HistReNeRF: Historic Image Relocalisation within Contemporary Neural Radiance Field Reconstructions

Benjamin T. Hughes and Stuart James

Durham University, Durham, UK
stuart.a.james@durham.ac.uk

Abstract. Relocalising archival photographs within a contemporary scene model is challenging because historic and modern views can differ in photographic appearance, visible objects, and spatial layout. Therefore, we present *HistReNeRF*, a framework that estimates the 6-DoF pose of a historic photograph by matching adapted DINOv2 patch features to candidate rays sampled from a contemporary Neural Radiance Field (NeRF) reconstruction. The continuous representation of a NeRF provides a queryable scene interface from which candidate rays can be sampled and matched, enabling domain adaptation between historic photography and contemporary images directly in the feature representation used for localisation. We evaluate embedding-space-based domain adaptation against pixel-space methods on a new cross-temporal dataset comprising 10,545 contemporary street-level images and 230 archival photographs from three European landmarks. Embedding-space adaptation reduces translation and rotation errors by an average of 11% and 16%, respectively, across the three scenes. These results show that neural scene relocalisation provides a natural interface for feature-space adaptation, reducing cross-temporal appearance shift without modifying the query image. Code and dataset at https://github.com/ARTUROLab/HistReNeRF.

Keywords: Historic Photograph · 3D Relocalisation · Neural Radiance Fields (NeRFs)

1 Introduction

Historic photographs provide an important record of past urban environments, architectural heritage, and visual culture. Relocalising such images within contemporary three-dimensional scene reconstructions would support the comparison of historical and present-day views, the documentation of urban change, and the spatial interpretation of archival collections. However, historic image relocalisation differs substantially from conventional visual localisation. Archival photographs may vary from contemporary imagery in photographic medium, film grain, contrast, resolution, and tonal range, while the depicted scene may have changed through demolition, renovation, vegetation growth, or the replacement of street furniture and surrounding structures. These combined appearance and

structural changes make it difficult to establish reliable correspondences between a historic query image and a modern scene model.

Classical localisation pipelines typically estimate pose by matching local image descriptors, such as SIFT [13], to a Structure-from-Motion (SfM) reconstruction [23]. Their success depends on repeatable local structure and sufficient photometric consistency between the query and reference images. Both assumptions are weakened in the historic setting where photographic changes affect the visibility and appearance of local features, while scene evolution can remove or alter the structures on which correspondences depend. Consequently, a contemporary point cloud may provide unreliable support for relocalising.

Neural scene representations offer an alternative reference model by coupling scene geometry and appearance within a continuous, queryable representation. A Neural Radiance Field (NeRF) [19] models a scene as a volumetric function that can be rendered from arbitrary viewpoints, allowing localisation to be formulated against a learned scene model rather than a sparse set of 3D descriptors. Early relocalisation work, iNeRF [31], used an inversion-based formulation, where it recovered camera pose by iteratively rendering a pretrained NeRF from candidate viewpoints and minimising its photometric difference from the query image. While this demonstrates the localisation potential of neural scenes, repeated rendering makes inference computationally expensive and dependent on a suitable pose initialisation. IFFNeRF [5] instead formulates localisation through ray–query correspondence. Candidate rays sampled from the neural scene are matched to DINOv2 [22] image features through cross-attention, and the pose is recovered by weighted least-squares ray minimum-distance estimation. This formulation is particularly relevant for historic imagery because it replaces direct low-level descriptor matching with a learned representation. However, historic queries remain separated from contemporary scene imagery by a substantial cross-temporal appearance gap.

To address the cross-temporal appearance gap, we present *HistReNeRF*, a framework for relocalising historic photographs within contemporary TensoRF reconstructions. HistReNeRF matches DINOv2 patch features extracted from a historic query image to candidate rays sampled from the neural scene. We introduce an embedding-space CycleGAN that maps historic patch features towards the contemporary feature distribution before ray matching. We evaluate on a new cross-temporal dataset across three scenes: Arc de Triomphe, Brandenburg Gate, and Piazza del Duomo.

The contributions of this paper are as follows: *i)* We present *HistReNeRF*, a cross-temporal neural scene relocalisation framework that applies unpaired domain adaptation directly to the query representation used for ray matching, and compares feature-space transfer against pixel-space translation. *ii)* We introduce a new dataset for historic image relocalisation with validated reference poses for historic queries across three European landmark scenes. *iii)* We provide a systematic evaluation of embedding-space and pixel-space adaptation for cross-temporal relocalisation, showing that feature-space transfer yields consistent gains.

2 Related Work

We review scene reconstruction techniques in Secs. 2.1 and relocalisation in 2.2, followed by domain adaptation methods addressing the appearance gap between historic and contemporary imagery in Sec. 2.3. Finally, we review cross-domain and localisation approaches in Sec. 2.4.

2.1 Scene Reconstruction

Classical Structure-from-Motion methods reconstruct camera poses and sparse three-dimensional point clouds from overlapping image collections, establishing the geometric basis for large-scale landmark reconstruction from crowd-sourced imagery [1, 27]. These reconstructions support conventional localisation through image-to-point correspondences, but their sparse structure provides limited support when the query differs substantially in appearance from the images used to build the model. Similarly, 3D Gaussian Splatting represents the scene through an explicit set of optimised Gaussian primitives and achieves real-time rendering through differentiable rasterisation [10].

Alternatively, neural scene representations extend this geometric reference with a dense model of scene appearance. Neural Radiance Fields (NeRFs) represent a scene as a volumetric function that maps position and viewing direction to density and colour, enabling novel views to be synthesised from arbitrary camera poses [19]. Subsequent work has made these representations more practical for real scenes primarily focusing on improving the speed of learning the function. TensoRF [6] factorises the radiance field into compact tensor components, substantially reducing the computational cost of reconstruction and rendering. In contrast, Instant-NGP [21] accelerated neural-field optimisation through multiresolution hash encodings. Extensive surveys have been conducted on NeRFs [29]; however, beyond generalising to the wild with unconstrained contemporary camera imagery [17], archival photography remains under-explored.

2.2 Visual Relocalisation in Neural Scenes

Classical visual relocalisation estimates the 6-DoF camera pose of a query image by matching local features against a three-dimensional reconstruction. Handcrafted descriptors such as SIFT [13] and learned alternatives such as SuperPoint [7] can provide correspondences when the query and reference imagery share sufficient local structure. Beyond correspondence-based approaches, diffusion models have been explored for probabilistic camera pose estimation, either by modelling distributions over camera parameters [28] or over distributed ray-based camera representations [32]. However, correspondence-based relocalisation remains sensitive to appearance changes that alter the visibility, texture, or contrast of local regions, making it difficult to apply directly to archival photographs.

Neural scene representations have introduced alternatives to descriptor-based localisation. iNeRF [31] recovers pose by iteratively rendering a pretrained NeRF from candidate viewpoints and minimising photometric error against the query

image. Alternatively, VRS-NeRF [30] reduces the cost of this analysis-by-synthesis formulation for sparse neural scenes, while retaining its dependence on optimisation against the rendered scene.

A complementary direction uses dense learned features from neural scene representations to establish correspondences more directly. CrossFire [20] derives self-supervised features from an implicit scene representation for camera relocalisation, while pNeRF-Loc [33] and NeRFect Match [34] investigate the use of NeRF-based representations for visual localisation and correspondence estimation. IFFNeRF [5] develops this direction into a feed-forward formulation that matches query-image features to candidate rays sampled from a neural scene and recovers pose without iterative inversion. Overall, these methods show that neural scenes can support localisation through learned correspondence representations, but they generally assume that query images remain visually compatible with the imagery used to reconstruct the scene.

2.3 Unpaired Domain Adaptation

Domain adaptation addresses the problem of applying a model across visual domains with different appearance distributions. In the unpaired setting, i.e. without one-to-one correspondences, CycleGAN learns mappings between source and target domains through adversarial training and cycle consistency, without requiring corresponding image pairs [35]. This is particularly relevant to historic imagery, where equivalent historic and contemporary photographs are rarely available from the same 6DoF viewpoint.

Relaxing the cycle consistency, methods such as UNIT [12] model the two domains through a shared latent representation, while MUNIT [8] separates content from style to support multiple plausible translations from a common source image. CUT [24] instead uses patch-wise contrastive learning to preserve local content while learning a one-directional translation between unpaired domains. These methods provide different mechanisms for altering image appearance while preserving the scene structure required by a downstream task.

Pixel-space adaptation has been used to improve performance under domain shift in perception and localisation. Style transfer has reduced the synthetic-to-real gap for monocular depth estimation, while night-to-day translation has been explored for retrieval-based visual localisation [2, 3]. In each case, adaptation is applied to pixels before downstream features are extracted. This creates a potential limitation for historic image relocalisation, where a translation model must alter photographic appearance without changing the architectural and spatial evidence required for pose estimation.

2.4 Domain Adaptation and Historic Visual Data

Visual localisation benchmarks have increasingly examined changing environmental conditions. Cambridge Landmarks [9] provides outdoor landmark scenes for camera-pose estimation, while Aachen Day-Night [26] evaluates localisation

across pronounced illumination changes and RobotCar Seasons extends this setting across weather and seasonal variation [14]. These benchmarks are important for evaluating appearance robustness, but they retain a relatively stable photographic process and scene structure compared with the century-scale change encountered in archival imagery.

Historic visual datasets have addressed related but distinct problems. Maiwald [15] constructs a benchmark for evaluating feature matching on historical architectural photographs, directly exposing the difficulty of establishing correspondences across temporal appearance change. Komorowicz et al. [11] use archival photographs for neural reconstruction of historical monuments, addressing degraded imagery and incomplete visual evidence from the historical domain. These works demonstrate the importance of historic imagery for geometric and visual analysis, but they do not consider the task of localising an archival photo against a separately reconstructed contemporary neural scene, which we address.

3 HistReNeRF

Given a historic query photograph I_q and a contemporary TensoRF neural radiance field reconstruction f_θ of the same scene, HistReNeRF estimates the 6-DoF camera pose $\hat{\mathbf{P}} = (\hat{\mathbf{R}}, \hat{\mathbf{t}})$ in the coordinate frame of the reconstruction. The method uses an IFFNeRF-style ray–query localiser trained on contemporary imagery. For the ray–query localiser, candidate rays are sampled from the TensoRF density field and matched to DINOv2 patch features extracted from the historic query through cross-attention. The highest-scoring rays estimate the camera centre $\hat{\mathbf{t}}$ through a weighted least-squares minimum-distance estimate, while their directions and the scene up-vector determine the camera orientation $\hat{\mathbf{R}}$. To reduce the historic-to-modern appearance gap, HistReNeRF applies unpaired domain adaptation directly to the DINOv2 patch representation used for ray–query matching. A CycleGAN maps historic patch embeddings towards the distribution of embeddings extracted from selected contemporary images after feature extraction and before cross-attention matching. The DINOv2 backbone, TensoRF reconstruction, ray–query localiser, and historic query image remain unchanged throughout adaptation. Figure 1 provides an overview of the pipeline.

3.1 Neural Scene Relocalisation

We adopt the candidate-ray relocalisation formulation of IFFNeRF [5]. Given a query image and a contemporary TensoRF reconstruction, samples of likely surface locations are extracted from the neural scene; from these point locations, multiple rays are cast. The query image is matched against this ray set, allowing the most relevant rays to be identified and combined into a pose estimate.

Surface-point sampling. We first extract G candidate ray origins from the density field of the contemporary TensoRF reconstruction. Let $\hat{\sigma}(\mathbf{x})$ denote the density predicted at position $\mathbf{x}$. The procedure is initialised by uniformly sampling G

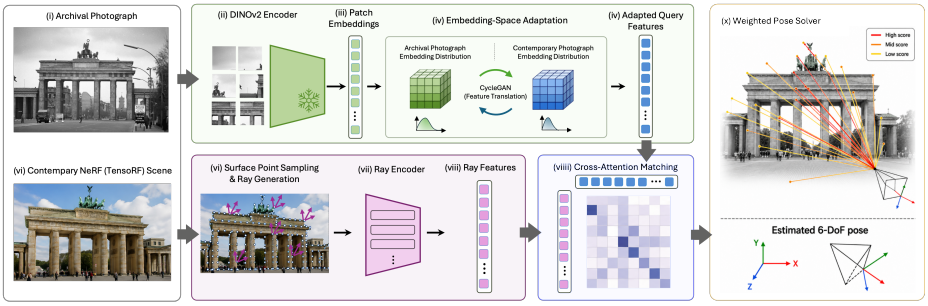

Fig. 1: Overview of *HistReNeRF*. Given a historic query image (i), it is encoded by frozen DINOv2 (ii) into patch embeddings (iii), which are mapped towards the contemporary embedding distribution by embedding-space adaptation (iv) to produce adapted query features (v). In parallel, a contemporary TensoRF reconstruction (vi) is used to sample candidate surface points and generate normal-oriented candidate rays, and encode them as ray features (viii). Cross-attention matching (viiii) scores the compatibility between adapted query features and candidate rays. The weighted pose solver (x) aggregates the highest-scoring rays to estimate the 6-DoF camera pose of the historic photograph within the contemporary neural scene.

points within the scene bounding box. These points are then refined through an iterative Metropolis–Hastings procedure. At iteration i, a new random position is proposed for each point. The proposal for point g is accepted only when its predicted density satisfies $\hat{\sigma}(\mathbf{u}_g^{i+1}) \geq \tau$, where the density threshold τ is determined from the empirical cumulative density distribution of the selected points in the preceding iteration. Specifically, the procedure retains proposals associated with the upper portion of this distribution, using the 60th-percentile criterion. The points retained after the final iteration form the set $\mathcal{U} = \{\mathbf{u}_g\}_{g=1}^{G}$, where g indexes the G sampled points; these provide the ray origins for the following ray-generation stage.

Ray generation. Each sampled point provides the origin for a set of rays. The neural scene is queried for the surface normal by the density gradient at $\mathbf{u}_g$, and ray directions are oriented by this normal to avoid casting rays towards the unobserved interior of the reconstructed surface. The deterministic isocell distribution [4,18] is applied, which partitions the unit sphere into equal-area cells and uses their centres to obtain approximately uniform directions, where $V = 27$ rays are generated for each surface point, as in IFFNeRF. Each ray is represented by its origin $\mathbf{o}_j$, direction $\mathbf{d}_j$, and rendered colour $\mathbf{c}_j$. The colour is obtained by volumetric rendering along the ray within the TensoRF reconstruction. The resulting set contains $N = GV$ candidate rays

$$\mathcal{R} = \{(\mathbf{o}_j, \mathbf{d}_j, \mathbf{c}_j)\}_{j=1}^{N}. \tag{1}$$

This ray bundle spans possible camera locations and viewing directions within the reconstructed scene.

Ray–query matching. The historic query image I_q is encoded using a frozen DINOv2 ViT-S/14 backbone [22]. For a 224×224 input image, the encoder produces $\mathbf{F}_q = \phi(I_q) \in \mathbb{R}^{K \times 384}$, where $K = 256$. The geometric and appearance parameters of each candidate ray are encoded separately. A ray encoder applies positional encoding to the ray origin, direction, and rendered colour before mapping them to ray features, $\mathbf{F}_r = \psi_r(\mathcal{R}) \in \mathbb{R}^{N \times d}$. Positional encoding expands small differences in ray origin and direction, allowing the encoder to distinguish rays with similar rendered appearance but different geometric configurations.

The query-image features act as attention queries and the ray features act as attention keys. Their compatibility is represented by an attention matrix

$$\mathbf{A}\left(\mathbf{F}_q, \mathbf{F}_r\right) \in \mathbb{R}^{K \times N}, \tag{2}$$

where $\mathbf{A}$ denotes the learned attention module. Each entry $\mathbf{A}_{k,j}$ measures the compatibility between query-image patch k and candidate ray j. Summing attention over the image patches gives a score for each candidate ray

$$\hat{s}_j = \sum_{k=1}^{K} A_{k,j}. \tag{3}$$

Geometric supervision. The ray scorer is trained on contemporary images with known camera centres. For a training image with camera centre $\mathbf{t}^*$, the supervision signal measures how closely each candidate ray passes to that camera centre. The closest point on ray j is $\mathbf{p}_j = \mathbf{o}_j + \max\left((\mathbf{t}^* - \mathbf{o}_j)^\top \mathbf{d}_j, 0\right) \mathbf{d}_j$, where the maximum restricts the projection to the forward ray half-line. The distance between the ray and the camera centre is then $\delta_j = \left|\mathbf{p}_j - \mathbf{t}^*\right|_2$. This distance is converted into a target score by

$$\tilde{s}_j = 1 - \tanh\left(\frac{\delta_j}{\lambda}\right), \tag{4}$$

where λ controls the range of target scores. Following IFFNeRF, the scores are normalised to sum to K, matching the scale of the aggregated attention scores:

$$s_j = \frac{K \tilde{s}_j}{\sum_{\ell=1}^{N} \tilde{s}_\ell}. \tag{5}$$

The ray scorer is trained by minimising the ℓ_2 discrepancy between predicted and target scores.

Pose recovery. At inference, the highest-scoring rays define a set $\mathcal{J}$. The estimated camera centre is the weighted least-squares minimum-distance estimate of these rays

$$\hat{\mathbf{t}} = \left(\sum_{j \in \mathcal{J}} \hat{s}_j \left(\mathbf{I}_3 - \mathbf{d}_j \mathbf{d}_j^\top\right)\right)^{-1} \sum_{j \in \mathcal{J}} \hat{s}_j \left(\mathbf{I}_3 - \mathbf{d}_j \mathbf{d}_j^\top\right) \mathbf{o}_j. \tag{6}$$

The selected ray bundle provides the corresponding camera-pose estimate within the TensoRF coordinate frame. HistReNeRF retains this localiser unchanged and adapts only the historic DINOv2 representation before ray–query matching.

3.2 Contemporary Domain Relevance Filtering

Contemporary imagery used to reconstruct the neural scene often consists of dash-cam or street-level views, including images dominated by roads, vehicles, or foreground street furniture despite being geographically aligned with the landmark and reconstruction. While these images provide useful geometric coverage for TensoRF reconstruction, many are unsuitable targets for cross-temporal adaptation. Such views can differ substantially from the architectural content represented in the historic collection. Training a domain-transfer model on the full contemporary image set may therefore encourage appearance changes that are unrelated to the historic-to-modern shift relevant for relocalisation.

We select a contemporary target subset using global CLIP image embeddings [25]. CLIP filtering is used only for within-scene view selection for domain adaptation, rather than for landmark identification or camera localisation. Let $\mathbf{e}_i^h$ denote the embedding of historic training image I_i^h and let $\mathbf{e}^m$ denote the embedding of a candidate contemporary image I^m. A contemporary image is retained when it lies sufficiently close to at least one historic image in CLIP embedding space. Specifically, its nearest historic embedding must be no more than 1.5 times the maximum pairwise cosine distance observed between images in the historic collection. This defines a scene-specific threshold that excludes visually dissimilar contemporary views while allowing for the expected historic-to-modern appearance shift.

$$\min_i d * \cos(\mathbf{e}^m, \mathbf{e}_i^h) \leq \alpha \max i, j d * \cos(\mathbf{e}_i^h, \mathbf{e}_j^h), \tag{7}$$

where $d * \cos$ is cosine distance and α controls the permitted historic-to-modern variation. This selection retains contemporary views that are visually compatible with the architectural content of the historic collection while preserving sufficient variation to represent the contemporary domain. The selected images define the target distribution used to train the embedding-space adaptation model.

3.3 Embedding-Space Adaptation

The contemporary subset selected in Sec. 3.2 defines the target visual distribution for adaptation. Rather than modifying the historical photograph itself, HistReNeRF aligns the DINOv2 representation used by the ray–query matching module. Given an image I, the frozen DINOv2 encoder encodes the image expressed as $\mathbf{E} = \phi(I) \in \mathbb{R}^{256\times384}$, which can be reshaped to $16 \times 16 \times 384$ to preserve the original patch layout of the image. This representation retains the spatial organisation of the historic query while expressing its content in the semantic feature space used by the localisation model.

Let $\mathbf{E}_h \sim \mathcal{D}_h$ and $\mathbf{E}_m \sim \mathcal{D}_m$ denote embedding grids extracted from historic and selected contemporary images. We train an unpaired CycleGAN [35] to learn mappings between these distributions. The model contains two U-Net-style convolutional generators, $G_{h\to m}$ and $G_{m\to h}$, together with two PatchGAN

discriminators, D_m and D_h. The generators operate over the spatial embedding grid, allowing the translation to account for local feature structure while maintaining the patch arrangement inherited from the query image.

The model is trained with least-squares adversarial losses [16], cycle consistency, and identity preservation. The complete objective is

$$\mathcal{L}_{\text{emb}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}. \tag{8}$$

Here, $\mathcal{L}_{\text{adv}}$ follows the standard least-squares GAN objective and aligns translated historic embeddings with the contemporary embedding distribution. Cycle consistency encourages a translated representation to remain recoverable after mapping back to its original domain

$$\begin{aligned}\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\mathbf{E}_h\sim\mathcal{D}_h}[\|G_{m\to h}(G_{h\to m}(\mathbf{E}_h)) - \mathbf{E}_h\|_1] \\ + \mathbb{E}_{\mathbf{E}_m\sim\mathcal{D}_m}[\|G_{h\to m}(G_{m\to h}(\mathbf{E}_m)) - \mathbf{E}_m\|_1].\end{aligned} \tag{9}$$

The identity term penalises unnecessary changes when an embedding is already sampled from the target domain

$$\begin{aligned}\mathcal{L}_{\text{id}} = \mathbb{E}_{\mathbf{E}_m\sim\mathcal{D}_m}[\|G_{h\to m}(\mathbf{E}_m) - \mathbf{E}_m\|_1] \\ + \mathbb{E}_{\mathbf{E}_h\sim\mathcal{D}_h}[\|G_{m\to h}(\mathbf{E}_h) - \mathbf{E}_h\|_1].\end{aligned} \tag{10}$$

We use $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{id}} = 5$. The DINOv2 encoder and neural-scene relocalisation model remain frozen throughout training, ensuring that the learned mapping aligns the historic representation with the existing contemporary localisation space rather than changing the representation used by the localiser.

At inference time, a historic query image is encoded by DINOv2, translated by $G_{h\to m}$, and restored to the patch-token representation. The adapted features $\widetilde{\mathbf{F}}_q$ replace $\mathbf{F}_q$ in the cross-attention module described in Sec. 3.1. The historic image, the contemporary TensoRF reconstruction, and the ray-based pose solver are not modified.

4 Evaluation

We evaluate HistReNeRF on cross-temporal relocalisation within contemporary neural scene reconstructions. Our experiments assess whether embedding-space adaptation improves pose estimation for historic query images, how it compares with pixel-space translation methods, and how performance varies across scenes with different viewpoint and appearance gaps. We first introduce the *HistScene* dataset in Sec. 4.1, before reporting quantitative and qualitative comparisons between no adaptation, pixel-space adaptation, and the proposed embedding-space approach in Sec. 4.2.

Table 1: HistScene statistics. Pose-validated images form the historic evaluation set. Remaining historic photographs are used as unpaired samples for domain adaptation training.

Scene	Contemporary	Historic	Pose-validated (Test)	Training
Arc de Triomphe	2,371	92	25	67
Brandenburg Gate	5,000	96	18	78
Piazza del Duomo	3,174	42	15	27
Total	10,545	230	58	172

Table 2: Reference-pose validation metrics averaged per scene. Gradient NCC and Edge F1 assess structural alignment between TensoRF renders and query images; higher is better. Grayscale SSIM is a photometric sanity check and is expected to be lower for historic images.

	Gradient NCC		Edge F1		SSIM (gray)	
Scene	Modern	Historic	Modern	Historic	Modern	Historic
Arc de Triomphe	0.124	0.047	0.240	0.637	0.499	0.268
Brandenburg Gate	0.044	-0.001	0.116	0.217	0.375	0.351
Piazza del Duomo	0.089	-0.004	0.092	0.099	0.413	0.418

4.1 HistScene Dataset

The HistScene dataset is a cross-temporal relocalisation dataset comprising 10,545 contemporary street-level images and 230 archival photographs across three locations: Arc de Triomphe in Paris, Brandenburg Gate in Berlin, and Piazza del Duomo in Milan. The three scenes provide complementary challenges. Arc de Triomphe exhibits substantial viewpoint mismatch between historic and contemporary imagery and strong architectural symmetry. Brandenburg Gate contains predominantly ground-level views with more comparable historic and contemporary camera distributions. Piazza del Duomo provides a wider architectural context that includes the cathedral facade and surrounding piazza.

For each scene, contemporary Mapillary imagery is used to recover a COLMAP reconstruction and train a TensoRF scene representation. Historic photographs were collected from Europeana and Wikimedia Commons and span a substantial range of photographic processes, image conditions, viewpoints, and scene content. Selected images were based on their Creative Commons licence. The historical collection is divided based on whether a reliable reference pose can be recovered in the contemporary COLMAP reconstruction. Successfully registered images, as explained below, form the evaluation set, while images without validated registration serve as unpaired historical training samples for domain adaptation. Dataset statistics are reported in Table 1.

Ground Truth Pose Verification. We recover reference poses for historic images by registering them against the existing COLMAP sparse reconstruction. SIFT correspondences between the historic query and reconstructed three-dimensional points are used within PnP-RANSAC to estimate camera pose and intrinsics. A query is retained for quantitative pose evaluation only when registration succeeds and a manual inspection confirms that reprojected scene points align with visible architectural structure. To achieve this, sparse points are re-

Fig. 2: Example images from the proposed HistScene dataset facing the monument with left archival photography and right Mapillary street view.

projected onto a neural-scene rendering from the COLMAP camera pose. This verification is performed independently of HistReNeRF predictions.

Table 2 reports gradient-normalised cross-correlation, edge F1, and grayscale SSIM between each historic query and a TensoRF rendering from its recovered pose. These quantities are used as supporting structural diagnostics rather than pose-selection criteria, since cross-temporal differences in photographic appearance make photometric agreement inherently unreliable. Together, geometric reprojection and visual inspection provide the basis for the validated reference poses used in the following experiments.

4.2 Cross-Temporal Relocalisation Evaluation

We evaluate cross-temporal relocalisation on the 58 pose-validated historic queries in HistScene. We report mean translation error, MTE, in metres and mean angular error, MAE, in degrees, where lower values indicate more accurate pose estimation. All configurations use the same contemporary TensoRF reconstruction, candidate-ray cache, frozen DINOv2 backbone, ray–query matching module, and pose solver. The comparison therefore isolates the representation at which cross-temporal adaptation is applied. We compare the following configurations as well as the proposed HistReNeRF:

- **No adaptation:** DINOv2 patch features are extracted directly from the historic query and matched to candidate rays from the contemporary neural scene.
- **Img. CycleGAN:** Translates the historic query image towards the contemporary domain using adversarial training and cycle consistency [35]. DINOv2 features are then extracted from the translated image.

Table 3: HistScene relocalisation MTE is reported in metres and MAE in degrees (lower is better). Best result per scene and metric is shown in bold.

		Arc		Brandenburg		Duomo	
Method	Setting	MTE	MAE	MTE	MAE	MTE	MAE
IFFNeRF	*Historic — no adaptation*	6.624	71.7°	4.712	23.4°	1.783	85.5°
IFFNeRF + iNeRF	*Historic — no adaptation*	6.818	81.5°	5.279	35.7°	2.234	71.0°
HistReNeRF	*Historic — domain transfer*	**6.608**	**69.6°**	**3.912**	**23.4°**	**1.589**	**58.6°**
HistReNeRF + iNeRF	*Historic — domain transfer*	6.608	76.4°	4.708	29.2°	1.902	68.9°
Img. CycleGAN	*Historic — domain transfer*	7.179	88.6°	4.809	25.6°	1.799	72.3°
Img. CycleGAN + iNeRF	*Historic — domain transfer*	6.966	103.5°	5.703	42.6°	1.973	64.0°
Img. CUT	*Historic — domain transfer*	7.007	90.9°	4.543	28.2°	1.812	80.7°
Img. CUT + iNeRF	*Historic — domain transfer*	7.079	106.1°	5.323	38.8°	1.961	90.9°
Img. UNIT	*Historic — domain transfer*	7.337	112.5°	5.427	30.8°	1.984	85.0°
Img. UNIT + iNeRF	*Historic — domain transfer*	7.483	111.0°	6.108	39.6°	2.062	80.7°
Img. MUNIT	*Historic — domain transfer*	8.548	118.1°	7.608	33.9°	3.071	121.8°
Img. MUNIT + iNeRF	*Historic — domain transfer*	8.686	124.7°	7.441	38.3°	3.238	119.1°

Table 4: Cumulative localisation recall (%) at angular in degrees and translation thresholds. Cell colour indicates recall level: darker green indicates higher recall.

		Angular (%)						Translation (%)						
Scene	Method	5°	15°	30°	45°	60°	90°	0.5	1.0	2.5	4.0	5.0	7.5	10.0
Arc	No DT	3.8	34.6	50.0	57.7	57.7	57.7	0.0	0.0	0.0	15.4	23.1	73.1	80.8
	Img. CycleGAN	0.0	19.2	26.9	34.6	38.5	46.2	0.0	0.0	0.0	11.5	26.9	65.4	80.8
	Img. CUT	3.8	15.4	26.9	34.6	34.6	42.3	0.0	0.0	0.0	0.0	23.1	61.5	92.3
	Img. UNIT	0.0	15.4	15.4	15.4	23.1	26.9	0.0	0.0	0.0	3.8	15.4	53.8	92.3
	Img. MUNIT	0.0	11.5	23.1	23.1	23.1	34.6	0.0	0.0	7.7	11.5	11.5	34.6	61.5
	Emb. CycleGAN	3.8	38.5	57.7	57.7	57.7	69.2	0.0	0.0	3.8	19.2	34.6	65.4	88.5
Brandenburg	No DT	0.0	22.2	77.8	94.4	94.4	100.0	0.0	0.0	16.7	55.6	77.8	88.9	100.0
	Img. CycleGAN	0.0	22.2	77.8	94.4	94.4	100.0	0.0	0.0	16.7	55.6	66.7	100.0	100.0
	Img. CUT	0.0	27.8	66.7	88.9	94.4	100.0	0.0	0.0	0.0	50.0	77.8	100.0	100.0
	Img. UNIT	0.0	0.0	50.0	94.4	94.4	100.0	0.0	0.0	11.1	33.3	44.4	94.4	94.4
	Img. MUNIT	0.0	0.0	44.4	77.8	88.9	100.0	0.0	0.0	0.0	0.0	5.6	66.7	94.4
	Emb. CycleGAN	0.0	22.2	72.2	94.4	94.4	100.0	0.0	0.0	27.8	50.0	72.2	94.4	100.0
Duomo	No DT	0.0	6.7	6.7	26.7	33.3	66.7	0.0	13.3	80.0	100.0	100.0	100.0	100.0
	Img. CycleGAN	0.0	6.7	20.0	26.7	40.0	66.7	13.3	26.7	73.3	100.0	100.0	100.0	100.0
	Img. CUT	0.0	6.7	13.3	20.0	33.3	46.7	0.0	26.7	86.7	100.0	100.0	100.0	100.0
	Img. UNIT	0.0	0.0	6.7	6.7	13.3	46.7	6.7	6.7	60.0	80.0	100.0	100.0	100.0
	Img. MUNIT	0.0	0.0	0.0	0.0	6.7	26.7	0.0	0.0	33.3	73.3	93.3	100.0	100.0
	Emb. CycleGAN	0.0	0.0	26.7	33.3	53.3	73.3	0.0	26.7	93.3	93.3	100.0	100.0	100.0

- **Img. CUT:** Translates the historic image through a one-directional contrastive objective that encourages preservation of local image patches [24].
- **Img. UNIT:** Translates the historic image through a shared latent representation learned across historic and contemporary image domains [12].
- **Img. MUNIT:** Decomposes image content and style to generate plausible contemporary-domain translations from the historic query [8].

Each adaptation model is trained separately for each scene using the same historic training images and selected contemporary target images. Apart from the adaptation stage, all configurations share the same neural scene, feature encoder, ray–query matching module, and pose solver. We additionally report iNeRF refinement initialised from the ray-based pose estimate.

Table 3 reports the results where HistReNeRF achieves the lowest translation error in all three scenes. On Arc de Triomphe, embedding-space adaptation produces a small reduction in MTE from 6.624m to 6.608m and reduces MAE

Fig. 3: Qualitative results across scenes. Each row shows a historic query image (col. 1), the domain-transferred output from the best pixel-space translation method (col. 2), the novel view rendered by TensoRF at the HistReNeRF estimated pose (col. 3), the render at the HistReNeRF + iNeRF refined pose (col. 4), and the ground-truth render at the registered pose (col. 5). Misalignment between cols. 3–4 and col. 5 reflects localisation error. The renders are visually imprecise across all methods, consistent with the high MTE and MAE values reported in Table 3.

from 71.7° to 69.6°. The modest translation gain is consistent with the substantial viewpoint mismatch and rotational symmetry of this scene. On Brandenburg Gate, HistReNeRF reduces MTE by 17%, from 4.712m to 3.912m, while retaining the baseline MAE of 23.4°. The greatest improvement occurs on Piazza del Duomo, where MTE decreases from 1.783,m to 1.589,m and MAE decreases from 85.5° to 58.6°. In contrast, pixel-space translation does not provide a comparable improvement. Across Arc de Triomphe and Piazza del Duomo, all pixel-space methods increase translation error relative to no adaptation. CUT produces a small translation improvement on Brandenburg Gate, reducing MTE from 4.712,m to 4.543,m, but increases MAE from 23.4° to 28.2°. HistReNeRF is therefore the only adaptation approach that improves or preserves both pose measures consistently across the three scenes. The following subsection examines the scene-dependent behaviour of pixel-space adaptation and the factors underlying this difference.

Table 4 provides complementary cumulative recall results at angular and translation thresholds. These results show that the benefit of embedding-space adaptation is not limited to mean error. On Arc de Triomphe, angular recall at 90° increases from 57.7% to 69.2%. On Piazza del Duomo, translation recall at 2.5,m increases from 80.0% to 93.3%, while angular recall at 30° increases from 6.7% to 26.7%.

The magnitude of improvement depends on contemporary scene coverage. Arc de Triomphe exhibits the smallest gain from embedding-space adaptation and the strongest degradation from pixel-space translation. As shown in Fig. 3, historic queries include distant, central, and elevated viewpoints that are underrepresented in the street-level Mapillary collection. Appearance adaptation can reduce the historic-to-contemporary feature gap but cannot recover architectural evidence that is absent from the contemporary reconstruction. This mismatch is damaging for pixel-level translation, which can introduce modern street-level textures that are inconsistent with the structure visible in historic views.

5 Conclusion

We presented HistReNeRF for relocalising historic photographs within contemporary neural scene reconstructions. The method adapts DINOv2 patch embeddings towards a contemporary feature distribution before ray–query matching, leaving the query image and scene representation unchanged. Experiments on HistScene show that embedding-space adaptation achieves the lowest translation error across all three landmarks while improving or preserving angular accuracy; pixel-space translation is inconsistent and often degrades localisation. The results indicate that adapting the matching representation is more reliable than translating query pixels under cross-temporal appearance change. Future work could evaluate additional site types, including less-photographed and archaeological sites, compare classical localisation baselines, investigate alternatives such as diffusion, and explore richer archival case studies with broader viewpoint and scene coverage.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 72–79 (2009). https://doi.org/10.1109/ICCV.2009.5459148
2. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L.: Night-to-day image translation for retrieval-based localization. In: 2019 International conference on robotics and automation (ICRA). pp. 5958–5964. IEEE (2019)
3. Atapour-Abarghouei, A., Breckon, T.P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2800–2810 (2018). https://doi.org/10.1109/CVPR.2018.00296
4. Beckers, B., Beckers, P.: Fast and accurate view factor generation (09 2016)
5. Bortolon, M., Tsesmelis, T., James, S., Poiesi, F., Del Bue, A.: IFFNeRF: Initialisation free and fast 6dof pose estimation from a single image and a nerf model. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 1985–1991. IEEE (2024)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: TensoRF: Tensorial radiance fields. In: European conference on computer vision. pp. 333–350. Springer (2022)
7. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). pp. 337–33712. IEEE (2018)
8. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: European conference on computer vision. pp. 179–196. Springer (2018)
9. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
11. Komorowicz, D., Sang, L., Maiwald, F., Cremers, D.: Coloring the past: Neural historical monuments reconstruction from archival photography. In: Cremers, D., Lähner, Z., Moeller, M., Nießner, M., Ommer, B., Triebel, R. (eds.) GCPR (2). Lecture Notes in Computer Science, vol. 15298, pp. 55–71. Springer (2024), http://dblp.uni-trier.de/db/conf/dagm/gcpr2024-2.html#KomorowiczSMC22
12. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Advances in neural information processing systems **30** (2017)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94, https://doi.org/10.1023/B:VISI.0000029664.99615.94
14. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017). https://doi.org/10.1177/0278364916679498, https://robotcar-dataset.robots.ox.ac.uk/
15. Maiwald, F.: Generation of a benchmark dataset using historical photographs for an automated evaluation of different feature matching methods. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLII-2/W13**, 87–94 (2019). https://doi.org/10.5194/isprs-archives-XLII-2-W13-87-2019, https://isprs-archives.copernicus.org/articles/XLII-2-W13/87/2019/

16. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)
17. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7210–7219 (2021)
18. Masset, L.: Partition of the circle in cells of equal area and shape (2011), https://api.semanticscholar.org/CorpusID:3234073
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
20. Moreau, A., Piasco, N., Bennehar, M., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Crossfire: Camera relocalization on self-supervised features from an implicit representation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 252–262. IEEE (2023)
21. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt
23. Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: European Conference on Computer Vision. pp. 58–77. Springer (2024)
24. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., et al.: Benchmarking 6dof outdoor visual localization in changing conditions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8601–8610. IEEE (2018)
27. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (Jul 2006). https://doi.org/10.1145/1141911.1141964, https://doi.org/10.1145/1141911.1141964
28. Wang, J., Rupprecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9739–9749. IEEE (2023)
29. Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. In: Computer graphics forum. vol. 41, pp. 641–676. Wiley Online Library (2022)
30. Xue, F., Budvytis, I., Olmeda Reino, D., Cipolla, R.: VRS-NeRF: Visual relocalization with sparse neural radiance field. In: European Conference on Computer Vision. pp. 86–102. Springer (2024)

31. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: iNeRF: Inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1323–1330. IEEE (2021)
32. Zhang, J., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. In: International conference on learning representations. vol. 2024, pp. 23345–23366 (2024)
33. Zhao, B., Yang, L., Mao, M., Bao, H., Cui, Z.: PNeRFLoc: Visual localization with point-based neural radiance fields. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7450–7459 (2024)
34. Zhou, Q., Maximov, M., Litany, O., Leal-Taixé, L.: The nerfect match: Exploring nerf features for visual localization. In: European Conference on Computer Vision. pp. 108–127. Springer (2024)
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV). pp. 2242–2251. Ieee (2017)