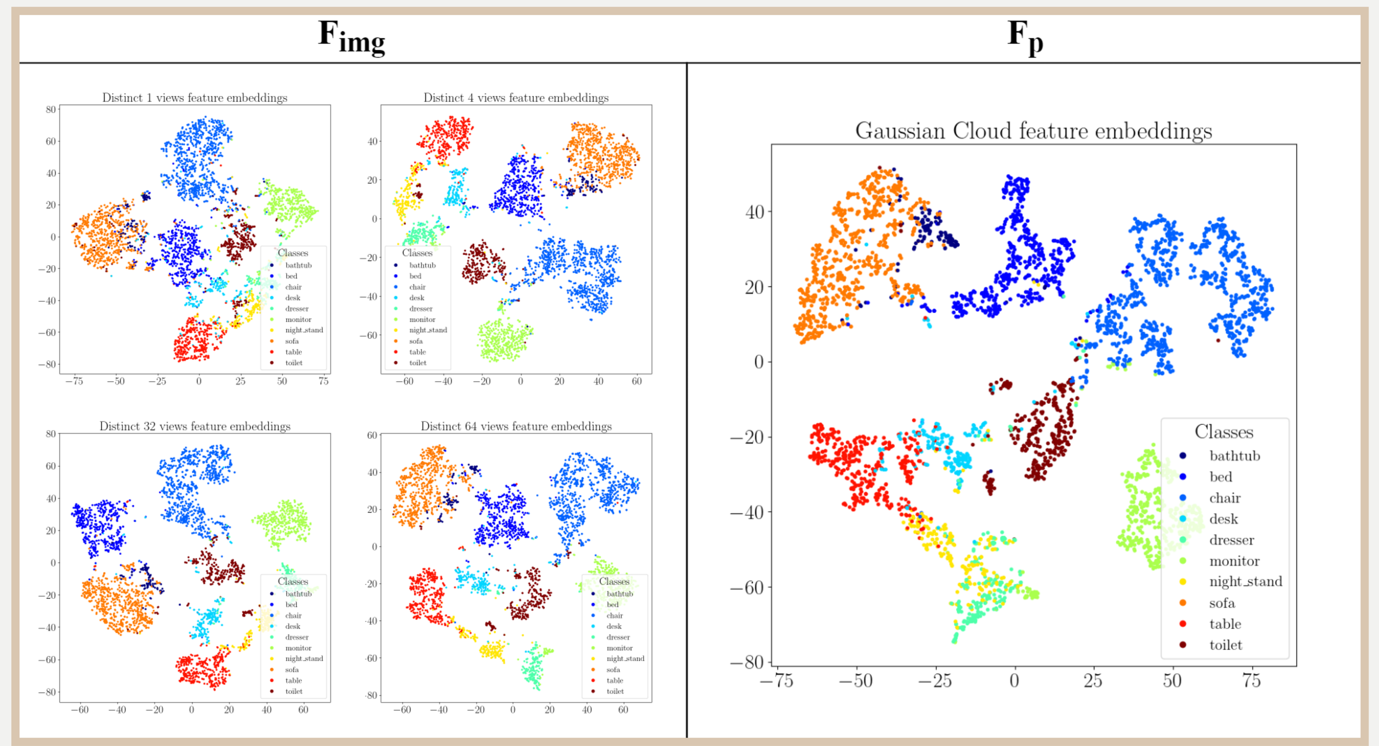

Figure 7: TSNE [3] visualization of the learned features on the image and point sub-networks. For the image sub-network, multiple views are considered. Class clusters are forming, meaning that the model has learned shape recognition and retrieval without any explicit signals. As the number of views increases, the clusters are more clearly defined, and approach the ones formed in the point sub-network.

Features	# Views	Accuracy (%)	
		PointNet	PointNet++
Positions	1	0.86	0.86
	4	0.93	0.93
	32	0.95	0.95
	64	0.95	**0.96**
	64*	**0.96**	**0.96**
+ Scale & Rotation	1	0.87	0.84
	4	0.93	0.92
	32	0.95	0.95
	64	**0.96**	0.95
	64*	**0.96**	0.95
+ Sh. Harmonics	1	0.85	0.83
	4	0.93	0.91
	32	**0.96**	0.95
	64	**0.96**	0.95
	64*	**0.96**	0.95

Table 3: Performance comparison for the 2D shape recognition accuracy for the image sub-network. * indicates that the evaluation is performed on reconstructed (remembered) views.

Training Data	Modality	Network	Accuracy (%)
100 %	Points	PointNet [10]	0.93
		PointNet++ [11]	0.95
	Gaussians	F_p*	0.89
		$F_{p++}*$	0.9
	Images	MVCNN [14]	**0.98**
		$F_{img}*$	0.96
		$F_{img++}*$	0.96
50 %	Points	PointNet [10]	0.92
		PointNet++ [11]	0.93
	Gaussians	F_p*	0.88
		$F_{p++}*$	0.9
	Images	MVCNN [14]	**0.96**
		$F_{img}*$	0.95
		$F_{img++}*$	0.95

Training Data	Modality	Network	Accuracy (%)
10 %	Points	PointNet [10]	0.9
		PointNet++ [11]	0.91
	Gaussians	F_p*	0.87
		$F_{p++}*$	0.89
	Images	MVCNN [14]	0.91
		$F_{img}*$	**0.93**
		$F_{img++}*$	0.92
1 %	Points	PointNet [10]	0.6
		PointNet++ [11]	0.75
	Gaussians	F_p*	0.76
		$F_{p++}*$	0.75
	Images	MVCNN [14]	0.49
		$F_{img}*$	**0.77**
		$F_{img++}*$	0.76

Table 5: Classification accuracy comparison with SOTA models on ModelNet10 [4] dataset, under different amounts of training data available. The proposed methods (marked with *) are trained using only the Gaussian positions, with FPS. ++ indicates that the *PointNet++* backbone was used during SSL and/or fine-tuning.

Features	Accuracy (%)	
	PointNet	PointNet++
Positions	0.89	**0.9**
+ Scale & Rotation	0.87	0.88
+ Sh. Harmonics	0.88	0.87

Table 4: Performance comparison for the 3D shape recognition accuracy for the point sub-network. Scaling up the network leads to better performance.

5. Conclusion

- The self-supervised networks achive **very high performance on the two pretext tasks it was trained on.** The TSNE [3] analysis on the learned features indicate that the **model learns shape recognition and retrieval tasks without explicit supervision**.
- Experimental results on the ModelNet10 [4] dataset indicate that Gaussian-based models **perform better** when considering **only the Gaussian positions as input.**
- **FPS preserves a better geometrical approximation** of the objects, which leads to a **higher 3D shape recognition accuracy.**
- Gaussian-based models exhibit a **performance boost when the point sub-network is up scaled.**
- The *Memory-based Vision* task fascilitates **lossless 2D reasoning about a previously observed scene**. The Gaussian representation doubles as *a memory-module* which unlocks a family of possible tasks ranging from **enhanced navigation and path planning to increased human-agent collaboration**.

6. Limitations

- **All models in the dataset have the same texture, are lit in exactly the same way**, and thus have similar view-dependent colors. The scale does not contribute significantly since **all models have been resized to identical dimensions**. Thus, **the extra features used do not aid the model in learning a better representation of the underlying input space**.
- The analysis of the *Memory-based Vision* task was performed on simple scenes (just one object in perfect lighting), and thus **the reconstruction loss of the rendered views is minimal and has little impact the 2D recognition accuracy.**

Project

Multi View Learning through 3D Gaussian Splatting

Author

Andrei Simionescu
asimionescu@tudelft.nl

Supervisor

Dr. Xucong Zhang

References

[1] Kerbl, Bernhard, et al. "3d gaussian splatting for real-time radiance field rendering." ACM Transactions on Graphics 42.4 (2023): 1-14.
[2] Jing, Longlong, Ling Zhang, and Yingli Tian. "Self-supervised feature learning by cross-modality and cross-view correspondences." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[3] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).
[4] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1912–1920, 2015.

TU Delft