Latent shape representation for tactile exploration

Felix Laarmann
Technical University Munich
Munich, Germany
felix.laarmann@tum.de

Fabian Heinrich
Technical University Munich
Munich, Germany
fabian.heinrich@tum.de

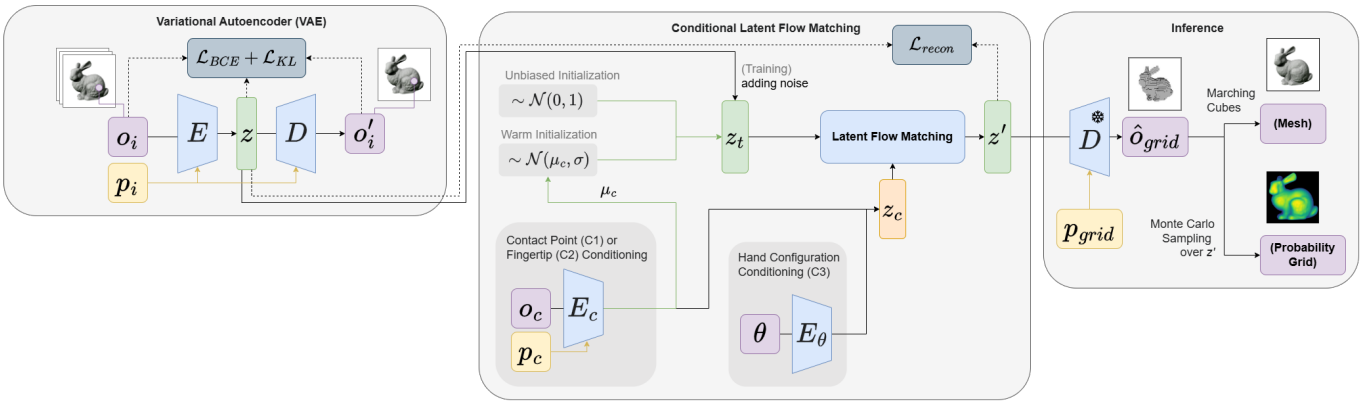

Fig. 1. We train a Variational Autoencoder (VAE) as occupancy network that encodes occupancy values o_i' for sample positions p_i to obtain compact latent shape representations z. Those are used to train a Flow Matching model - conditioned tactile information such as sparse contact point clouds (C1), fingertip positions (C2) or an encoding of the hand configuration (C3). As alternative to the objective of recovering valid latent codes from gaussian noise, we explore parametrizing the source distribution mean μ_c with a contact point encoding. During inference, we query the decoder D with a grid of positions p_{grid} to obtain an occupancy grid $\hat{o}_{grid}$. Meshes can optionally be extracted using the Marching Cubes algorithm. By averaging over multiple decoded occupancy grids, occupancy probabilities can be approximated.

***Abstract*—Tactile exploration estimates shapes based on collision measurements. Efficiently encoding and processing geometry is key for enabling compute-efficient applications but needs to operate in a sparse input domain of measurements as well as in the dense, high frequency output domain of shape estimates. We present a Latent Flow Matching pipeline that enables memory and compute efficient shape estimation. A Variational Autoencoder (VAE) encodes occupancy values of 3D coordinates into compact latent representations while incorporating a prior on expected shapes. Within this latent space we train a Flow Matching model to complete shapes based on grasp measurements such as contact points, fingertip positions and hand configurations. We introduce task-specific metrics and compare architecture variants.**

I. Objective

Many robotic applications require an efficient representation of geometry for tasks such as grasping, collision avoidance, and object recognition. However, explicit geometric descriptions like voxel grids or meshes scale poorly with resolution, making high-frequency geometric detail expensive to store and process [1, 2]. This becomes especially challenging when training generative models. We represent shapes in a continuous, low-dimensional latent space to enable the efficient training of generative models. As a downstream application we focus on shape estimation in tactile exploration.

II. Related Work

The success of generative models for 2D images motivated their application in the 3D domain. Since voxel representations as a natural 3D lifting of pixel representations scale cubically with resolution, more efficient variants such as octrees [3] have been explored. For most applications 3D data is inherently sparse, encouraging sparse representations, i.e. point clouds. PointNet, PointNet++ [4, 5] and subsequent point-based methods extract features from unstructured point clouds to represent geometry in compact feature vectors but face difficulties in encoding and decoding complex topologies. Similar challenges arise for meshes, so respective methods often rely on the use of template meshes [6]. Neural Fields such as Signed Distance Fields [7] and Occupancy Networks [8] enable continuous representations and adaptive memory allocation in the weights of small MLPs and have been used to train Variational Autoencoders (VAEs) in order to enable 3D latent diffusion models such as Diffusion-SDF [9]. The novel generative paradigm of Flow Matching [10] promises improved fidelity at reduced compute efforts. It was explored in combination with occupancy networks for 3D shape synthesis [11] and completion of buildings [12]. We transfer the approach to the domain of robotic tactile exploration.

III. Method

Our training and inference architecture is visualized in Figure 1. We train a Variational Autoencoder (VAE) on the task of encoding occupancy values at query positions to implicitly describe 3D shapes. Using the VAE, we encode a dataset of 3D shapes into latent codes. We condition a Flow Matching model on grasp measurements to predict corresponding shape latent codes that can be decoded to an occupancy grid representing a completed shape using the VAE decoder.

A. Dataset

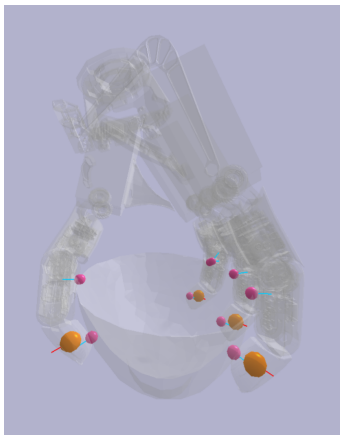

Fig. 2. Our training dataset is compiled from grasp simulations. We calculate hand-mesh collision centers (pink) and normals (blue) and fingertip positions (yellow) and orientations (red).

We base our dataset on simulated grasps of the DLR-Hand II on meshes from a subset of ShapeNet classes [13]. Following the preprocessing of the Acronym dataset [14] we avoid flipped normals and surface gaps by preprocessing the meshes with Manifold [15] resulting in $108,014$ grasps on $14,507$ watertight shapes. For each grasp a score quantifies the grasping success. We filter for successful grasps by setting a threshold of 10 and compute collision center points and normals between the hand geometry and the mesh to simulate tactile sensor measurements. We also extract the positions and orientations of the fingertips and the hand configuration consisting of the hand base transformation and joint angles. For each watertight mesh we uniformly sample 512 points within a cube around the object. We additionally sample 512 points distributed around the surface. For each query point we store position and occupancy.

B. Variational Autoencoder

We train a series of VAE models, evaluating suitable bottleneck dimensions, positional encoding and MLP-based architectures. Following the PointNet design [4] we explore max pooling over the feature dimensions. We compare sinusoidal (default), Fourier, and learnable encoding or no positional encoding to a raw input of the point coordinates without observing significant performance differences. The training objective follows the standard VAE loss of Binary Cross Entropy (BCE) of occupancy values and a Kullback Leibler Divergence (KL). To ensure convergence we schedule the influence of the KL-Term with a linear ramp. We evaluate the benefits of a reprojection error loss inspired by LeJEPA [16] and a cross-view training that samples different point sets for encoding and decoding. Although training time significantly increased, we did not obtain better results.

To evaluate design choices, we compare BCE and KL after convergence, investigate latent space statistics and measure prediction confidence with entropy. Additionally we evaluate latent space interpolations (Figure 3) and t-SNE projections. Encoder and decoder of the best-performing variant consist of MLPs with four hidden layers of dimension 256 that are projected to a bottleneck dimension of 32. We apply max pooling over features and use sinusoidal positional encoding.

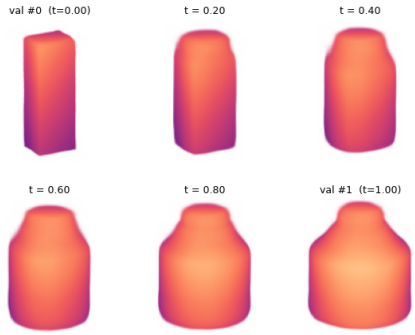

Fig. 3. Decoding of the latent interpolation between the first two val set items. Color encodes depth, voxel opacity is modulated with the predicted occupancy probability.

TABLE I
Architecture Comparison: VAE

Run	BCE ↓	KL ↓	std(σ/dim) ↓	entropy ↓
pos enc: none	0.1891	3.4282	**0.0776**	0.0241
pos enc: Fourier	0.1865	3.1454	0.0845	0.0229
pos enc: learnable	0.1766	3.2638	0.0931	0.0202
bottleneck = 32 (default)	0.1614	3.1843	0.1119	0.0166
bottleneck = 64	0.1598	**2.8301**	0.1096	0.0165
crossview	0.1655	2.9485	0.1449	0.0185
4-layer architecture	**0.1218**	3.2260	0.1169	**0.0097**

C. Unconditional Latent Denoising

We encode the occupancy dataset using the selected VAE to obtain a latent dataset that is used to train a denoising model. As architecture candidates we investigate different scales of MLP, MLP with residual connections (ResMLP), 1D UNet and DiT [17] by tokenizing per dim or per latent (flat).

We compare training with the objectives of Diffusion Frameworks DDPM [18] with cosine noise schedule [19] to predict added noise ϵ or the noise velocity v, the continuous SDE variant EDM [20] to predict a denoised estimate x_0 and Flow Matching [10] to predict a denoising velocity v under optimal transport constraint.

For all architectures we employ sinusoidal time embedding (dim 32) that we either concatenate to the input latent or that we use to modulate features following FiLM [21].

To evaluate performance, we calculate the kernel-based dissimilarity measure **Maximum Mean Discrepancy (MMD²)**, **Coverage** - the fraction of real samples whose ϵ-ball contains at least one generated sample, and **Density** - the average number of generated samples that fall inside the ϵ-ball of a real sample.

As the best-performing architecture, we identify Flow Matching using a ResMLP with a hidden dimension of 512, conditioned on the timestep using FiLM.

TABLE II
ARCHITECTURE COMPARISON: UNCONDITIONAL DENOISING

Run	MMD^2 ↓	Cov. ↑	Dens. ↑
MLP Baseline - ϵ-pred	0.403502	0.0000	0.0000
DiT - per latent tokenization	0.413730	0.0000	0.0000
DiT - per dim tokenization	0.026261	0.0103	0.0017
1D UNet - ϵ-pred	0.016977	0.7834	0.2870
ResMLP - ϵ-pred	0.443958	0.0014	0.0001
ResMLP - v-pred	0.001201	0.8400	0.3737
ResMLP - EDM ($\sigma = 10$)	0.001155	0.7793	0.3447
ResMLP - FM	**0.000426**	0.8434	0.3643
ResMLP (FiLM) - FM	0.000701	0.8586	0.3611
ResMLP (FiLM) - FM (h512)	0.000853	**0.9648**	**0.5309**

D. Conditional Latent Flow Matching

We expand the latent dataset to investigate strategies to condition the selected Flow Matching model on grasp information. In all versions we expand the FiLM conditioning by adding a linear projection of encoded grasp information.

Contacts - We consider two variants of encoding contact point coordinates by using a dedicated PointNet-style encoder that is trained with the Flow Matching model or our frozen pretrained VAE encoder and a learned linear mapping to the condition dimension. Since the VAE Encoder expects a fixed point count, we fill the remaining positions with randomly sampled positions labeled as *not occupied*. For both variants we additionally explore exploiting the contact normal information. In the case of the PointNet-style encoding we concatenate points and corresponding normals (resulting in 6 dimensions per contact) before extracting features. The VAE encoder expects position-occupancy tuples, so we calculate the position of an empty point with a fixed offset along the normal and add this set of points to the encoder input.

Warm initialization - When encoding contact positions with the pretrained VAE encoder, we receive vectors that live in the VAE latent space, but not on the manifold of valid shape latent codes. We leverage this fact to train Flow Matching models on the task of predicting velocity fields that approximate the optimal transport from initial latent to ground truth latent instead of denoising from Gaussian noise. We do so by initializing the mean of the source distribution with the VAE mean prediction.

Fingertips - To simulate applications without tactile sensors, we explore conditioning on the positions of fingertips and hand base. Here the count of points is fixed, so in addition to the PointNet and VAE Encoder approaches, we directly learn a linear mapping to the condition dimension. We again explore the benefits of normal information by optionally providing the fingertip and hand base orientation.

Hand configuration - The previously described approaches required knowledge of the forward kinematics. We investigate an edge case that solely considers the measurements of hand position and orientation as well as the three joint angles of each of the four fingers, resulting in an 18-dimensional array for each pose that is mapped to the conditioning dimension by an MLP.

E. Metrics

To mimic the frequent haptic exploration application of collision anticipation, we introduce a **Ray Visibility Error (RayV)** metric that computes the difference between the lengths of rays that are cast from the box surfaces until colliding with occupied voxels, and **Exterior Distance Field Error (extDF)** that measures differences in the free-space distance fields outside two shapes, ignoring interior filling.

The **Intersection over Union (IoU)** quantifies the agreement between two occupancy grids by computing the ratio between the number of voxels jointly predicted as occupied and the total number of voxels predicted as occupied by either grid. **Chamfer Distance (CD)** measures the average bidirectional distance between generated and reference surfaces. Consistent with real-world scenarios, our dataset contains both solid and hollow shapes. Therefore, a well-trained model may generate either representation for a given conditioning. However, IoU and Chamfer Distance penalize predictions that differ in volumetric occupancy from the ground truth (e.g., generating a solid shape for a hollow object or vice versa), despite these differences being irrelevant for tactile exploration, which primarily depends on the outer surface geometry. To stay compatible with further shape completion research, we nevertheless report these canonical metrics.

IV. RESULTS

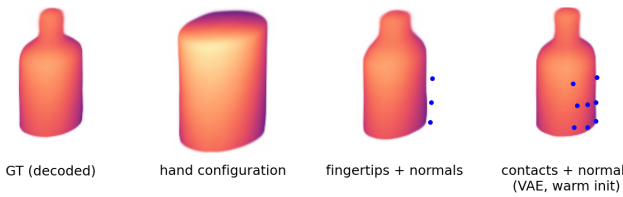

Fig. 4. Shape estimation for val set item 266 conditioned on hand config, fingertips+normals (embedded with linear mapping) and contacts+normals (embedded with VAE encoder and warm init)

All conditioning variants produce plausible shapes for most evaluation scenarios. As expected, providing more detailed information such as contact points instead of fingertip positions and normals in addition to positions results in higher similarity to ground truth samples. Even when conditioning with the

TABLE III
ARCHITECTURE COMPARISON: CONDITIONAL FLOW MATCHING (BEST CHECKPOINTS)

Run	MMD² ↓	Cov. ↑	Dens. ↑	IoU_{mesh} ↑	CD_{mesh} ↓	$extDF_{mesh}$ ↓	$rayV_{mesh}$ ↓
FM FiLM, cond: hand config	0.000423	0.9474	0.6891	0.3820	0.09879	0.10177	0.14327
FM FiLM, cond: fingertips	0.000418	0.9420	0.6678	0.4284	0.09637	0.09023	0.11908
FM FiLM, cond: fingertips (PointNet)	0.000360	0.9455	0.6800	0.3877	0.09749	0.09525	0.13356
FM FiLM, cond: fingertips (VAE)	0.000690	0.9435	0.6749	0.4325	0.09123	0.08681	0.11960
FM FiLM, cond: fingertips+normals	0.000374	0.9419	0.6704	0.4577	0.09305	0.08563	0.11053
FM FiLM, cond: fingertips+normals (PointNet)	**0.000077**	0.9025	0.5391	0.3961	0.10393	0.09836	0.13171
FM FiLM, cond: contacts (PointNet)	0.000400	0.9439	0.6973	0.4413	0.09015	0.08433	0.11607
FM FiLM, cond: contacts+normals (PointNet)	0.000207	**0.9576**	**0.7085**	0.5012	0.08592	0.07697	**0.09607**
FM FiLM, cond: contacts (VAE)	0.000517	0.9460	0.6736	0.4943	0.08183	0.07432	0.10064
FM FiLM, cond: contacts+normals (VAE)	0.000522	0.9501	0.6694	0.4944	0.08169	0.07434	0.09936
FM FiLM, warm init, cond: contacts+normals (VAE)	0.004743	0.8869	0.5670	**0.5119**	**0.06037**	**0.05978**	0.10065

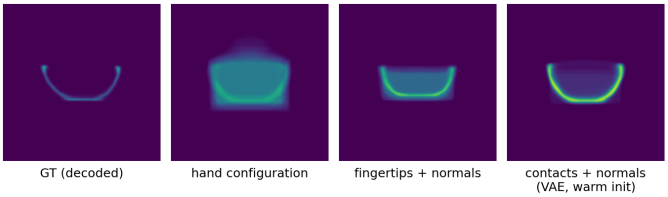

Fig. 5. XZ-slice of mean probabilities over 24 decoded samples (val set)

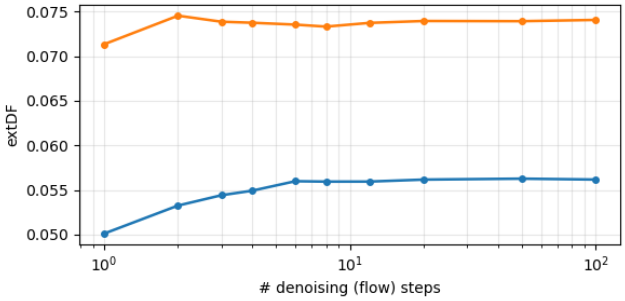

Fig. 6. Exterior distance-field error (extDF) over the validation set when sampling with 1 to 100 (training objective) steps: warm-init (blue) vs denoising from gaussian noise (orange). Lower is better.

hand configuration, plausible shapes are decoded, suggesting that the model successfully learned the forward kinematics.

For occupancy probability estimation, the mean over a batch of samples can be computed, resulting in probability maps that successfully respect uncertainty about object fillings and exact geometry far from conditioning information (Fig. 5).

When adjusting the flow matching objective to learn the optimal transport from encoded conditioning to a valid shape (warm-init), we observe a significant improvement of extDF at the cost of impaired generative metrics (MMD², Cov, Dens). Each model is trained to denoise in 100 steps, but sampling can extrapolate to fewer steps. Figure 6 visualizes accelerated denoising at superior similarity using warm initialization.

V. DISCUSSION

Our pipeline effectively estimates 3D shapes from sparse tactile data, producing plausible reconstructions. Notably, while providing explicit contact points and normals yields

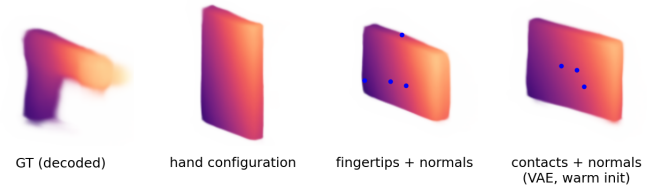

Fig. 7. Out-of-distribution performance on unseen geometry.

the highest similarity to ground truth samples, the model successfully decodes plausible shapes even when conditioned solely on the hand configuration. The architectural evaluations revealed that a ResMLP utilizing FiLM conditioning outperformed spatially biased models, such as DiT and 1D UNet, for unconditional denoising. This is likely because the continuous, compact latent space produced by the VAE lacks the rigid spatial structures that Transformers or UNets natively exploit. Heavy reliance on the learned VAE prior may limit generalization to novel, out-of-distribution objects (Figure 7). Employing warm initialization accelerates inference and improves surface alignment, though it naturally constrains generative sampling diversity. Finally, task-specific surface metrics (extDF, RayV) offer a far more faithful evaluation for robotic collision avoidance than canonical volumetric metrics, which penalize irrelevant internal geometry.

VI. CONCLUSION

Generally, the underlying hypothesis of predicting shapes leveraging latent generative models conditioned on sparse tactile data proved successful. However, the high prediction accuracy might heavily rely on a strong shape prior encoded by the VAE.

VII. FUTURE WORK

Our dataset contains samples from only few ShapeNet classes, mostly rotationally symmetric. A more diverse dataset would allow to transfer the approach to scenarios with higher shape fidelity and variance. Some potential architectural improvements, such as alternative conditioning approaches in addition to FiLM and constrained VAE architectures, like quantized VAEs, could improve the latent space structure.

References

[1] L. Zhou, Y. Du, and J. Wu, "3D Shape Generation and Completion through Point-Voxel Diffusion," Aug. 2021, arXiv:2104.03670 [cs]. [Online]. Available: http://arxiv.org/abs/2104.03670

[2] Z. Wang, D. Li, Y. Wu, T. He, J. Bian, and R. Jiang, "Diffusion Models in 3D Vision: A Survey," Apr. 2025, arXiv:2410.04738 [cs]. [Online]. Available: http://arxiv.org/abs/2410.04738

[3] B. Xiong, S. Wei, X. Zheng, Y. Cao, Z. Lian, and P. Wang, "OctFusion: Octree-based Diffusion Models for 3D Shape Generation," *Computer Graphics Forum*, vol. 44, no. 5, p. e70198, Aug. 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/cgf.70198

[4] C. R. Qi, L. Yi, H. Su, and L. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[5] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," Apr. 2017, arXiv:1612.00593 [cs]. [Online]. Available: http://arxiv.org/abs/1612.00593

[6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end Recovery of Human Shape and Pose," Jun. 2018, arXiv:1712.06584 [cs]. [Online]. Available: http://arxiv.org/abs/1712.06584

[7] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 165–174. [Online]. Available: https://ieeexplore.ieee.org/document/8954065/

[8] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy Networks: Learning 3D Reconstruction in Function Space," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 4455–4465. [Online]. Available: https://ieeexplore.ieee.org/document/8953655/

[9] G. Chou, Y. Bahat, and F. Heide, "Diffusion-SDF: Conditional Generative Modeling of Signed Distance Functions," Mar. 2023, arXiv:2211.13757 [cs]. [Online]. Available: http://arxiv.org/abs/2211.13757

[10] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," Feb. 2023, arXiv:2210.02747 [cs.LG]. [Online]. Available: http://arxiv.org/abs/2210.02747

[11] J. Xiang, X. Chen, S. Xu, R. Wang, Z. Lv, Y. Deng, H. Zhu, Y. Dong, H. Zhao, N. J. Yuan, and J. Yang, "Native and Compact Structured Latents for 3D Generation," Dec. 2025, arXiv:2512.14692 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2512.14692

[12] J. Sui, R. Liu, and H. Zhang, "OCCDiff: Occupancy Diffusion Model for High-Fidelity 3D Building Reconstruction from Noisy Point Clouds," Dec. 2025, arXiv:2512.08506 [cs]. [Online]. Available: http://arxiv.org/abs/2512.08506

[13] D. Winkelbauer, R. Triebel, and B. Bäuml, "A Learning-based Controller for Multi-Contact Grasps on Unknown Objects with a Dexterous Hand," Sep. 2024, arXiv:2409.12339 [cs]. [Online]. Available: http://arxiv.org/abs/2409.12339

[14] C. Eppner, A. Mousavian, and D. Fox, "ACRONYM: A Large-Scale Grasp Dataset Based on Simulation," Nov. 2020, arXiv:2011.09584 [cs.RO]. [Online]. Available: http://arxiv.org/abs/2011.09584

[15] J. Huang, H. Su, and L. Guibas, "Robust Watertight Manifold Surface Generation Method for ShapeNet Models," Feb. 2018, arXiv:1802.01698 [cs.CG]. [Online]. Available: http://arxiv.org/abs/1802.01698

[16] R. Balestriero and Y. LeCun, "LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics," Nov. 2025, arXiv:2511.08544 [cs.LG]. [Online]. Available: http://arxiv.org/abs/2511.08544

[17] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," Mar. 2023, arXiv:2212.09748 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2212.09748

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Dec. 2020, arXiv:2006.11239 [cs.LG]. [Online]. Available: http://arxiv.org/abs/2006.11239

[19] A. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," Feb. 2021, arXiv:2102.09672 [cs.LG]. [Online]. Available: http://arxiv.org/abs/2102.09672

[20] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the Design Space of Diffusion-Based Generative Models," Oct. 2022, arXiv:2206.00364 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2206.00364

[21] E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," Dec. 2017, arXiv:1709.07871 [cs.CV]. [Online]. Available: http://arxiv.org/abs/1709.07871