ReassembleNet: Learnable Keypoints and Diffusion for 2D Fresco Reconstruction

Supplementary Material

A. Introduction

In this supplementary martial, we present details on: the experimental details (sec. B; a detailed description of the diffusion process (sec. C); the evaluation over the semi-synthetic dataset (sec. D), metric formulation (sec. E); the keypoint selector (sec. F); an ablation study on the features (sec. G); and qualitative results for synthetic (sec. H) and RePAIR (sec. I) datasets.

B. Experiment Details

Hardware. The experiments were conducted on four machines, each equipped with an NVIDIA A100 GPU (40GB), 380GB of RAM, and two Intel(R) Xeon(R) Silver 4210 CPUs (2.20GHz, Sky Lake architecture).

C. The Diffusion Process

Forward Process. We define the forward process as a fixed Markov chain that adds noise following a Gaussian distribution to each input, i.e., each keypoints, $\mathbf{x}_{i0}^m$ to obtain a noisy version, $\mathbf{x}_{it}^m$, at timestep t. Following [16], we adopt the variance β_t according to a cosine scheduler and define $q(\mathbf{x}_{it}^m|\mathbf{x}_{i0}^m)$ as:

$$q(\mathbf{x}_{it}^m|\mathbf{x}_{i0}^m) = \mathcal{N}(\mathbf{x}_{it}^m; \sqrt{\overline{\alpha}_t}\mathbf{x}_{i0}^m, (1-\overline{\alpha}_t)\mathbf{I}), \tag{4}$$

where $\overline{\alpha}_t = \prod_{c=1}^{t}(1-\beta_c)$ and $\mathbf{I}$ is the identity matrix.

Reverse Process. The reverse process iteratively recovers the initial poses for the set of elements $\hat{X}_{t-1}$ using the current (noisy) poses $X_t = \{X_t^m\}_{m=1}^M$ and the features $H = \{H^m\}_{m=1}^M$, where each $H^m = \{\mathbf{h}_i^m\}_{i=1}^K$ is the set of features for the keypoints in each piece. The recovered poses $\hat{X}_{t-1}$ are computed as:

$$\hat{X}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(X_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(X_t, H, t)\right), \tag{5}$$

where $\alpha_t = 1-\beta_t$, and $\epsilon_\theta(X_t, H, t)$ is the estimated noise output by ReassembleNet that has to be removed from $\hat{X}_t$ at timestep t to recover $\hat{X}_{t-1}$.

The reverse (denoising) step adds a stochastic term $\sigma_t\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$, which governs the randomness injected at each timestep t (see Eq. (11) in [31]). By setting $\sigma_t = 0$, the reverse diffusion becomes fully deterministic.

D. Semi-Synthetic Dataset Evaluation

We compare ReassembleNet on this dataset with learnable methods. We train ReassembleNet using geometric, local, and global features in three different configurations: (i) ReassembleNet-conf. 1, which has *no Learnable KP selection*, (ii) ReassembleNet-conf. 2, which uses *Frozen Learnable KP selection*, and (iii) ReassembleNet-conf. 3, which incorporates *Learnable KP selection*.

Method	RMSE ($\mathcal{R}^\circ$) ↓	RMSE ($\mathcal{T}_{mm}$) ↓
DiffAssemble [26]	122.92	73.79
PairingNet [40]	60.11	266.84
ReassembleNet-conf. 1	40.43	16.91
ReassembleNet-conf. 2	36.02	14.69
ReassembleNet-conf. 3	35.79	15.58

Table 4. Results on Semi-Synthetic dataset.

Results. Table 4 presents the results on the Semi-Synthetic Dataset. As shown, ReassembleNet outperforms the second-best method across all the metrics. This result demonstrates that representing irregular objects as 2D points, as done by ReassembleNet, is more effective than treating them as squared images with padding, as done by DiffAssemble, to achieve a regular shape.

E. Metrics Explanation

To evaluate the performance of the methods, we use three different metrics: RMSE for translation, RMSE for rotation and the Q_{pos}.

The RMSE for translation and rotation are defined as:

$$\text{RMSE}(\mathcal{T}_{mm}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\|\mu_{\hat{\mathbf{t}}}^m - \mu_{\mathbf{t}}^m\|_2}, \tag{6}$$

$$\text{RMSE}(\mathcal{R}^\circ) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\|\,\mu_{\hat{r}}^m - \mu_r^m\|_2}, \tag{7}$$

where μ_t^m denotes the mean ground truth translations, and μ_R^m denotes the corresponding mean ground truth rotations for the m-th piece.

We also evaluate the performance of the methods using the Q_{pos} metric [36], which quantifies the overlap between

Category	Method	Global Feats	Local Feats	Geom Feats	$Q_{pos} \uparrow$	RMSE ($\mathcal{R}^\circ$) $\downarrow$	RMSE ($\mathcal{T}_{mm}$) $\downarrow$
No Transfer Learning	no Learnable KP selection	X	X	X	0.18	64.51	80.19
	Frozen Learnable KP selection	X	X	X	0.23	62.45	33.82
	no Learnable KP selection	X	X	V	0.17	59.76	17.76
	Frozen Learnable KP selection	X	X	V	0.22	43.11	22.03
	no Learnable KP selection	V	X	X	0.14	63.24	32.30
	Frozen Learnable KP selection	V	X	X	0.22	62.95	24.42
	no Learnable KP selection	X	V	X	0.27	64.08	19.75
	Frozen Learnable KP selection	X	V	X	0.28	49.58	23.11
	no Learnable KP selection	V	V	V	0.35	55.01	16.12
	Frozen Learnable KP selection	V	V	V	0.39	51.96	26.67
	Learnable KP selection	V	V	V	0.27	47.61	19.16
Transfer Learning	no Learnable KP selection)	X	X	X	0.27	53.47	28.74
	Frozen Learnable KP selection)	X	X	X	0.15	51.95	21.63
	no Learnable KP selection	X	X	V	0.21	56.27	17.76
	Frozen Learnable KP selection	X	X	V	0.15	41.74	20.92
	no Learnable KP selection	V	X	X	0.19	59.07	23.43
	Frozen Learnable KP selection	V	X	X	0.13	58.97	23.26
	no Learnable KP selection	X	V	X	0.20	61.83	17.64
	Frozen Learnable KP selection	X	V	X	0.15	45.46	16.68
	no Learnable KP selection	V	V	V	0.16	42.98	18.11
	Frozen Learnable KP selection	V	V	V	0.17	39.12	18.41
	Learnable KP selection	V	V	V	0.21	32.91	17.18

Table 5. Ablation on ReassembleNet settings.

the ground truth fragment poses (translation and rotation) and the reconstructed solution. To ensure that the metric is invariant to rigid motions—preventing good solutions from being penalized due to differing global rotations—we first apply a rigid transformation to align the largest reconstructed fragment (referred to as the *anchor*) with its corresponding ground truth fragment in both translation and rotation. To compute Q_{pos}, we first define the area of a fragment, denoted as $A(m)$. In 2D, the shared area can be determined in two different ways: *i)* by comparing the non-transparent pixels of two large canvases containing all fragments, or *ii)* by computing the area intersection of the registered 2D point clouds. Additionally, fragments are weighted based on their area, emphasizing the impact of errors on larger fragments. The metric is formally defined as:

$$Q_{pos} = \sum_{m=1}^{M} w_m \cdot \frac{|A(m \cap \tilde{m})|}{|A(\tilde{m})|}, \tag{8}$$

where $w_m = \frac{|A(m)|}{\sum_{k=1}^{M} |A(k)|}$ represents the weight of each fragment, and $A(\tilde{m})$ denotes the area of the fragments with predicted rotation and translation.

F. Keypoints Selector

As detailed in Section 3.2, our approach involves selecting k keypoints. To achieve this, we employ our learnable keypoint selection module, which is pre-trained to improve its effectiveness. During the pre-training phase, we utilize the RePAIR dataset, treating each piece independently. This dataset enables the model to learn to identify salient keypoints in a diverse and representative context.

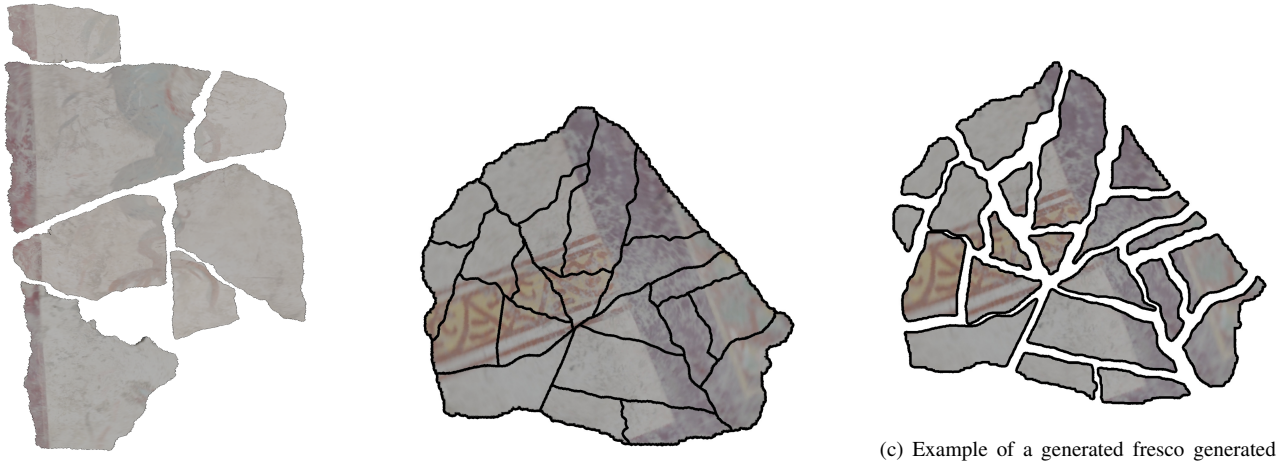
We then optimize the module using the two loss functions defined in Equation (2), with $\lambda_{area} = 1$ and $\lambda_{per} = 1$. These losses work together to enforce geometric precision and structural consistency, while also mitigating selection bias toward task-specific nodes.

G. Ablation Study on Multimodal Features

Table 5 presents a comprehensive ablation study assessing the impact of the final configuration used for ReassembleNet. The results clearly demonstrate that incorporating all features and leveraging transfer learning are crucial for tackling this challenging task. By utilizing the full set of features, our model gains both geometric awareness of the object and semantic understanding through local and global image representations. This injected bias enhances the network's ability to learn effectively.

H. Qualitative comparison on Semi-Synthetic Dataset Creation Process

In this section, we are reporting a visual representation of the semi-synthetic dataset created following [40] and the final results of the semi-synthetic dataset we were able to create by adding the random erosion of the borders with a slight random rotation and translation. Each fragment undergoes morphological erosion using a 3×3 kernel, with between 1-5 iterations randomly simulating varying degrees of degradation. Then, each fragment is randomly augmented with



(a) (b) (c) Example of a generated fresco generated using our modified algorimth

Figure 6. An illustration of (a) an example of a RePAIR fresco, (b) a synthetic fresco generated using the algorithm proposed by [40], and (c) a synthetic fresco generated using our modified algorithm. The black contour is intentionally added to highlight the borders of the pieces in b and c.

rotation (±3°) or translation (±3 pixels in x and y), applied via affine transformations to introduce geometric variability. These augmentations ensure diversity and realism in the generated dataset.

Figure 6 shows the visual differences in the creation of the semi-synthetic dataset. As can be seen, our proposed algorithm (Figure 6c) exhibits a certain similarity to Figure 6a, which is taken from the real-world dataset RePAIR. In contrast, Figure 6b clearly shows that the puzzle generated using the algorithm in [40] deviates significantly from the characteristics present in RePAIR: the pieces are assembled to align perfectly without gaps, ensuring a seamless matching between the pieces.

I. More Qualitative Results on RePAIR Dataset

We report some more qualitative results on the RePAIR dataset. In particular, we report with Figure 7 some failure cases where it can be seen that the model is learning the complexity of groundtruth data. We also provide baseline comparison results in Fig. 8.

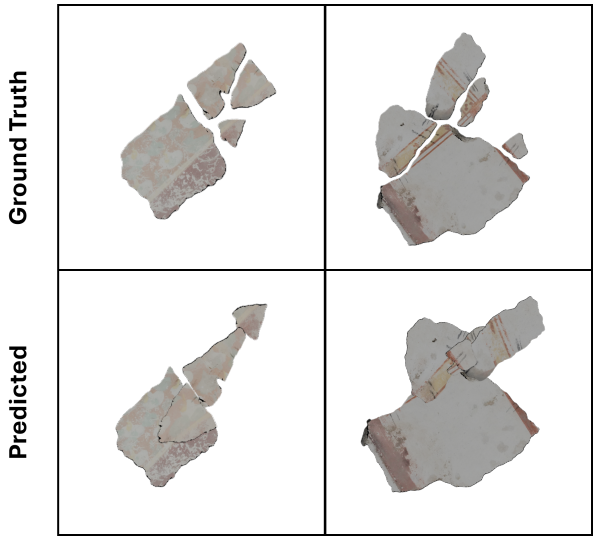


Figure 7. Qualitative results.

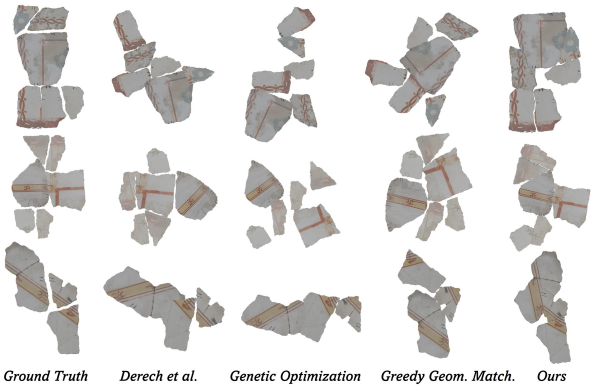


Figure 8. Qualitative comparison.