IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Reliable and Fast Humans Removed Visual Scene Representation

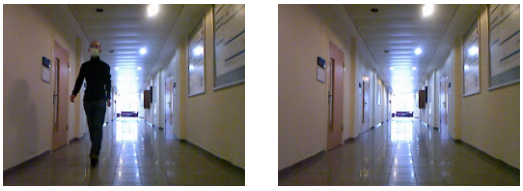
Serhat İşcan and H. Işıl Bozma

***Abstract*—This paper introduces a reliable and fast method for scene representation from a single RGB frame, even with human occlusion. Our goal is to enhance vision-based spatial reasoning in dynamic environments where human presence varies over time. Once humans are detected, the method addresses two key challenges: estimating the level of visual obstruction and generating a scene descriptor with humans removed. The first is handled via a novel visual obstruction measure that prevents descriptor generation under high occlusion. The second is addressed by adapting the previously presented bubble descriptor so that surface regions corresponding to detected humans are deformed using a modified spherical interpolation method—eliminating the need for inpainting or reconstruction and enabling rapid computation. We validate our approach through extensive comparisons across multiple datasets, including two new datasets collected using both stationary and mobile robots. Results show comparable representation quality with a 14-44× reduction in computation time.**

***Index Terms*—Scene descriptors, dynamic entity removal, place learning and recognition, spatial reasoning, visual odometry, mapping.**

I. Introduction

SCENE representation plays a critical role in vision-based spatial reasoning tasks such as mapping, place learning, and recognition. However, most existing representations are designed for static environments and their performance degrades in the presence of humans, who occlude the underlying static scene. Human presence, however, is inevitable in many real-world settings, particularly in domains such as social and service robotics. For example, a robot's ability to recognize a location as a corridor from an RGB image (Fig.1a) depends on the similarity between the resulting representation and that of the unobstructed scene (Fig.1b). Given that human presence varies over time, robust scene representation requires invariance to such dynamic elements, and it is achieved by encoding only the static scene. A common strategy is to detect and remove humans prior to descriptor construction (Fig. 2a) [1]. While this enables more reliable representations, the associated preprocessing (typically involving inpainting or reconstruction) is computationally expensive and delays downstream spatial reasoning. Furthermore, to the best of our knowledge, existing methods do not consider the degree of visual obstruction prior to human removal. As occlusion increases, the informativeness of the RGB frame diminishes; beyond a certain threshold, the frame may cease to be a reliable basis for scene representation.



(a) With human presence (b) Static scene

Fig. 1: RGB data of a scene with human and without human

This paper addresses the problem of scene representation with humans present and proposes a novel approach that is both reliable and fast. First, we introduce a visual obstruction measure that can be used by the robot to process only the data in which the humans do not obstruct too much of the scene. Its formulation depends on the number of humans detected and their corresponding geometry on the image. Next, we formulate what we refer to as the 'humans removed scene descriptor'. The descriptor is directly constructed from a single RGB frame, even when humans are present in the scene. It adapts bubble descriptors [2] by deforming regions associated with detected humans using a modified spherical interpolation algorithm, enabling human-removed scene representation without preprocessing steps such as inpainting or reconstruction. Motivated by the fact that many spatial reasoning tasks require only a static scene representation rather than full reconstruction, this approach enhances view association regardless of human presence while significantly reducing computational overhead by eliminating preprocessing. Our approach is validated experimentally along with a comparative study of two state-of-the-art inpainting/image completion methods. In summary, our key contributions are:

- A visual obstruction measure that can be used to define how much humans obstruct the scene.
- A humans removed scene descriptor that can be constructed directly from data once humans are detected.
- Two human-occupied scene datasets from static and navigating robot scenarios that can be used to assess similarity performance.

It is demonstrated that the proposed approach enables reliable and comparatively fast representation of RGB scene data with humans removed. While this work focuses on human-related data removal, the method is readily applicable to any detected dynamic entity.

Manuscript received: August, 14, 2025; Revised November, 11, 2025; Accepted December, 10, 2025.

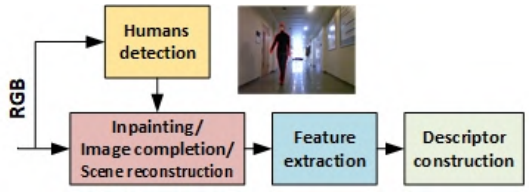
This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments. This work has been supported in part by TUBITAK EEEAG-118E857 and in part by ROYAL CB SBB 2019K12-149250.

S. İşcan and H.I. Bozma are with the Intelligent Systems Laboratory,Electrical and Electronics Eng. Dept., Boğaziçi University, Istanbul, Turkey. serhat_iscan@hotmail.com , bozma@bogazici.edu.tr

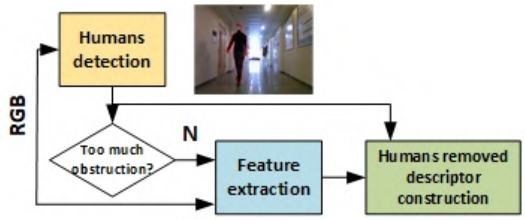
Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

The outline of the paper is as follows: A brief discussion on the related work is given in Section II. The obstruction measure is formulated in Section III. Scene representation based on bubble descriptors is reviewed briefly in Section IV for completeness. The construction of the humans removed scene descriptor is explained in Section V. Experimental results along with two comparative studies are discussed in Section VII. The paper concludes with a brief summary and discussion of future work.



(a) Previous approaches



(b) Proposed approach

Fig. 2: Previous approaches rely on preprocessing to remove humans, whereas the proposed method evaluates the obstruction level and directly constructs the scene descriptor when obstruction is acceptable.

II. Related Work

Scene representation is generally based on constructing descriptors that capture the 'distinct' features of the RGB data. If there are humans in the scene, the problem is treated as a particular case of representing scenes containing dynamic objects. The common strategy consists of i) detecting the segments corresponding to humans in the RGB frame [3]; ii) removing them from the data and then iii) constructing the scene descriptor as shown in Fig. 2a.

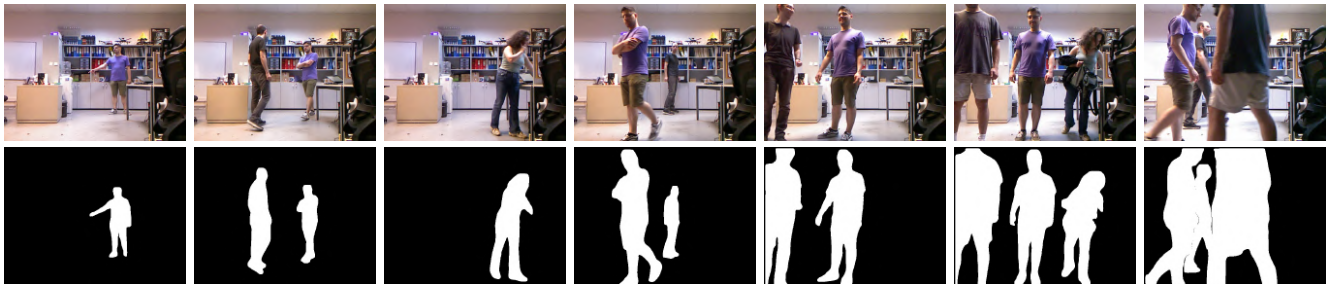
A. Scene Descriptors

Scene descriptors can be categorized as hand-crafted or learned, and as local or global. Hand-crafted descriptors, designed by experts, balance representativeness with computational efficiency [4], [5]. Local descriptors—such as corners, SIFT [6], ORB [7], SURF [8], and geometric features — provide sparse, low-level features that lack semantic meaning and global context, limiting their effectiveness for comprehensive scene representation. They often yield hundreds of features per image, making direct matching inefficient and unreliable [9]. Additionally, representing complex scenes requires many descriptors, increasing computational load and redundancy. Global descriptors like bag-of-words (BOW) [10] and VLAD [11] reduce computational complexity but are less flexible than local descriptors in handling pose variations [12], [13], and rely on learned codebooks. Alternatively, deep learning–based methods such as convolutional neural networks (CNNs) provide powerful and generic visual descriptors [14]. However, their reliance on large labeled datasets [15], [16] limits their applicability in mobile robotics, where continuous adaptation to novel and dynamic environments is essential. Self-supervised approaches like SuperPoint mitigate the need for labeled data and offer relatively lightweight solutions. [17], yet their inference speed and computational demands remain significantly higher than those of handcrafted descriptors—particularly on power- and resource-constrained embedded platforms. Moreover, as an inherently local descriptor, SuperPoint inherits the typical limitations of local representations, including limited semantic awareness and poor robustness to large viewpoint or appearance changes, as discussed above. In prior work, bubble descriptors were introduced as flexible handcrafted features combining local and global representations [2]. Their key strengths include rotational invariance, incremental and lightweight computation, and robust performance, as demonstrated in extensive place recognition benchmarks and comparative studies with state-of-the-art descriptors [2], [18], as well as in place learning [19] and autonomous spatial cognition [20]. This paper further highlights an additional advantage: the efficient adaptability of bubble descriptors for reliable and fast humans removed scene representation.

B. Humans' Removal

Regions occluded by dynamic objects are typically addressed using either single-frame methods— such as inpainting or image completion — or multi-frame approaches based on static scene reconstruction. Single-frame methods often aim to synthesize missing pixel data [21], generating plausible texture and structure in the occluded regions [22], with extensions to depth completion as well [23]. These techniques, however, are computationally intensive. Alternatively, semantic reconstruction methods retrieve similar reference images from a database to fill in missing areas [21], [24], though the search process is also costly [25]. Multi-frame approaches reconstruct the static background by detecting and tracking dynamic objects across time [26]–[30]. While generally reliable, these methods rely on temporal data, limiting their applicability in real-time or single-frame scenarios. Recent advances using CNNs and GANs frame inpainting and view synthesis as conditional generation problems [31], [32], and deformation-based techniques such as D-NeRF [33] and Nerfies [34] map scenes to canonical frames. However, these are restricted to small-motion or object-centric scenes, or require prior knowledge of canonical views. Detection-based methods instead identify dynamic regions and use trained networks to regress missing pixel values [35]–[40]. These approaches often suffer from decreased performance when test-time occlusions differ significantly from training data, and their computational demands limit real-time use in robotics. Importantly, to the best of our knowledge, none of the aforementioned methods assess the level of scene obstruction prior to processing. Moreover, while they aim to generate realistic reconstructions, this may be unnecessary for many robotics tasks — where it suffices to derive a representation similar to that of the unobstructed static scene. Our proposed method addresses both limitations, as illustrated in Fig. 2b. First, the robot evaluates the visual obstruction and discards overly occluded frames. Second, it directly constructs the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) N_h=N'_h=1, $\mathcal{C}_o$=0.096 (b) N_h=N'_h=2, $\mathcal{C}_o$=0.251 (c) N_h=N'_h=1, $\mathcal{C}_o$=0.267 (d) N_h=N'_h=2, $\mathcal{C}_o$=0.525 (e) N_h=N'_h=2, $\mathcal{C}_o$=0.969 (f) N_h=N'_h=3, $\mathcal{C}_o$=1.461 (g) N_h=3,N'_h=1 $\mathcal{C}_o$=2.659

Fig. 3: Sample scenes with human presence, detected human regions and obstruction levels

scene descriptor—without inpainting, image completion, or reconstruction—enabling efficient and reliable representation from a single RGB frame.

III. Visual Obstruction Levels

Amount of human occlusion directly impacts RGB frame informativeness, limiting the robot's ability to generate reliable scene representations. To quantify this effect, a visual obstruction metric is computed following human detection, capturing the extent of occlusion present in each frame.

Suppose the robot detects N_h humans in an incoming $N_1 \times N_2$ image with area $A = N_1 N_2$ with each human's region represented by a binary mask $\mathcal{S}_i^h$ $i = 1, \ldots, N_h$. Let $N'_h \leq N_h$ represent the number of connected human masks after merging overlapping or adjacent masks via a connected components algorithm. This merging accounts for scenarios in which small inter-human distances or adjacent positions cause individual masks to merge into a single contiguous occlusion region as illustrated in Fig.3g. Let each resulting mask be characterized by its width w_i, height h_i, and area A_i. To quantify the degree of visual obstruction caused by the detected humans, we introduce a visual obstruction measure $\mathcal{C}_o$ defined as:

$$\mathcal{C}_o = \sum_{i=1}^{N'_h} \frac{A_i}{A} e^{\frac{w_i}{N_1}} e^{\frac{h_i}{N_2}} \tag{1}$$

This formulation captures three key factors contributing to visual obstruction: i) Relative mask area: The ratio $\frac{A_i}{A}$ quantifies the proportional coverage of the scene by each obstructed region (see Fig.3a and Fig.3c). Larger contiguous occlusions contribute more to visual obstruction; for a fixed total occluded area, obstruction complexity increases with the size of the largest connected region, as fragmented masks impose a lower impact; ii) Spatial Extent Penalty: The exponential terms $e^{\frac{w_i}{N_1}}, e^{\frac{h_i}{N_2}}$ emphasize the normalized spatial extent of occluded regions along the image width and height, respectively. This formulation penalizes elongated or large bounding regions more strongly, reflecting their greater potential to obstruct critical visual information; iii) Aggregation over connected components: Summation over all connected human regions captures both the count and spatial distribution of occlusions, with the measure increasing as the number of detected humans rises or as proximity causes mask merging into larger obstructed areas (Fig. 3a and Fig. 3b or Fig. 3c and Fig. 3d). As individuals approach the camera and occupy a larger portion of the field of view, $\mathcal{C}_o$ increases correspondingly. As shown in Fig. 3b and Fig. 3c, a single nearby person can thus produce occlusion levels comparable to those caused by multiple individuals at greater distances. This spatial dependence also explains the variability observed in Fig. 3d–3e, where identical human counts yield different $\mathcal{C}_o$ values due to positional differences. $\mathcal{C}_o$ ranges from 0 (no occlusion) to a theoretical maximum of ≈ 7.3891 (complete occlusion), governed chiefly by the robot–human proximity. In practice, significant occlusion occurs once $\mathcal{C}_o$ exceeds about 2.7, corresponding to distances below roughly 50 cm, where the human almost fully blocks the camera's field of view. In this case, since frames are insufficiently informative for reliable descriptor extraction, frames with $\mathcal{C}_o > 2.7$ are designated as *severely obstructed* and excluded from descriptor computation. Non-severely obstructed frames ($\mathcal{C}_o \leq 2.7$) are classified into *low* ($0 \leq \mathcal{C}_o \leq 0.5$), *medium* ($0.5 < \mathcal{C}_o \leq 1$), and *high* ($1 < \mathcal{C}_o \leq 2.7$) obstruction levels based on empirical thresholds. This can be used to modulate downstream descriptor processing relative to occlusion severity.

IV. Scene Representation

Scene representation is based on bubble descriptors [2]. A concise overview is provided for completeness; a comprehensive analysis and comparison with state-of-the-art methods is available in the original publication. Notably, their formulation and experimental results with benchmark data focus primarily on static scenes, whereas the present work addresses the more challenging and realistic scenario of dynamic scenes with human presence. Each incoming RGB frame is first represented by a set of N_v bubble surfaces $B_v(x) : \mathcal{F} \to \mathbb{R}^{\geq 0}$ $(v = 1, \ldots, N_v)$ depending on the robot's position $c \in \mathbb{R}^2$ and heading $\alpha \in S^1$. Here, $x = \begin{bmatrix} c & \alpha \end{bmatrix}^T$, N_v refers to the number of different features and $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \subset S^2$ refers to the viewing directions defined in spherical coordinates. At each x, they are all initialized to be S^2 spheres with radius $\rho_v(b, 0) = \rho_0$. For each feature v and for each viewing direction $f \in \mathcal{F}$ with $f = \begin{bmatrix} f_1 & f_2 \end{bmatrix}^T$, if there is an observed feature value $q_v(b)$, the corresponding bubble surface $B_v(x)$ is deformed at f accordingly as:

$$B_v(x) = \left\{ \begin{bmatrix} f \\ \rho_0 + q_v(b) \end{bmatrix} \mid \forall f \in \mathcal{F} \text{ and } b = \begin{bmatrix} x \\ f \end{bmatrix} \right\}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Thus, each bubble surface $B_v(x)$ is a robo-centric compact representation of the observed values of v^{th} feature while also encoding their relative S^2-geometry. Bubble descriptors are vector representation of bubble surfaces $B_v(x)$ as obtained using double trigonometric Fourier series (DTFS) [41]:

$$\rho_v(b) \cong \sum_{h_1=0}^{H_1-1} \sum_{h_2=0}^{H_2-1} \lambda_{h_1h_2} z_{vh_1h_2}^T(x) e_{h_1h_2}(f). \tag{2}$$

The parameters H_1 and H_2 are positive-valued integers that correspond to the number of harmonics. Note that with $H_1 = H_2 = \infty$, the representation is exact. In practice, with $H_1, H_2 << \infty$, bubble surfaces are approximated. The number of harmonics can be set on the application requirements and the sensor used [2]. The parameters $\lambda_{h_1h_2}$ are defined as: 0.25 (h_1= 0, h_2= 0), 0.5 ($h_1 > 0$, $h_2 = 0$ or $h_1 = 0$, $h_2 > 0$),1 (otherwise). For each pair (h_1, h_2) of harmonics, the vector $e_{h_1h_2}(f) \in \mathbb{R}^4$ is a vector of an orthonormal set of trigonometric basis functions:

$$e_{h_1h_2}(f) = \begin{bmatrix} \cos(h_1f_1)\cos(2h_2f_2) \\ \sin(h_1f_1)\cos(2h_2f_2) \\ \cos(h_1f_1)\sin(2h_2f_2) \\ \sin(h_1f_1)\sin(2h_2f_2) \end{bmatrix}, \quad f \in \mathcal{F}.$$

These functions have period $0 \leq f_1 \leq 2\pi$, period $0 \leq f_2 \leq \pi$ and are orthogonal on the corresponding rectangle. The set of vectors $z_{vh_1h_2}(x) \in \mathbb{R}^4$, $h_1 = 0, \ldots, H_1 - 1$, $h_2 = 0, \ldots, H_2 - 1$ is comprised of DTFS coefficients defined as:

$$z_{vh_1h_2}(x) = \frac{2}{\pi^2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{-\pi}^{\pi} \rho_v(b) e_{h_1h_2}(f) df_1 df_2 \tag{3}$$

The *bubble descriptor* $I(x) \in \mathbb{R}^d$ is obtained via the vertical concatenation of H_1H_2-dimensional vectors $I_v(x)$, $v = 1, \ldots, N_v$ that are obtained from the DTFS coefficients of each bubble surface $B_v(x)$:

$$I_v(x) = \left[I_{v00}(x), \ldots, I_{v(H_1-1)(H_2-1)}(x)\right]^T \tag{4}$$

where $I_{vh_1h_2}(x) = z_{vh_1h_2}^T(x) z_{vh_1h_2}(x)$. Hence, it is a d-dimensional vector with $d = N_vH_1H_2$

$$I(x) = \left[I_{1,1}(x), I_{1,2}(x), \ldots, I_{N_v,H_1H_2}(x)\right]^T. \tag{5}$$

Note that if robot's position x is not known, then $I(x)$ can be simply referred as I.

Scene similarity is assessed by comparing bubble descriptors I^C and I^R. Standard metrics such as Euclidean or χ^2 distances are inadequate due to scale variability among descriptor elements, while element-wise normalization is impractical in continual and incremental learning. Additionally, asymmetric similarity assessment is preferred to explicitly capture deviations from a fixed reference. To address these challenges, we propose a dissimilarity measure that evaluates each element individually:

$$\xi(I^C, I^R) = \frac{1}{H_1H_2} \sum_{f=1}^{N_v} \|\xi_f(I^C, I^R)\| \tag{6}$$

where $\xi_f(I^C, I^R) = \left[\ \xi_{f,1}(I^C, I^R) \quad \ldots \quad \xi_{f,H_1H_2}(I^C, I^R)\ \right]$ with each entry of this term defined as follows:

$$\xi_{f,i}(I^C, I^R) = \frac{e^{|I^C_{f,i} - I^R_{f,i}|} - 1}{e^{|I^R_{f,i}|}} \tag{7}$$

Thus, as two descriptors become similar, the dissimilarity measure $\xi(I^C, I^R) \to 0$.

V. Humans Removed Scene Representation

Humans removed scene representation is based on modifying the original bubble descriptor formulation as to encode the scene while removing humans related data - referred to as humans removed scene descriptor (HRSD). This is motivated by the fact that the deformation of the bubble surface region $\mathcal{F}_i^h \subset \mathcal{F}$ associated with each detected human region $\mathcal{S}_i^h$, $i = 1, \ldots, N_h$ in the RGB data will differ from that of the scene without human presence - since those parts of the scene are obstructed by the humans. Hence, the deformation of these regions should not be based on observed values, but rather an estimation $\hat{\rho}_v(b)$ of visual features values corresponding to the obstructed scene regions. This can be achieved using bubble surface interpolation. The advantage of this approach is that interpolation can be done directly as part of descriptor construction. Thus, no a priori preprocessing step is required. This is in contrast to using inpainting, image completion, or scene reconstruction methods prior to the construction of the descriptor.

Bubble surface interpolation is based on adapting the spherical k-nearest neighbors interpolation algorithm [42]. In our adaptation, in order to take full advantage of the segmentation mask, the interpolated value at each $f \in \mathcal{F}_i^h$ is computed from the neighbors' set $\mathcal{N}(f; L) \subset \mathcal{F}$ on the bubble surface. It is defined by four neighboring bubble surface regions $\mathcal{N}(f; L) = \cup_{m=1}^4 N_L(f^m)$, $m = 1, \ldots, 4$. The points $f^m \in \mathcal{F}_i$ $m = 1, \ldots, 4$ correspond to four bubble surface points just outside the segment boundary along each of the two orthogonal geodesic lines that pass through f and parallel to $\mathcal{F}_1$ and $\mathcal{F}_2$ axes. Each point f^m is then taken as an anchor point around which a corresponding neighbors' set $N_L(f^m)$ is set as $N_L(f^m) = N_{(2L+1)\times(2L+1)}(f^m) \setminus F_i^h$ - namely subset of $(2L+1) \times (2L+1)$ neighborhood around f^m that is outside the detected human region $\mathcal{F}_i^h$. The parameter L determines the interpolation neighborhood size. For $L \geq 0$, the cardinality $\mid N_L(f^m) \mid$ depends on the shape of the respective human segment $\mathcal{F}_i^h$ and the position of the corresponding anchor point in the image. For example, consider a detected human region $\mathcal{F}^h$as projected onto bubble surface as shown in Fig. 4. For a given $f \in \mathcal{F}^h$, the corresponding anchor points f^m, $m = 1, \ldots, 4$ are as shown by the four points on the boundary. For $L = 1$, the neighbors' set $N_1(f^m)$ is then determined from the $3\times3\setminus F^h$ region around each point f^m as the bubble surface point. Following, the proximal weight - namely the relevance - of each neighbor point $f' \in \mathcal{N}(f; L)$ is defined by Normalized Inverse Squared (NIS) distance based on orthodromic distance as:

$$NIS(f, f') = \frac{1}{\sum_{f'' \in \mathcal{N}(f,L)} \frac{1}{\delta(f,f'')^2 + \epsilon}} \frac{1}{\delta(f, f')^2 + \epsilon} \tag{8}$$

where the orthodromic distance $\delta(f, f')$ is defined as:

$$\delta(f, f') = 2\rho_0 \arctan 2(\sqrt{(d(f, f'))}, \sqrt{(1 - d(f, f'))}) \tag{9}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

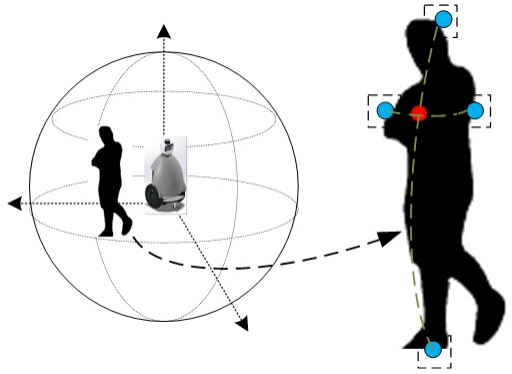


Fig. 4: Sample bubble surface interpolation at a point f (red dot) in a detected human region F^h is done using the neighbors $N_L(f^m)$, $m = 1, \ldots, 4$ consisting of $N_{(2L+1)\times(2L+1)}(f^m) \setminus F^h$ region around each anchor point f^m (blue dots).

and $\epsilon > 0$ is a small number. The orthodromic distance estimates the relevance of each neighborhood based on the haversine formula:

$$d(f, f') = \min\big(1, \max(0, \sin((f_2' - f_2)/2)^2 + \cos(f_2)\cos(f_2')\sin((f_1' - f_1)/2)^2)\big) \quad (10)$$

The arc length between two points is computed in radians and is used in conjunction with the sphere's deformation to end up with the orthodromic distance between the two points. The last step is neighborhood distribution debiasing. This enables the re-weighting of the proximal weights based on the bias of the spatial distribution of visual feature observations. For this, the neighbor centroid is calculated as:

$$\bar{f} = \frac{1}{|\,\mathcal{N}(f; L)\,|} \sum_{f' \in \mathcal{N}(f;L)} f' \quad (11)$$

Next, centroid distances are used to re-normalize the prior proximal weights:

$$w(f, f') = \frac{NIS(f, f')\delta(f', \bar{f})}{\sum_{f'' \in \mathcal{N}(f;L)} NIS(f, f'')\delta(f'', \bar{f})} \quad (12)$$

Finally, the surface deformation $\hat{\rho}(b)$ associated with humans-removed scene can be estimated as:

$$\hat{\rho}(b) = \sum_{f' \in \mathcal{N}(f;L)} w(f, f')\rho(b') \text{ where } b' = \begin{bmatrix} x \\ f' \end{bmatrix} \quad (13)$$

VI. Human-Occupied Scene Datasets

To measure performance in scenes containing people, two benchmark datasets[1] have been generated. The need for these datasets stems from the lack of ground truth static images in publicly available human-occluded image sets. Without the static scene (namely, no human presence), the effectiveness of human removal cannot be observed. To collect data, a mobile robot with a camera located on top at a height of 74cm is utilized. Differing from available datasets, for each scene, both the static scene image and images with varying human obstruction levels (ranging from low to high, with a maximum of 2.7) are included so that a fair assessment of how effective the human removal action is can be made.

The ISL-BU-S dataset comprises approximately 10,000 RGB Kinect images of size 640×480 captured by a stationary robot observing scenes with varying levels of human-induced visual obstruction. Data were collected across five distinct environments, with 1–4 individuals moving at varying distances in front of the robot. Representative examples are shown in Fig. 5, and the distribution of images across low, medium, and high obstruction levels per scene is summarized in Table I.

TABLE I: ISL-BU-S dataset: Distribution of images wrt $\mathcal{C}_o$

	#images	**Low $\mathcal{C}_o$**	**Medium $\mathcal{C}_o$**	**High $\mathcal{C}_o$**
Scene 1	2378	731	983	664
Scene 2	2094	1141	645	308
Scene 3	2223	538	1162	523
Scene 4	1812	575	800	437
Scene 5	1431	339	801	291

The second dataset, ISL-BU-D, is designed for both single human-guided and solo navigation scenarios. To isolate the effect of motion on similarity assessment, only one human is present in the guided cases—reflecting common conditions in home service applications, especially in single or elderly households. The dataset includes scene sequences captured by the same robot (but now with a ZED2 camera) navigating two distinct indoor routes, SC_1 and SC_2, each lasting approximately 10 minutes and covering corridors and rooms across two floors. For each route, three scenarios are recorded: in the first two, the robot follows different individuals at separate times from a distance of 2–3 meters, resulting in obstruction levels ranging from low to medium. In the third scenario (SC_3), the robot navigates the route autonomously, serving as a baseline. Sample frames from human-following and solo navigation scenarios are shown in Fig. 6. Each sequence is downsampled to approximately 500 images of size 640×480.

VII. Experiments

This section presents experimental results using the proposed humans removed scene descriptor (HRSD) on RGB data as obtained by a mobile robot. Three different experiments are conducted: 1) Stationary robot; 2) Dynamic robot; 3) Scene Recognition. In all, humans are detected as segments using YOLO network [43] - specifically the pretrained segmentation model named 'yolo11l-seg' as it is known to be robust in cluttered backgrounds, bad illumination, or varying human postures. As part of the ablation study, a baseline scenario without human removal (NHR) is first considered. For the first two experimental sets, we also evaluate performance when preprocessing RGB frames with image inpainting/completion before computing the corresponding scene descriptors (SD). The preprocessing is performed using three highly cited, state-of-the-art methods, all of which have publicly available implementations. The first method is a classic texture-synthesis approach[2] by Criminisi et al [22]. The second is a deep learning–based method (DeepFill v2) [40] trained on indoor scenes from the dataset [44]. The third, LaMa, is another deep

[1]Collected ISL-BU-S and ISL-BU-D datasets can be accessed via this link: `github.com/islboun/human-occupied-scene-datasets`

[2]While there are also other more recent works, these are primarily variations of this method and furthermore, their codes are not publicly available.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

(a) Scene 1 (b) Scene 2 (c) Scene 3 (d) Scene 4 (e) Scene 5

Fig. 5: ISL-BU-S: Sample images

(a) Following a human (b) Solo navigation

Fig. 6: ISL-BU-D: Sample images along one route

learning generative method [45] trained on the same dataset [44]. and employed within the proposed coding architecture [46]. The 100-dimensional descriptors are constructed using the original bubble descriptor formulation, encoding intensity features with parameters[3] $H_1 = H_2 = 10$. The dissimilarity measure is as defined in Section IV whose value decreases with increased similarity.

A. Stationary Robot Experiments

The first set of experiments evaluates the effectiveness of human removal in scenes observed by a stationary robot, using the ISL-BU-S dataset. Resulting descriptors are compared to corresponding static scenes using average dissimilarity, as summarized in Table II. All methods significantly lower dissimilarity compared to the no-human-removal baseline, with greater improvements at higher obstruction levels. Notably, the proposed descriptor achieves the lowest dissimilarity, and its performance gains amplify as visual obstruction increases. We also examine the effect of neighbors' set size $L \in \{1, \ldots, 5\}$ on performance, observing improvement as the interpolation neighborhood grows.

TABLE II: Stationary robot experiments: Average dissimilarity

Method		Low $\mathcal{C}_o$	Medium $\mathcal{C}_o$	High $\mathcal{C}_o$	Overall
(NHR) SD		0.0510	0.1446	0.1817	0.1216
Criminisi et al.[22]+SD		0.0159	0.0286	0.0557	0.0304
Deepfill v2 [40]+SD		0.0248	0.0406	0.0722	0.0424
LaMa [45]+SD		0.0205	0.0364	0.0672	0.0380
HRSD	$L = 1$	0.0168	0.0218	0.0482	0.0260
	$L = 2$	0.0157	0.0211	0.0394	0.0234
	$L = 3$	0.0153	0.0207	0.0377	0.0227
	$L = 4$	0.0150	0.0205	0.0365	0.0223
	$L = 5$	0.0149	0.0204	0.0365	0.0221

The main advantage of the proposed approach lies in its processing times. Comparative processing times, including color feature computation but excluding human detection, are presented in Table III. All methods run on an Intel Core i9 processor. For the no-human-removal case, the processing time is 17 ms; thus, additional times for each method can be approximated by subtracting this baseline.For the proposed

[3]For the choice of these parameters, please kindly refer to [2].

TABLE III: Processing times per RGB scene

Method		Processing time (sec)
(NHR) SD		0.017
Criminisi et al.[22]+SD		0.22-11
Deepfill v2 [40]+SD		0.67
LaMa [45]+SD		0.94
HRSD	$L = 1$	0.021-0.173
	$L = 2$	0.022-0.229
	$L = 3$	0.023-0.330
	$L = 4$	0.026-0.395
	$L = 5$	0.027-0.486

method and [22], processing time scales with occlusion level, whereas it remains constant for [40] and [45]. As expected, the proposed method is significantly faster, since it does not require preprocessing steps like inpainting, image completion, or scene reconstruction. For $L = 1$, speedup ranges from 10–45× at low obstruction levels and 4–64× at high obstruction, yielding average gains of 14–44× depending on the method. For $L = 5$, average gains vary between 10-18x. Due to minimal similarity improvements but increased processing times with higher L, we use $L = 1$ for all remaining experiments.

B. Dynamic Robot Experiments

The second set of experiments evaluates the effectiveness of human removal in dynamic scenes captured by a moving robot, using the ISL-BU-D dataset. For each sequence, the resulting descriptors are compared against those from other sequences along the same route. Unlike the stationary setup, direct frame-to-frame comparison is infeasible due to the absence of exact temporal correspondence, caused by variations in robot poses, start positions, and trajectories. To address this, the dissimilarity measure ξ is extended to a temporal dissimilarity measure $\tilde{\xi}$. For each k in the first sequence, it is defined as $\tilde{\xi}(I^k, I^{k'}) = \min_{k'} \frac{1}{2n_l+1} \sum_{j=k-n_l}^{k+n_l} \xi(I^{k+j}, I^{k'+j})$ where the search for the best matching frame index k' in the second sequence is restricted to the interval $k' \in \left[(k - n_s)\frac{M_2}{M_1}, (k + n_s)\frac{M_2}{M_1}\right]$ with M_1 and M_2 representing the lengths of the first and second sequences, respectively. The parameter n_l and n_s are chosen based on the robot's velocity profile along each route; here $n_l = 3$ and $n_s = 20$.

Comparative results in Tables IVa–IVd show all four methods similarly reduce dissimilarity. As expected, the reduction relative to baseline is smaller due to obstruction levels fluctuating between low and medium while the robot follows a person at 2–3 meters. Swapping the order of comparison does not necessarily yield symmetric results because frame sequences lack 1:1 correspondence. Average temporal dissimilarity between sequences with different humans is reported in Tables IVe–IVf. Here, the proposed method consistently

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV: Dynamic robot experiments: Average temporal dissimilarity

(a) Human 1 present scenes vs static scenes: SC_1-SC_3

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0527	0.0501	0.0499	0.0499	0.0500
2^{nd} Route	0.0307	0.0244	0.0238	0.0234	0.0239

(b) Human 2 present scenes vs static scenes: SC_2-SC_3

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0556	0.0534	0.0523	0.0516	0.0524
2^{nd} Route	0.0429	0.0405	0.0412	0.0406	0.0410

(c) Static scenes vs human 1 present scenes: SC_3-SC_1

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0497	0.0445	0.0445	0.0450	0.0435
2^{nd} Route	0.0382	0.0283	0.0280	0.0277	0.0267

(d) Static scenes vs human 2 present scenes: SC_3-SC_2

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0469	0.0438	0.0444	0.0436	0.0428
2^{nd} Route	0.0450	0.0391	0.0378	0.0384	0.0374

(e) Human 1 present scenes vs human 2 present scenes: SC_1-SC_2

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0490	0.0483	0.0490	0.0481	0.0478
2^{nd} Route	0.0363	0.0346	0.0330	0.0329	0.0332

(f) Human 2 present scenes vs human 1 present scenes: SC_2-SC_1

	(NHR) SD	[22]+SD	[40]+SD	[45]+SD	HRSD
1^{st} Route	0.0573	0.0521	0.0510	0.0505	0.0478
2^{nd} Route	0.0454	0.0414	0.0410	0.0400	0.0400

outperforms the baseline (no human removal), yielding lower dissimilarity values.

C. Scene Recognition

The final experiments evaluate the proposed method's effectiveness for scene recognition as a downstream task in scenes containing humans, using the ISL-BU-S dataset comprising five distinct environments with varying levels of human obstruction, as previously discussed. For each pair of scenes, average dissimilarity was computed over all frame pairs — first without human removal, then with human removal using the proposed method. The results are normalized on a per-row basis by dividing all entries in a row by the minimum dissimilarity value in that row and are presented in Tables Va and Vb, respectively. In both cases, dissimilarity is lowest within the same scene, confirming the effectiveness of the bubble descriptors, both with and without human removal. Notably, using descriptors from the proposed method yields nearly a three-fold increase in mean off-diagonal dissimilarity, indicating enhanced scene discrimination.

To assess the scene discrimination performance of the proposed method against a state-of-the-art baseline, the same experimental setup is repeated, substituting bubble descriptors with embeddings from a pretrained DINOv3 model [47]-specifically ViT-L/16 model that is trained with LVD-1689M dataset. Embeddings are generated from both raw (NHR) and LaMa-inpainted [45] images. The results are again normalized with respect to the smallest average dissimilarity, are reported in Tables VIa and VIb. Without human removal, dissimilarity

TABLE V: Average scene dissimilarity using bubble descriptors

(a) No human removal (NHR)

	Scene 1	**Scene 2**	**Scene 3**	**Scene 4**	**Scene 5**
Scene 1	1.0000	2.3173	2.0251	2.2942	2.3464
Scene 2	1.4083	1.0000	2.1211	2.4022	1.3037
Scene 3	1.2481	1.4613	1.0000	1.5004	1.7044
Scene 4	2.4413	2.8109	1.8047	1.0000	2.0593
Scene 5	1.8104	1.9131	2.0914	2.5158	1.0000

(b) Proposed approach (HRSD)

	Scene 1	**Scene 2**	**Scene 3**	**Scene 4**	**Scene 5**
Scene 1	1.0000	4.4470	8.1258	4.8907	4.0033
Scene 2	3.5056	1.0000	7.2641	8.1400	2.4018
Scene 3	4.6773	5.7980	1.0000	4.7980	6.2365
Scene 4	8.7359	8.6667	6.8874	1.0000	6.6537
Scene 5	6.1086	7.8778	8.0995	8.8100	1.0000

is lowest within the same scene, making DINOv3 an effective baseline for this task. While the comparison of Table Va and Table VI shows an increase in average off-diagonal dissimilarity, from 1.9790 to 3.0854, it should be noted that such cross-descriptor comparisons of dissimilarity magnitudes are not necessarily indicative of better performance. This is because differences in descriptor characteristics, such as value ranges, can affect absolute dissimilarity scores even when underlying relationships are unchanged; therefore, assessments should be made within each descriptor individually.

Interestingly, unlike the proposed approach, DINOv3-based descriptors are adversely affected by the removal of human regions as shown in Table VIb. A possible explanation for this observation is the texture noise introduced by the inpainting process. These results reveal an additional advantage of the proposed approach for privacy-preserving representations: once human removal is enforced, LaMa+DINOv3 performance degrades substantially, whereas HRSD remains effective. To mitigate such effects, it may be beneficial to train the off-the-shelf DINOv3 model on inpainted images as well. However, this would require additional effort and training, which is not necessary for the proposed approach. The inference time for DINOv3-based descriptors is 160 ms, whereas applying LaMa inpainting and generating DINOv3-based descriptors takes approximately $0.92 + 0.16 = 1.08$s. This represents a substantial increase in processing time compared with bubble descriptors.

TABLE VI: Average scene dissimilarity: DINOv3-based descriptors

(a) No human removal (NHR) + DINOv3 [47]

	Scene 1	**Scene 2**	**Scene 3**	**Scene 4**	**Scene 5**
Scene 1	1.0000	2.0846	2.5036	2.5924	2.7609
Scene 2	2.4658	1.0000	2.8059	2.8801	3.2239
Scene 3	5.4928	5.3781	1.0000	5.5320	6.1906
Scene 4	2.3274	2.2048	2.3670	1.0000	2.2804
Scene 5	2.1877	2.1524	2.2781	1.9990	1.0000

(b) LaMa [45] + DINOv3 [47]

	Scene 1	**Scene 2**	**Scene 3**	**Scene 4**	**Scene 5**
Scene 1	1.0000	1.7890	2.1188	2.2268	2.4043
Scene 2	1.9188	1.0000	2.2316	2.2999	2.6438
Scene 3	3.3253	3.2503	1.0000	3.4621	3.9009
Scene 4	1.7079	1.6525	1.7315	1.0000	1.7436
Scene 5	2.0140	2.0004	2.0919	1.8874	1.0000

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

VIII. Conclusion

We have presented a reliable and fast scene representation approach — effective even in the presence of multiple humans — by adapting the previously introduced bubble descriptors. In this approach, the robot first computes the visual obstruction level in the RGB data. In case of no obstructedness, it forms the humans removed scene descriptor directly without a preprocessing stage such as inpainting, image completion, or scene reconstruction. The proposed approach is validated through both stationary and mobile robot experiments, demonstrating comparable or superior similarity to static scene descriptors at significantly reduced computational cost. Although this work focuses on human removal, the proposed approach is generalizable to any detectable dynamic entity.

References

[1] B. Bescos, J. M. Facil, J. Civera, and J. Neira, "DynaSLAM: tracking, mapping, and inpainting in dynamic scenes," *IEEE RAL*, vol. 3, no. 4, p. 4076–4083, Oct 2018.

[2] O. Erkent and H. I. Bozma, "Bubble space and place representation in topological maps," *The Int. J. Rob. Res.*, vol. 32, no. 6, pp. 672–689, 2013.

[3] N. Sayez and C. De Vleeschouwer, "Accelerating the creation of instance segmentation training sets through bounding box annotation," 2022.

[4] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct 2012.

[5] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, Jul. 2008.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[9] S. Lowry, N. Sanderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. on Rob.*, vol. 32, no. 1, pp. 1–19, 2016.

[10] A. Angeli, D. Filliat, S. Doncieux, and J. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, Oct 2008.

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conf. on Comp. Vis. and Patt. Recog.*, 2010, pp. 3304–3311.

[12] J. Vogel, A. Schwaninger, C. Wallraven, and H. Bulthoff, "Categorization of natural scenes: Local versus global information and the role of color," *ACM Trans. on Applied Perception*, vol. 4, no. 3, p. 19, 2007.

[13] N. M. Elfiky, J. Gonzàlez, and F. X. Roca, "Compact and adaptive spatial pyramids for scene recognition," *Image and Vision Computing*, vol. 30, no. 8, pp. 492 – 500, 2012.

[14] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] T. Schmidt, R. Newcombe, and D. Fox, "Self-supervised visual descriptor learning for dense correspondence," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 420–427, April 2017.

[16] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *IEEE/RSJ Int'l Conf. on Intell. Rob. Sys.*, Sep. 2017, pp. 9–16.

[17] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[18] H. Karaoğuz, Ö. Erkent, and H. I. Bozma, "Rgb-d based place representation in topological maps," *Machine Vision and Applications*, vol. 25, no. 8, pp. 1913–1927, 2014.

[19] O. Erkent, H. Karaoguz, and H. I. Bozma, "Hierarchically self-organizing visual place memory," *Adv. Rob.*, vol. 31, no. 16, pp. 865–879, 2017.

[20] H. Karaoğuz and H. I. Bozma, "An integrated model of autonomous topological spatial cognition," *Auton. Robots*, vol. 40, no. 8, pp. 1379–1402, 2016.

[21] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. Graph.*, vol. 26, no. 3, p. 4–es, Jul. 2007.

[22] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *CVPR*, vol. 2, 2003, pp. II–II.

[23] A. Atapour-Abarghouei and T. P. Breckon, "A comparative review of plausible hole filling strategies in the context of scene depth image completion," *Computers & Graphics*, vol. 72, pp. 39–58, 2018.

[24] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. on Graphics*, vol. 28, no. 3, 2009.

[25] K. He and J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE Trans. PAMI*, vol. 36, no. 12, pp. 2423–2435, 2014.

[26] I. A. Barsan, P. Liu, M. Pollefeys, and A. Geiger, "Robust dense mapping for large-scale dynamic environments," in *IEEE ICRA*, 2018, pp. 7510–7517.

[27] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual slam towards dynamic environments," in *IEEE/RSJ IROS*, 2018, pp. 1168–1174.

[28] D.-H. Kim and J.-H. Kim, "Effective background model-based rgb-d dense visual odometry in a dynamic environment," *IEEE TRO*, vol. 32, no. 6, pp. 1565–1573, 2016.

[29] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss, "Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals," in *IEEE/RSJ IROS*, 2019, pp. 7855–7862.

[30] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *IEEE ICRA*, 2015, pp. 75–82.

[31] X. Zhang, D. Zhai, T. Li, Y. Zhou, and Y. Lin, "Image inpainting based on deep learning: A review," *Information Fusion*, vol. 90, pp. 74–94, 2023.

[32] W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka, "Deep learning-based image and video inpainting: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2401.03395

[33] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *IEEE/CVF CVPR*, 2021, pp. 10 313–10 322.

[34] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *IEEE/CVF ICCV*, 2021, pp. 5845–5854.

[35] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017.

[36] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *IEEE/CVF Conf. on CVPR*, 2019, pp. 5785–5794.

[37] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *IEEE CVPR*, 2017, pp. 4076–4084.

[38] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated contextual transformations for high-resolution image inpainting," in *Arxiv*, 2020, pp. –.

[39] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *IEEE/CVF CVPR*, 2021, pp. 2855–2864.

[40] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *IEEE/CVF CVPR*, 2019, pp. 4471–4480.

[41] G. P. Tolstov, *Fourier Series*. Prentice-Hall, 1962.

[42] P. Trempe, "Spherical k-nearest neighbors interpolation," *arXiv preprint arXiv:1910.00704*, 2019.

[43] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[44] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. PAMI*, 2017.

[45] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of IEEE/CVF Winter Conference on Applications of computer vision*, 2022, pp. 2149–2159.

[46] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint anything: Segment anything meets image inpainting," *arXiv preprint arXiv:2304.06790*, 2023.

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