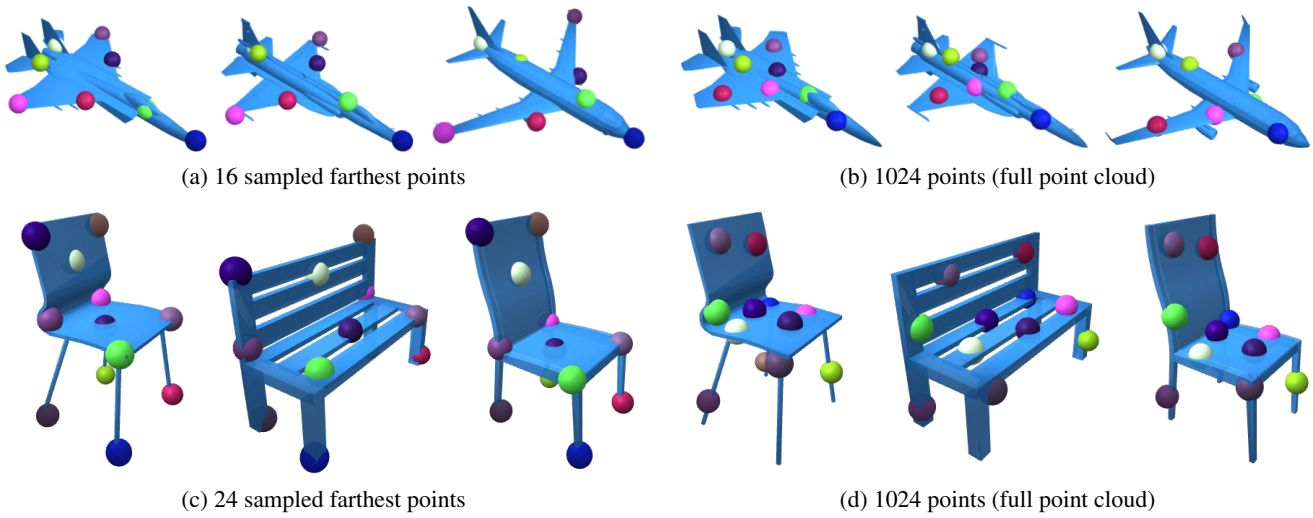
Appendix: Unsupervised 3D Keypoint Discovery for Shape Control

Tomas Jakab[1,4*], Richard Tucker[4], Ameesh Makadia[4], Jiajun Wu[3], Noah Snavely[4], Angjoo Kanazawa[2,4]
[1]University of Oxford, [2]UC Berkeley, [3]Stanford University, [4]Google Research

This supplemental document provides an ablation study (Appendix A), extensive results (Appendices B and C), and further implementation details (Appendix D). Please also refer to our video on our project page[1] that demonstrates our method in action.

A. Ablations



(a) 16 sampled farthest points (b) 1024 points (full point cloud)

(c) 24 sampled farthest points (d) 1024 points (full point cloud)

Figure 1: **Farthest Point Keypoint regularizer ablation.** We investigate the influence of the number J of sampled farthest points $\boldsymbol{q}$ used for the keypoint regularizer (Section 3.2 of the main paper) on the quality of discovered unsupervised keypoints. We show unsupervised 3D keypoints trained using two versions of regularization. First, we set the number of sampled farthest points to double of the number of keypoints **(a, c)**. This is the setup that we use throughout the paper. Second, we set the number of sampled farthest points to the number of points in the point cloud representing the shape. This essentially results in a regularizer that is minimizing the Chamfer distance between the unsupervised keypoints and the object point cloud. Although the learned unsupervised keypoints have a good coverage **(b, d)** they are not as equally spaced and characteristic of the shape as **(a, c)**.

Varying number of regularizing points. We examine the importance of the number of sampled farthest points $\boldsymbol{q}$ on the quality of keypoint regularization (Section 3.2 of the main paper). Figure 1 shows the effect of different numbers of sampled farthest points on the discovered keypoints. Using a high number of sampled farthest points in the regularization fails to learn keypoints that are equally spaced and characteristic of the underlying shape.

Varying number of keypoints. We vary the number of unsupervised keypoints discovered by our method. Figure 2 shows that our keypoints remain semantically consistent for different numbers of discovered keypoints.

* Work done while interning at Google Research.
[1] http://tomasjakab.github.io/KeypointDeformer

B. Shape Control via Unsupervised 3D Keypoints

We show user guided interactive shape control in our supplementary video on our project page. Figure 5 shows frames captured from user-guided interactive shape editing. Editing using our keypoints is fast and intuitive while preserving the character and details of the original shape.

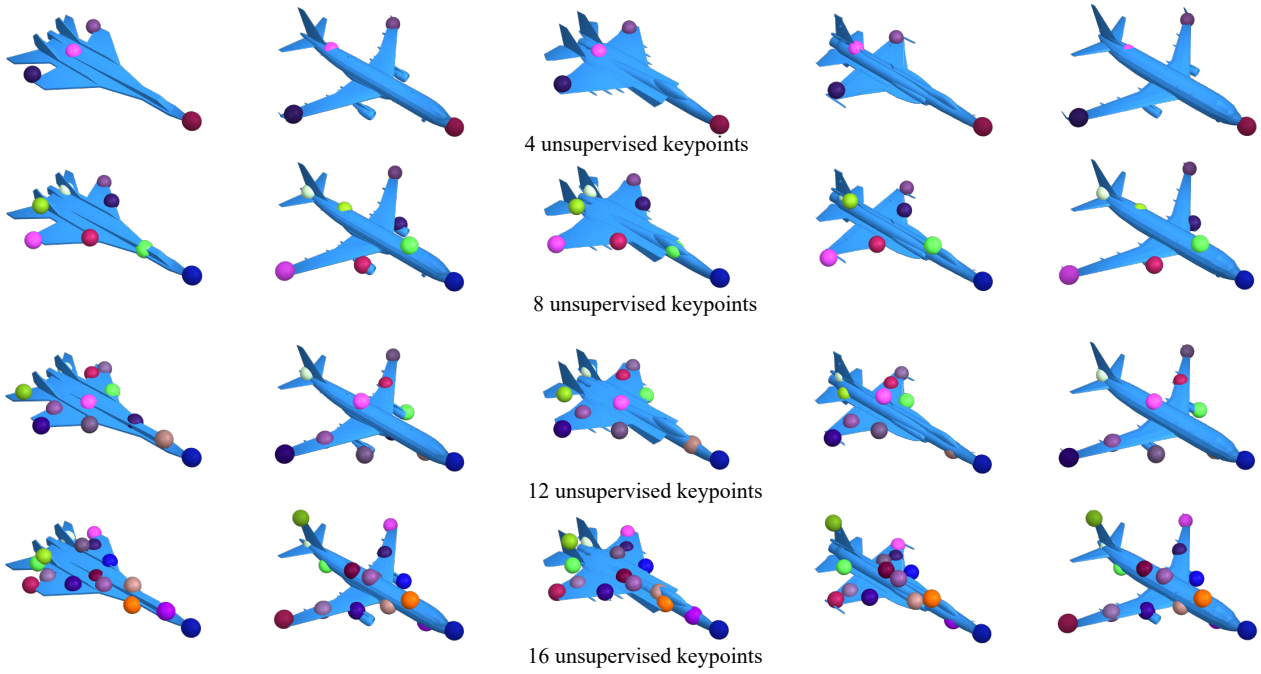
C. Unsupervised 3D Keypoints

We show extended quantitative results for semantic part correspondence experiment and detailed correlation tables in Figure 3 for the ShapeNet Car category. Figures 6 to 10 show extensive *randomly* sampled qualitative test results for our unsupervised 3D keypoints.

D. Implementation Details

Our model assumes that the shapes are aligned (in the same orientation). The initial cage is a 42-vertex icosphere. We limit the influence matrix W to influence at most M nearest cages vertices (Section 3.1) per each keypoint, with $M = \lfloor C/K \rfloor$, where C is the number of cage vertices and K is the number of discovered keypoints. We use a learning rate of 0.001. The scalar loss coefficients (Section 3.2) α_{kpt} and α_{inf} are set to 1.0 and 10^{-6} respectively. Figure 4 shows detailed description of network architectures used for the keypoint predictor Φ and the influence predictor Γ.

Datasets. KeypointNet [7] dataset contains semantic 3D keypoint annotations for ShapeNet dataset [1]. Some models in KeypointNet are missing full keypoint annotations, therefore we use a subset of annotated keypoints that are contained in at least 80% of the models. KeypointNet also does not follow the standard training and testing splits from ShapeNet. We resample the KeypointNet dataset splits to make it compatible with the original ShapeNet splits.

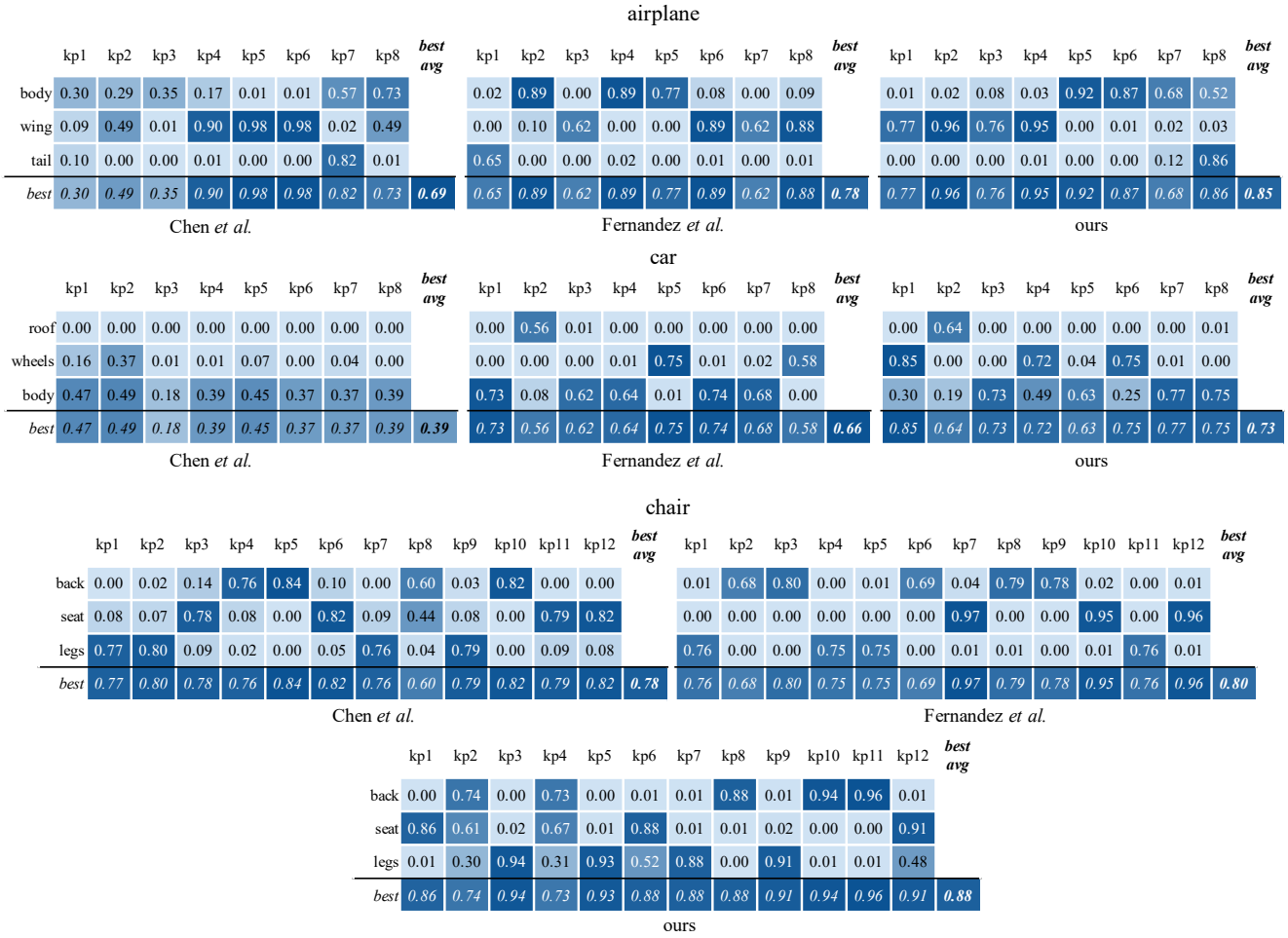


(a) predicted unsupervised keypoints

# of unsupervised keypoints	4	8	12	16
PCK@0.05	0.56	0.61	0.71	0.71

(b) quantitative evaluation

Figure 2: **Varying number of keypoints.** The figure (a) shows the effect of different number of discovered keypoints (4, 8, 12, 16 from the top). The results are shown on randomly sampled results for ShapeNet Airplane category. Results in the table (b) are in terms of PCK@0.05 on airplanes from the KeypointNet dataset.



(a) unsupervised keypoints correlation

	airplane	cap	car	chair	guitar	knife	laptop	motorbike	mug	skateboard	table	pistol	bag	rocket	earphone	lamp
Chen *et al.* [2]	0.69	0.24	0.39	0.78	0.97	0.94	0.95	0.91	0.50	0.89	0.75	0.78	0.35	0.56	0.30	0.50
Fernandez *et al.* [3]	0.78	0.45	0.66	0.80	0.93	0.92	0.85	0.90	0.78	0.92	0.85	0.60	0.72	0.61	0.24	0.40
ours	**0.85**	**0.71**	**0.73**	**0.88**	**0.99**	**0.96**	**0.96**	**0.93**	**0.94**	**0.96**	**0.92**	**0.91**	**0.85**	**0.90**	**0.72**	**0.53**

(b) average unsupervised keypoints correlation

Figure 3: **Semantic part correspondence.** We evaluate semantic part correspondence for the ShapeNet Car category. The tables (a) shows the frequency of each unsupervised keypoint **[kp*]** being associated with a given object part. Chen *et al.* [2] show worse performance in this task because that methods tends to predict keypoints inside the object far from the annotated object surface. We also report the average unsupervised keypoints correlation for each category (b). ↑ is better.

Operation	Kernel	Output channels	Output size	Activation
Input $\boldsymbol{x}$	-	3	1024	-
Conv 1D	1	64	1024	ReLU
Conv 1D	1	128	1024	ReLU
Conv 1D	1	256	1024	None
Max. pool	1024	256	1	-
Squeeze	-	256	-	-
Linear	-	256	-	LReLU
Linear	-	512	-	LReLU
Linear	-	256	-	LReLU
Linear	-	$3 \cdot K$	-	None
Reshape	-	3	K	-

(a) Keypoint predictor Φ

Operation	Kernel	Output channels	Output size	Activation
Input $\boldsymbol{x}$	-	3	1024	-
Conv 1D	1	64	1024	ReLU
Conv 1D	1	128	1024	ReLU
Conv 1D	1	256	1024	None
Max. pool	1024	256	1	-
Squeeze	-	256	-	-
Linear	-	$C \cdot K$	-	LReLU
Linear	-	$C \cdot K$	-	LReLU
Linear	-	$C \cdot K$	-	None
Reshape	-	C	K	-

(b) Influence predictor Γ

Figure 4: **Network architectures.** The network architectures are based on a PointNet encoder [5, 6]. K is the number of discovered keypoints, C is the number of cage vertices. LReLU stands for Leaky ReLU with 0.1 negative slope.

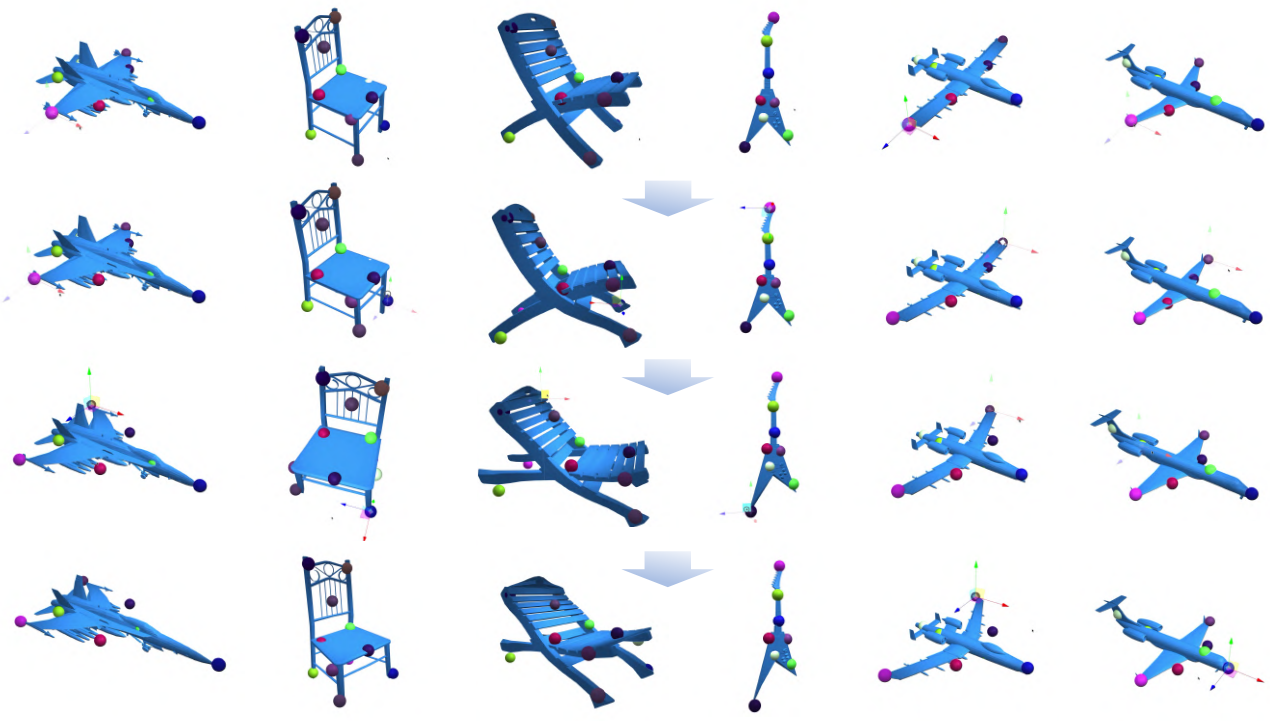


Figure 5: **Shape control via unsupervised 3D keypoints.** We show steps from shape editing using our unsupervised 3D keypoints. Please refer to our supplementary video on our project page to see the editing in action.

Figure 6: **Unsupervised 3D keypoints on real-world 3D scans.** Randomly sampled results with 8 unsupervised keypoints on real-world 3D scans of shoes from Google Scanned Objects dataset [4].

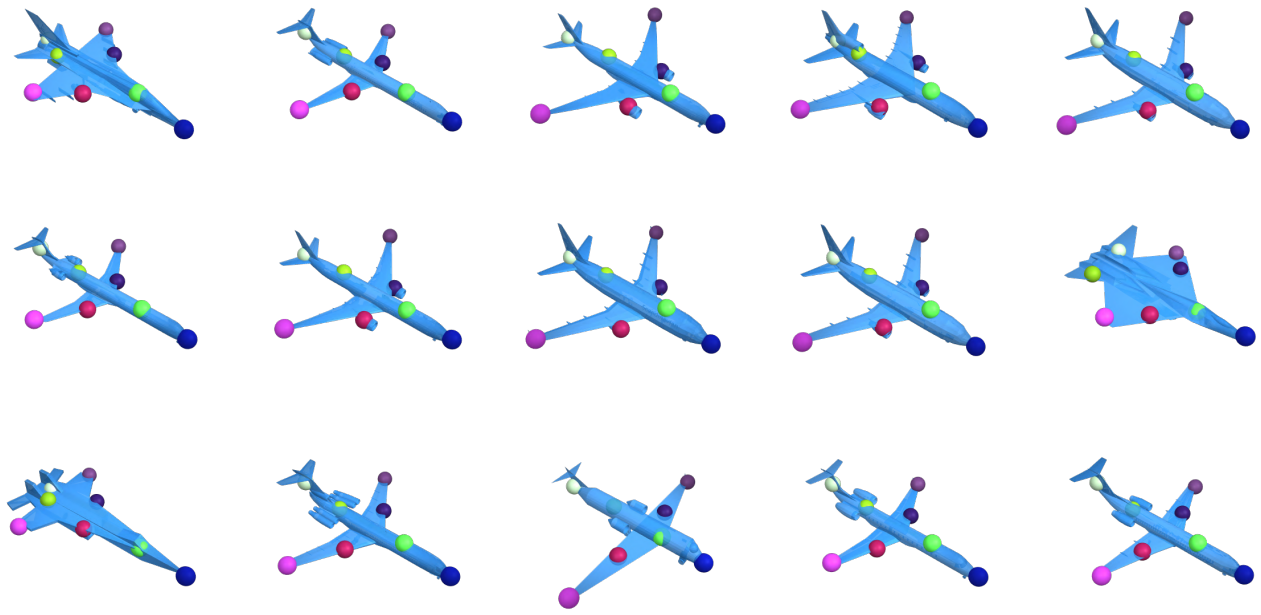


Figure 7: **Unsupervised 3D keypoints.** Randomly sampled results with 8 unsupervised keypoints for ShapeNet Airplane category.

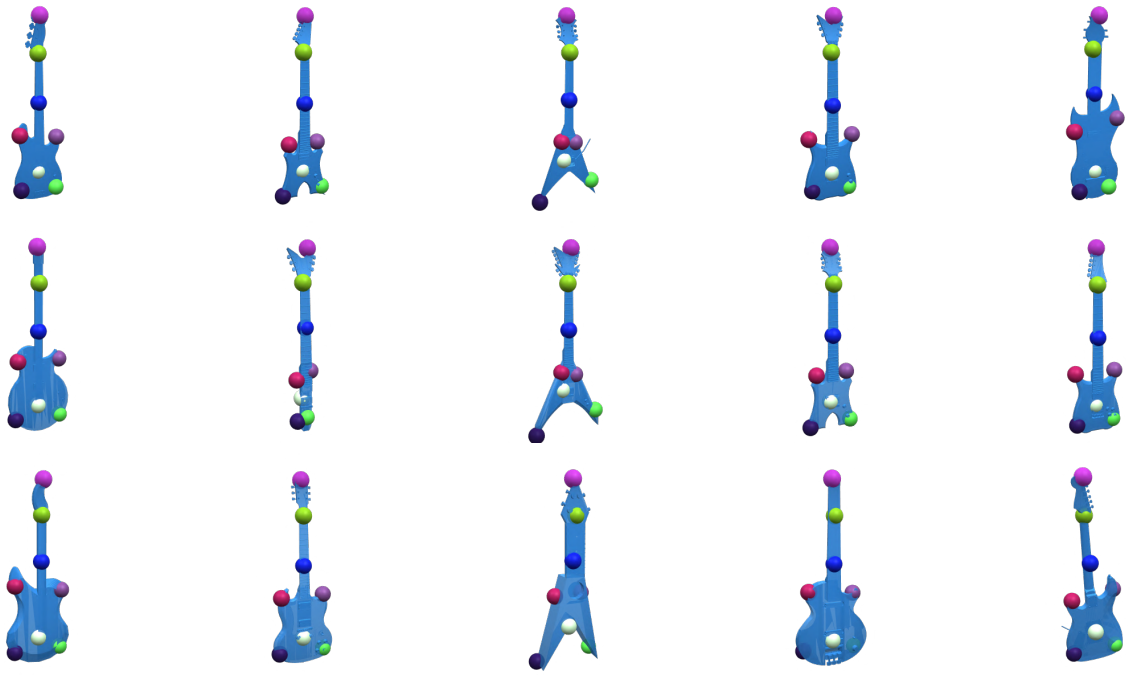


Figure 8: **Unsupervised 3D keypoints.** Randomly sampled results with 8 unsupervised keypoints for ShapeNet Guitar category.

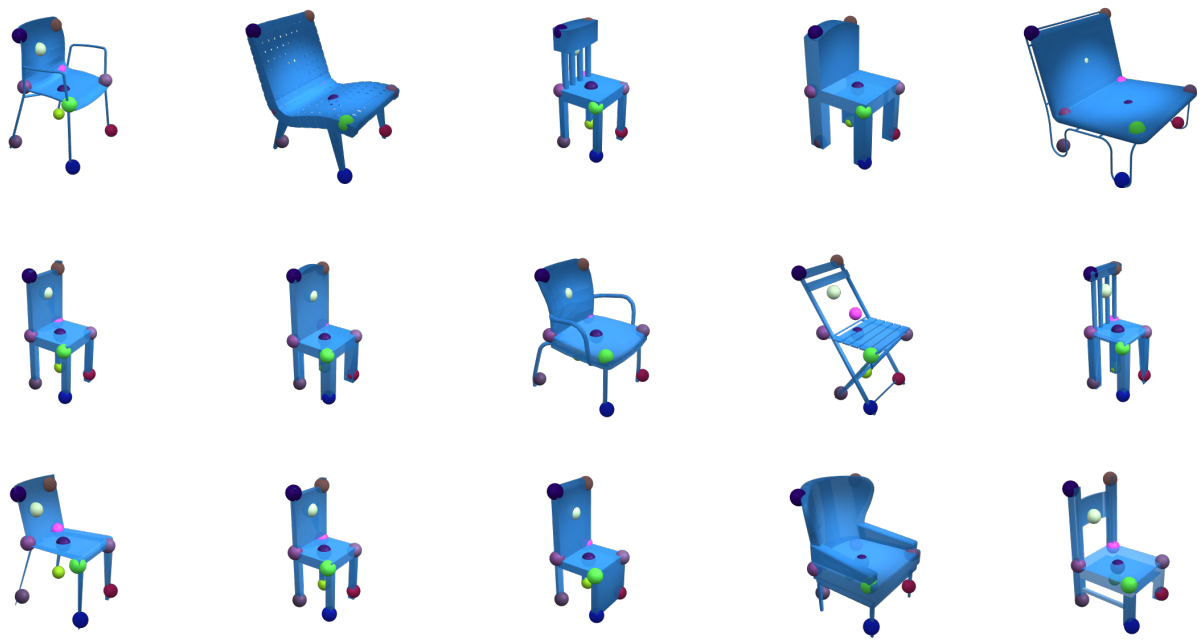


Figure 9: **Unsupervised 3D keypoints.** Randomly sampled results with 12 unsupervised keypoints for ShapeNet Chair category.

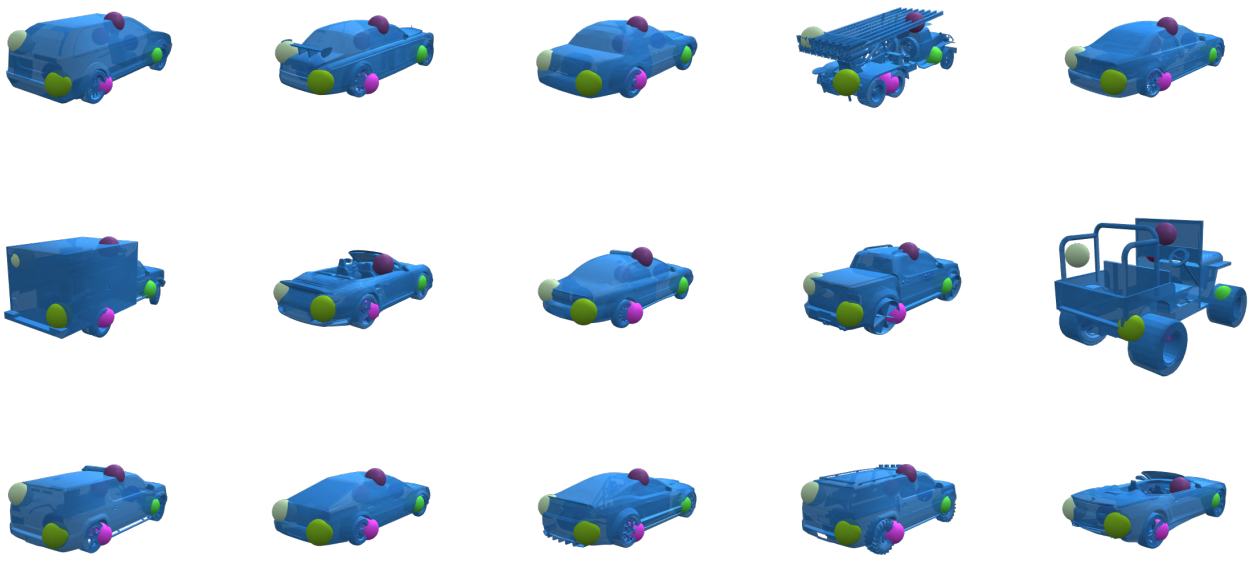


Figure 10: **Unsupervised 3D keypoints.** Randomly sampled results with 8 unsupervised keypoints for ShapeNet Car category.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[2] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9121–9130, 2020. 4

[3] Clara Fernandez-Labrador, Ajad Chhatkuli, Danda Pani Paudel, Jose J Guerrero, Cédric Demonceaux, and Luc Van Gool. Unsupervised learning of category-specific symmetric 3d keypoints from point sets. *European Conference on Computer Vision (ECCV)*, 2020. 4

[4] GoogleResearch. Google scanned objects. https://fuel.ignitionrobotics.org/1.0/GoogleResearch/fuel/collections/Google%20Scanned%20Objects, September 2020. 6

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5

[6] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 75–83, 2020. 5

[7] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13656, 2020. 2