TSDFFilter: Content-Aware Communication Planning for Remote 3D Reconstruction

Xu-Qiang Hu
Tsinghua University
huxq@outlook.com

Yu-Ping Wang
The Beijing Institute of Technology
wyp_cs@bit.edu.cn

Zi-Xin Zou
Tsinghua University
zouzx19@mails.tsinghua.edu.cn

Dinesh Manocha
University of Maryland
dmanocha@gmail.com

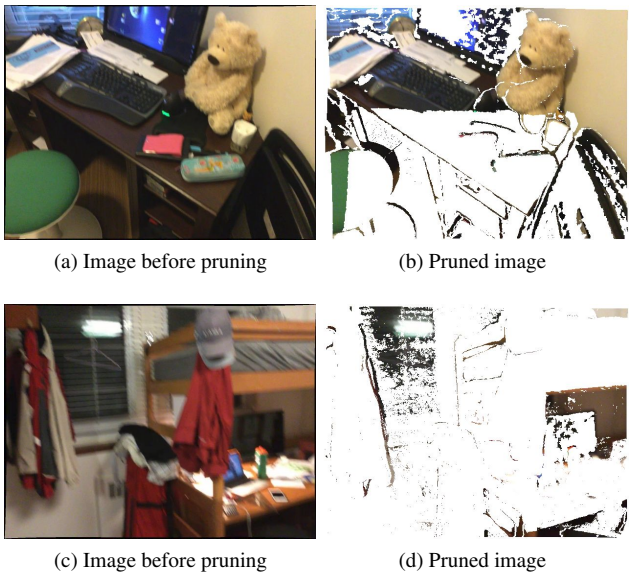
Abstract

We present a novel solution, TSDFFilter, for remote 3D reconstruction to relieve the high bandwidth requirement problem. Our approach is designed for scenarios where agents are used to collect data using an RGB-D camera and then transmit the information over the regular network to a high-performance server, where a global, dense, and volumetric model of a real-world scene is reconstructed. Our approach uses a content-aware communication planning framework in which agents can prune the gathered RGB-D information according to the transmission policy generated by the server. To generate the transmission policy, we introduce a confidence value to estimate how much each RGB-D pixel contributes to the reconstruction quality, and present an algorithm to find the confidence value. As a result, agents can transmit less RGB-D information without blindly compromising the reconstruction quality as the key-frame method and down-sampling method do. We implement our TSDFFilter framework to achieve real-time agent-assisted 3D reconstruction. Extensive evaluations show that comparing with the key-frame and down-sampling methods, our TSDFFilter framework can reduce the bandwidth requirement by up to 36% with similar reconstruction Chamfer distance, and reduce the reconstruction Chamfer distance by up to 78% with similar bandwidth requirement.

Keywords: Communication Planning, Remote 3D Reconstruction, TSDF, Transmission Policy.

1. Introduction

Reconstructing dense, volumetric models of real-world 3D scenes is an important research topic in Visual Media [4,36,42,43]. With the wide usage of consumer RGB-D cameras, gathering visual information for 3D reconstruction has become affordable, but data collection is still time-consuming. Thus, offline 3D reconstruction, where reconstruction is done after data collection, is time-consuming. With the widely usage of RGB-D cameras on mobile devices, and the development of robotics and drones, agents are being used to automatically collect RGB-D data at remote locations [6, 11]. Such a 3D scene reconstruction framework is also called multi-agent collaborative dense scene reconstruction, or remote 3D reconstruction for short.



(a) Image before pruning (b) Pruned image

(c) Image before pruning (d) Pruned image

Figure 1: Examples of how the pruning works. (a) and (c) are RGB images before pruning. (b) and (d) are the pruned RGB images (white pixels are pruned), and the required bandwidth can be reduced about 50% and 90% respectively. For (a) and (b), the camera is moving up. For (c) and (d), the camera is moving to the right.

Under the remote 3D reconstruction framework, each

agent is responsible for capturing part of the scene. For the reason of costs and energy, agents are not equipped with powerful GPUs which are usually needed for real-time reconstruction. Therefore, agents transmit the gathered RGB-D information to a high-performance server on which the global 3D model is reconstructed. However, this framework raises the requirements for high network bandwidth. In [11], the authors stated that the network bandwidth required to transmit the RGB-D images captured by a Kinect (640×480) is roughly 250Mbps, or 25Mbps if the same sequence is compressed, but their WiFi router can only provide a bandwidth of about 8Mbps. The situation was even worse when there were multiple RGB-D cameras, since they competed for the bandwidth of the WiFi router and the server's network card. Due to this issue, in practice, offline 3D reconstruction is usually preferred than remote 3D reconstruction in order to obtain more precise 3D model. Therefore, reducing the bandwidth requirements while retaining precision is essential for remote 3D reconstruction to scale to more agents and higher-resolution cameras.

Similar bandwidth problems also arise in remote simultaneous localization and mapping (SLAM) systems [9, 24, 33, 34]. These systems also gather and transmit high-resolution visual information to build a map and achieve localization. However, the main purpose of SLAM algorithms is localization, and building a sparse map of the scene is sufficient for accurate localization [8]. Therefore, not all collected points are valuable to SLAM algorithms. Transmitting only the feature points is sufficient and can relieve the bandwidth problem for remote SLAM systems [29]. However, this kind of solution cannot be used for remote 3D reconstruction frameworks.

In 3D reconstruction algorithms, all collected points are potentially valuable. They improve the reconstruction quality by either providing new information or reducing errors [27]. To relieve the bandwidth problem, remote 3D reconstruction systems currently have only two options: selecting key-frames or down-sampling the gathered images [6, 11]. Selecting key-frames results in the loss of some information provided only by the dropped frames and down-sampling the gathered images causes the details to be ignored. Reducing bandwidth requirements without compromising reconstruction quality is still a challenging problem for remote 3D reconstruction frameworks.

Main Results: In this paper, we present a communication planning algorithm for remote 3D reconstruction, TSDFFilter. Instead of totally dropping some of the RGB-D frames, our idea is to drop some RGB-D pixels. We introduce a ***confidence*** value for each RGB-D pixel and theoretically show that it can represent how much the RGB-D pixel contributes to the reconstruction quality. Based on the confidence value, the server can generate a transmission policy to the agents. Further, based on the transmission policy, the agents then transmit only the pixels that contribute more to the reconstruction quality, and thus the bandwidth requirement is reduced. We test our TSDFFilter framework to achieve real-time 3D scene reconstruction. Experimental results show that comparing with the key-frame and down-sampling methods, our TSDFFilter framework can reduce the bandwidth requirement by up to 36% with similar reconstruction Chamfer distance, and reduce the reconstruction Chamfer distance by up to 78% with similar bandwidth requirement. The main contributions of this paper include:

(1) We present a communication planning framework for remote 3D reconstruction, TSDFFilter, which can reduce the bandwidth requirement while retaining more useful details.

(2) We present the confidence value for each RGB-D pixel to estimate how much it contributes to the reconstruction quality, and an efficient algorithm to generate the confidence value.

(3) We apply our TSDFFilter framework to practical TSDF-based reconstruction system, InfiniTAM [17, 18], and implement it based on ROS [32], which is the de-facto robotic middleware.

(4) Our TSDFFilter framework is extensively evaluated with data from three datasets (the Scannet RGB-D dataset [3], the TUM RGB-D dataset [38] and the Cow & Lady RGB-D dataset [28] and show significant results. When the resulting reconstruction Chamfer distance is similar, our TSDFFilter framework can reduce the bandwidth requirement by up to 36% comparing with key-frame and down-sampling methods. When the bandwidth is similar, our TSDFFilter framework can reduce the reconstruction Chamfer distance by up to 78% comparing with key-frame and down-sampling methods.

2. Related Work

2.1. 3D Reconstruction

3D reconstruction is a common research interest in Computer Vision, Robotics, and Multimedia. Since the emergence of commodity RGB-D sensors (e.g., Microsoft's Kinect) and modern GPU programming frameworks (e.g., NVidia's CUDA), real-time dense 3D reconstruction has become feasible on commodity hardware. KinectFusion [16, 26] was one of the first real-time volumetric reconstruction frameworks. To handle large-scale scenes, the state-of-the-art solutions [17, 27] organize voxels with a hash map. To achieve high reconstruction quality, the state-of-the-art frameworks [17, 18] employ the truncated signed distance function (TSDF) [2] as the data structure to store integrated depth images. Since the task of updating the TSDF value of each voxel is suitable for data-parallel algorithms [46], these frameworks rely on GPU to achieve real-time perfor-

mance [39]. Our work is designed for frameworks employing voxel hashing and TSDF. Prior arts [1] consider pixel confidence maps in RGB-D reconstruction. Our work uses pixel confidence maps to generate transmission policy.

2.2. Remote SLAM and 3D Reconstruction

Widely used mobile phone and commodity UAVs have provided more convenient ways to capture sensor data. Research has shown that these sensor data can be used for SLAM and 3D reconstruction. However, when these devices cannot provide the high performance required for real-time algorithms, the sensor data must be stored and processed off-line [49]. To process the sensor data in an online manner, remote SLAM [9, 24, 34, 45] and remote 3D reconstruction [6, 11] frameworks have been proposed. Such frameworks take advantage of the computing power provided by high-performance central server(s), and captured sensor data are transmitted to the server(s).

For remote SLAM frameworks, a major issue is what data representation should be transmitted. Opdenbosch et al. [29] proposed a solution where fast feature extraction is performed at the agent, and the server performs the following SLAM by collecting the features. These features are sufficient to generate a sparse 3D map, but far from sufficient to reconstruct a dense volumetric model. Therefore, this kind of solution works well for remote SLAM, but is not suitable for remote 3D reconstruction.

For remote 3D reconstruction, the only options are selecting key-frames and down-sampling every frame [6, 11]. These two solutions can reduce the bandwidth requirement significantly, but they also damage the reconstruction quality. The key-frame methods treat each image as a whole and drop some full images without distinguishing which pixels might be valuable. The down-sampling methods consider all pixels to be of equal value, which is not always true.

Our idea is to distinguish which pixels are more valuable to the global model. This is difficult, if not impossible, without knowing the status of the global map. Our solution is inspired by Giamou et al. [10]. This work aims to detect inter-robot loop closures for remote SLAM under a constrained network bandwidth. In this work, agents exchange some meta-data before actually exchanging visual information. Based on these meta-data, agents can design a policy regarding which data should be exchanged. Our solution to the remote 3D reconstruction also allows the server to give some hints to the agent, so the agent can distinguish which pixels are more valuable.

2.3. Video Streaming

Dynamic Adaptive Streaming over HTTP (DASH) [37] is the de-facto standard for video streaming. By using the DASH technique, each video is partitioned into segments and each segment is encoded in multiple bitrates. Which bitrate to use for the next segment is determined by Adaptive Bitrate (ABR) algorithm. Generally, video streaming services aim to improve the Quality of Experience (QoE) by increasing the average video bitrate, reducing rebuffering events, and/or increasing video bitrate smoothness. These factors cannot be satisfied at the same time, and existing solutions have made significant improvements in balancing these factors [7, 15, 22, 31, 35, 44, 47, 48].

Volumetric video streaming, such as point cloud-based volumetric video [12, 20] or depth image (i.e., RGBD)-based volumetric video [21, 40], is more in line with application for 3D reconstruction, where remote agents transmit captured color and depth information to the server. Some researches extend DASH technique towards volumetric video streaming [13, 41].

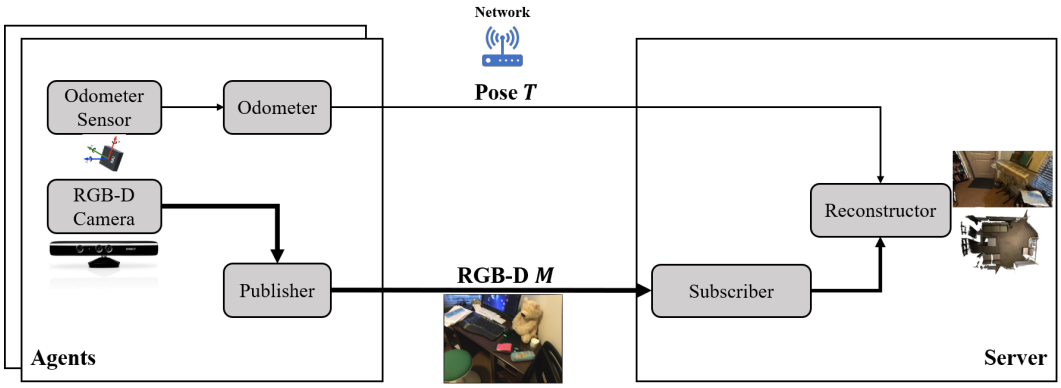
However, the DASH technique is not well suited for 3D reconstruction services for two main reasons. On the one hand, down-sampling method used in DASH when bad network situation would lead to worse 3D reconstruction quality. On the other hand, some of content in sequence are redundant which will not be contributed to improve 3D reconstruction, e.g., for those area of already being well reconstructed. Our TSDFFilter algorithm only transmits the content which is useful for 3D reconstruction.

There are also some works [19] employ video compression techniques to transmit RGB-D streams and achieve competitive results. In all experiments of our framework and the comparing frameworks, we employ the same image compression algorithms for the convenience of implementation based on ROS [32]. It is feasible to change the compression algorithms to video compression algorithms, but we consider it orthogonal to our improvements.

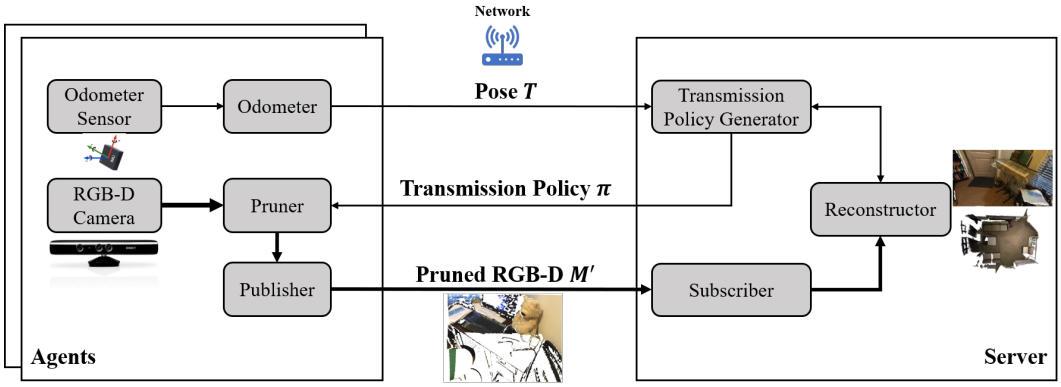
3. Our Approach

Figure 2(b) illustrates our TSDFFilter framework. In our framework, the agent transmits the pose T of the equipped camera to the server, the server generates a transmission policy $\pi(T, G)$ based on the pose T and the global status G of the server, and the agent prunes its RGB-D data from M to $M'(M, \pi)$ based on the transmission policy. In contrast, in the common framework (shown in Figure 2(a)), the agent does not know any information from the server and can only blindly transmit its poses and RGB-D data to the server. Down-sampling and selecting key-frames can only reduce the bandwidth requirement without knowing the needs of the server.

Here, we assume that each agent knows its own pose. This can be achieved separately by either low-cost SLAM algorithms [25], low-bandwidth remote SLAM frameworks [5], or other sensors (e.g., IMU) equipped on the agent. Therefore, we can focus on how to transmit the information needed by the integration routines of 3D reconstruction frameworks.



(a) Illustration of the common remote 3D reconstruction framework, where the agent blindly transmits its poses and RGB-D data to the server without knowing which data the server would prefer.



(b) Illustration of our TSDFFilter framework, where the server generates a transmission policy based on the pose and the agent can prune its RGB-D data based on the transmission policy.

Figure 2: An overview of the frameworks for remote 3D reconstruction. In our experiments, the overall bandwidth requirements are compared.

3.1. Generating Transmission Policy

The key challenge is how to express and generate a transmission policy that can conform to the following principles:

- The pruned RGB-D data and the original RGB-D data should contribute similarly to the reconstruction quality.
- The transmission policy should be generated efficiently.
- The transmission policy should help reduce the overall bandwidth.

The first principle requires us to mine the properties of the reconstruction algorithm. As we have stated, to achieve high-quality 3D reconstruction, TSDF has become a de-facto expression of high-quality 3D models. TSDF is a volumetric representation of a scene for integrating depth images. When integrating a new depth image, new voxels that have never been captured before are allocated. Then, the TSDF value $TSDF(x)$ and the weight $W(x)$ of each associated voxel x are updated with Equation (1). t is the truncation parameter, $w_i(x)$ is the weight for each round of update, which is usually set as 1, $d_i(x)$ is the depth captured by depth camera, and $z_i(x)$ is the depth of the voxel.

$$\begin{aligned} TSDF_i(x) &= \frac{TSDF_{i-1}(x) \cdot W_{i-1}(x) + tsdf_i(x) \cdot w_i(x)}{W_{i-1}(x) + w_i(x)} \\ tsdf_i(x) &= max(-1, min(1, \frac{d_i(x) - z_i(x)}{t})) \\ W_i(x) &= W_{i-1}(x) + w_i(x) \end{aligned} \tag{1}$$

For each round of updates, the captured depth $d_i(x)$ is an approximate value of the depth from the view point to the real-world surface, $s_i(x)$. We can assume that $d_i(x)$ is a random variable that follows a normal distribution with mean $s_i(x)$, and each $d_i(x)$ is independent (see Figure 3). The variance σ_i^2 reflects the error introduced by the RGB-D camera. Thus, we can deduce in Equation (2) that the

$TSDF_i(x)$ also follows a normal distribution after $W_i(x)$ rounds of update. In this normal distribution, the expectation reflects the real-world $TSDF(x)$. Note that the variance becomes smaller as the update round $W_i(x)$ increases. Specifically, when all σ_j is the same, the variance becomes $\frac{\sigma^2}{t^2 \cdot W_i(x)}$ which is inversely proportional to $W_i(x)$.

$$
\begin{aligned}
d_i(x) &\sim N(s_i(x), \sigma_i^2) \\
tsdf_i(x) &\sim N(\frac{s_i(x) - z_i(x)}{t}, \frac{\sigma_i^2}{t^2}) \\
TSDF_i(x) &\sim N(\frac{\Sigma(s_j(x) - z_j(x))}{t \cdot W_i(x)}, \frac{\Sigma \sigma_j^2}{t^2 \cdot W_i^2(x)})
\end{aligned}
\tag{2}
$$

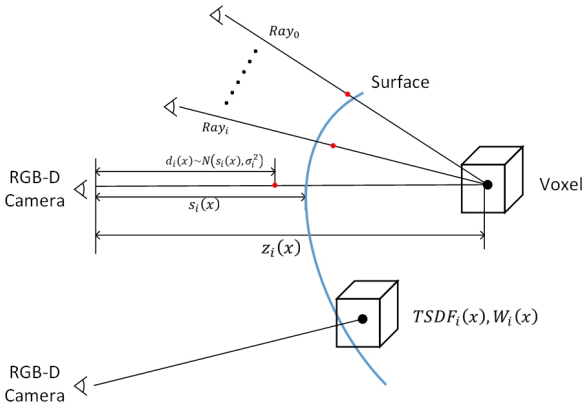


Figure 3: The depth captured by the RGB-D camera can be modeled as a normal distribution. For each voxel x, the $TSDF_i(x)$ also follows a normal distribution after $W_i(x)$ round of updates.

For the second and the third principles, our purpose is to express the transmission policy, which decides whether each RGB-D pixel needs to be transmitted. If we can find a voxel x on the ray corresponding to the RGB-D pixel with its $TSDF(x) = 0$, we can estimate the real-world $TSDF(x)$ with the normal distribution of voxel x. If we can estimate the real-world $TSDF(x)$ with enough confidence, we do not need further updates. Since smaller variance indicates more confidence, and the variance is inversely proportional to $W_i(x)$, we can set a W_{MAX} to indicate the confidence threshold. Thus, the error caused by the generated transmission policy follows a normal distribution in Equation (3), and we can expect the reconstruction Chamfer distance of the result dense model is about $\frac{\sigma^2}{t^2 \cdot W_{MAX}}$. Since σ^2 is a constant that reflects the error introduced by the RGB-D camera, and t is a constant set by the 3D reconstruction algorithm, the reconstruction Chamfer distance of the resulting dense model is expected to be inversely proportional to W_{MAX}.

$$
Error(x) \sim N(0, \frac{\sigma^2}{t^2 \cdot W_{MAX}}) \tag{3}
$$

Algorithm 1 Generate a Transmission Policy (on the server)

```
Global: model, the current 3D model
        W_MAX, the confidence threshold
Input: pose, the current pose of the robot.
Output: res, a binary image
 1 procedure gen_trans_policy(pose)
 2   for each pixel p of res
 3     loop ray-casting from p, pose
 4       x <- current voxel
 5       if (x is unexplored) then
 6         p <- 1
 7       else if (TSDF(x) > 0) then
 8         continue ray-casting
 9       else
10         interpolate a position with TSDF = 0
11         compute the average W at the interpolate position
12         if W >= W_MAX then
13           p <- 0 // this pixel need not to be transmitted
14         else
15           p <- 1
16         end if
17       end if
18     end loop
19   end for
20   return res
21 end procedure
```

Figure 4: The pseudo code of generating a transmission policy based on the current pose of the agent.

Overall, the algorithm for generating the transmission policy is shown in Figure 4. The transmission policy is expressed with a binary image with the same size of the RGB-D images. For each RGB-D pixel, we try to find a voxel x on the corresponding ray that satisfies $TSDF(x) = 0$ and $W(x) \geq W_{MAX}$. If such a voxel is found, the transmission policy decides not to transmit the RGB-D pixel or, otherwise, the transmission policy decides to still transmit the RGB-D pixel. This routine is highly parallelizable and can be accomplished quickly on the server with modern GPU, which meets the second principle. The binary image that expresses the transmission policy can be further compressed to reduce the required bandwidth. We will verify the third principle in Section 4.

3.2. Alternative Solutions

We can also theoretically analyze the error introduced by the key-frame method and the down-sample method.

For the key-frame method, let key-frame ratio $K \leq 100\%$ represent the ratio of preserved frames, e.g., $K = 25\%$ indicates that one out of every 4 frames is preserved. On average, this configuration is equivalent to the case where each frame has a probability of K to be dropped. Thus, for a voxel x that could be updated for $W(x)$ rounds, there are only $K \cdot W(x)$ rounds after key-frame selection. We showed in Equation (2), the error for a voxel x whose $TSDF(x) = 0$ is inversely proportional to $W(x)$. Key-frame method reduces every $W(x)$ by K times, and

therefore increase the error for every voxel by K^{-1} times. If $W(x)$ is large for all voxels originally, the key-frame method introduces fewer errors; however, if $W(x)$ is small, the key-frame method could introduce large errors. In extreme cases, for some voxel x whose $W(x) = 1$ originally, it could be reduced to $W(x) = 0$, which means the voxel x is completely lost. In this case, the error depends on the distance to the next nearest voxel, which is not predictable.

For the down-sampling method, let down-sampling ratio $R \leq 100\%$ represent the ratio of preserved pixels on each width and height, e.g., $R = 25\%$ indicates that a 1024×768 image will be down-sampled into a 256×192 image. On average, this configuration is equivalent to the case where each ray of each frame has a probability of R^2 to be dropped. Similar to the analysis of the key-frame method, we can expect that the down-sampling method to increase the error for every voxel by R^{-2} times. In practice, however, the down-sampled RGB-D images are up-sampled by interpolation back to the original resolution before being used to achieve 3D reconstruction. If some part of the scene is a plane, linear interpolation can accurately restore the depth of the plane. However, on other parts of the scene, interpolation introduces a large error in the depth of the observed depth $d_i(x)$.

3.3. Pruning RGB-D Data

When the agent receives the transmission policy from the server, it is used to prune the RGB-D data. Since we have represented the transmission policy with a binary image, we can simply prune each pixel of the RGB-D data when the corresponding pixel on the binary image is 0. The pseudo code of pruning the RGB-D data is shown in Figure 5.

Algorithm 2 Tailoring the RGB-D data (on the agent)

Input: ***rgb***, the RGB image.
depth, the depth image.
trans, the transmission policy (a binary image).

```
1 procedure tailor_rgbd(rgb, depth, trans)
2   for each pixel depth[i][j]
3     if trans[i][j] is 0 then
4       depth[i][j] <- 0
5       rgb[i][j] <- (0, 0, 0)
6     end if
7   end for
8 end procedure
```

Figure 5: The pseudo code of pruning the RGB-D data.

Figure 1 shows two examples of how the pruning works. Figure 1(a) shows an RGB image in the Scannet RGB-D dataset, and Figure 1(b) shows the image after our pruning routine. At run time, the camera is moving to the upper-right. At the rays corresponding to the upper-right side of the pose, there is not enough information, and the transmission policy of those pixels is 1. Therefore, we cannot drop those pixels. In contrast, the bottom-left side of the pose has already been updated enough times, and most of the pixels are pruned. Similarly, for Figure 1(c) and (d), the camera is moving to the right, and most pixels on the left are pruned.

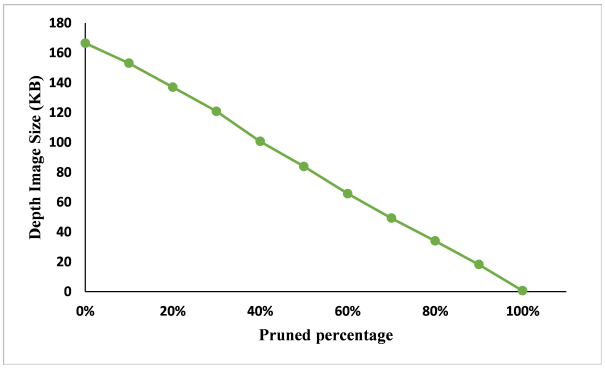


Figure 6: The effectiveness when image compression methods working on our pruned RGB-D images. As more pixels are pruned, the size of the compressed image decreases in a linear fashion.

Before the server can achieve the reconstruction task with the pruned RGB-D data, there is an issue about how to transmit pruned RGB-D data efficiently. In our pruned RGB-D images, many pixels are pruned and replaced with depth 0 and the color black (i.e., (R, G, B)=(0, 0, 0)). We find that ordinary compression methods for RGB images and depth images work well, since connected regions of the same value can be greatly compressed. In order to quantify the effectiveness when image compression methods working on our pruned RGB-D images, we conduct a study. We generate a series of images by pruning pixels from the same image, compress them with the same PNG compression level, and measure their sizes. The result is shown in Figure 6. As we can see, as more pixels are pruned, the size of the compressed image decreases in a linear fashion. The decrease in size will directly result in a reduction in bandwidth, which is exactly what our framework needs.

Finally, after the server receives the pruned RGB-D data, it can achieve the reconstruction task. Since the depths of the pruned pixels are 0, they will not contribute to the reconstruction.

4. Evaluation

To show practical results, our experiments are carried out completely online. We use three datasets, Scannet RGB-D dataset [3], TUM RGB-D SLAM dataset [38] and Cow & Lady real-world RGB-D dataset [28]. We select 6 sequences of different lengths from the three datasets, because we consider them representative to different scenarios. The detail and characteristic of these 6 sequences are shown in Table 1. Each data sequence is replayed on the agent side and transmitted to the server side for 3D reconstruction. On

the server side, we run the InfiniTAM framework [17, 18], a state-of-the-art 3D reconstruction framework. All software modules (i.e., 3D reconstruction and data sequence replaying) are connected with ROS middleware. For convenience, we use image compression algorithms employed by ROS in data transmission. The ROS version is Melodic on Ubuntu 18.04.

4.1. Comparison with Mobile Reconstruction Methods

In the remote 3D reconstruction scenario, agents transmit the collected data to the server to utilize the powerful computing resources of the server to complete the reconstruction task and generate a reconstruction model. However, we can also use the extremely limited computing resources on agents to complete the reconstruction task. We refer this kind of solutions as mobile reconstruction methods. Voxfield [30] is the state-of-the-art TSDF-based mobile reconstruction framework, which can complete reconstruction tasks without the support of high-performance GPU.

In order to show the difference between the remote reconstruction framework and the mobile reconstruction framework, we first conduct an experiment comparing the results of Voxfield and InfiniTAM. We run the experiment multiple times under varying voxel size settings. The processing time per frame is recorded to compare the execution efficiency, and the error is calculated to compare the quality of the resulting models generated by two frameworks. The error is calculated as the reconstruction Chamfer distance [23] between the generating reconstruction models and the ground-truth model provided by datasets. The Chamfer distance is calculated between the reconstructed mesh N and the ground-truth G is:

$$d_{\mathrm{CD}}\left(N, G\right)=\frac{1}{2N}\sum_{n\in N}\min_{g\in G}\|n-g\|_2^2+\frac{1}{2G}\sum_{g\in G}\min_{n\in N}\|g-n\|_2^2 \tag{4}$$

Table 2 shows the results. We can see that with the support of high-performance GPUs, the InfiniTAM method only takes about one percent of the processing time of Voxfield method to complete the reconstruction task with the same voxel size, and generate a 3D model with smaller error at the same time. Therefore, in order to obtain higher quality models and more efficient execution efficiency, we need the agent to transmit data to the server for 3D reconstruction. When the voxel size is small, it will be difficult to complete the reconstruction task only by relying on the agents' limited computing resources.

4.2. Numeric Comparison

The main purpose of our TSDFFilter framework is to reduce the bandwidth requirement while retaining more useful details. To measure detail retention, we use the *rostopic bw* tool on each ROS topic. The overall bandwidth requirement is the sum of transmitting the RGB-D data from the agent to the server and transmitting the transmission policy from the server to the agent. In order to measure detail retention capabilities, we measure the Chamfer distance between the resulting reconstruction models and the ground-truth model. The ground-truth model is achieved by running the same reconstruction algorithm in an off-line manner, because we would like to show the effect of our communication planning algorithm by showing the difference of its results with that of the off-line results.

There is a parameter in our TSDFFilter framework that can affect the result, i.e., the threshold W_{MAX}. We run the experiment multiple times with different values of W_{MAX} to show the influence of this value on the result. For comparison, we run the experiment with the key-frame method and the down-sampling method.

Since our TSDFFilter framework prunes the transmitted information based on its own features, it is expected that these results would differ when using different sequences. To make a fair and extensive comparison, we run the experiment with 6 sequences from 3 different datasets, shown in Table 1.

All of the results are organized into Figure 7. The figure shows the reconstruction errors that can be obtained by the above three methods under different bandwidth conditions. With the increase of W_{MAX}, our method can obtain smaller model error under the same bandwidth condition. Each figure includes the results of handling a different sequence in three methods, including down-sampling, key-frame and our TSDFFilter. For the key-frame method, each point has different a key-frame ratio K, resulting in different bandwidths and reconstruction Chamfer distance results; for the down-sampling method, each point has different down-sampling ratio R; for our TSDFFilter framework, each point has different threshold W_{MAX}. Each subfigure includes the normalized bandwidth requirement (the horizontal axis) and reconstruction Chamfer distance (the vertical axis). The bandwidth requirement is normalized by dividing by the bandwidth required to generate the ground-truth.

In general, with a lower key-frame ratio K, a lower down-sampling ratio R, or a lower threshold W_{MAX}, the bandwidth requirement becomes lower while the reconstruction Chamfer distance become higher. The bandwidth requirement when using the key-frame method or the down-sampling method is less affected by the content of different sequences. On the other hand, the bandwidth requirement when using our TSDFFilter framework differs when handling different sequences. The result of reconstruction Chamfer distance values verifies the analysis in Section 3 that the reconstruction Chamfer distance using our TSDFFilter is inversely proportional to W_{MAX}. In Figure 7, we

Dataset	Sequence	Characteristic	Description
Scannet	scene0709_00	The scene is a medium kitchen.	Captured with the camera moving around a kitchen.
	scene0710_00	The scene is a room and the camera aim at nearly the same direction.	Captured with the camera translating and aiming at nearly the same direction.
	scene0717_00	The scene is large and most voxels has a small W.	Captured with the camera moving in a big college dormitory.
TUM	fr1/xyz	The scene contain people and there is information overlap between the frames of this sequence.	Recorded laboratory desks and a student sitting in a chair.
	fr1/desk	The scene doesn't contain any people and there is large overlap between the frames of this sequence.	Recorded laboratory desks without people.
Cow & Lady	cow and lady	The scene is large and there are not many effective pixels measured by the depth camera.	Recorded an indoor scene with a cow, mannequin and a few other typical office accessories.

Table 1: Sequences used in experiments.

Voxel Size	Process Time (s)			Reconstruction Chamfer Distance (m)		
	Voxfield	InfiniTAM	V / I Ratio	Voxfield	InfiniTAM	V / I Ratio
0.080	0.0887	0.000597	148.7	0.0631	0.0526	120%
0.040	0.0959	0.000572	167.5	0.0341	0.0295	115%
0.020	0.1035	0.000609	169.9	0.0223	0.0188	119%
0.010	0.1379	0.000618	223.0	0.0234	0.0222	106%
0.005	0.4244	0.000636	667.5	0.0323	0.0256	126%

Table 2: Comparison between the state-of-the-art mobile 3D reconstruction framework (Voxfield with only CPU) and remote 3D reconstruction framework (InfiniTAM with GPU) under varying voxel sizes. The "V / I Ratio" column means the value of Voxfield divided by the corresponding value of InfiniTAM. With the same voxel size, the processing time of Voxfield is more than 100 times that of InfiniTAM, and the reconstruction Chamfer distance is also larger.

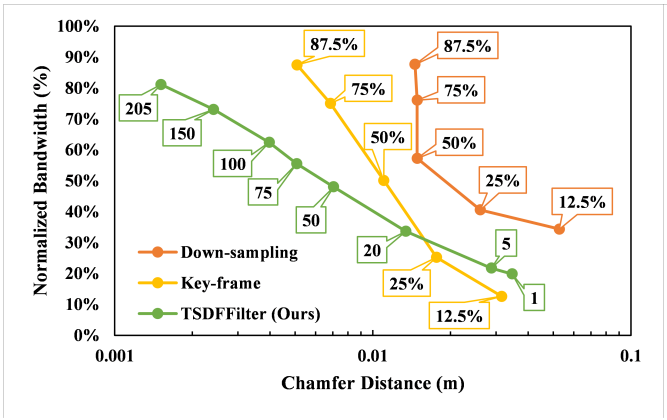
show that W_{max} values varies from 1 to 205. Specially, we show the extreme result when the threshold $W_{MAX} = 1$, which indicates a pixel will be pruned as long as it has been observed once. Another extreme result is when the threshold $W_{MAX} = 255$. This threshold is so high that our TSDFFilter framework can prune little pixels. In this case, the resulting normalized bandwidth requirement is about 80%, which also indicates that the bandwidth required for transmitting the transmission policy is low. However, the resulting reconstruction Chamfer distance is not zero because of the uncontrollable randomness. For $W_{MAX} = 50$ and 75, the results are comparable with the key-frame method and the down-sampling method. With similar bandwidth requirements, our TSDFFilter framework can achieve about 70% lower reconstruction Chamfer distance.

We should note that, for the *Cow & Lady* sequence, the bandwidth for our TSDFFilter framework is high even when $W_{MAX} = 1$. This is probably because of the characteristics of the *Cow & Lady* sequence. In this sequence, the camera moves along without turning back, and thus each frame is valuable.

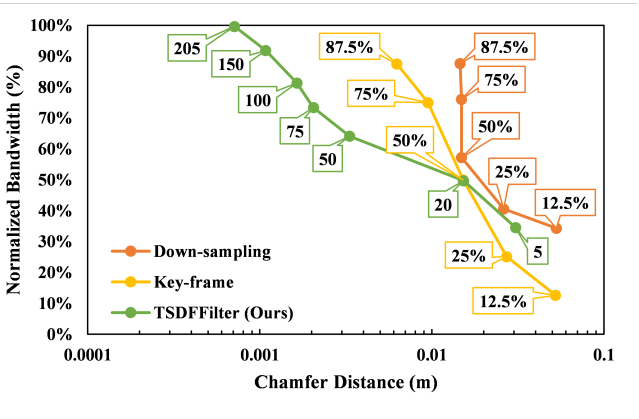
4.3. Visualized Reconstruction Results

In addition to the numeric results, we also visualize an example of the results to show the advantage of our TSDFFilter framework.

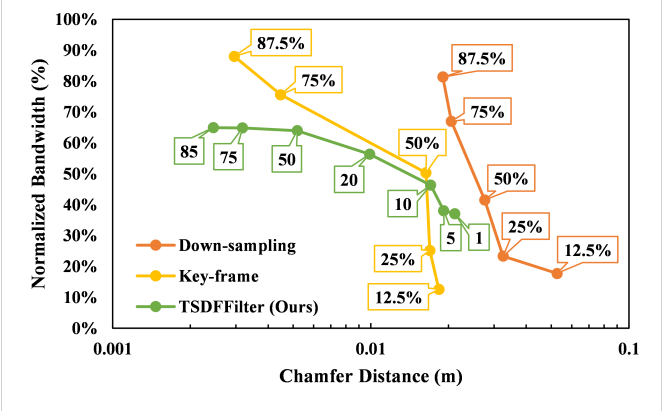
Figure 8 shows the visualized results of the Scannet scene0710_00 sequence. Figure 8(a) illustrates the reconstruction result with the key-frame method ($K = 0.75$); Figure 8(b) illustrates the reconstruction result with the



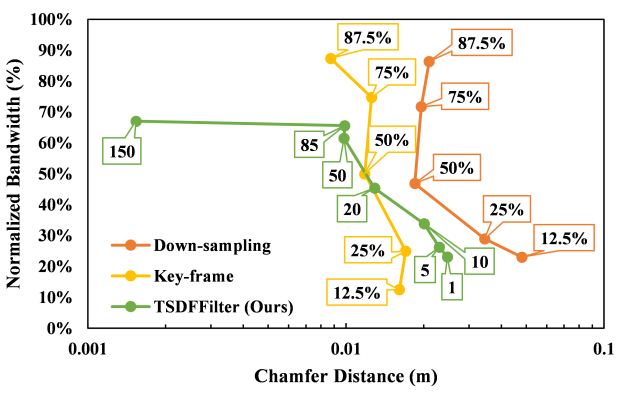
(a) TUM-fr1/xyz



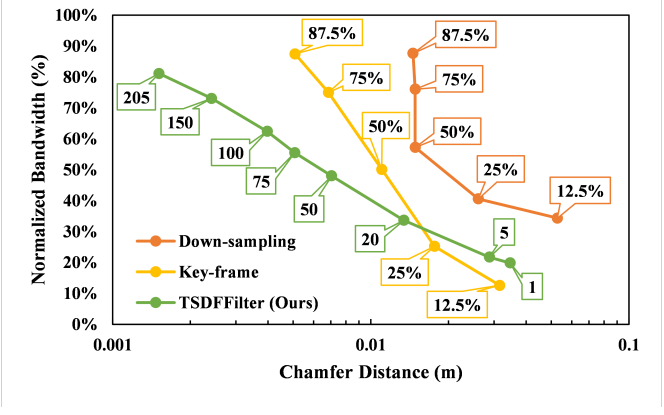
(b) TUM-fr1/desk



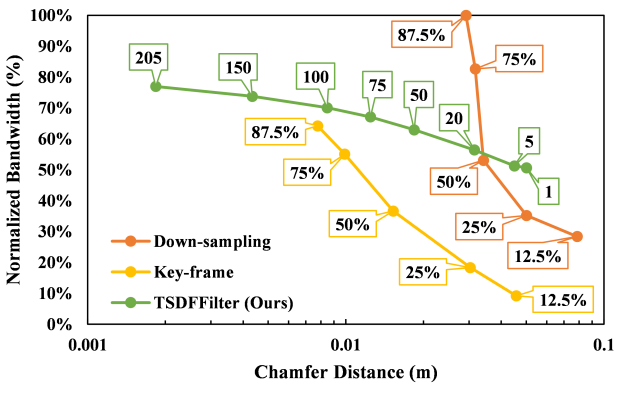
(c) Scannet-scene0709_00



(d) Scannet-scene0710_00



(e) Scannet-scene0714_00



(f) Cow & Lady

Figure 7: The numeric comparison results. Each figure includes the results of handling a different sequence with three method: down-sampling, key-frame and our TSDFFilter. The figure shows the bandwidth requirements that need with different the reconstruction errors using these three methods. With the increase of W_{MAX}, our method can work in less bandwidth requirement conditions with the same reconstruction errors.

down-sampling method ($R = 0.75$); Figure 8(c) illustrates the reconstruction result with our TSDFFilter framework ($W_{MAX} = 50$); Figure 8(d) illustrates the reconstruction result with off-line manner (i.e., the ground truth). The

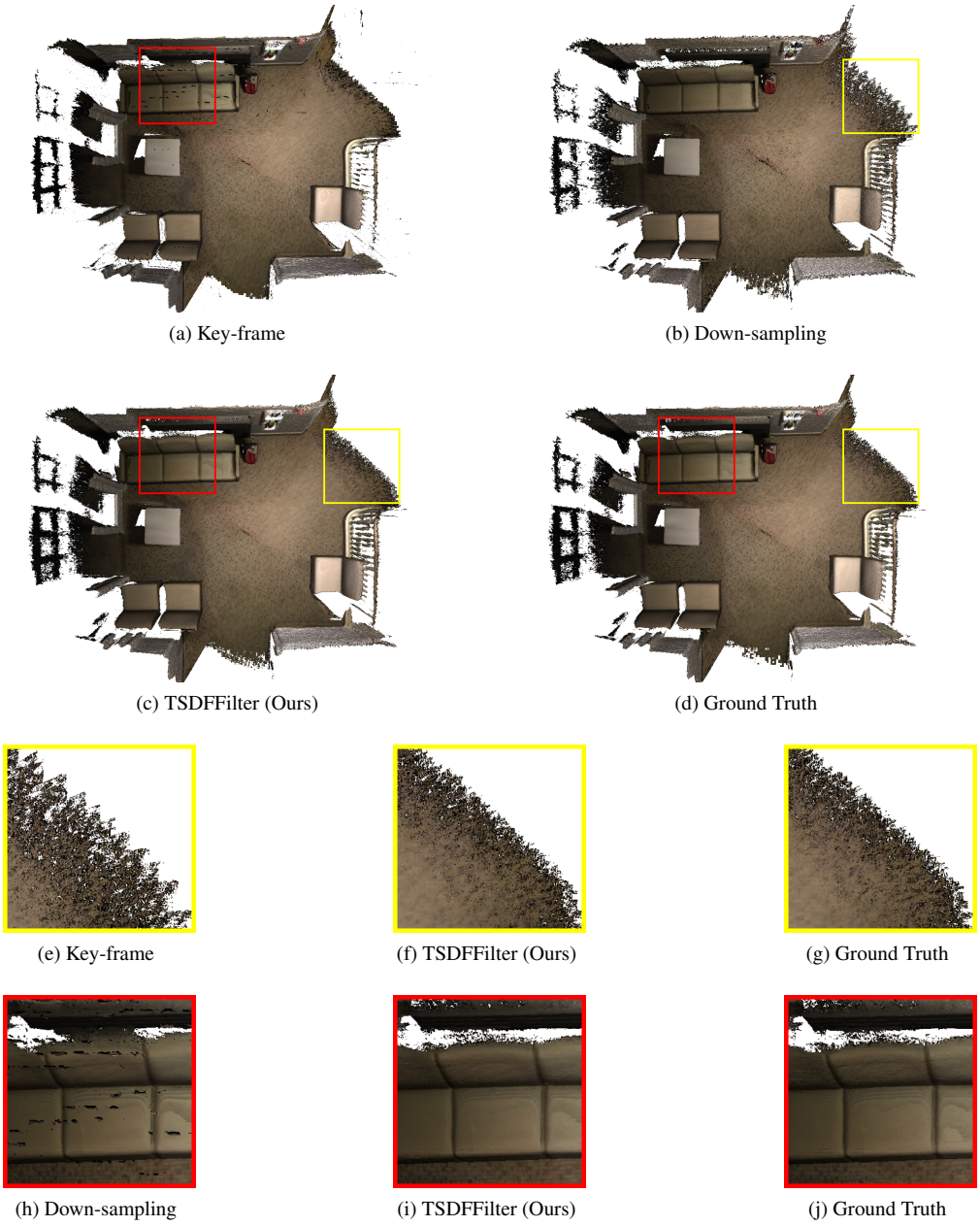
(a) Key-frame (b) Down-sampling

(c) TSDFFilter (Ours) (d) Ground Truth

Figure 8: Visualized reconstruction results of the Scannet scene0710_00 sequence. Each subfigure is generated by one of the four methods. In each subfigure, the first row illustrates higher views of the reconstructed models and the second row illustrates the detail within the yellow box and red box, which shows the detail on the door handle and sweeping robot. In the yellow boxes, we can clearly see the door handle, which indicates that our TSDFFilter framework successfully retains more details. The down-sampling method loses more detail because it has more depth values which are estimated by interpolation; the key-frame method, on the other hand, tends to smooth the result.

bandwidth requirements for these sets of results are similar. The first row of each subfigure illustrates higher views of the reconstructed models. The second row of each subfiugre illustrates the detail within the yellow and red boxes, which shows the detail on the door handle and sweeping robot. From the parts within the yellow boxes, we can clearly see the door handle, which indicates that our TSDFFilter framework successfully retains more details. Interestingly, our TSDFFilter framework successfully retains more details around the door knob (shown in the yellow boxes) than the ground truth. This is because our TSDFFilter framework stops updating voxels whose confidence value is high enough, and can probably avoid these voxels being damaged by future inaccurate scans.

Figure 9 shows the visualized results of the Scannet scene0714_00 sequence. Each subfigure illustrates the reconstruction result with one of the four methods. The first and second rows illustrate the full views of the reconstructed models. The third and fourth rows illustrate the detail within the red and yellow boxes, which shows the detail on the sofa and wall. From the parts within the yellow boxes, we can clearly see our TSDFFilter framework can obtain more points compapred to the key-frame method. From the parts within the red boxes, we can clearly see the sofa, which indicates that our TSDFFilter framework successfully retains more details and the down-sampling method introduce outliners. The down-sampling method loses more detail because it has more depth values that are estimated by inter-



(a) Key-frame (b) Down-sampling

(c) TSDFFilter (Ours) (d) Ground Truth

(e) Key-frame (f) TSDFFilter (Ours) (g) Ground Truth

(h) Down-sampling (i) TSDFFilter (Ours) (j) Ground Truth

Figure 9: Visualized reconstruction results of the Scannet scene0714_00 sequence. Each subfigure is generated by one of the four methods. The first and second rows illustrate higher views of the reconstructed models. The third and fourth rows illustrate the detail within the yellow boxes and red boxes. The key-frame method loses a lot of points,and the down-sampling method produces lots of outliers.

polation; the key-frame method, on the other hand, tends to smooth the result; and our TSDFFilter framework retains much of the detail.

5. Conclusion and Future Work

We have presented a communication planning framework for remote 3D reconstruction called TSDFFilter. In our TSDFFilter framework, agents do not blindly transmit their data but are instead able to prune their data according to the transmission policy generated by the server. To

generate the transmission policy, we present the confidence value for each RGB-D pixel to estimate how much it contributes to the reconstruction quality and an efficient algorithm to generate the confidence value. Experimental results show that our TSDFFilter framework can reduce the bandwidth requirement and overcome the disadvantages of down-sampling and key-frame methods.

As far as we know, this is the first remote 3D reconstruction framework that applies feedback from the server to guide the agents how to transmit data. Besides, our TSDFFilter framework focuses on TSDF-based reconstruction method, and cannot be directly applied on semantic-aware approaches, such as [14, 50]. But it would be an interesting future work.

References

[1] Y. Cao, L. Kobbelt, and S. Hu. Real-time high-accuracy three-dimensional reconstruction with consumer RGB-D cameras. *ACM Trans. Graph.*, 37(5):171, 2018. 3

[2] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In J. Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312. ACM, 1996. 2

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 6

[4] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3):24:1–24:18, 2017. 1

[5] X. Ding, Y. Wang, L. Tang, H. Yin, and R. Xiong. Communication constrained cloud-based long-term visual localization in real time. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*, pages 2159–2166. IEEE, 2019. 3

[6] S. Dong, K. Xu, Q. Zhou, A. Tagliasacchi, S. Xin, M. Nießner, and B. Chen. Multi-robot collaborative dense scene reconstruction. *ACM Trans. Graph.*, 38(4):84:1–84:16, 2019. 1, 2, 3

[7] Z. Duanmu, K. Ma, and Z. Wang. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1752–1760, 2017. 3

[8] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):611–625, 2018. 2

[9] C. Forster, S. Lynen, L. Kneip, and D. Scaramuzza. Collaborative monocular SLAM with multiple micro aerial vehicles. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, November 3-7, 2013*, pages 3962–3970. IEEE, 2013. 2, 3

[10] M. Giamou, K. Khosoussi, and J. P. How. Talk resource-efficiently to me: Optimal communication planning for distributed loop closure detection. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–9. IEEE, 2018. 3

[11] S. Golodetz, T. Cavallari, N. A. Lord, V. A. Prisacariu, D. W. Murray, and P. H. S. Torr. Collaborative large-scale dense 3d reconstruction with online inter-agent pose optimisation. *IEEE Trans. Vis. Comput. Graph.*, 24(11):2895–2905, 2018. 1, 2, 3

[12] B. Han, Y. Liu, and F. Qian. Vivo: visibility-aware mobile volumetric video streaming. In *MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking, London, United Kingdom, September 21-25, 2020*, pages 11:1–11:13. ACM, 2020. 3

[13] M. Hosseini and C. Timmerer. Dynamic adaptive point cloud streaming. In *Proceedings of the 23rd Packet Video Workshop*, pages 25–30, 2018. 3

[14] S.-S. Huang, H. Chen, J. Huang, H. Fu, and S.-M. Hu. Real-time globally consistent 3d reconstruction with semantic priors. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–1, 2021. 12

[15] T. Huang, R.-X. Zhang, C. Zhou, and L. Sun. Qarc: Video quality aware rate control for real-time video streaming based on deep reinforcement learning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1208–1216, 2018. 3

[16] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In J. S. Pierce, M. Agrawala, and S. R. Klemmer, editors, *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 559–568. ACM, 2011. 2

[17] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. H. S. Torr, and D. W. Murray. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Trans. Vis. Comput. Graph.*, 21(11):1241–1250, 2015. 2, 7

[18] O. Kähler, V. A. Prisacariu, J. P. C. Valentin, and D. W. Murray. Hierarchical voxel block hashing for efficient integration of depth images. *IEEE Robotics and Automation Letters*, 1(1):192–197, 2016. 2, 7

[19] J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong. Project starline: a high-fidelity telepresence system. *ACM Trans. Graph.*, 40(6):242:1–242:16, 2021. 3

[20] A. Q. Li, W. Cheung, R. Kawiak, D. Robbins, M. Chen, P. Quesada, T. Tran, J. Juang, and C. Jiang. An investigation of volumetric vod streaming compression technique. 2019. 3

[21] J. Lu, H. Benko, and A. D. Wilson. Hybrid HFR depth: Fusing commodity depth and color cameras to achieve high frame rate, low latency depth camera interactions. In

G. Mark, S. R. Fussell, C. Lampe, m. c. schraefel, J. P. Hourcade, C. Appert, and D. Wigdor, editors, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 5966–5975. ACM, 2017. 3

[22] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication, SIGCOMM 2017, Los Angeles, CA, USA, August 21-25, 2017*, pages 197–210. ACM, 2017. 3

[23] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 7

[24] G. Mohanarajah, V. Usenko, M. Singh, R. D'Andrea, and M. Waibel. Cloud-based collaborative 3d mapping in real-time with low-cost robots. *IEEE Trans Autom. Sci. Eng.*, 12(2):423–431, 2015. 2, 3

[25] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robotics*, 33(5):1255–1262, 2017. 3

[26] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, Basel, Switzerland, October 26-29, 2011*, pages 127–136, 2011. 2

[27] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6):169:1–169:11, 2013. 2

[28] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 2, 6

[29] D. V. Opdenbosch, M. Oelsch, A. Garcea, T. Aykut, and E. G. Steinbach. Selection and compression of local binary features for remote visual SLAM. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 7270–7277. IEEE, 2018. 2, 3

[30] Y. Pan, Y. Kompis, L. Bartolomei, R. Mascaro, C. Stachniss, and M. Chli. Voxfield: Non-projective signed distance fields for online planning and 3d reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022)*, 2022. 7

[31] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. D. Turck. Qoe-driven rate adaptation heuristic for fair adaptive video streaming. *TOMM*, 12(2):28:1–28:24, 2016. 3

[32] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. Ros: an open-source robot operating system. In *IEEE International Conference on Robotics and Automation, ICRA, Workshop on Open Source Software, Kobe, Japan*, 2009. 2, 3

[33] S. Saeedi, M. Trentini, M. Seto, and H. Li. Multiple-robot simultaneous localization and mapping: A review. *J. Field Robotics*, 33(1):3–46, 2016. 2

[34] P. Schmuck. Multi-uav collaborative monocular SLAM. In *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*, pages 3863–3870. IEEE, 2017. 2, 3

[35] S. Sengupta, N. Ganguly, S. Chakraborty, and P. De. Hotdash: Hotspot aware adaptive video streaming using deep reinforcement learning. In *2018 IEEE 26th International Conference on Network Protocols, ICNP 2018, Cambridge, UK, September 25-27, 2018*, pages 165–175. IEEE Computer Society, 2018. 3

[36] H. Shi, Z. Wang, J. Lv, Y. Wang, P. Zhang, F. Zhu, and Q. Li. Semi-supervised learning via improved teacher-student network for robust 3d reconstruction of stereo endoscopic image. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, and B. Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4661–4669. ACM, 2021. 1

[37] T. Stockhammer. Dynamic adaptive streaming over HTTP -: standards and design principles. In A. C. Begen and K. Mayer-Patel, editors, *Proceedings of the Second Annual ACM SIGMM Conference on Multimedia Systems, MMSys 2011, Santa Clara, CA, USA, February 23-25, 2011*, pages 133–144. ACM, 2011. 3

[38] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2, 6

[39] T. Takikawa, J. Litalien, K. Yin, K. Kreis, C. Loop, D. Nowrouzezahrai, A. Jacobson, M. McGuire, and S. Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. *CoRR*, abs/2101.10994, 2021. 3

[40] Z. Tang, X. Feng, Y. Xie, H. Phan, T. Guo, B. Yuan, and S. Wei. Vvsec: Securing volumetric video streaming via benign use of adversarial perturbation. In C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3614–3623. ACM, 2020. 3

[41] J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, and H. Hellwagner. Towards 6dof http adaptive streaming through point cloud compression. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2405–2413, 2019. 3

[42] J. van der Hooft, T. Wauters, F. D. Turck, C. Timmerer, and H. Hellwagner. Towards 6dof HTTP adaptive streaming through point cloud compression. In L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 2405–2413. ACM, 2019. 1

[43] Y. Wang, Y. Lu, Z. Xie, and G. Lu. Deep unsupervised 3d sfm face reconstruction based on massive landmark bundle adjustment. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, and B. Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1350–1358. ACM, 2021. 1

[44] Y. Wang, W. Wang, D. Liu, X. Jin, J. Jiang, and K. Chen. Enabling edge-cloud video analytics for robotics applications. *IEEE Transactions on Cloud Computing*, 2022. 3

[45] Y. Wang, Z. Zou, C. Wang, Y. Dong, L. Qiao, and D. Manocha. Orbbuf: A robust buffering method for remote visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pages 8706–8713. IEEE, 2021. 3

[46] D. Werner, A. Al-Hamadi, and P. Werner. Truncated signed distance function: Experiments on voxel size. In A. J. C. Campilho and M. S. Kamel, editors, *Image Analysis and Recognition - 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II*, volume 8815 of *Lecture Notes in Computer Science*, pages 357–364. Springer, 2014. 2

[47] H. Yeo, Y. Jung, J. Kim, J. Shin, and D. Han. Neural adaptive content-aware internet video delivery. In A. C. Arpaci-Dusseau and G. Voelker, editors, *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018*, pages 645–661. USENIX Association, 2018. 3

[48] B. Zhang, X. Jin, S. Ratnasamy, J. Wawrzynek, and E. A. Lee. Awstream: Adaptive wide-area streaming analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 236–252, 2018. 3

[49] H. Zhang, G. Wang, Z. Lei, and J. Hwang. Eye in the sky: Drone-based object tracking and 3d localization. In L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 899–907. ACM, 2019. 3

[50] T. Zheng, G. Zhang, L. Han, L. Xu, and L. Fang. Building fusion: semantic-aware structural building-scale 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 12