

E-M3RF: An Equivariant Multimodal 3D Re-assembly Framework

Adeela Islam^{1,2}, Stefano Fiorini¹, Manuel Lecha^{1,2}, Theodore Tsismelis¹,
Stuart James³, Pietro Morerio¹, and Alessio Del Bue¹

¹ Istituto Italiano di Tecnologia, Genoa, Italy

² University of Genoa, Genoa, Italy

³ Durham University, Durham, UK

{adeela.islam}@iit.it

Abstract. 3D reassembly is a fundamental geometric problem, and in recent years it has increasingly been challenged by deep learning methods rather than classical optimization. While learning approaches have shown promising results, most still rely primarily on geometric features to assemble a whole from its parts. This reliance limits their effectiveness in cases where geometric information is incomplete or degraded, such as for small or eroded fragments. Additionally, solutions do not impose physical constraints that explicitly prevent overlapping assemblies. To address these limitations, we introduce E-M3RF, an equivariant multimodal 3D reassembly framework that takes as input the point clouds, containing both point positions and color of fractured fragments, and predicts the transformations required to reassemble them, using $SE(3)$ flow matching. Each fragment is represented by both geometric and color features: i) 3D point positions are encoded as rotation-consistent geometric features using a rotation-equivariant encoder, while ii) the colors at each 3D point are encoded with a transformer. The two feature sets are then combined to form a multimodal representation. We experimented on five datasets: two synthetic datasets, Breaking Bad and Fantastic Breaks, and three real-world cultural heritage datasets, RePAIR, Presious, and FRACTURA, demonstrating that E-M3RF on the RePAIR dataset (the most challenging one) reduces rotation error by **23.1%** and translation error by **13.2%**, while Chamfer Distance decreases by **18.4%** compared to competing methods. Project page: <https://em3rf.github.io/>.

Keywords: Reassembly · Multimodal Learning · Equivariance · Frescoes · Cultural Heritage

1 Introduction

Reassembly tasks play a fundamental role in many domains, including reconstructing archaeological 3D artifacts [25], piecing together shredded documents [17], molecular docking [4], and solving jigsaw or 3D puzzles [6, 21].

Accurate reassembly of fractured fragments requires precise estimation of each piece’s $SE(3)$ transformation that rebuilds the whole 3D object shape. This

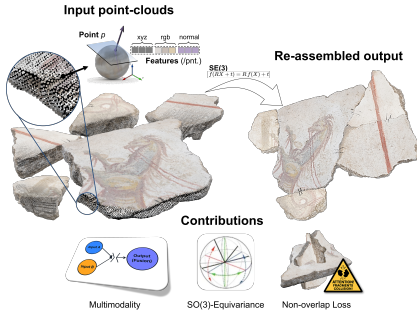


Fig. 1: E-M3RF overview. Fragments are represented as colored point clouds with per-point features $(x, y, z, \text{rgb}, \mathbf{n})$. A multimodal $SE(3)$ -equivariant backbone fuses geometry and color and predicts per-fragment rigid transforms (R, t) via flow-based estimation, while a differentiable non-overlap loss enforces physical plausibility. The predicted poses ensure coherent reassembly by aligning fracture geometry and color patterns.

estimation becomes challenging when geometric cues are weak or degraded, such as in eroded fragments. In recent years, 3D fracture reassembly has remained a fundamentally geometric optimization task but has increasingly been approached using deep learning [20, 21] rather than classical optimization [12, 31]. Learning-based methods have progressed along two complementary trends: improving fragment representations via powerful encoders and pre-training schemes that emphasize fine-scale geometry and surface appearance [8, 11], and reframing pose estimation as a generative task, where flow- or diffusion-based pipelines iteratively refine fragment alignments [20]. Despite these recent advances, geometry-centric designs have practical limitations: geometric information alone can be ambiguous [25], while real fragments often exhibit appearance details, such as paint, grain, or tool marks. These discriminative cues are ignored by geometry-only pipelines, overlooking an important source of information. Moreover, geometry-only methods often fail to account for the formal group structure of transformations acting on the pieces. This is compounded by state-of-the-art models’ struggle with physical constraints, which often results in intersecting fragments [8, 20].

To address the limitations of the geometry-only approaches and handle complex fractures, we introduce E-M3RF, a multimodal 3D reassembly framework that takes as input the colored point clouds of fractured fragments and predicts the corresponding $SE(3)$ transformations via $SE(3)$ flow matching [2, 3, 13], see Fig. 1. The model jointly leverages fragment color and geometric information: an $SO(3)$ -equivariant encoder extracts rotation-consistent geometric features from origin-centered point clouds (*i.e.*, no translation involved), while a parallel color branch encodes per-fragment appearance. The resulting features are concatenated into a joint multimodal embedding per fragment point. Both rotations and translations are predicted with a transformer-based architecture trained via a flow-matching objective on the $SE(3)$ manifold. A differentiable no-overlap loss further enforces physical plausibility by penalizing fragment intersections, as illustrated in Fig. 2.

We evaluate E-M3RF on a combination of synthetic and real fracture datasets, including the large-scale synthetic Breaking Bad (BB) [21] and Fantastic Breaks (FB) [9], as well as the real-world Presious [22], RePAIR [25] and FRACTURA [11]. In all five datasets, we consistently outperform geometry-only baselines, achieving higher part accuracy and lower rotation, translation, and Chamfer Distance

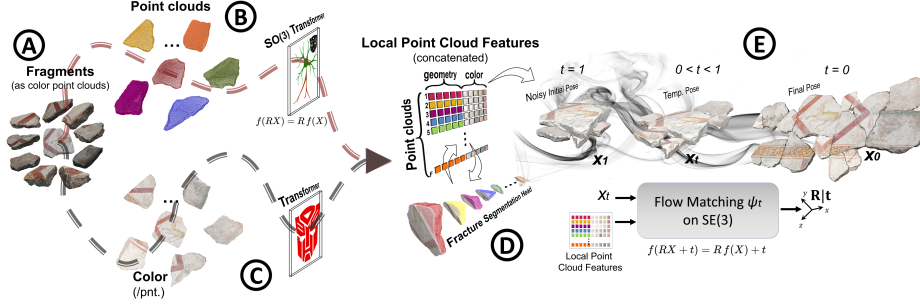


Fig. 2: E-M3RF’s pipeline. (A) A set of textured fragments (left) loaded as colored point clouds (w/ per-point RGB). Two transformer encoders extract (B) **fracture-aware geometric features**, which are geometrically equivariant, since they are processed by an $SO(3)$ -equivariant **Transformer** so representations transform consistently under motions (rotation/ translation) and (C) **color features** through a Transformer over per-point colors which extracts dense color descriptors. The streams are concatenated into (D) **local point-cloud features** (xyz/rgb/normals tokens); the geometry stream is guided by a fracture-segmentation head during pretraining. Using these fused features, we perform (E) **flow matching on $SE(3)$** : a time-dependent vector field ψ_t transports fragment poses from a noisy initialization x_1 through intermediate states to the assembled configuration x_0 , yielding per-fragment transforms (R, t) . During training only, a **non-overlap loss** penalizes interpenetration to enforce physically plausible assemblies (omitted from the illustration for clarity).

errors. To quantify the impact of our design choices, we perform ablations over each component.

The main contributions of this work are:

- An end-to-end multimodal fragment representation that using both geometric and color features, producing a joint appearance–geometry embedding that resolves ambiguities and leverages surface patterns when available.
- We leverage an inductive bias in the form of a rotation-equivariant geometric backbone. Specifically, $SO(3)$ -equivariant layers ensure that features transform consistently under rotations, improving robustness to arbitrary fragment orientations and enhancing generalization.
- We enforce physical plausibility via a differentiable no-overlap loss that penalizes fragment intersections, applied only during training.

2 Related Work

We review the state of the art in 3D reassembly and puzzle-solving problems, together with a deep dive into representation encoders commonly used in the solution of 3D tasks.

2.1 3D Fracture Assembly

3D puzzle solving and fractured-object reassembly have been approached through a broad range of techniques, spanning classical geometric registration, learning-based pose estimation, and more recent generative formulations. Classical pipelines rely on hand-crafted local descriptors, correspondence hypothesis testing, and ICP-style refinement [16, 26]. These methods provide interpretable matching criteria grounded in geometry and remain effective for clean, low-ambiguity fragments. Learning-based methods shift to data-driven feature extraction and direct pose prediction. Recent works learn per-point or per-fragment embeddings and estimate 6-DoF alignments using regression or transformer-based architectures [11, 15, 20, 28], often supported by large-scale synthetic fracture corpora such as BB [21]. Generative approaches, including diffusion or flow-based estimators [11, 14, 20, 28], treat assembly as sampling or transporting distributions over $SE(3)$, enabling iterative refinement and uncertainty-aware reasoning. These methods represent a growing direction for globally coherent reconstruction.

Although real-world fragments contain useful color cues, existing learning-based reassembly methods rely only on geometric information. E-M3RF addresses this limitation by fusing geometric and color features into a unified multimodal representation, enabling more discriminative fragment descriptors and more reliable alignment.

2.2 Representation (Encoder)

PointNet [18] introduced a shared-MLP encoder with global feature aggregation, proving that deep learning on point sets is effective for 3D tasks. PointNet++ [19] extended PointNet with hierarchical neighborhood grouping, allowing the network to capture local geometric features across multiple scales. Subsequent methods strengthened local feature modeling through convolutional and graph-based operators. EdgeConv [27] captures point-to-point relationships via dynamic graphs, while KPConv [24] employs kernel-point convolutions as a continuous 3D filtering mechanism, providing strong geometric inductive bias. Transformer-based encoders have recently gained prominence in 3D vision. Point Transformer [30] and PCT [7] use self-attention to capture long-range point dependencies, while hierarchical variants like Swin3D [29] apply windowed attention to 3D grids. However, these models are not inherently rotation-aware and do not ensure consistent features under arbitrary $SO(3)$ transformations. To address this issue, Vector Neurons [5] lift point features from scalars to 3D vectors, allowing rotations to act via matrix multiplication. Building on this representation, the VN-Transformer [1] extends self-attention to vector features, yielding a rotation-equivariant attention mechanism that preserves $SO(3)$ consistency throughout the network and allows reasoning over 3D geometry in a rotation-aware manner.

To effectively address reassembly scenarios where fragments undergo arbitrary rotations, we leverage this rotation-equivariant inductive bias so that the backbone inherits $SO(3)$ transformations directly into the learned features, pro-

viding E-M3RF with representations that remain geometrically consistent under all input orientations.

3 Problem Statement

We formalize the geometric object reassembly problem given a finite collection of disjoint fragments.

Fragments. Let $P = \{p_i\}_{i=1}^N$ be a finite set of points. A fragment is defined as the tuple

$$\mathcal{F} = (P, \mathcal{X}, \{\mathcal{G}_k\}_{k=1}^{K_{\text{geo}}}, \{\mathcal{A}_k\}_{k=1}^{K_{\text{app}}}). \quad (1)$$

The map $\mathcal{X} : P \rightarrow \mathbb{R}_{\text{xyz}}^3$ assigns spatial coordinates, *i.e.* the 3D coordinates for each point. Each geometric attribute $\mathcal{G}_k : P \rightarrow \mathbb{R}^3$ transforms under the standard vector representation of $SE(3)$ (*e.g.*, normals), while each appearance attribute $\mathcal{A}_k : P \rightarrow \mathbb{R}^{d_k}$ is invariant under rotations (*e.g.*, color). In general, there can be $K_{\text{geo}} \in \mathbb{N}_0$ geometric attributes and $K_{\text{app}} \in \mathbb{N}_0$ appearance attributes. Fixing an ordering of the points induces two structured feature spaces:

$$V_{\text{geo}} = \mathbb{R}^N \otimes \mathbb{R}^{K_{\text{geo}}+1} \otimes \mathbb{R}^3, \quad V_{\text{app}} = \mathbb{R}^N \otimes \mathbb{R}^{d_{\text{app}}}, \quad (2)$$

where the $K_{\text{geo}} + 1$ geometric channels include $\mathcal{X}$ as a dedicated vector channel, and $d_{\text{app}} = \sum_{k=1}^{K_{\text{app}}} d_k$ is the total dimensionality of all appearance channels. The full fragment representation lies in

$$V_F = V_{\text{geo}} \oplus V_{\text{app}}, \quad (3)$$

where $\oplus$ denotes feature concatenation and we write $F \in V_F$ for the feature tensor associated with a fragment. We denote its positional component by $F_{\text{xyz}} \in \mathbb{R}^N \otimes \mathbb{R}^3$.

Unassembled Object. An unassembled object is represented as a collection of M fragments $\mathcal{F}$, each recentered so that its centroid lies at the origin:

$$\hat{\mathcal{O}} = \left\{ \hat{F}_i \in V_F \mid \frac{1}{N_i} \mathbf{1}_{N_i}^\top \hat{F}_{\text{xyz},i} = \mathbf{0} \in \mathbb{R}^3 \right\}_{i=1}^M,$$

where $\mathbf{1}_{N_i}$ is the all-ones vector of length N_i .

Assembled Object. We assume the existence of a canonical assembled configuration of the object. Each centered fragment $\hat{F}_i$ is placed in this configuration by a rigid transformation $g_i \in SE(3)$, acting exclusively on its positional channel. The assembled object is then denoted by

$$\mathcal{O} = \{F_i = g_i \cdot \hat{F}_i\}_{i=1}^M.$$

Object Reassembly. Given any unassembled object $\hat{\mathcal{O}} = \{\hat{F}_i\}_{i=1}^M$, the goal is to recover rigid transformations $\tilde{g}_i \in SE(3)$ that approximate the unknown ground-truth g_i . We therefore seek a single model

$$\Phi : \hat{\mathcal{O}} \rightarrow SE(3)^M, \quad \Phi(\hat{\mathcal{O}}) \approx \{g_i\}_{i=1}^M,$$

that operates on arbitrary unassembled objects and whose predictions transform and assemble the fragments:

$$\Phi(\hat{\mathcal{O}}) \cdot \hat{\mathcal{O}} \approx \mathcal{O}.$$

4 Methodology

We identify two capabilities required for object reassembly: (1) a representation expressive enough to capture the geometric and appearance cues that govern how fragments fit together, and (2) a mechanism that uses this representation to predict the rigid transformations realizing the correct assembly. Having formalized the task in Section 3, our methodology is built to satisfy these requirements in a principled and rotation-aware manner, see Fig. 3. We first learn fragment-level descriptors using an $S_N \times SO(3)$ -equivariant encoder, ensuring that the learned representations respect intrinsic fragment permutations and 3D rotations (more details are provided in Appx. A of the suppl. material). These descriptors condition a geometric generative model, a conditional $SE(3)$ Riemannian flow-matching network, which predicts a coherent set of rigid transformations aligning fragments into their assembled configuration. The full pipeline couples equivariant representation learning with flow-based geometric reasoning, yielding a unified approach that maps any unassembled object $\hat{\mathcal{O}}$ to an assembled one.

4.1 Encoding Fragments

Colored Fragments. We consider fragments whose per-point attributes consist of positions, surface normals, and RGB color. Following the formulation in Eq. (1) of Sec. 3, each fragment is represented as

$$\mathcal{F} = (P, \mathcal{X}, \mathcal{N}, \mathcal{C}), \quad \mathcal{X}, \mathcal{N}, \mathcal{C} : P \rightarrow \mathbb{R}^3,$$

where $\mathcal{X}$ gives point coordinates, $\mathcal{N}$ gives surface normals, and $\mathcal{C}$ gives per-point color. Accordingly, the geometric and appearance feature spaces in Eq. (2) specialize to

$$V_{\text{geo}} = \mathbb{R}^N \otimes \mathbb{R}^2 \otimes \mathbb{R}^3, \quad V_{\text{rgb}} = \mathbb{R}^N \otimes \mathbb{R} \otimes \mathbb{R}^3.$$

where the geometric features consist of the 3D point coordinates $\mathcal{X}$ and the corresponding surface normals $\mathcal{N}$. Each fragment is represented by an element $F \in V_F$ in the joint space of Eq. (3), with canonical projections: $F_{\text{geo}} \in V_{\text{geo}}$ collecting the point coordinates and normals; $F_{\text{xyz}} \in \mathbb{R}^N \otimes \mathbb{R}^3$ giving the point coordinates; $F_{\text{n}} \in \mathbb{R}^N \otimes \mathbb{R}^3$ giving the normals; and $F_{\text{rgb}} \in V_{\text{rgb}}$ giving the per-point color vectors.

Fragment Transformations. Since each fragment $\hat{\mathcal{O}}$ is a rigid body processed in isolation, the fragment’s representation is invariant under translations. We therefore center each fragment at the origin [11, 20]. Consequently, its geometric variations are fully captured by global rotations $SO(3) \subset SE(3)$. Additionally, the ordering of points induces discrete index permutations described by the group S_N . The resulting transformation group acting on geometric features is

$$\mathbf{G} := S_N \times SO(3),$$

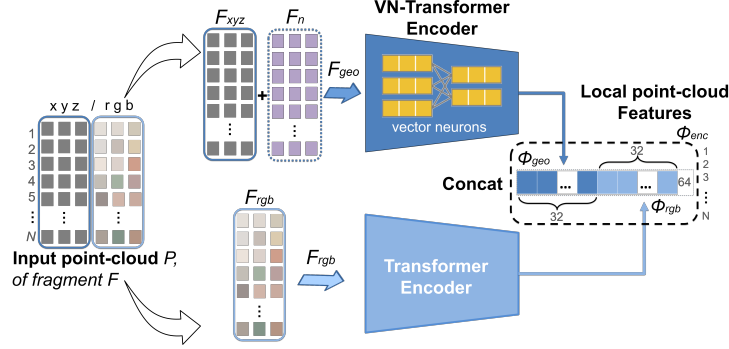


Fig. 3: Geometry/color feature extraction and fusion. The input point cloud P (per-point xyz , rgb and normals F_n), of fragment $\mathcal{F}$ is split into F_{geo} and F_{rgb} . A **VN-Transformer encoder** ($SO(3)$ -equivariant via vector neurons) produces geometric features Φ_{geo} , while a standard **Transformer encoder** produces color features Φ_{rgb} . The features are concatenated per point to yield $\Phi_{enc} = [\Phi_{geo} \parallel \Phi_{rgb}]$, which serve as the local point-cloud representation for downstream pose estimation.

where S_N permutes point indices and $SO(3)$ acts on each geometric $\mathbb{R}^3$ channel. For $\sigma \in S_N$, let P_σ denote the associated permutation matrix, and let I_2 denote the identity on the two geometric channels. The induced action on $F_{geo} \in V_{geo}$ is

$$(\sigma, R) \cdot F_{geo} = (P_\sigma \otimes I_2 \otimes R) F_{geo}, \quad (\sigma, R) \in \mathbf{G}, \quad (4)$$

where $\otimes$ is the Kronecker product implementing the tensor-product representation. In pointwise notation, for $i \in \{1, \dots, N\}$,

$$((\sigma, R) \cdot F_{geo})_i = \begin{bmatrix} R(F_{xyz})_{\sigma^{-1}(i)} \\ R(F_n)_{\sigma^{-1}(i)} \end{bmatrix} \in \mathbb{R}^2 \otimes \mathbb{R}^3,$$

where the first component corresponds to permuted and rotated coordinates and the second to permuted and rotated normals.

For color features $F_{rgb} \in V_{rgb}$, the $SO(3)$ rotations act trivially. Using the same tensor-product structure, the induced action is

$$(\sigma, R) \cdot F_{rgb} = (P_\sigma \otimes I_1 \otimes I_3) F_{rgb}, \quad (\sigma, R) \in \mathbf{G}, \quad (5)$$

reflecting that RGB values are invariant under rotations and only affected by the pointwise ordering.

Geometric Encoder Φ_{geo} . The geometric branch implements a $\mathbf{G}$ -equivariant map using a VN-Transformer [1],

$$\Phi_{geo} : V_{geo} \rightarrow \mathbb{R}^N \otimes \mathbb{R}^{C_{geo}} \otimes \mathbb{R}^3,$$

which outputs C_{geo} equivariant vector channels, each with values in $\mathbb{R}^3$. The $\mathbf{G}$ -action on this output space is obtained by lifting the action in Eq. (4) from

the two input geometric channels to the C_{geo} learned channels. As a result, Φ_{geo} satisfies

$$\Phi_{\text{geo}}((P_{\sigma} \otimes I_2 \otimes R) F_{\text{geo}}) = (P_{\sigma} \otimes I_{C_{\text{geo}}} \otimes R) \Phi_{\text{geo}}(F_{\text{geo}}),$$

for all $(\sigma, R) \in \mathbf{G}$. Here $I_{C_{\text{geo}}}$ denotes the identity in $\mathbb{R}^{C_{\text{geo}}}$.

Color Encoder Φ_{rgb} . Color features are encoded by a Transformer

$$\Phi_{\text{rgb}} : V_{\text{rgb}} \rightarrow \mathbb{R}^N \otimes \mathbb{R}^{C_{\text{rgb}}} \otimes \mathbb{R}^3,$$

which operates on rotation-invariant channels. Consequently, the only nontrivial component of the $\mathbf{G}$ -action is the permutation action of S_N . Thus, the $SO(3)$ factor of $\mathbf{G}$ acts trivially on V_{rgb} , and the only nontrivial group action arises from permutations of point indices. Lifting the permutation action of Eq. (5) to the output space, Φ_{rgb} satisfies

$$\Phi_{\text{rgb}}((P_{\sigma} \otimes I_1 \otimes I_3) F_{\text{rgb}}) = (P_{\sigma} \otimes I_{C_{\text{rgb}}} \otimes I_3) \Phi_{\text{rgb}}(F_{\text{rgb}}),$$

for all $\sigma \in S_N$.

Fused Encoder Φ_{enc} . Let $C = C_{\text{geo}} + C_{\text{rgb}}$. We define the fused encoder

$$\Phi_{\text{enc}} : V_F \longrightarrow \mathbb{R}^N \otimes \mathbb{R}^C \otimes \mathbb{R}^3, \text{ by } \Phi_{\text{enc}}(F) = [\Phi_{\text{geo}}(F_{\text{geo}}) \parallel \Phi_{\text{rgb}}(F_{\text{rgb}})],$$

where $\parallel$ denotes concatenation along the feature-channel axis. For each point i , the geometric block $\Phi_{\text{geo}}(F_{\text{geo}})_i \in \mathbb{R}^{C_{\text{geo}}} \otimes \mathbb{R}^3$ and the color block $\Phi_{\text{rgb}}(F_{\text{rgb}})_i \in \mathbb{R}^{C_{\text{rgb}}}$ are merged into a single C -channel tensor. Since both branches respect the $\mathbf{G}$ -action on their respective representation spaces, the fused encoder Φ_{enc} is $\mathbf{G}$ -equivariant by construction.

4.2 Fracture-Boundary Segmentation Pretraining

The encoder is pretrained via an auxiliary fracture-boundary segmentation task, which encourages the learned fragment representations to capture the geometric signatures of fracture interfaces. For each fragment $\hat{F}_i$, the encoder produces the pointwise embedding

$$H_i = \Phi_{\text{enc}}(\hat{F}_i) \in \mathbb{R}^{N_i} \otimes \mathbb{R}^C \otimes \mathbb{R}^3. \quad (6)$$

A lightweight segmentation head, using a two-layer MLP

$$\Psi_{\text{seg}} : H_i \longrightarrow \hat{y}_i, \quad \hat{y}_i \in [0, 1]^{N_i},$$

predicts per-point boundary probabilities, supervised by binary ground-truth masks. The pretraining objective optimizes the composition $\Psi_{\text{seg}} \circ \Phi_{\text{enc}}$ to correctly identify fracture boundaries, forcing Φ_{enc} to learn geometry-appearance interactions characteristic of fracture morphology. This procedure produces embeddings that maintain consistent geometric information, which are then used as inputs to the subsequent flow-matching model on $SE(3)$.

4.3 Fragment Assembly

Given the encoded fragment features, $H_i = \Phi_{\text{enc}}(\hat{F}_i)$, we enrich each fragment with geometric priors using standard Fourier positional encoding. Specifically, we concatenate multi-frequency sinusoidal embeddings $\text{PE}(\cdot)$ [23] of point cloud coordinates $\hat{F}_{\text{xyz}}$, normals $\hat{F}_{\text{n}}$, and scale information $s_i \in \mathbb{R}$ along the channel dimension. Let $\mathcal{H} := \mathbb{R}^{N_\star} \otimes \mathbb{R}^{C_{\text{enc}}} \otimes \mathbb{R}^3$, where $N_\star$ is a fixed maximum number of points. The enriched fragment representation becomes:

$$H_i := f_{\text{shape}} \left([H_i \parallel \text{PE}((\hat{F}_{\text{xyz}})_i) \parallel \text{PE}((\hat{F}_{\text{n}})_i) \parallel \text{PE}(s_i)] \right),$$

where $f_{\text{shape}}(\cdot)$ is the shape embedding function. This provides high-frequency geometric cues that improve pose estimation. We additionally introduce a learned *anchor token* that designates the reference fragment around which all poses are predicted. The anchor serves as a global reference frame and resolves the intrinsic gauge ambiguity of assembly, since the set of fragment poses is only defined up to a shared rigid transformation in $SE(3)$. Fixing one fragment removes this degenerate symmetry and yields a well-defined target for learning relative poses. Following prior work, the anchor fragment is randomly selected during training to improve robustness across different reference choices, while at inference time we use the largest fragment as the anchor for a stable geometric reference [11]. Conditioned on H_i , we define a conditional Riemannian flow-matching model on the Lie group $SE(3)$ which learns a continuous velocity field whose integration yields the rigid transformation $g_i = (R_i, \beta_i) \in SE(3)$ required to place the fragment i in its correct location within the assembled object. Formally, the model parametrizes a time-dependent velocity field

$$v : SE(3) \times [0, 1] \times \mathcal{H} \longrightarrow \mathfrak{se}(3) = \mathfrak{so}(3) \times \mathbb{R}^3,$$

which associates to each element $g \in SE(3)$, time $t \in [0, 1]$, and conditioning features $H_i \in \mathcal{H}$ a tangent vector $v(g, t; H_i) \in \mathfrak{se}(3)$. During training, the flow matching objective aligns this vector field with the target probability path interpolating between an initial distribution $p_0(g) = \mathcal{U}(SO(3)) \otimes \mathcal{N}(\mathbf{0}, I_3)$ and the empirical distribution of assembled configurations. At inference time, we utilize Euler integration for flow sampling where the learned vector field is integrated over the unit interval as

$$\dot{g}_i^t = v(g_i^t, t; H_i), \quad g_i^0 = (I, \mathbf{0}),$$

yielding the final transformations g_i^1 reassembling the object:

$$\mathcal{O} = \{ g_i^1 \cdot \hat{F}_i \}_{i=1}^M.$$

This formulation defines a continuous flow on $SE(3)$ that preserves geometric consistency while transporting fragments from their initial centered configuration to globally consistent positions in the assembled object. In detail, the conditional

trajectory between an initial pose g_i^0 and the target pose g_i^1 follows a geodesic interpolation in rotation and a linear interpolation in translation:

$$R_i^t = \exp_{R_i^0}(t \log_{R_i^0}(R_i^1)), \quad \beta_i^t = (1-t)\beta_i^0 + t\beta_i^1.$$

Accordingly, let us denote the rotational and translational velocity fields as

$$v_R^{(i)}(g_i^t, t) \in \mathfrak{so}(3), \quad v_\beta^{(i)}(g_i^t, t) \in \mathbb{R}^3.$$

Following the flow-matching formulation, these predicted velocities are trained to match the target geodesic residuals scaled by $(1-t)^{-1}$. Hence, the objective is expressed as:

$$\begin{aligned} \mathcal{L}_{\text{flow}} = & \mathbb{E}_{t, p_1(g_i^1), p_t(g_i^t|g^1)} \left[\frac{1}{N} \sum_{i=1}^N \left(\left\| v_\beta^{(i)}(g_i^t, t) - \frac{\beta_i^1 - \beta_i^t}{1-t} \right\|_2^2 \right. \right. \\ & \left. \left. + \left\| v_R^{(i)}(g_i^t, t) - \frac{1}{1-t} \log_{R_i^t}(R_i^1) \right\|_2^2 \right) \right], \end{aligned} \quad (7)$$

where $\log_{R_i^t}(R_i^1) \in \mathfrak{so}(3) \cong \mathbb{R}^3$ denotes the Lie-algebra residual from R_i^t to R_i^1 . This formulation ensures that predicted fragment transformations follow the correct geodesic trajectories on $SE(3)$, maintaining smooth and consistent rotational and translational flows.

4.4 No-Overlap Loss

To prevent implausible fragment intersections, we introduce a differentiable no-overlap loss. For each pair of fragments (i, j) , we compute soft occupancy masks $M_i(x), M_j(x) \in [0, 1]$, which denote the soft occupancy values of fragments i and j at volume cell (grid location) x in a regular volumetric grid. We define the pairwise IoU:

$$\text{IoU}_{ij} = \frac{\sum_x \min(M_i(x), M_j(x))}{\sum_x \max(M_i(x), M_j(x)) + \varepsilon}, \quad (8)$$

where $\varepsilon > 0$ ensures numerical stability. The no-overlap loss aggregates these pairwise terms:

$$\mathcal{L}_{\text{no-overlap}} = \frac{1}{|P|} \sum_{(i,j) \in P} \text{IoU}_{ij}. \quad (9)$$

Minimizing $\mathcal{L}_{\text{no-overlap}}$ enforces non-overlap between fragments, improving visual quality and physical plausibility.

The full loss combines the flow-matching term and the no-overlap penalty:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \alpha \mathcal{L}_{\text{no-overlap}}, \quad (10)$$

where α balances pose accuracy with overlap prevention. Together, the two-stage training and differentiable no-overlap loss enable the network to produce $SE(3)$ predictions that are both accurate and physically consistent.

5 Experimental Evaluation

We evaluate our method on five different datasets. The datasets, evaluation metrics, and competing methods are described in Section 5.1, while Section 5.2 reports the performance of all methods in various ablation studies, including generalization, along with some qualitative results. Section B of the supplementary material provides additional ablations and analysis as well as technical details on computational efficiency and model parameters.

5.1 Dataset and Evaluation Metrics

Datasets. Following the evaluation protocol in [11], we evaluate on both large-scale synthetic and real scanned datasets, enabling assessment of complementary challenges in 3D assembly. For the colorless (geometry-only) setting, we use: (i) Breaking Bad [21], a synthetic dataset of over one million fractured instances derived from about 10k base shapes (PartNet/Thing10k), providing clean ground-truth poses and diverse geometry for controlled ablations and large-scale pre-training; (ii) Fantastic Breaks [9], with paired scans of intact and broken real objects capturing realistic micro-fracture morphology and scanning artifacts for high-fidelity geometric evaluation; and (iii) FRACTURA [11], a diverse dataset curated with domain experts that includes real-world fracture types across ceramics, bones, eggshells, and lithics. For the colored fragments (multimodal setting), we evaluate on RePAIR [25] and the Presious [22] datasets. RePAI is a collection of archaeological fresco fragments with high-resolution RGB imagery and 3D geometry, and Presious is a set of several 3D cultural heritage fragment groups. Both are well suited to assessing color-based cues and multimodal fusion challenges, like erosion and missing parts. Due to the size of the Presious dataset, *i.e.* only six sets, we use it only for testing the generalization accuracy of E-M3RF. Accordingly, we report the all “with color” results only on the RePAIR and Presious datasets, while geometry-only ablations and generalization studies are conducted on all aforementioned datasets. Following [11, 28], we sample 5k points per object, ensuring uniform per-fragment density.

Metrics. We report four measures: **RMSE** ($\mathcal{R}^\circ$) is the root-mean-square rotation error (in degrees); **RMSE** ($\mathcal{T}_{\text{mm}}$) is the root-mean-square translation error (in millimeters); **PA** (Part Accuracy) is the percentage of fragments whose per-fragment Chamfer Distance to the ground-truth placement falls below 0.01%; and **CD** is the Chamfer Distance between the assembled object point cloud and the ground-truth assembly. We keep the same definitions and threshold as in [11] to ensure a fair comparison across datasets and baselines. The best performance metric is highlighted in **bold**, while the second-best competitor is underlined.

5.2 Results

Table 1 compares E-M3RF against baselines on the color-enabled RePAIR and Presious datasets. On RePAIR, our complete model outperforms GARF [11],

reducing rotation and translation errors by 23.1% and 13.2%, and Chamfer Distance by 18.4%, while improving Part Accuracy by 15.3%. This lead persists even when GARF is augmented with naive color concatenation to point coordinates. These results underscore that geometry-only pipelines overlook critical disambiguating information available in color-enabled data.

Table 1: Quantitative results on RePAIR and Presious datasets. See Fig. 4 for corresponding qualitative results.

Method	RMSE ($\mathcal{R}^\circ$) $\downarrow$	RMSE ($\mathcal{T}_{mm}$) $\downarrow$	PA (%) $\uparrow$	CD $\downarrow$
<i>Tested on the RePAIR dataset</i>				
DiffAssemble [20]	69.54	67.96	17.92	4.18
PMTR [10]	76.72	71.40	12.24	5.23
PF++ [28]	67.75	70.49	20.82	18.97
GARF [11]	49.31	32.19	31.14	2.66
GARF [11] (w/ color)	<u>46.56</u>	<u>31.82</u>	<u>31.92</u>	<u>2.25</u>
E-M3RF (w/o color)	40.17	29.01	33.71	2.98
E-M3RF (w/o non-overl. loss)	45.81	25.01	32.91	3.83
E-M3RF (w/ rotation-inv.)	43.27	41.60	32.50	3.71
E-M3RF (w/o rotation-equiv.)	43.22	29.28	32.26	3.07
E-M3RF	37.91	27.93	35.91	2.17
<i>Presious: Generalizing from RePAIR</i>				
DiffAssemble [20]	56.65	176.41	24.72	37.73
PMTR [10]	80.01	68.33	14.27	18.21
PF++ [28]	70.61	146.84	24.79	41.58
GARF [11]	43.33	14.24	50.27	18.01
GARF [11] (w/ color)	<u>42.05</u>	<u>14.11</u>	<u>51.95</u>	17.05
E-M3RF (w/o color)	35.12	13.85	52.91	21.18
E-M3RF (w/o non-overl. loss)	36.12	11.91	50.05	21.43
E-M3RF (w/ rotation-inv.)	41.26	45.01	47.12	24.62
E-M3RF (w/o rotation-equiv.)	39.67	15.72	46.11	23.71
E-M3RF	30.26	12.85	57.49	20.95

E-M3RF exploits its superiority mainly via the dense per-point color features fused with fracture-aware geometry, while its $SO(3)$ -equivariant backbone makes the learned features transform consistently under pose, improving rotation estimates in particular. The non-overlap loss further suppresses physically implausible interpenetrations, reflected in the lower CD despite tighter placement (higher PA). This is corroborated by the ablations that we have conducted, where we examine the effect of each of our contributions to the overall pipeline:

- **w/o color:** Removing the color branch causes E-M3RF to rely solely on geometric features. PA decreases by 2.2 percentage points (pp), CD increases by 37%, rotation error increases by 6%, and translation error increases by 3.9% in comparison to the full model. These results indicate that color provides complementary cues to geometry, helping the model distinguish similar surfaces and resolve ambiguities that geometry alone cannot.

- **w/o non-overlap loss:** Removing the non-overlap loss causes the largest increase in CD (76%) and a decrease in PA by 3.0 pp. Interestingly, translation RMSE decreases (27.93 $\rightarrow$ 25.01 mm). However, this improvement is misleading: without the non-overlap loss, fragments interpenetrate, yielding closer centroids but worse CD and PA. These results confirm that the non-overlap loss is essential

for plausible placements and accurate surface alignment, as it can be clearly seen in Fig. 4 where in none of our solutions there is an overlap.

- **w/ rotation-inv.:** We replace the rotation-equivariant backbone with a rotation-invariant one, removing the inductive bias toward learning orientation-consistent features. We achieve rotation invariance by appending a VN-Invariant that converts rotation-equivariant features into normalized, rotation-invariant representations (*i.e.* an invariant projection). Rotation RMSE increases by 14.1%, translation RMSE rises by 48.9%, PA decreases by 9.5 pp, and CD worsens by 71.0%. These results show that rotation-invariance degrades performance, as it removes orientation cues essential for alignment and thus the model cannot distinguish rotated inputs, resulting in larger errors.

- **w/o rotation-equiv.:** We replace the rotation-equivariant backbone with a standard architecture lacking any inductive bias toward orientation consistency. CD increases most (29%), while rotation error increases by 12%, PA decreases by 3.6 pp, and translation error increases by 5%. These results highlight that rotation-equivariant features act as a crucial inductive bias for this task, enabling consistent reasoning under arbitrary orientations and improving alignment stability and generalization.

Table 2: Quantitative results on Breaking Bad, Fantastic Breaks and FRACTURA datasets.

Methods	RMSE ($\mathcal{R}^\circ$) $\downarrow$	RMSE ($\mathcal{T}_{mm}$) $\downarrow$	PA (%) $\uparrow$	CD $\downarrow$
<i>Breaking Bad: Tested on the Everyday Subset</i>				
Jigsaw [15]	42.19	6.85	68.89	8.22
PMTR [10]	31.57	9.95	70.60	5.56
PF++ [28]	35.61	6.05	76.17	2.78
GARF-mini [11]	6.68	1.34	94.77	0.25
GARF [11]	<u>6.10</u>	<u>1.22</u>	<u>95.33</u>	<u>0.22</u>
E-M3RF (w/o color) - - - -	5.31	1.14	96.20	0.18
<i>Breaking Bad: Tested on the Artifact Subset</i>				
Jigsaw [15]	43.75	7.91	65.12	8.50
PF++ [28]	47.03	10.63	57.97	8.24
GARF-mini [11]	7.67	1.77	93.34	0.81
GARF [11]	<u>5.82</u>	<u>1.27</u>	<u>95.04</u>	<u>0.42</u>
E-M3RF (w/o color) - - - -	4.63	1.07	96.80	0.20
<i>Fantastic Breaks: Generalizing from Breaking Bad</i>				
Jigsaw [15]	26.30	6.43	73.64	10.47
PF++ [28]	20.68	4.37	83.33	6.68
GARF [11]	<u>10.62</u>	<u>2.10</u>	<u>91.00</u>	<u>2.12</u>
E-M3RF (w/o color) - - - -	7.96	2.02	92.01	2.19
<i>FRACTURA: Generalizing from Breaking Bad on the Real Subset</i>				
Jigsaw [15]	62.25	12.21	29.71	36.32
PF++ [28]	66.55	15.26	32.75	25.02
GARF [11]	<u>45.03</u>	<u>1.18</u>	<u>46.36</u>	<u>1.14</u>
E-M3RF (w/o color) - - - -	37.36	0.15	52.90	0.91

Table 2 compares performance on Breaking Bad, Fantastic Breaks, and FRACTURA. Even when using only the geometry stream due to lack of color data

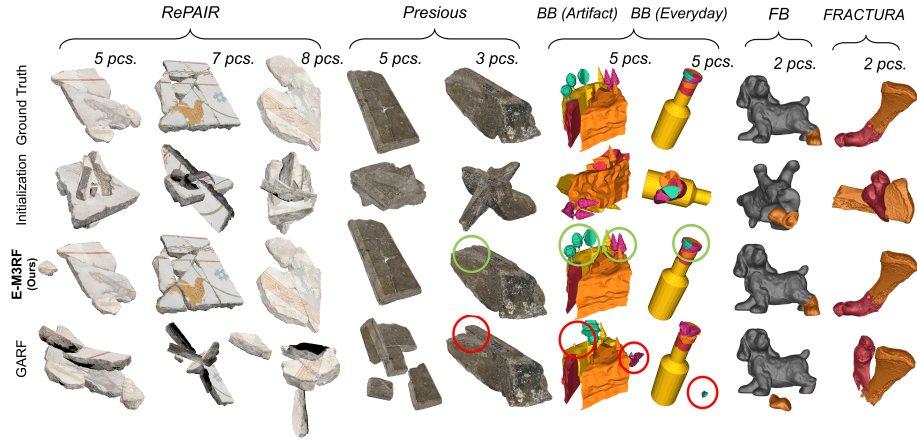


Fig. 4: Qualitative Comparisons on the RePAIR, Presious, Breaking Bad (BB), Fantastic Breaks (FB) and FRACTURA. E-M3RF consistently produces more accurate re-assemblies. It generalizes well to unseen objects, particularly on Presious, FB, and FRACTURA scenes. Green circles highlight fine, ambiguous contact regions correctly recovered by our method. See Appx. C in the suppl. mat. for more visualizations.

in these datasets, E-M3RF outperforms all baselines. On Everyday objects, it reduces rotation error by 13.0%, translation by 6.6%, and CD by 18.2%. On Artifact subset, it yields even larger gains, reducing CD by 35.5%. Averaged across subsets, our method improves rotation, translation, CD, and PA by approximately 15%, 11%, 27%, and 1.2 pp respectively over the second best, GARF.

Zero-shot Generalization: We evaluate E-M3RF’s zero-shot generalization on Presious (color+geometry), Fantastic Breaks and FRACTURA (both geometry-only), and fine-tune from RePAIR and Breaking Bad for each setting respectively. Without any training or fine-tuning on Presious (bottom of Table 1), E-M3RF achieves the lowest rotation and translation errors (30.26° and 12.85 mm) and the highest part accuracy (57.49%), demonstrating strong zero-shot generalization to the new dataset. The full model consistently outperforms all variants on every metric except Chamfer Distance, where it is a close second, and each ablation degrades performance as expected. Removing color (w/o color) lowers PA and increases CD, as appearance cues no longer disambiguate matches; removing the non-overlap loss (w/o non-overl. loss) increases CD by allowing overlap even for plausible poses; replacing the rotation-equivariant backbone with a rotation-invariant one (w/ rotation-inv.) harms both rotation and slightly translation; removing rotation equivariance (w/ rotation-equiv.) primarily degrades rotation accuracy and slightly destabilizes translation. On Fantastic Breaks (Table 2), E-M3RF’s equivariant geometry backbone generalizes beyond the training distribution: despite missing color, it outperforms [11] by 25.0% in rotation and translation, and matches it on other metrics. On FRACTURA (bottom of Table

2), E-M3RF demonstrates significant generalization capabilities. It outperforms GARF by 16.9% in rotation and 87.3% in translation, while achieving superior performance in both precision (PA) and Chamfer Distance (CD).

Geometry-only baselines such as GARF lag behind the full E-M3RF, showing that our fusion of color with an equivariant backbone transfers robustly to unseen data without dataset-specific training, demonstrating strong out-of-distribution generalization (see Fig. 4).

Efficiency and Computational Cost. E-M3RF is a substantially more lightweight and computationally efficient architecture. This is reflected in the number of trainable parameters during pretraining, requiring only 1.9M parameters compared to 12.7M for the baseline. Additionally, E-M3RF achieves approximately 17% faster inference and reduces memory consumption by around 58% on average (Table 3). Additional analysis is provided in Appx. B.

Table 3: Inference time and GPU memory comparison on the RePAIR dataset under varying numbers of fragments. Lower values indicate better efficiency.

# Frags.	E-M3RF (Ours)		GARF	
	Time (ms)↓	Mem. (GB)↓	Time (ms)↓	Mem. (GB)↓
2	182.0	2.08	198.0	4.92
7	187.0	2.13	212.0	5.09
15	200.0	2.22	245.0	5.35
40	223.0	2.27	298.0	5.38

6 Conclusion

We introduce E-M3RF, a multimodal 3D reassembly framework that fuses fracture-aware geometry via an $SO(3)$ -equivariant backbone with color cues and enforces physical plausibility through a non-overlap loss. Evaluated on two real-world data collections with colored fractures and three synthetic/real-world without color benchmarks, E-M3RF consistently outperforms all competing methods. These results highlight the potential of E-M3RF for robust, generalizable 3D reconstruction in both synthetic and real-world settings.

Limitations & Future Work. While point density improves accuracy (see suppl. material, Appx. B), it increases computational costs. To scale, we will explore linearized attention and efficient architectures. We also plan to move beyond point clouds to a mesh-first approach, utilizing triangle meshes and textures to capture high-resolution details and connectivity.

Acknowledgements

This work is part of the RePAIR project that has received funding from the European Union’s Horizon 2020 research and innovation programme under the grant agreement No. 964854.

References

1. Assaad, S., Downey, C., Al-Rfou', R., Nayakanti, N., Sapp, B.: VN-transformer: Rotation-equivariant attention for vector neurons. *Transactions on Machine Learning Research* (2023)
2. Bose, J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C., Nica, A., Korablyov, M., Bronstein, M., Tong, A.: Se (3)-stochastic flow matching for protein backbone generation. In: *International Conference on Learning Representations*. vol. 2024, pp. 22590–22621 (2024)
3. Chen, R.T., Lipman, Y.: Flow matching on general geometries. In: *International Conference on Learning Representations*. vol. 2024, pp. 47922–47945 (2024)
4. Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T.: Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)* (2023)
5. Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., Guibas, L.J.: Vector neurons: A general framework for so (3)-equivariant networks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12200–12209 (2021)
6. Freeman, H., Garder, L.: Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *IEEE Transactions on Electronic Computers* (2), 118–127 (1964)
7. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. *Computational visual media* **7**(2), 187–199 (2021)
8. Islam, A., Fiorini, S., James, S., Morerio, P., Del Bue, A.: Reassemblenet: Learnable keypoints and diffusion for 2d fresco reconstruction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9048–9057 (2025)
9. Lamb, N., Palmer, C., Molloy, B., Banerjee, S., Banerjee, N.K.: Fantastic breaks: A dataset of paired 3d scans of real-world broken objects and their complete counterparts. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4681–4691 (2023)
10. Lee, N., Min, J., Lee, J., Kim, S., Lee, K., Park, J., Cho, M.: 3d geometric shape assembly via efficient point cloud matching. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2024)
11. Li, S., Jiang, Z., Chen, G., Xu, C., Tan, S., Wang, X., Fang, I., Zyskowski, K., McPherron, S.P., Iovita, R., et al.: Garf: Learning generalizable 3d reassembly for real-world fractures. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5711–5721 (2025)
12. Liao, S.h., Xiong, C., Liu, S., Zhang, Y.q., Peng, C.l.: 3d object reassembly using region-pair-relation and balanced cluster tree. *Computer Methods and Programs in Biomedicine* **197**, 105756 (2020)
13. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: *The Eleventh International Conference on Learning Representations* (2023)
14. Lu, J., Hua, G., Huang, Q.: Jigsaw++: Imagining complete shape priors for object reassembly. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6704–6714 (2025)
15. Lu, J., Sun, Y., Huang, Q.: Jigsaw: Learning to assemble multiple fractured objects. *Advances in Neural Information Processing Systems* **36**, 14969–14986 (2023)
16. Niu, X., Wang, Q., Liu, B., Zhang, J.: An automatic chinaware fragments reassembly method framework based on linear feature of fracture surface contour. *ACM Journal on Computing and Cultural Heritage* **16**(1), 1–22 (2022)

17. Paixao, T.M., Berriel, R.F., Boeres, M., Koerich, A.L., Badue, C., Souza, A.F.D., Oliveira-Santos, T.: Fast (er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14343–14351 (2020)
18. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
19. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2017)
20. Scarpellini, G., Fiorini, S., Giuliari, F., Moreiro, P., Del Bue, A.: Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 28098–28108 (2024)
21. Sellán, S., Chen, Y.C., Wu, Z., Garg, A., Jacobson, A.: Breaking bad: A dataset for geometric fracture and reassembly. *Advances in Neural Information Processing Systems* **35**, 38885–38898 (2022)
22. T., T., Papaioannou, G.: PRESIOUS 3d cultural heritage fragments (2013), <https://www.ntnu.edu/presious/3d-data-sets>, accessed: 2026-06-27
23. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems* **33**, 7537–7547 (2020)
24. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019)
25. Tsesmelis, T., Palmieri, L., Khoroshiltseva, M., Islam, A., Elkin, G., Shahar, O.I., Scarpellini, G., Fiorini, S., Ohayon, Y., Alali, N., et al.: Re-assembling the past: The repair dataset and benchmark for real world 2d and 3d puzzle solving. *Advances in Neural Information Processing Systems* **37**, 30076–30105 (2024)
26. Wang, J., Zhang, C., Wang, P., Li, X., Cobb, P.J., Theobalt, C., Wang, W.: Batch-based model registration for fast 3d sherd reconstruction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 14473–14483 (2022)
27. Wang, Y., Solomon, J.: Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* **38**(5), 1–12 (2019)
28. Wang, Z., Chen, J., Furukawa, Y.: Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. In: *International Conference on Learning Representations*. vol. 2025, pp. 27584–27607 (2025)
29. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *Computational Visual Media* **11**(1), 83–101 (2025)
30. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021)
31. Zheng, S., Huang, R., Li, J., Wang, Z.: Reassembling 3d thin fragments of unknown geometry in cultural heritage. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2**, 393–399 (2014)