SuperPrimitive: Scene Reconstruction at a Primitive Level

Kirill Mazur, Gwangbin Bae, Andrew J. Davison
Dyson Robotics Lab, Imperial College London
{k.mazur21, g.bae, a.davison}@imperial.ac.uk

Abstract

Joint camera pose and dense geometry estimation from a set of images or a monocular video remains a challenging problem due to its computational complexity and inherent visual ambiguities. Most dense incremental reconstruction systems operate directly on image pixels and solve for their 3D positions using multi-view geometry cues. Such pixel-level approaches suffer from ambiguities or violations of multi-view consistency (e.g. caused by textureless or specular surfaces).

We address this issue with a new image representation which we call a SuperPrimitive. SuperPrimitives are obtained by splitting images into semantically correlated local regions and enhancing them with estimated surface normal directions, both of which are predicted by state-of-the-art single image neural networks. This provides a local geometry estimate per SuperPrimitive, while their relative positions are adjusted based on multi-view observations.

We demonstrate the versatility of our new representation by addressing three 3D reconstruction tasks: depth completion, few-view structure from motion, and monocular dense visual odometry. Project page: https://makezur.github.io/SuperPrimitive/

1. Introduction

Enriching monocular incremental reconstruction with prior world knowledge is essential for resolving visual ambiguities. This issue is particularly prevalent in scenarios with scarce data observations available: a notable example would be monocular visual SLAM, where images are being streamed from a camera into the system in real-time.

When a monocular vision system encounters a new scene region, it must estimate the region's geometry based on a very limited number of observations. Without this, continuous camera motion tracking would not be possible. Once the scene region is thoroughly observed, the initial geometry estimate should be refined to better explain the multi-view information.

This naturally leads to a question: what sort of priors are effective in both providing reliable initial geometry estimates and supporting multi-view consistency? Geometric priors generally fall into one of two categories: local and global. Local priors, such as smoothness assumption [30] or surface normal regularisation [52], impose additional constraints within a small neighbourhood. Global priors, on the other hand, aim to impose constraints on a larger scale, such as depth prediction [2, 11].

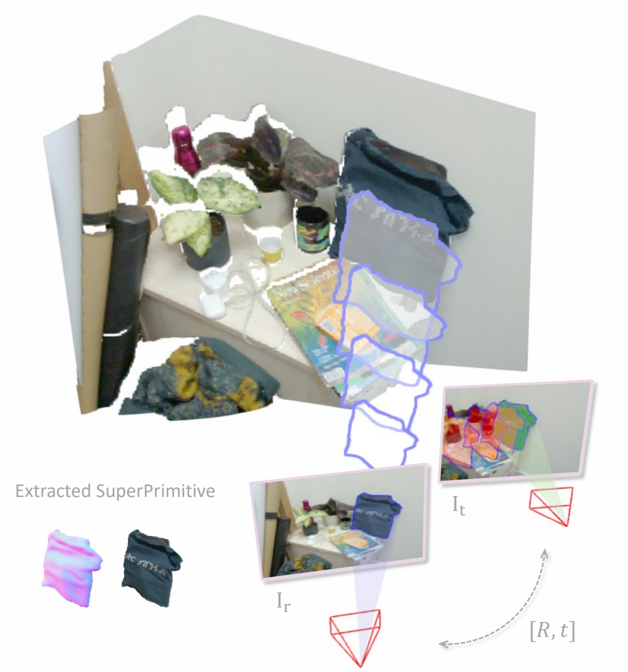

Figure 1. **Multi-View Geometry with SuperPrimitives.** SuperPrimitives are extracted from an input frame by dividing it into image segments equipped with estimated surface normal directions (bottom-left). Each SuperPrimitive induces a dense reconstruction within the corresponding image segment up to *a priori unknown* scale. Different possible reconstructions are shown in light blue. The scales are then jointly optimised together with a relative camera pose to fit multi-view photometric constraints (visualised in green and red). The resulting dense reconstruction of the reference frame is shown in the top.

Our key observation is that some of the geometrical correlations are more reliable than the others, and therefore could be safely "locked in" together within local regions based on a single-view prediction. Points belonging to the same rigid body are strongly correlated, making it unnecessary to determine their depth independently. In contrast, distinct and unrelated objects can be placed arbitrarily in the scene. As the number of objects increases, learning a reliable global prior on their relative positions becomes an increasingly complex problem.

In this work, we show that *purely local but strong priors* are enough to achieve excellent performance across a variety of geometric vision tasks. For that purpose, we introduce a novel representation, *SuperPrimitives*. A SuperPrimitive represents a local image segment, coupled with a dense shape estimate which is determined up to a scale factor. The scale factor can be further adjusted based on information observed from other views or additional measurements.

We show that SuperPrimitives can be efficiently constructed using a *front-end* which consists of two single-image neural networks, extracting image segmentation and surface normal prediction. Effectively, our neural front-end predicts whether adjacent pixels belong to the same geometrical entity through image segmentation and estimates a surface normal at this point, thereby providing an infinitesimal geometry estimate.

We delegate global, scene-level alignment to our streamlined multi-view, iterative, optimisation-based *back-end*. The resulting front-end / back-end tandem combines the flexibility of multi-view based optimisation methods with the observation efficiency common in prior-driven systems.

Our new representation showcases its versatility in three key applications, where single-view ambiguity is resolved via additional measurements or viewpoints:

- Firstly, it adeptly handles zero-shot depth completion tasks in real-world scenarios, matching the performance of state-of-the-art methods tailored for depth completion;
- Secondly, it facilitates joint pose and depth estimation using a limited set of unstructured images, surpassing its nearest competitor even in the absence of global priors;
- Thirdly, our method outperforms previous monocular visual odometry systems on the challenging TUM dataset, and exhibits robustness across various domains.

2. Related Work

Monocular Reconstruction. Both offline monocular reconstruction systems, such as COLMAP [43] and [16], or online systems, such as MonoSLAM [9] and DSO [22] track, filter, and reconstruct only well-constrained points with high and reliable photometric information. This often involves fine-grained and sophisticated point management to reliably resolve visual ambiguities and filter unreliable visual observations. DTAM [30] demonstrated feasibility of incremental dense reconstruction in the monocular scenario. DTAM employed a hand-crafted local smoothness prior to handle regions with poor texture. Subsequently, other local priors [32, 54] have been extensively explored to regularise multi-view geometry estimation problems.

Global Priors. We are interested in exploring the space of possible geometric priors from single-view networks. Depth prediction [12, 37, 38] is the most obvious choice but leads to a rigid per-image reconstruction which is difficult to feed into multi-view optimisation. In the recent years, deep-learning based approaches sought to replace explicit multi-view geometry estimating with learning-based methods [3, 42, 48]. These methods, however, assume known poses and are, therefore, not suitable for joint pose and geometry estimation. Notably, CodeSLAM [2] introduced depth prediction conditioned on latent codes, which are then optimised to achieve cross-view consistency. However, CodeSLAM still struggles with out-of-domain data, as depth prediction networks are known to struggle with generalisation [51].

Higher-Level Mapping. Introducing parametric primitives, such as lines [36, 50], planes [6, 17, 19, 25, 26, 45] or even high-order algebraic shapes [7, 23, 31] to better constrain multi-view geometry problems have been thoroughly explored over the last few decades. They all however use assumptions which may not often hold for all 3D scenes, especially for dense reconstruction.

Besides parametric algebraic primitives object-level mapping has been explored in the last decade. SLAM++ [41] represented a map with a set of CAD models retrieved from an existing database and tracked camera position against these models. This method, however, could not be applied in a variety of settings due to the limited size of the CAD databases. Even objects of the same class can vary in their geometric appearance from instance to instance. This was later approached by [15, 47], who also represent their map via a set of objects, but learn a latent space of possible geometric variations per object class. Fusion++ [29] extended this approach into a more versatile per object depth fusion using an RGB-D sensor and mask proposals from a pre-trained neural network. These methods rely on an RGB-D inputs and assume a pre-defined set of object classes, therefore lacking generality.

3. Method

Firstly, we introduce the concept of a SuperPrimitive and explain how an image is processed into a set of its SuperPrimitives, referring to this part of our method as the *front-end*. In the second part of this section, we describe how multi-view geometry problems can be reformulated at the level of SuperPrimitives instead of pixels. We refer to this stage of SuperPrimitive alignment as the *back-end*.

The core of our method proposes splitting a given image into a set of (possibly overlapping) minimal segments, image regions which are likely to have strongly correlated

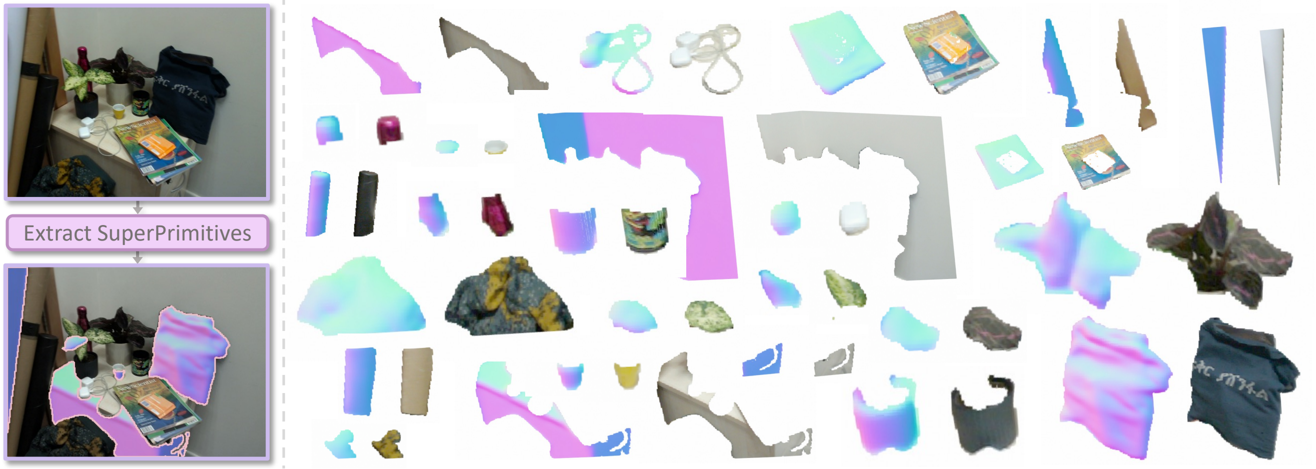

Figure 2. **SuperPrimitves Extraction. (left)** Our *front-end processor* extracts SuperPrimitivies from an image by dividing it into a set of image regions with surface normal directions estimated for each image pixel within the segment. **(right)** Highlighted SuperPrimitives extracted from the image are visualised by showing their estimated normal and colour maps side by side. Note some of them are scaled either up or down for better viewing. While some of the SuperPrimitives are akin to object-level segmentation, the others tend to represent more low-level image segments.

geometry. We repurpose the recent state-of-the-art Segment Anything (SAM) [21] model into predicting these minimal segments.

Our key idea is to estimate local geometry within each segment from a single view, while leaving the relative positioning of the segments to be estimated via multi-view photometric consistency optimisation. We refer to these geometrically enhanced minimal image segments as *SuperPrimitives*, since they are inspired by both superpixels [39] and geometric 3D primitives.

For per-segment local geometry estimation we employ an off-the-shelf surface normal prediction network [1], to estimate infinitesimal geometry for each image pixel. The surface normals within an image segment can be used to estimate its depth via simple integration *up to a scale factor*. We set these scale factors — *depth scales* — to be optimisiable parameters, which are either optimised via multi-view cues or explicit depth measurements for the depth completion experiments.

Thus, our method combines the strong priors provided by state-of-the-art neural networks in the front-end with the flexibility and consistency offered by multi-view optimisation.

3.1. Conventions

Unless stated otherwise, we use lowercase letters for scalar values, bold letters for vectors, and uppercase letters for matrices. We consider images $\mathrm{I} \in \mathbb{R}^{3\times h\times w}$ of height h and width w captured by a camera with a known calibration matrix $\mathrm{K} = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$. Image pixels are parameterised as $\boldsymbol{u} = (u, v) \in [0, h-1] \times [0, w-1]$. Given a per pixel depth function $z(\boldsymbol{u})\colon \mathrm{I} \to \mathbb{R}^{+}$, we define image's unprojection onto the 3D space to be $\pi^{-1}(\boldsymbol{u}) = z(\boldsymbol{u})\mathrm{K}^{-1}\dot{\boldsymbol{u}}$, where dot is a homogenisation operator $\dot{\boldsymbol{p}} = (\boldsymbol{p}, 1)$. Conversely, for a set of 3D points $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$ we define the projection $\pi(\boldsymbol{x}) = \gamma\left(\frac{1}{z}K\boldsymbol{x}\right)$, where the γ function drops the last coordinate $\gamma((x, y, z)^T) = (x, y)^T$.

We represent camera poses as matrices $\mathrm{T}_{WC} \in \mathrm{SE}(3)$ in the camera-to-world coordinates unless stated otherwise. During pose optimisation we store linearised pose increments $\xi \in \mathfrak{se}(3)$ as Lie algebra elements. These increments are used then to update current camera pose estimates, see [13, 14] for more detail.

3.2. SuperPrimitives

We split the input image I into a set of possibly overlapping connected regions (segments), each of which is equipped with its local 3D geometry. More formally, a *SuperPrimitive* $P = (\Omega, \mathfrak{D})$ is a connected image region $\Omega \subseteq \mathrm{I}$ which is also equipped with an *unscaled* depth map $\mathfrak{D}\colon \Omega \to [0, \infty)$ for each pixel within the region Ω. Here, unscaled means that each SuperPrimitive's depth $\mathfrak{D}$ differs from its ground truth depth D by an *a priori unknown* scalar. In other words, there exists a single scalar s such that:

$$D = s \cdot \mathfrak{D} \qquad \text{for all pixels p } \in \Omega \tag{1}$$

We say that an image I is *primified* into super-primitives $\mathcal{P}(\mathrm{I}) = \{P_i = (\Omega_i, \mathfrak{D}_i)\}$ if it has n, *possibly overlapping* primitives that lie within the image I, i.e $\bigcup \Omega_i \subseteq \mathrm{I}$.

For brevity we use the words *primitive* and *SuperPrimitive* interchangeably throughout the rest of the paper.

Representing Geometry with SuperPrimitives. A set of primitives $\mathcal{P}$ itself is not enough to extract dense image geometry due to scale/depth ambiguity. Therefore, we introduce the concept of optimisable *depth scales* which anchor the correct depth scale for each primitive.

Given a scalar depth scale s for a super-primitive P, one can infer its depth as $D(p) = s \cdot \mathfrak{D}(p)$ for every point $p \in \Omega$. In practice, we employ a log-depth representation to represent depth throughout the whole system. This means that we store optimisable log-depth scales $\log s$ and unscaled log-depth $\log \mathfrak{D}$, and so depth inference reduces to a simple shift operation $\log D = \log s + \log \mathfrak{D}$.

In contrast to other dense methods, we estimate the dense geometry $\mathcal{G}$ of the image in the form of a point cloud rather than a depth image. This choice is driven by the fact that multiple primitives' supporting segments Ω_i and Ω_j might overlap, which would lead to different depth estimates within their intersection $\Omega_i \cap \Omega_j$.

Given a set of *depth scaled primitives* $\mathbb{P} = \{(s_i, P_i)\}$, the geometry $\mathcal{G}$ of the image I can be estimated as follows:

$$\mathcal{G} := \bigcup_i \pi^{-1}(s_i \cdot \mathfrak{D}_i) \tag{2}$$

To obtain estimated depth $\hat{D}$, we average depth values along the corresponding camera rays.

Converting an Image into SuperPrimitives. Next we explain how SuperPrimitives are obtained in practice. An input image I is first split into a set of minimal image segments Ω_i produced by an image segmentation model. We use the Segment Anything [21] model in this work, since it captures low-level semantic correlations highly accurately. Given the extracted segments, we independently estimate the local geometry $\mathfrak{D}_i$ within each segment by integrating surface normals predicted by another state-of-the-art neural network [1] within this region.

Note that this approach is premised on the assumption of geometry being continuous within each predicted image segment Ω_i, hence its unscaled depth values could be obtained by integrating its surface normals. Although this is not guaranteed *a priori*, we observed that a more fine-grained mask selection yields a compelling correlation with geometrical continuity. We discuss this in more detail in the following section.

Image Segment Retrieval. An ideal neural network which estimates image segments $\{\Omega_i\}$ should predict regions of geometric continuity, where depth could be obtained via surface normal integration. However, to the best of our knowledge, such a network does not exist. Our approach aims to approximate this behaviour by utilising the Segment Anything model, coupled with a specialised mask selection process. Specifically, for each SAM query, we strive to select the smallest predicted mask surrounding that point. While such over-segmentation could potentially increase the dimensionality of the multi-view optimisation problem, under-segmentation might lead to incorrect geometry estimate within the primitive. Since we only adjust the scale of a primitive after the normal integration is completed, any incorrect geometry estimates within a primitive cannot be compensated at this stage.

The segmentation model employed in this work outputs three binary segmentation masks given a query point $q \in \mathrm{I}$. For each query q we first filter predicted masks using the post-processing introduced in [21], such as stability filtering and Non-Maximum-Suppression (NMS). Even though our method allows redundant regions, we employed NMS to remove similar segments and save compute in the image alignment stage. With filtering done, the mask with the smallest area is selected. If the filtered set is empty, we discard the query point.

We query the SAM backbone feature extractor once per image. Then we first sample 300 query points randomly across the image, followed by the filtering discussed above. Then the image coverage mask is calculated and an additional 100 mask query points are actively sampled in the uncovered regions.

Normal Integration. To estimate each segment's local geometry, we first pass the image through a surface normal estimation network. This network effectively predicts the derivative of desired depth values. Then, we "integrate" these surface normal vectors within each image segment to obtain its unscaled depth map.

For each image pixel $\boldsymbol{u} = (u, v)^T \in \mathrm{I}$ we estimate its surface normal vector $\boldsymbol{n} = (n_x, n_y, n_z)^T$ with a pretrained state-of-the-art CNN [1]. As was demonstrated in [5], the log-depth $\tilde{z} = \log(z(\boldsymbol{u}))$ satisfies the following PDEs within a segment Ω:

$$\tilde{n}_z \partial_u \tilde{z} + n_x = 0 \qquad \text{and} \qquad \tilde{n}_z \partial_v \tilde{z} + n_y = 0 \tag{3}$$

where $\tilde{n}_z = n_x(u - c_u) + n_y(v - c_v) + n_z f$. The depth values within the segment Ω can therefore be obtained via minimising the following functional:

$$\min_z \iint_\Omega (\tilde{n}_z \partial_u \tilde{z} + n_x)^2 + (\tilde{n}_z \partial_u \tilde{z} + n_y)^2 \; dudv \tag{4}$$

Note that this leads to *a family of solutions* $\tilde{z} + Const$ which differ by a shift constant. This is due to the fact that only partial derivatives of $\tilde{z}$ are used in the functional. That means that the actual depth will be estimated up to scale, since $z = \exp(\tilde{z})$ by definition.

This ambiguity prompted us to introduce the notion of depth scales, to allow each segment to be adjusted towards its true depth. We implemented batched normal integration, efficiently solving all optimisation problems as a single sparse linear system using conjugate gradient method [18]. This implementation enables the integration of approximately 100 to 300 segments within a total time frame of $\sim$100ms.

3.3. Primitive-based Image Alignment

Now we explain how our new representation can be used as a building block for dense multi-view geometry and pose estimation, combining an optimisation-based mindset with learned single view priors. To make optimisation computationally tractable, we design our image alignment to *fully abstract away any knowledge about the neural networks involved in the image primification stage*. Hence, even though there has been early evidence that SAM is independently capable of establishing (coarse) mask correspondences across neighbouring frames, we rely on photometric information only during the image alignment stage. We believe that more semantic primitive association would be "heavier" and coarser since it would not provide per pixel correspondences, and would therefore *not be a good fit to be used in a multi-view optimisation loop*. However, we speculate that such a method could be employed to further enhance our system, e.g. to improve occlusion handling, particularly as these models become more computationally affordable.

3.3.1 Two-view SfM on SuperPrimitives

For the sake of clarity, we formulate our method in the simplest case of two frames observing the scene, although it could be trivially extended into a setting with a higher number of views. At the core of our Structure-from-Motion (SfM) approach lies a *per-primtive photometric alignment*, which contrasts with widely accepted per-pixel photometric alignment [14, 30] techniques. Informally speaking, our proposition is to treat each image primitive as a "rigid" piece, which is only allowed to be scaled up or down. This degree of freedom is due to the scale ambiguity present for each primitive. Thus, instead of estimating a depth value per pixel, we only estimate a depth-scale per primitive, which is illustrated in Fig. 1. This greatly reduces the dimensionality of the optimisation problem, especially in the case of an unknown relative pose. Note that our method does not require the target image to be primified nor does it not require any pre-established correspondences (e.g. primitive to primitive).

We assume a primified reference image I_{ref} with its set of primitives $\mathcal{P}(\mathrm{I}) = \{(\Omega_i, \mathfrak{D}_i)\}$. Given an *unposed* target image I_{t}, our goal is to *jointly* estimate its relative pose T_{tr} with respect to the reference image I_{ref} as well as the dense geometry $\mathcal{G}_{\text{ref}} = \bigcup_i \pi^{-1}(s_i \cdot \mathfrak{D}_i)$ of the reference frame.

While unscaled depths $\mathfrak{D}_i$ are given by a per-image preprocessor after the primification of the reference image, the set of depth scales s_i are yet to be estimated. In our case, we jointly estimate both depth scales s_i and a relative image pose T_{tr} by solving a photometric consistency optimisation problem. First, we warp each depth-scaled primitive (s_i, P_i) from the reference frame I_{ref} into the target frame I_{t}:

$$\hat{P}_i[\boldsymbol{u}] = \pi\left(\mathrm{T}_{tr}\pi^{-1}\left(\boldsymbol{u}, s_i \cdot \mathfrak{D}_i\right)\right) \tag{5}$$

Then a per-segment photometric residual for the primitive $P_i = (\Omega_i, \mathfrak{D}_i)$ is defined by averaging all photometric reprojection ℓ^1 errors across every pixel $\boldsymbol{u} \in \Omega_i$:

$$r(P_i, s_i, \mathrm{I}_{\text{t}}, \mathrm{T}_{tr}) = \frac{1}{|\Omega_i|} \sum_{\boldsymbol{u} \in \Omega_i} \left\| \mathrm{I}_{\text{ref}}(\boldsymbol{u}) - \mathrm{I}_{\text{t}}(\hat{P}_i[\boldsymbol{u}]) \right\|_1 \tag{6}$$

The resulting photometric cost, aggregated across all depth-scaled primitives $(s_i, P_i) \in \mathcal{P}(\mathrm{I})$, is:

$$E_{\text{photo}} = \frac{1}{|\mathcal{P}(\mathrm{I})|} \sum_{P_i = (\Omega_i, \mathfrak{D})} r(P_i, s_i, \mathrm{I}_{\text{t}}, \mathrm{T}_{tr}) \tag{7}$$

Note that we abstain from explicit occlusion handling in our photometric alignment, as our per-segment alignment is by design robust to pixel-level occlusion. We, however, expect our system could further be refined with explicit primitive correspondence checks.

Finally, to obtain the relative pose and depth scales, we minimise the photometric cost using the Adam [20] optimiser:

$$\{s_i^{\text{opt}}, \mathrm{T}_{tr}^{\text{opt}}\} = \operatorname{argmin}_{s_i, \mathrm{T}_{tr}} \ E_{\text{photo}} \tag{8}$$

Our approach trivially extends to multiple reference and target views via photometric cost summation, and we can jointly solve all depth scales and poses using the resulting cost function.

3.3.2 Monocular Visual Odometry

To demonstrate that our representation is suitable for *simultaneous geometry and pose estimation*, we design a novel monocular visual odometry system which operates directly on primitives. Informally, we incrementally build a local 3D map out of primitives and then track new incoming frames against this map, also in a per-primitive manner.

Visual Odometry (VO) consist of three components: initialisation, tracking, and mapping. Below, we cover each of these components. We interleave tracking and mapping in a single thread.

We adopt a keyframe-based approach [22], which means that the 3D map is represented by a set of posed keyframes $\{\mathrm{I}_{\text{kf}}^i, \mathrm{T}_{\text{kf}}^i, \mathcal{G}^i\}$. Note that each keyframe also has its estimated dense geometry $\mathcal{G}^i$, which results in a three-dimensional local map of the scene. We map in a sliding window fashion, with a window size of 5 keyframes. When the window is full, the earliest keyframe is popped from the window.

Initialisation. Initialisation of a VO system typically involves estimating the geometry of the first keyframe and its pose with respect to the subsequent few frames. Initialisation of a monocular incremental SfM or SLAM system is often hard, because one has to solve a chicken-and-egg problem: poses are required to reliably estimate geometry and vice versa. Therefore many VO / SLAM systems employ a wide range of heuristics in order to initialise their system. However, using our new representation no special treatment is required in this case. In order to bootstrap our system, we simply create the two first keyframes and then employ the Structure-from-Motion method proposed in Sec. 3.3.1. We keep the pose of the first keyframe fixed to fix gauge freedom.

Tracking. Given a map, the goal of tracking is to estimate the pose $\mathrm{T}_{track} \in \mathrm{SE}(3)$ of a new incoming frame I_{track}. We adapt Lucas-Kanade [28] tracking onto our per-primitive formulation. At each step we track a new incoming frame against the latest keyframe in the odometry sliding window.

We solve the same photometric cost formulated in Eq. (8), but for the pose only. The latest keyframe serves as the reference frame I_{ref} whereas the new incoming image is set to be the target frame I_{t}. The depth scales of the latest keyframe have already been estimated at the mapping stage.

Since target images are not assumed to be primified, our tracking is essentially equivalent to the classical Lucas–Kanade method and hence could be implemented efficiently on modern hardware. Our proof-of-concept implementation does tracking at 2–3 FPS, but we expect high performance gains from employing standard machinery, such as Gauss-Newton optimisation [24].

Mapping. The mapping stage ensures geometric consistency within the keyframe sliding window of size n. Concretely, our mapping refines depth scales $\{s_j\}$ and a pose T^i_{kf} for each keyframe I^i_{kf} where $0 \leq i \leq n$.

For each keyframe, we define a set of target frames against which the photometric cost is being optimised with a *connectivity function* $\mathcal{M}(t) = \{\mathrm{I}_s, \mathrm{T}_s\}$. In our implementation, $\mathcal{M}(t)$ includes temporally neighbouring keyframes $\mathrm{I}^{t-1}_{\text{kf}}$ and $\mathrm{I}^{t+1}_{\text{kf}}$ (if they exist) used as supporting frames, as well as 4 additional supplementary views, for which only the pose is being estimated. We employ these views to better constrain the geometry and observed that increasing the number of supplementary views does not induce high computational cost.

Thus, the mapping stage is done by solving a joint photometric cost for all keyframes:

$$E_{\text{mapping}} = \sum_{t=0}^{n} \sum_{\mathrm{I}_s \in \mathcal{M}(t)} \sum_{P_i \in \mathcal{P}(\mathrm{I}^t_{kf})} r(P_i, s_i, \mathrm{I}_s, \mathrm{T}_s^{-1}\mathrm{T}_{kf}) \tag{9}$$

4. Experiments

4.1. Sparse Depth Completion

Our SuperPrimitives representation could be seamlessly applied to depth completion with no pre-training required, thereby solving it in a zero-shot manner. For each primitive $P_i \in \mathcal{P}(\mathrm{I})$ we adjust its depth scale to minimise depth discrepancy $\|s_i \cdot \mathfrak{D}_i - \hat{D}\|$ with given ground truth sparse depth $\hat{D}$ across all valid depth points within the segment Ω_i. If a segment lacks valid depth measurements, it is discarded. For image regions not covered by any valid primitives, we generate a dense depth prediction by simply bilinearly interpolating depth values.

Many existing depth completion studies evaluate their methods in artificial scenarios, such as selecting input depth points randomly from a known ground truth. In contrast, our approach is tested using the real-world VOID benchmark [57]. The benchmark provides video sequences captured with a RealSense D435i camera together with sparse metric depth measurements acquired from an external visual-inertial SfM system. This setup exhibits noise and biases that might be present in the sparse depth inputs. These measurements are already provided by the dataset itself and are therefore shared across all the methods being compared.

We compare our results with a recent state-of-the-art model [55], which significantly outperforms previous depth completion methods, particularly in zero-shot generalisation contexts. Other methods evaluated in [55] are also reported in the table. Our depth completion method also operates in a zero-shot manner both task-wise and dataset-wise. That means our method was not trained for the depth completion task and neither of our surface normal prediction network nor segmentation model were trained on the VOID dataset.

We focus on the most challenging “150 points” density setting, characterised by minimal sparse depth points per image. The depth is estimated in full 480×640 image resolution, following [55].

In Tab. 1 our method quantitatively performs on par with a recent state-of-the-art method, which is enabled by a monocular depth predictor pre-trained on a vast mixture of datasets [37, 38] and fine-tined on the VOID train set. In the zero-shot setting, we outperform VI-Depth [55], which uses a DPT-Hybrid pre-trained backbone, on three out of four metrics. Note that no training was done the for depth completion task for our method. Compared to the ground truth depth maps obtained via a noisy sensor, our predictions show (Fig. 3) sharper object boundaries and are better at preserving straight lines and perpendicular structures.

4.2. Few-View Structure-from-Motion

Contrasted with depth completion, where global geometry structure is roughly given, we also test our method on few view Structure-from-Motion. Given a set of *unposed* images $\{\mathrm{I}_{\text{ref}}, \mathrm{I}^0_s, \dots, \mathrm{I}^{(n-1)}_s\}$ captured within a small timeframe, the goal is to estimate the depth of the reference image I_{ref}. In that setup, the set of unposed supplementary views $\{\mathrm{I}^0_s, \dots, \mathrm{I}^{(n-1)}_s\}$ will be leveraged for multi-view geometry estimation.

We choose the test set of the ScanNet dataset for this evaluation. In each test sequence we select every 200-th

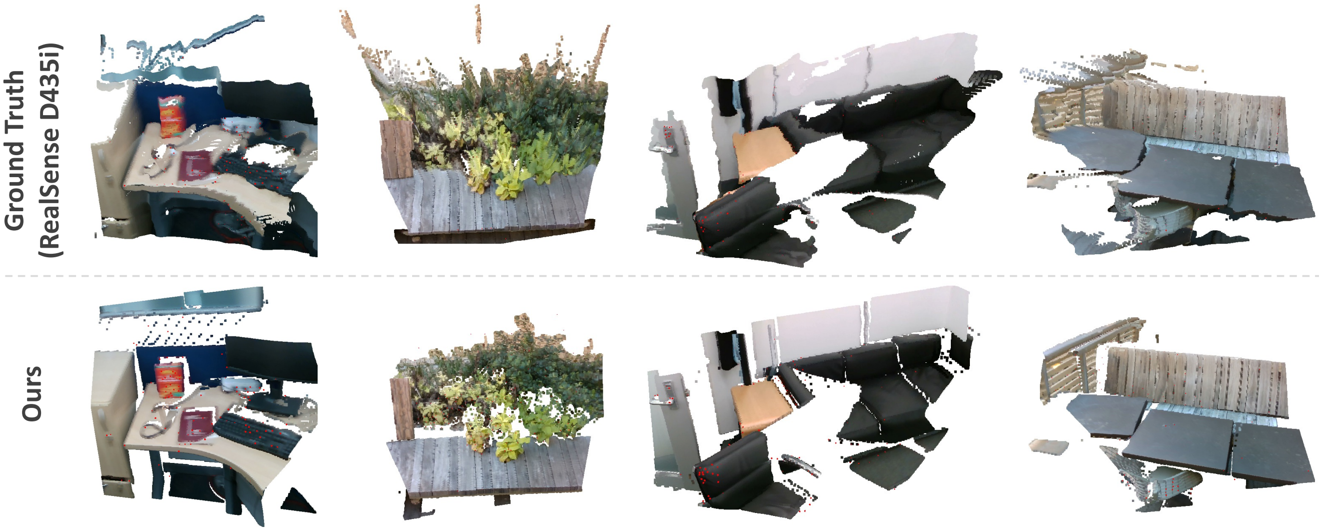

Figure 3. **Depth Completion on VOID.** We visualise the coloured unprojections of ground truth depth maps provided by a sensor (top row) and the geometry estimated by our method (bottom row). Sparse depth input points are visualised as red dots (electronic zoom-in recommended). Qualitatively, we achieve sharper geometry estimates than from a commodity depth sensor.

Method	MAE	RMSE	iMAE	iRMSE
VOICED [57]	174.04	253.14	87.39	126.3
NLSPN [34]	-	-	143	238.10
ScaffNet [58]	150.65	255.08	80.79	133.33
KBNet [56]	131.54	263.54	66.84	128.33
MonDi [27]	104.96	225.60	48.44	96.79
VI-Depth (DPT-Hybrid + NYUv2)	-	-	55.90	85.20
Ours	109.0	204.15	47.32	83.40
VI-Depth* (MiDaS) [55]	113.27	193.38	53.86	84.82
VI-Depth* (DPT-Hybrid) [55]	**97.03**	**167.82**	**46.62**	**74.67**

Table 1. **Depth Completion on VOID.** We report four most widely used metrics for depth completion on the VOID benchmark. The methods which did not use the VOID dataset for training are in the top section. Methods trained on VOID are marked with an asterisk*. Our method demonstrates superior performance on three out of the four metrics within the zero-shot group. It is second-best on two out of the four metrics overall.

frame to be the reference frame I_{ref}. Supporting frames are gathered from the neighbouring frames. Then we discard the frame sets with not enough motion to remove mostly static video clips, where SfM could not be performed. This resulted in ~500 reference frames. Note that in these experiments we use a surface normal neural network pre-trained only on the synthetic HyperSim dataset [40].

Since the multi-view depth estimation problem has scale ambiguity, we employ median-scaling to align estimated depth to the metric scale for evaluation. We report iMAE and iRMSE in Fig. 6 for our method with varying number of supplementary views. We compare our depth estimation quality against the method closest to ours, DeepV2D [49], which can also estimate depth together with supplementary frames poses. We demonstrate that our geometry quickly saturates after observing as little as 2 supporting views. Unlike DeepV2D we do not use any external tracking or initial relative pose estimation, yet still our method demonstrates consistent improvement over DeepV2D. Additionally, our method does not have any global prior on relative object positions, unlike DeepV2D. Our approach therefore could be used as an VO / SLAM initialisation mechanism, estimating joint relative poses and geometry.

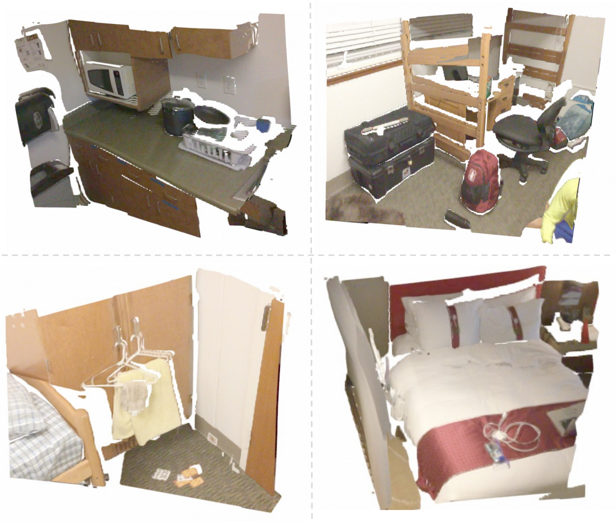

Figure 4. **3-View SfM on ScanNet.** We provide the visualisations of unprojected reference frame depth maps predicted by our method for few-view SfM using one reference and 2 supplementary views. Note that we used surface normal prediction network which was only pretrained on HyperSim [40] for this experiment.

4.3. Monocular Visual Odometry

Monocular Visual Odometry requires both accurate pose and geometry estimation to successfully track camera motion across long trajectories. Minor pose estimation inaccuracies can accumulate over time resulting in what is called

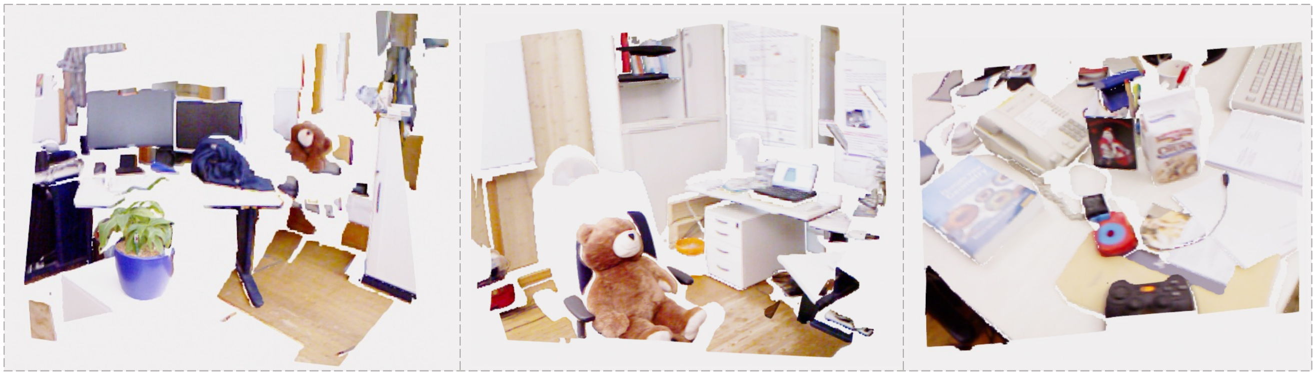

Figure 5. **TUM Reconstruction Results.** Examples of reconstructions produced by our monocular VO system on the TUM dataset. Each image shows a coloured point cloud of the geometry estimated on an odometry keyframe.

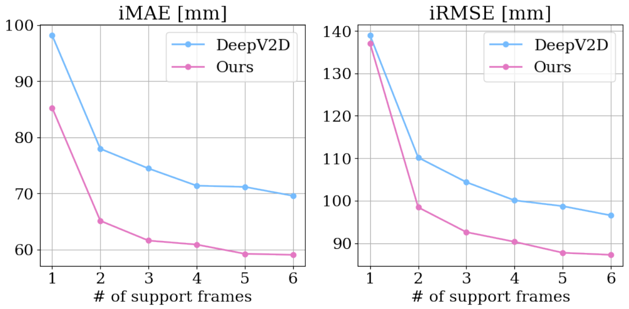

Figure 6. **Few-View SfM Depth Estimation Quality.** We evaluate the quality of our depth estimation method from an unstructured set of images on the ScanNet dataset. On the x-axis the number of supporting views is shown, while on the y-axis the corresponding depth reconstruction quality metric is reported. We show that the quality of our depth estimation quickly saturates and consistently outperforms its closest competitor, DeepV2D.

drift. With geometry being incorrectly estimated, accurate pose tracking becomes impossible, and vice versa. Our SuperPrimitive representation allows estimating both pose and geometry, enabling us to build a simple monocular VO system which performs better even in hard conditions.

We evaulate our monocular odometry on the TUM RGB-D [46] dataset. The dataset was captured with a handheld camera in indoor scenes. It is renowned for being incredibly challenging (especially for dense reconstruction systems), due to motion blur, rolling shutter artefacts, and abundance of pure rotational motion. We show that thanks to the strong priors encapsulated in our SuperPrimitives, our simple monocular odometry system can handle the TUM dataset without any special treatment (e.g. our method does not involve any special motion blur handling).

We compare against other VO systems that do not have global bundle adjustment, following the protocol of [10] and evaluate on 8 sequences from the Frieburg 1 split. We use only RGB images as the input to our system and downsample them to 120×160 for efficiency purposes. Since our method is purely monocular, the estimated trajectory lacks global scale and we first use $\mathrm{Sim}(3)$ alignment to the ground truth scale, following standard practice [4]. Tab. 2 shows that, despite the simplicity of our odometry system, it outperforms all other methods in terms of Average Trajectory Error (ATE) [46], averaged across all trajectories. Additionally, our VO is either the best or second best on five out of eight sequences. Besides quantitative evaluation, we also demonstrate reconstruction results in Fig. 5.

Seq.	TartanVO [53]	DeepV2D [49]	DeepFactors [8]	DepthCov [10]	Ours
360	0.178	0.243	0.159	**0.128**	0.173
desk	0.125	0.166	0.170	**0.056**	0.085
desk2	0.122	0.379	0.253	**0.048**	0.108
plant	0.297	0.203	0.305	0.261	**0.153**
room	0.333	**0.246**	0.364	0.257	0.363
rpy	0.049	0.105	**0.043**	0.052	0.055
teddy	0.339	0.316	0.601	0.475	**0.253**
xyz	0.062	0.064	**0.035**	0.056	0.036
mean	0.188	0.215	0.241	0.167	**0.153**

Table 2. **Trajectory Estimation Error on TUM.** Average Trajectory Error (ATE) is compared against other monocular odometry systems on the TUM Frieburg 1 split. The best and second best results are highlighted in **bold** and underscored correspondingly. Our method outperforms others in terms of the ATE averaged across all trajectories.

5. Conclusion

We presented a new representation, SuperPrimitive, which demonstrates how recent advances in building strong single image priors could be incorporated into pose and dense geometry estimation problems. We show that incorporating these priors alleviates the need for sophisticated handcrafted heuristics and paves the way into monocular reconstruction with relative ease.

Acknowledgments. Research presented in this paper was supported by Dyson Technology Ltd. The authors would like to thank Hidenobu Matsuki, Eric Dexheimer, Riku Murai, and other members of the Dyson Robotics Lab for countless fruitful discussions.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[2] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison. CodeSLAM — learning a compact, optimisable representation for dense visual SLAM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 12

[3] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Proc. Neural Information Processing Systems (NeurIPS)*, 2021. 2

[4] Carlos Campos, Richard Elvira, Juan J. Gomez, Jose M. M. Montiel, and Juan D. Tardos. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 2021. 8

[5] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[6] Alejo Concha and Javier Civera. Using superpixels in monocular SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014. 2

[7] G. Cross and A. Zisserman. Quadric Reconstruction from Dual-Space Geometry. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1998. 2

[8] J. Czarnowski, T. Laidlow, R. Clark, and A. J. Davison. Deepfactors: Real-time probabilistic dense monocular SLAM. In *IEEE Robotics and Automation Letters*, 2020. 8, 12

[9] A. J. Davison, N. D. Molton, I. Reid, and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2007. 2

[10] Eric Dexheimer and Andrew J. Davison. Learning a depth covariance function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems (NeurIPS)*, 2014. 1

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Neural Information Processing Systems (NeurIPS)*, 2014. 2

[13] Jakob Engel, Thomas Schoeps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[14] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017. 3, 5

[15] Jiahui Fu, Yilun Du, Kurran Singh, Joshua B Tenenbaum, and John J Leonard. Neuse: Neural se (3)-equivariant embedding for consistent spatial understanding with objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 2

[16] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards Internet-scale multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[17] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[18] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 1952. 4

[19] M Kaess. Simultaneous localization and mapping with infinite planes. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015. 2

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 4

[22] G. Klein and D. W. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007. 2, 5

[23] T. Laidlow and A. J. Davison. Simultaneous localisation and mapping with quadric surfaces. In *International Conference on 3D Vision (3DV)*, 2022. 2

[24] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research (IJRR)*, 2014. 6

[25] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] Tianlin Liu, Parth T. Agrawal, Allison Chen, Byung-Woo Hong, and A. Wong. Monitored distillation for positive congruent depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 7

[28] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1981. 6

[29] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger. Fusion++:volumetric object-level slam. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018. 2

[30] R. A. Newcombe, S. Lovegrove, and A. J. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1, 2, 5

[31] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 2019. 2

[32] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[33] Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 13

[34] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*, 2019. 13

[36] Albert Pumarola, Alexander Vakhitov, Antonio Agudo, Alberto Sanfeliu, and Francese Moreno-Noguer. Pl-slam: Real-time monocular visual slam with points and lines. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 2

[37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 6, 12

[38] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2, 6

[39] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003. 3

[40] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[41] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[42] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[43] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[44] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998. 12

[45] Jingjia Shi, Shuaifeng Zhi, and Kai Xu. Planerectr: Unified query learning for 3d plane recovery from a single view. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2

[46] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2012. 8, 13

[47] E. Sucar, K. Wada, and A. J. Davison. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2

[48] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[49] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 7, 8

[50] Alexander Vakhitov, Jan Funke, and Francesc Moreno-Noguer. Accurate and linear time pose estimation from points and lines. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[51] Tom van Dijk and Guido C.H.E. de Croon. How do neural networks see depth in single images? In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[52] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[53] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. TartanVO: A Generalizable Learning-based VO. In *Conference on Robot Learning (CoRL)*, 2020. 8

[54] C. S. Weerasekera, Y. Latif, R. Garg, and I. Reid. Dense monocular reconstruction using surface normals. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 2

[55] Wofk, Diana and Ranftl, René and Müller, Matthias and Koltun, Vladlen. Monocular Visual-Inertial Depth Estimation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 6, 7, 12

[56] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[57] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 2020. 6, 7

[58] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 2021. 7

[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *Proceedings of SIGGRAPH*, 2018. 12