CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

Point-MF: Mean Flowsを用いた高速な単一画像点群生成

馬場 雄大[1,a)]　柳井 啓司[1,b)]

概要

単一画像からの点群再構成では, 1 枚の RGB 画像のみから, 遮蔽領域を含む完全な 3 次元形状を推定する必要がある. 拡散モデルベースの手法は高精度である一方, 多数のデノイジング反復を要するため, 推論が低速かつ高コストになりやすい. 本研究では, 低 NFE な単一画像点群再構成のための Mean Flow ベース手法 Point-MF を提案する。Point-MF は, VAE ベースの潜在表現を用いず, 点群空間上で平均速度場を直接学習することで, 1 回のネットワーク関数評価（1-NFE）によるワンステップ再構成を実現する. さらに、大きな時間区間のジャンプに対応するため, 凍結した DINOv3 特徴を軽量トークンアダプタで導入し, 区間・時刻条件付けを備えた Diffusion Transformer を採用する. また, 予測速度場から得られる denoised-space 推定値 x_θ に対し, 集合距離に基づく補助損失 Denoised Space Anchor を導入することで, 大ステップ生成を安定化し, 外れ値や点密度の不均一性を抑制する. ShapeNet-R2N2 および Pix3D における実験により, Point-MF は, 多段拡散ベースラインおよび feedforward 型手法と比較して, 再構成品質と推論速度の良好なバランスを達成し, ミリ秒オーダーの低遅延で高品質な点群を生成できることを示す.

1. はじめに

単一画像からの点群再構成は，1 枚の画像に含まれる可視情報のみから，不可視領域を含む完全な 3 次元形状を推定する必要があるため，本質的に不良設定な問題である．この課題は，AR/VR における空間理解や視覚提示，ならびに把持やナビゲーションを含むロボティクスの環境認識において重要であり，実運用では低遅延な推論が求められる．

点群は，前処理が比較的容易で，出力解像度を柔軟に制御できることから，3 次元再構成の中間表現として有用である [3], [16], [20], [25], [52].

既存の 3 次元再構成手法は，大きく feedforward 型手法 [2], [8], [15], [17], [26], [38], [48], [49] と拡散モデルベース手法 [6], [7], [18], [21], [24], [30], [31], [32], [35], [36], [40], [51], [53] に分類される．前者は高速であり，点群再構成では RGB2Point [26] が強力なベースラインであるが，不可視領域に対する複数の妥当な解を表現しにくく，過度に平滑化された再構成を生じやすい．一方，後者は確率的生成により不確実性を扱えるが，反復的サンプリングに伴う多数のネットワーク関数評価（network function evaluation, NFE）を要し，推論コストが高い（図 1-(a)）.

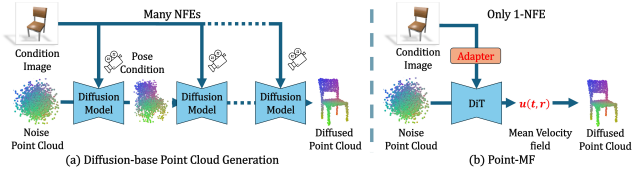


図 1　単一画像からの点群再構成手法の比較．代表的な拡散モデルベース手法 (a), Point-MF (b)

Mean Flow [11] は，Flow Matching [29] に基づく平均速度場を直接予測することで，NFE の削減を目指す手法である．しかし，その主な実証は事前学習済み VAE による潜在空間上で行われてきた．点群の潜在オートエンコーディングは，近傍ベース演算 [9], [34], [42], [50] やスパース 3 次元畳み込みを含む複雑な構造 [13], [53], [55], [57] に依存することが多く，実装・最適化の負担が大きい．加えて，このような動的演算はエッジデバイス向け形式への変換も容易ではない.

これらの課題に対し，本研究では，VAE ベースの潜在表現に依存せず，点群空間上で生成過程を直接定式化する Mean Flow ベースの低 NFE 単一画像点群再構成手法 **Point-MF** を提案する（図 1-(b)）．VAE を用いない場合，速度場からの生成では形状多様体からの逸脱や外れ値が生じやすい．そこで，速度場から誘導される denoised estimate x_θ に集合距離ベースの補助損失を導入し，生成を安定化する．合成・実画像データセットでの実験により，提案手法の有効性を示す．主な貢献は以下の通りである.

- 我々の知る限り初めて，単一画像からの点群再構成に対する Mean Flow ベースの低 NFE 生成フレームワークを導入し，latent-free な条件付き点群生成手法 **Point-MF** を提案する.
- 少数ステップ生成で生じやすい形状逸脱を抑制するた

[1] 電気通信大学
[a)] baba-y@gmm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

め，denoised estimate x_θ に対する集合距離ベースの補助損失を導入し，データ空間における幾何学的一貫性を直接促進する．

- ShapeNet-R2N2 および Pix3D における実験を通して，提案する pose-free な手法が，既存の feedforward 型手法および反復的な拡散モデルベース手法と比較して，再構成品質と推論効率の両立に有効であることを示す．

2. 関連研究

2.1 画像からの 3 次元再構成

画像に基づく 3 次元再構成では，メッシュ，ボクセル，点群，陰関数場など多様な表現が用いられてきた [8], [37], [38], [41], [48]．中でも点群は，構造が比較的単純で解像度制御が容易なため，単一画像条件付き再構成の表現として広く用いられている．RGB2Point [26] は feedforward による高速な点群生成を実現する一方，PC^2 [36] や BDM [51] などの拡散ベース手法は高品質な再構成が可能である反面，多数の NFE を要する．本研究は，この両者の間を埋めるべく，latent-free かつ低 NFE な点群生成を目指す．

2.2 拡散モデルおよびフローベースモデル

拡散モデルおよびフローベース生成モデルは，事前分布からデータ分布への写像を連続時間の軌道として定式化し，速度場やスコア関数を学習する [14], [29], [46]．高品質な生成が可能である一方，推論時には数値積分に基づく反復更新を必要とする．近年では，復元された clean output x_0 に対する補助損失が拡散モデルや少数ステップ生成の改善に有効であることが報告されており [1], [28]，本研究ではこの考え方を点群生成に拡張する．

2.3 少数ステップ生成モデル

標準的な拡散モデルおよびフローベースモデルは，多段の数値積分を前提としている．これに対し，低 NFE 生成を目的として，大きな時間更新を直接学習する少数ステップ生成モデルが提案されている [10], [11], [22], [45], [56]．

Mean Flow [11] は，2 時刻間の平均速度場を直接パラメータ化する手法であり，その後も改良が提案されている [12], [27], [33], [54]．一方，Pixel Mean Flows [33] は，latent-free な高次元空間では単純な平均速度場の直接予測が難しいことを示している．点群も N 点集合として $\mathbb{R}^{3N}$ に属するため，同様の課題を抱えうる．そこで本研究では，Mean Flow を点群空間へ導入し，denoised estimate を集合距離損失（DSA）で拘束することで，1 ステップ生成を安定化する．

3. 予備知識: Mean Flows

Mean Flow (MF) [11] は，Flow Matching における瞬間速度場 v の代わりに，2 時刻間の平均速度場 u を直接学習することで低 NFE 生成を可能にする．区間 $[r, t]$ における平均速度場は

$$u(x_t, r, t) = \frac{1}{t-r} \int_r^t v(x_\tau, \tau)\, d\tau \tag{1}$$

で定義され，対応する更新は $x_r = x_t - (t-r)u(x_t, r, t)$ となる．本研究では，単一画像条件付き点群再構成のために classifier-free guidance (CFG) を組み込んだ条件付き Mean Flow を用い，学習には次の目的関数を用いる：

$$\mathcal{L}_{\text{MF-CFG}}(\theta) = \mathbb{E}\left[\|u_\theta(x_t, r, t \mid c) - \text{sg}(u_{\text{tgt}})\|_2^2\right]. \tag{2}$$

詳細な導出および u_{tgt} の構成は Supplementary Sec. A に示す．

4. 提案手法

4.1 アーキテクチャ

アーキテクチャの詳細は Supplementary Sec. B に譲り，本論文では主要な設計のみを述べる．全体概要を図 2 に示す．本手法では，凍結した DINOv3 [4], [44] により入力画像からグローバル特徴 $\mathbf{z}_{\text{img}}$ とパッチ特徴 Z_{ctx} を抽出し，それらを Diffusion Transformer (DiT) に入力して平均速度場 $u_\theta(x_t, r, t \mid c)$ を予測する．グローバル特徴は時刻埋め込み $e_t(t)$ および区間長埋め込み $e_{dt}(dt)$ とともに AdaLN-Zero の条件ベクトルとして用い，パッチ特徴は軽量アダプタを介して Cross-Attention のコンテキストとして与える．

DiT は Self-Attention，Cross-Attention，MLP からなるブロックを積層した構造をとる．また，点群は順序を持たないため，インデックス依存の位置埋め込みを用いず，置換等変性を保ったまま各点に対する速度場を予測する．さらに，凍結 DINOv3 特徴を点群生成へ適応させるため，画像側に軽量な Post-MHSA Adapter (PMA) を導入する．PMA は画像パッチ特徴間の相互作用を一度だけ計算し，各 DiT ブロックに適した形へ再射影することで，条件表現を強化する．

4.2 損失関数

4.2.1 Denoised Space Anchor

Mean Flow では区間平均速度場を用いて大きな時間更新を行うため，予測速度の誤差が最終的な位置誤差へ増幅されやすい．また，点群は順序を持たない集合であるため，速度場のみの回帰では，最終的な再構成における幾何学的一貫性が必ずしも保証されない．そこで本研究では，図 3 に示すように，予測された平均速度場から denoised-space

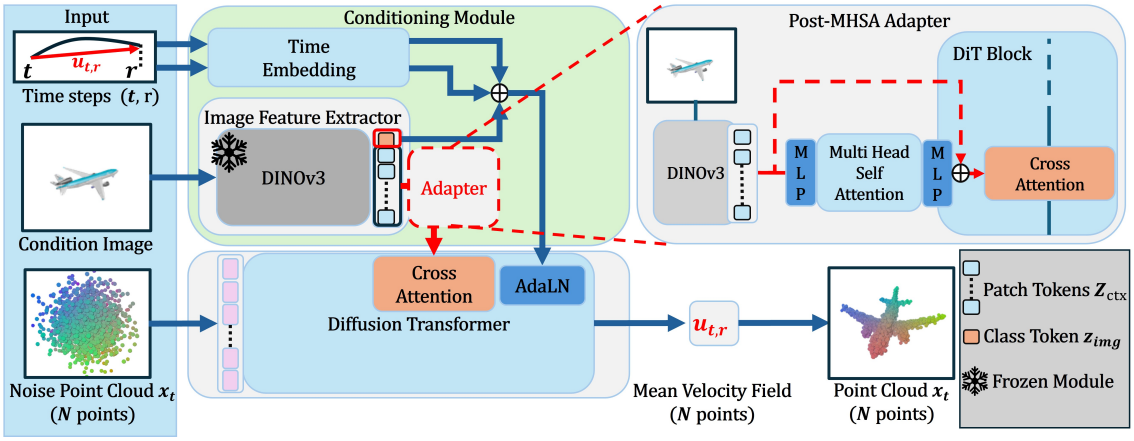


図 2　Point-MF アーキテクチャの全体像.

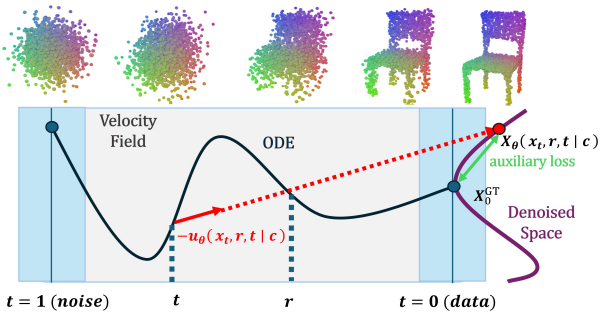


図 3　Denoised Space Anchor (DSA) の概念図.

上の再構成点群を $x_r = x_t - (t-r)u(x_t, r, t)$ にて $r = 0$ として,

$$x_\theta(x_t, r, t \mid c) = x_t - t\, u_\theta(x_t, r, t \mid c) \tag{3}$$

を構成し，これを正解点群 x_0^{GT} に直接整合させる補助損失 **Denoised Space Anchor (DSA)** を導入する．点群は順序を持たないため，本研究では集合距離として点間の 1 対 1 の対応を確率的に計算する, Adaptive Probabilistic Matching Loss (APML) [43] を用いる．DSA は次式で定義される：

$$\mathcal{L}_{\mathrm{DSA}} = L_{\mathrm{APML}}\big(x_\theta(x_t, r, t \mid c),\, x_0^{\mathrm{GT}}\big)\,. \tag{4}$$

最終的な学習目的関数は，Mean Flow の主損失に DSA を加えた

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\text{MF-CFG}} + s\lambda(t)\mathcal{L}_{\mathrm{DSA}} \tag{5}$$

とする．ここで，$\lambda(t)$ の具体的な設計，スケール調整係数 s，および実際の学習時の詳細は Supplementary Sec. C に示す．この設計により，点群空間における clean-data 側の幾何学的一貫性を直接促進する．

5.　実験

本節では，提案手法の再構成品質，推論時間，および VRAM 使用量を評価する．追加結果，限界，アブレーションは Supplementary Sec. D に示す．

5.1　データセット

単一画像条件付き点群生成を，ShapeNet [5] および Pix3D [47] で評価する．ShapeNet については，先行研究 [8], [39] で用いられている 13 カテゴリからなる ShapeNet-R2N2 の分割に従う．実画像と物体形状の対応を含む Pix3D については，Xu et al. [51] の分割に従う．

5.2　評価指標

生成点群 $\hat{X}$ と，同じ点数を持つ正解点群 X との間の L1 Chamfer Distance（CD）および Earth Mover's Distance（EMD）を用いて，再構成品質を評価する．CD は計算効率に優れる一方，多対一対応に起因して点密度の偏りに対する感度が低いため，本研究では CD と EMD を相補的な指標として用いる．

すべての手法は，ShapeNet-R2N2 では 1024 点，Pix3D では 2048 点で評価する．低 NFE 生成の実用性を評価するため，推論時間およびピーク VRAM 使用量も比較する．推論は単一の NVIDIA RTX A4000 GPU 上で実施し，平均と標準偏差を報告する．

5.3　定量評価

表 1 に ShapeNet-R2N2 における CD および EMD の結果を示し，Pix3D における結果は表 2 に示す．

表 1 に示すように，Point-MF は，1-NFE のみを用いるにもかかわらず，多段サンプリングを必要とする拡散ベース手法を上回る性能を示す．RGB2point は CD を学習目的として訓練されているため，学習目標と評価指標が一致しており，CD 指標において有利である．一方で，PC^2, BDM，Point-MF のような確率的生成手法と比較すると，EMD の観点では劣る．さらに，実画像を含む Pix3D データセットにおいても（表 2），Point-MF は 1-NFE のみで総合的に最も優れた性能を達成した．

表 3 に推論速度の比較を示す．Point-MF は確率的生成モデルであるにもかかわらず，フィードフォワード型手法

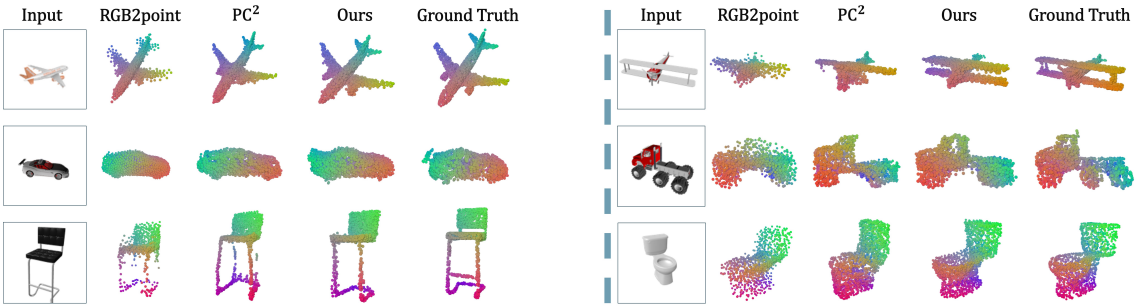


図 4 ShapeNet の定性評価.

表 1 ShapeNet-R2N2 における CD, EMD での定量評価.

	CD×100 ↓				EMD×100 ↓			
Method	Car	Chair	Air	Avg.	Car	Chair	Air	Avg.
Self-Sup. [39]	5.48	10.91	7.11	7.11	4.95	14.93	11.07	10.31
DIFFER [23]	6.35	9.78	5.67	7.27	6.03	16.21	9.90	10.71
ULSP [19]	5.40	9.72	5.91	7.01	4.78	10.18	7.66	7.54
RGB2point [26]	**4.22**	5.43	**2.70**	4.33	5.63	9.53	6.86	7.83
PC² [36] (256-NFE)	4.95	5.60	3.37	4.84	4.69	5.36	4.27	4.84
BDM [51] (276-NFE)	5.12	5.02	3.12	4.64	4.25	4.53	3.92	4.28
Ours (1-NFE)	**4.22**	**4.29**	2.76	**3.92**	**3.82**	**4.29**	**3.04**	**3.82**
w/o PMA	4.58	4.29	3.01	4.12	4.02	4.62	3.54	4.13
w/o DSA	5.21	6.16	3.71	5.23	4.89	6.08	4.12	5.16

太字 は最良値を示し, 下線 は次点を示す. (アブレーションは除外).

表 2 Pix3D における CD, EMD の定量評価.

	CD×100 ↓				EMD×100 ↓			
Method	Chair	Sofa	Table	Avg.	Chair	Sofa	Table	Avg.
PC² (256-NFE)	7.07	5.87	11.39	7.82	7.13	6.14	10.72	7.75
BDM (276-NFE)	6.62	5.55	8.35	7.12	7.05	**4.98**	8.79	7.12
RGB2point	6.62	5.40	9.70	7.06	9.34	7.11	12.69	9.59
Ours (1-NFE)	**6.48**	**5.02**	**8.20**	**6.53**	**6.74**	5.06	**7.94**	**6.61**
w/o PMA	7.58	5.58	8.45	7.28	7.23	5.52	8.34	7.06
w/o DSA	8.82	6.58	10.1	8.58	8.88	7.02	9.76	8.62

と同程度のオーダーで動作し，拡散ベース手法と比べて大幅に高速である.

5.4 定性評価

ShapeNet-R2N2（図 4）および Pix3D（図 5）の両方において定性評価を行い，生成点群の形状忠実性および点密度分布を視覚的に評価した．全体として，Point-MF は形状の一貫性を保ちながら，点密度の不均衡を抑えた点配置を生成する．この傾向は，表 1 において Point-MF が一貫して良好な EMD を示しているという定量結果とも整合的である．一方，Chamfer Distance を直接最小化する RGB2point は，点が表面へ強く引き寄せられるため，輪郭が鋭くなる傾向があるが，局所的な点密度の偏りや，特定部位への点の集中が見られる場合がある．Pix3D では，Point-MF はテーブル下面や複雑な椅子脚のような難しい構造も再構成できている.

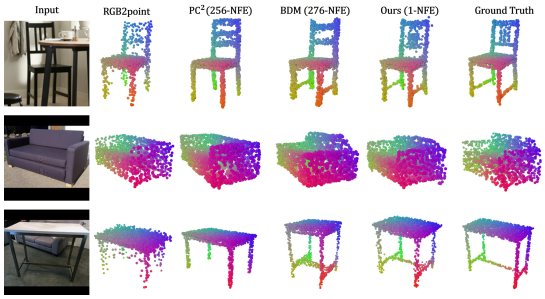


図 5 Pix3D の定性評価.

表 3 推論時間とピーク VRAM (RTX A4000).

Method	Time [ms/sample]	VRAM [GiB]
PC² (256-NFE)	22000 ± 200	1.730
BDM (276-NFE)	28000 ± 200	1.970
RGB2point	28.39 ± 2.18	0.818
Ours (1-NFE)	63.45 ± 0.25	1.282

6. 結論と考察

本論文では，低 NFE サンプリングの下で幾何学的一貫性の高い点群を生成することを目的とした．我々の知る限り，Mean Flow に基づく平均速度回帰を単一画像条件付き点群生成に適用したのは本研究が初めてであり，ポーズ情報を必要としない DiT アーキテクチャと，denoised space における幾何学的一貫性を促進する **Denoised Space Anchor（DSA）** を提案した.

ShapeNet-R2N2, Pix3D において, 提案手法は CD, EMD の各指標で既存手法を上回り，再構成された点群の被覆性および密度整合性の改善を示した．さらに，実行時間の比較により，提案手法は拡散ベース手法よりも大幅に高速な推論を実現しており，実用性を示している．以上の結果は，3 次元再構成における Mean Flow の有効性を示すものである.

今後の課題としては，Improved Mean Flows [12] で提案された推論時の可変ガイダンススケールの導入や，Pixel Mean Flow [33] で示唆されているように，主目的関数を x_θ 上で直接定義するような変換の導入が有望である.

参考文献

[1] An, J., Yang, Z., Wang, J., Li, L., Liu, Z., Wang, L. and Luo, J.: Bring metric functions into diffusion models, *Proc. of the International Joint Conference on Artificial Intelligence*, pp. 578–586 (2024).

[2] Boss, M., Huang, Z., Vasishta, A. and Jampani, V.: Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16240–16250 (2025).

[3] Boulch, A. and Marlet, R.: Poco: Point convolution for surface reconstruction, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6302–6314 (2022).

[4] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A.: Emerging properties in self-supervised vision transformers, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660 (2021).

[5] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L. and Yu, F.: ShapeNet: An Information-Rich 3D Model Repository, *arXiv:1512.03012* (2015).

[6] Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L. and Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2416–2425 (2023).

[7] Chou, G., Bahat, Y. and Heide, F.: Diffusion-sdf: Conditional generative modeling of signed distance functions, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 2262–2272 (2023).

[8] Choy, C. B., Xu, D., Gwak, J., Chen, K. and Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, *Proc. of the European Conference on Computer Vision*, pp. 628–644 (2016).

[9] Du, B., Hu, W. and Liao, R.: Multi-scale latent point consistency models for 3d shape generation, *arXiv preprint arXiv:2412.19413* (2024).

[10] Frans, K., Hafner, D., Levine, S. and Abbeel, P.: One Step Diffusion via Shortcut Models, *Proc. of the International Conference on Learning Representations* (2025).

[11] Geng, Z., Deng, M., Bai, X., Kolter, J. Z. and He, K.: Mean flows for one-step generative modeling, *Advances in Neural Information Processing Systems*, Vol. 38 (2025).

[12] Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Kolter, J. Z. and He, K.: Improved mean flows: On the challenges of fastforward generative models, *arXiv preprint arXiv:2512.02012* (2025).

[13] Graham, B.: Sparse 3D convolutional neural networks, *Proc. of the British Machine Vision Conference*, pp. 150–1 (2015).

[14] Ho, J., Jain, A. and Abbeel, P.: Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851 (2020).

[15] Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T. and Tan, H.: LRM: Large Reconstruction Model for Single Image to 3D, *Proc. of the International Conference on Learning Representations* (2024).

[16] Huang, Z., Boss, M., Vasishta, A., Rehg, J. M. and Jampani, V.: Spar3d: Stable point-aware reconstruction of 3d objects from single images, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16860–16870 (2025).

[17] Huang, Z., Jampani, V., Thai, A., Li, Y., Stojanov, S. and Rehg, J. M.: ShapeClipper: Scalable 3D Shape Learning from Single-View Images via Geometric and CLIP-Based Consistency, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12912–12922 (2023).

[18] Huang, Z., Johnson, J., Debnath, S., Rehg, J. M. and Wu, C.-Y.: Pointinfinity: Resolution-invariant point diffusion models, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 10050–10060 (2024).

[19] Insafutdinov, E. and Dosovitskiy, A.: Unsupervised Learning of Shape and Pose with Differentiable Point Clouds, *Advances in Neural Information Processing Systems*, Vol. 31 (2018).

[20] Jignasu, A., Balu, A., Sarkar, S., Hegde, C., Ganapathysubramanian, B. and Krishnamurthy, A.: SDFConnect: Neural Implicit Surface Reconstruction of a Sparse Point Cloud with Topological Constraints, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5271–5279 (2024).

[21] Jun, H. and Nichol, A.: Shap-e: Generating conditional 3d implicit functions, *arXiv preprint arXiv:2305.02463* (2023).

[22] Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y. and Ermon, S.: Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion, *Proc. of the International Conference on Learning Representations* (2024).

[23] L Navaneet, K., Mandikal, P., Jampani, V. and Babu, V.: DIFFER: Moving Beyond 3D Reconstruction with Differentiable Feature Rendering, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019).

[24] Lan, Y., Hong, F., Yang, S., Zhou, S., Meng, X., Dai, B., Pan, X. and Loy, C. C.: ln3Diff: Scalable Latent Neural Fields Diffusion for Speedy 3D Generation, *Proc. of the European Conference on Computer Vision*, Springer, pp. 112–130 (2024).

[25] Lan, Y., Zhou, S., Lyu, Z., Hong, F., Yang, S., Dai, B., Pan, X. and Loy, C. C.: GaussianAnything: Interactive Point Cloud Latent Diffusion for 3D Generation, *Proc. of the International Conference on Learning Representations* (2025).

[26] Lee, J. J. and Benes, B.: Rgb2point: 3d point cloud generation from single rgb images, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2952–2962 (2025).

[27] Lee, K., Yu, S. and Shin, J.: Decoupled meanflow: Turning flow models into flow maps for accelerated sampling, *Proc. of the International Conference on Learning Representations* (2026).

[28] Lee, S., Lin, Z. and Fanti, G.: Improving the training of rectified flows, *Advances in Neural Information Processing Systems*, Vol. 37, pp. 63082–63109 (2024).

[29] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. and Le, M.: Flow Matching for Generative Modeling, *Proc. of the International Conference on Learning Representations* (2023).

[30] Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J. and Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion, *Proc. of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10072–10083 (2024).

[31] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S. and Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object, *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309 (2023).

[32] Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T. and Wang, W.: SyncDreamer: Generating Multiview-consistent Images from a Single-view Image, *Proc. of the International Conference on Learning Representations* (2024).

[33] Lu, Y., Lu, S., Sun, Q., Zhao, H., Jiang, Z., Wang, X., Li, T., Geng, Z. and He, K.: One-step Latent-free Image Generation with Pixel Mean Flows, *arXiv preprint arXiv:2601.22158* (2026).

[34] Lyu, Z., Wang, J., An, Y., Zhang, Y., Lin, D. and Dai, B.: Controllable mesh generation through sparse latent point diffusion models, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 271–280 (2023).

[35] Ma, B., Han, Z., Liu, Y.-S. and Zwicker, M.: Neural-Pull: Learning Signed Distance Functions from Point Clouds by Learning to Pull Space onto Surfaces, *Proc. of the International Conference on Machine Learning*, Vol. 139 (2021).

[36] Melas-Kyriazi, L., Rupprecht, C. and Vedaldi, A.: Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12923–12932 (2023).

[37] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S. and Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470 (2019).

[38] Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V. and Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images, *Proc. of the European Conference on Computer Vision*, pp. 414–431 (2020).

[39] Navaneet, K., Mathew, A., Kashyap, S., Hung, W.-C., Jampani, V. and Babu, R. V.: From image collections to point clouds with self-supervised shape and pose networks, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1132–1140 (2020).

[40] Nichol, A., Jun, H., Dhariwal, P., Mishkin, P. and Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts, *arXiv:2212.08751* (2022).

[41] Park, J. J., Florence, P., Straub, J., Newcombe, R. and Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174 (2019).

[42] Qi, C. R., Yi, L., Su, H. and Guibas, L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space, *Advances in Neural Information Processing Systems*, Vol. 30 (2017).

[43] Sharifipour, S., Casado, C. Á., Sabokrou, M. and López, M. B.: APML: Adaptive Probabilistic Matching Loss for Robust 3D Point Cloud Reconstruction, *Advances in Neural Information Processing Systems*, Vol. 38 (2025).

[44] Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M. et al.: Dinov3, *arXiv:2508.10104* (2025).

[45] Song, Y., Dhariwal, P., Chen, M. and Sutskever, I.: Consistency models, *Proc. of the International Conference on Machine Learning* (2023).

[46] Song, Y. and Ermon, S.: Generative modeling by estimating gradients of the data distribution, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).

[47] Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J. B. and Freeman, W. T.: Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2974–2983 (2018).

[48] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W. and Jiang, Y.-G.: Pixel2mesh: Generating 3d mesh models from single rgb images, *Proc. of the European Conference on Computer Vision*, pp. 52–67 (2018).

[49] Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H. and Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model, *Proc. of the European Conference on Computer Vision*, pp. 57–74 (2024).

[50] Wu, W., Qi, Z. and Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630 (2019).

[51] Xu, H., Lei, Y., Chen, Z., Zhang, X., Zhao, Y., Wang, Y. and Tu, Z.: Bayesian diffusion models for 3d shape reconstruction, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10628–10638 (2024).

[52] Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K. and Neumann, U.: Point-nerf: Point-based neural radiance fields, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448 (2022).

[53] Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S. and Kreis, K.: LION: latent point diffusion models for 3D shape generation, *Advances in Neural Information Processing Systems*, pp. 10021–10039 (2022).

[54] Zhang, H., Siarohin, A., Menapace, W., Vasilkovsky, M., Tulyakov, S., Qu, Q. and Skorokhodov, I.: Alphaflow: Understanding and improving meanflow models, *Proc. of the International Conference on Learning Representations* (2026).

[55] Zhou, C., Zhong, F., Hanji, P., Guo, Z., Fogarty, K., Sztrajman, A., Gao, H. and Oztireli, C.: Frepolad: Frequency-rectified point latent diffusion for point cloud generation, *Proc. of the European Conference on Computer Vision*, pp. 434–453 (2024).

[56] Zhou, L., Ermon, S. and Song, J.: Inductive Moment Matching, *Proc. of the International Conference on Machine Learning* (2025).

[57] Zhu, D., Di, Y., Gavranovic, S. and Ilic, S.: Sealion: Semantic part-aware latent point diffusion models for 3d generation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11789–11798 (2025).