INSTITUT FÜR INFORMATIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

Master's Thesis

Pixel-based 2 DoF Synthesis of 360° Viewpoints Using Flow-based Interpolation

Rosalie Kletzander

Draft from January 26, 2021

INSTITUT FÜR INFORMATIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

Master's Thesis

Pixel-based 2 DoF Synthesis of 360° Viewpoints Using Flow-based Interpolation

Rosalie Kletzander

Aufgabensteller:	Prof. Dr. Dieter Kranzlmüller
Betreuer:	Prof. Dr. Jean-François Lalonde (Université Laval, Kanada) Markus Wiedemann
Abgabetermin:	2. Februar 2021

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 26. Januar 2021

. . . . . .
(Unterschrift des Kandidaten)

Abstract

Hier steht eine kurze Zusammenfassung der Arbeit. Sie darf auf gar keinen Fall länger als eine Seite sein, ca. eine drittel bis eine halbe Seite ist optimal.

Contents

1. Introduction

Over the past decade, Virtual Reality (VR) technology has experienced a resurgence in popularity due to the development of a number of affordable, consumer-quality head-mounted displays[1]. These displays allow a user to experience and interact with a virtual environment in 3D, for example by playing games, or by taking virtual tours of cities, historical landmarks, or remote locations in nature.

Some of these environments are modeled meticulously in 3D, while others use 360° images taken at the locations. Often, the 3D-modeled environments allow a lot of freedom of movement, enabling the user to "walk" around and inspect elements of the scene at will. Unfortunately, modeling a real environment by hand to be viewed interactively requires an enormous amount of effort and even then it is very unlikely to achieve photo-realism. An alternative is to capture the location with a 360° camera, which records the entire surroundings in a single image. Viewing these images offers photo-realism (as they are actual photos), but often awkward navigation, for example forcing the user to "jump" from one image location to the next, instead of being able to "walk" smoothly around the scene. Nonetheless, the ease of capturing 360° images with modern 360° cameras, along with their significant advantage of photo-realism, makes this an attractive method for creating immersive VR environments.

The difficulties of navigating an environment captured by way of 360° cameras are not a new phenomenon. Virtual tours also exist outside of the realm of VR, viewed on regular computer or smartphone screens, for example interactive tours of museums, real-estate, or other locations of interest. A prominent example is Google's Street View, which allows users to navigate streets and monuments around the world by way of 360° images.

Whether a scene is viewed stereoscopically with a head-mounted display, or monoscopically on a flat screen, the greatest obstacle at the moment is the problem of interactive, user-driven navigation. Ideally, a user would be able to go anywhere they liked in the scene, and view anything they wanted from any angle and at any level of detail. Unfortunately, for environments captured by way of a 360° camera, this type of interaction would require a prohibitive amount of data, as well as being impossible to manually execute, since a separate image would have to be captured at every possible viewing position.

An alternative to capturing all of these viewpoints manually is to generate them digitally. This requires capturing a much smaller subset of images and using these to generate the rest of the viewpoints. Generating new images based on already captured images is generally known as *image-based rendering* (IBR), or image-based synthesis. There are many different approaches to synthesizing new images from captured ones, and they can generally be categorized by the area where new images can be synthesized, as well as the type and amount of information extracted from the captured images.

The area and the degrees of freedom (DoF) that are necessary usually depend on the goal and requirements of the application: A virtual tour on a pre-defined path for example, would

[1] The price of a consumer-quality VR headset is between approximately 20 and 1000 USD as of January 2021, including headsets for use with a smartphone (`https://www.tomsguide.com/us/best-vr-headsets,review-3550.html`, *accessed Jan 13, 2021*)

only require generating intermediate views between existing viewpoints (i.e. synthesis with 1 DoF) for a smooth transition. For a less constrained virtual tour that enables the user to navigate freely, generating viewpoints with 2 DoF at eye-height could be sufficient. The user could move around and look in any direction but not change their viewing height. Some applications may require three degrees of freedom, for example in order to enable the user to closely inspect certain objects from all angles, but restrict them from moving away from the object.

Depending on the type of scene, the required fidelity, and real-time requirements, different IBR techniques leverage different amounts and types of information extracted from the input images. On one end of the scale, there are approaches that extract as much geometry information as possible from the set of images and try to reconstruct the 3D geometry of the scene as closely as possible. These approaches can suffer from the problem of extracting accurate geometry information, since errors in geometry can lead to unappealing results. Then, there are approaches that try to extract some information from the image, such as feature correspondences, or motion vectors (optical flow), which enables interpolation between pairs of images, i.e. a smooth transition with 1 DoF. Approaches at the other end of the scale use no semantic image information whatsoever. They may use color values, simple proxies for the scene geometry, or information about the relative location of captures, but they tend to operate on pixel-level. There is no distinct name for this, so the term used in this thesis is *pixel-based synthesis*.

One very basic form of pixel-based synthesis is to use a simple proxy ("stand-in") geometry, for example a sphere, in place of the real scene geometry[2]. This allows for resampling the captured images in order to create a new viewpoint any location within the scene, without having to estimate or record the actual scene geometry. However, the potentially drastic difference between the proxy and the actual geometry can lead to severe distortions and artefacts in the resampled images. Although this basic form of pixel-based rendering using proxy geometry may be unsuitable for scenes where the real geometry differs greatly from the proxy geometry, it may be possible to improve these results by combining the basic technique with a 1 DoF interpolation method.

Problem Statement

A number of 1 DoF interpolation techniques exist, many of them using some form of feature correspondence. Flow-based interpolation is one of these techniques that has already been used successfully for 360° images. It uses motion vectors ("optical flow") between pairs of images to interpolate new viewpoints between them. The goal of this thesis is to answer the following research questions:

1. Can flow-based interpolation be used in pixel-based synthesis with proxy geometry?
2. If it can, does this improve the accuracy of the results?

In order to measure the accuracy of the different results, the synthesized images are compared with the ground truth images using two different error metrics, as well as being assessed visually.

[2]The term "proxy geometry" is used in this thesis as a term for the model used in place of the real scene geometry. It can be anything from a simple geometric shape such as a sphere, to a simplified version of the real scene geometry

Scope

The number of possible different environments is infinite, as well as the positions at which images can be captured. In order to reduce the parameter space for testing and evaluation, several restrictions are made. First of all, only indoor scenes are examined. The scenes are assumed to be static, meaning the objects within the scene do not change their positions. To reduce the complexity of the implementation, as well as the number of necessary input images to be captured, only 2 DoF synthesis is considered. This means that all captured and synthesized viewpoints are located on a plane at approximate eye-height parallel to the ground.

The parameters that will be examined are:

Density of captured viewpoints For the evaluation in this thesis, the viewpoints used for synthesis (the "captured viewpoints") are arranged on a regular grid. The effect of different densities on the accuracy of the results is explored, specifically, grids of between 3m and 30cm distance are used.

Location of synthesized points relative to captured points Since the captured viewpoints are arranged on a grid, there are many positions where a synthesized point can be on a line between two captured viewoints. In this case, the problem is reduced to 1 DoF interpolation. This is compared to the case where the synthesized point is not located on any line, which requires 2 DoF synthesis.

Difference between the scene geometry and the proxy geometry The proxy geometry used in the approach is a sphere of the approximate size of the scene. Scenes of different shapes are tested, in order to gauge the effect of the difference between the proxy model and the scene. The "difference" of the model to the sphere is difficult to quantify in a meaningful way, so it is not measured but done intuitively. For example, a scene with the basic shape of a cube is more similar to a sphere than a narrow, long scene. Only simple basic scene shapes are examined, for example a spherical scene, a scene in the shape of a cube, and a scene in the shape of an oblong rectangular cuboid.

Location of synthesized points within the scene Finally, the location of the synthesized points within the scene in relation to objects in the scene is examined. Again, this is difficult to quantify and will be done by synthesizing points regularly throughout the scene and visually comparing the error values to the proximity to different objects.

These parameters will not be examined exhaustively, as this is impossible, given the infinite possibilities of scene shapes and object placements, alone. Instead, for each parameter to be examined, a scenario is designed that tests this parameter in a reasonable context.

Methodology

In order to implement and evaluate an algorithm that combines basic pixel-based synthesis and flow-based interpolation, this thesis follows a methodology made up of two distinct phases: implementation and evaluation (Figure 1.1). The implementation consists of first developing a basic pixel-based synthesis algorithm using a proxy geometry. Then, based

on existing 1 DoF flow-based interpolation algorithms (presented in Section 2.2), the extended pixel-based algorithm using flow-based interpolation is developed. The approach and implementation are presented in Chapter 3.

The evaluation step is further divided into three phases. In the preparation phase, based on the parameter space described in detail in Section 4.1, data for testing and evaluation is acquired. This data includes virtual scenes rendered using animation software, and real scenes captured with a 360° camera. Additionally, the error metrics are adapted for 360° images (Section 4.2). Then, in the synthesis and testing phase, the images are synthesized using the implementations developed in the implementations step. Using the result images and the ground truth data generated during the data acquision, the error calculation is performed with the adapted metrics. Finally, using the calculated error values, along with the synthesized images, the results are analyzed. The analysis is made up of two parts: First, an analysis is performed on the virtual data that explores the limits of the chosen parameters (Section 4.3). Then, the results using the real data are examined as a proof-of-concept (Section 4.4).

Results

The results of this thesis are:

- A basic proof-of-concept implementation of a 2 DoF pixel-based synthesis algorithm with flow-based interpolation
- An evaluation of the results of this algorithm with a comparison of the pixel-based algorithm by itself versus with flow-based interpolation, proving that flow-based interpolation does improve the results of the basic algorithm in the majority of the examined cases

The results of this thesis can be the basis of various future work. Since the implementation is a first attempt at combining the two methods, the problems and limitations discovered in the evaluation can be used to improve future versions. Alternative proxy geometries could be examined, potentially based on sparse scene geometry, which would improve the basic results and could lead to a distinct improvement of the flow-based results. Furthermore, parallelization and optimization could be used in order to achieve real-time framerates.

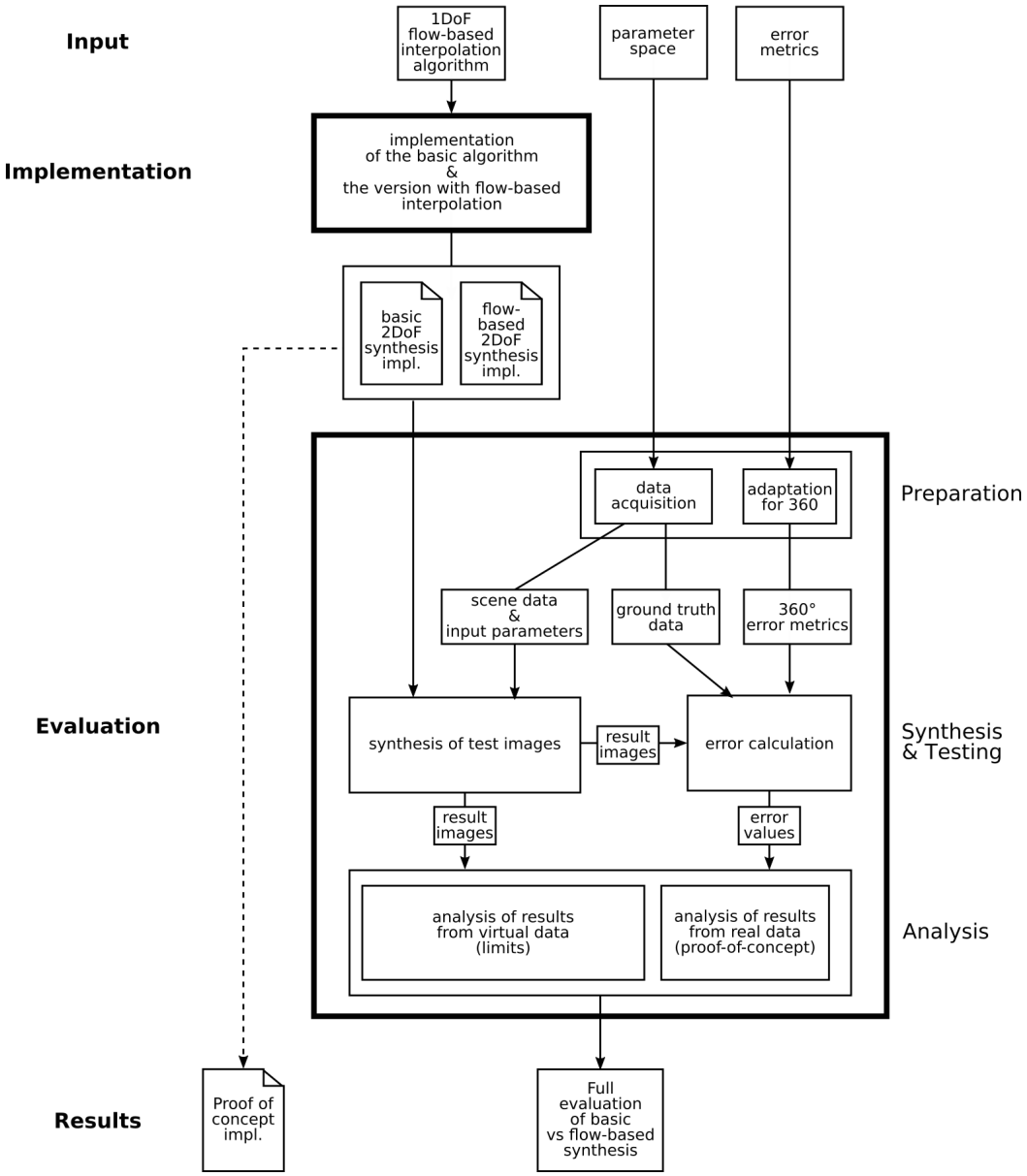


Figure 1.1.: Methodology

2. Background and Related Work

Before diving into the details of the pixel-based rendering approach with flow-based interpolation that is the focus of this thesis, it is important to outline the basic concepts and methods used in these techniques (Section 2.1), as well as to understand the state-of-the-art of image-based rendering techniques (Section 2.2).

2.1. Fundamentals

The images used in this thesis, as well as many other approaches, are 360° images. Since 360° images differ significantly from "regular" images in how they are captured and visualized, it is important to understand how 360° cameras capture their surroundings, how the captured data can be mapped to a planar surface, and what the most common mappings for 360° images are (Section 2.1.1). Also, the concept of optical flow is introduced, as it is a prerequisite for a number of image-based rendering techniques, including the flow-based interpolation used in this thesis.

2.1.1. 360° Images

Capturing an image with a 360° camera differs significantly from capturing an image with a regular camera. A regular camera captures incoming rays of light with a limited field of view. The sensors on the camera (or the film, for analog cameras), are arranged on a plane and register the wavelengths of incoming rays. This process represents a projection of the scene onto a plane. The measured light values can then be stored directly as a planar image (Figure 2.1a).

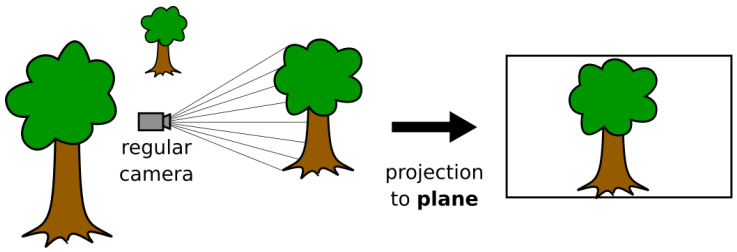
A 360° camera[1], on the other hand, captures light rays with an field of view of 360°. This means that the sensors are arranged in a way that captures light rays from the entire surroundings. For the sake of simplicity, this can be pictured as a number of sensors in the form of a sphere[2]. The camera must then perform an additional conversion in order to transform the light values captured on a sphere to a planar image (Figure 2.1b). [SI14]

The projection onto a flat surface is necessary, since image data is generally stored in 2D, and the majority of viewing devices are planar (e.g. computer or smartphone screens). The process of translating data from a 3D model to a 2D image and vice versa is well known in computer graphics and is called *uv mapping* or *texture mapping*.

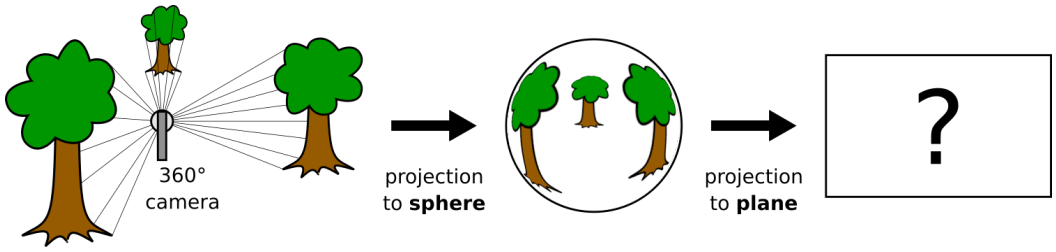
Specifically, the process of uv mapping for spherical geometry is needed for mapping the data from the sphere to a planar image. This process describes a bijective operation in

[1]The term "omnidirectional camera" is also used informally, however, this term is less exact, since an omnidirectional camera can also be a camera that captures a single hemisphere, instead of the full scene [SI14].

[2]The sensors cannot actually take the form of a perfect sphere, since the camera needs to have some form of casing. Instead, several lenses are usually used ("polydioptric cameras" [SI14]), and the image stitched together in software.



(a) Image capturing process with a regular camera: The sensors are arranged on a plane, so capturing the rays corresponds to a projection to a plane.



(b) Image capturing process with a 360° camera: The sensors are arranged in an approximation of a sphere, which means an additional step is needed to convert the captured data to a planar representation.

Figure 2.1.: Capturing an image with a regular camera compared to a 360° camera

which the points (x,y,z) on the sphere (described by *unit directions* which are unit vectors) are associated with pixel positions in image coordinates (u,v). Figure 2.2 shows an example mapping between the unit sphere and a planar image. In this example, the poles of the sphere are mapped to the entire top and bottom pixel row of the image, and the equator is mapped to the row of pixels in the vertical center of the image. This means that the areas near the poles are *oversampled*, which indicates that the mapping function is not *equal area* [RWPD05, p.450] i.e. it does not conserve how much area a pixel value occupies.

In the case of a 360° camera mapping the captured light rays to a planar image for storage, the image values are some type of color values. However, other kinds of information can also be uv-mapped to a shape, for example illumination data, depth values (bump mapping), and more.

The most common mappings for 360° images are the *cube map*, the *ideal mirrored sphere*, the *angular map*, and the *equirectangular map* [RWPD05, p. 535]. The image data can be projected using any of these mappings with only minimal data loss (by interpolation). These mappings are briefly presented in the following paragraphs.

Cube Map The cube map is a mapping that splits the image data into six separate square views, one in each direction (top, front, left, right, back, bottom). This is the equivalent of capturing the surroundings with six different cameras with a field of view of 90° each, and then stitching the resulting images into a shape that can be "folded" into a cube (see Figure 2.3a), which also gives this mapping its name.

Due to the projection of a spherical surface to a plane, there is some distortion towards the edges of each face. However, this distortion is comparable to the distortion at the edges of a regular, planar image, which is a significant advantage compared to other mappings (see Figure2.3b,d,f,h[3]). The disadvantage is that each face is projected separately, which

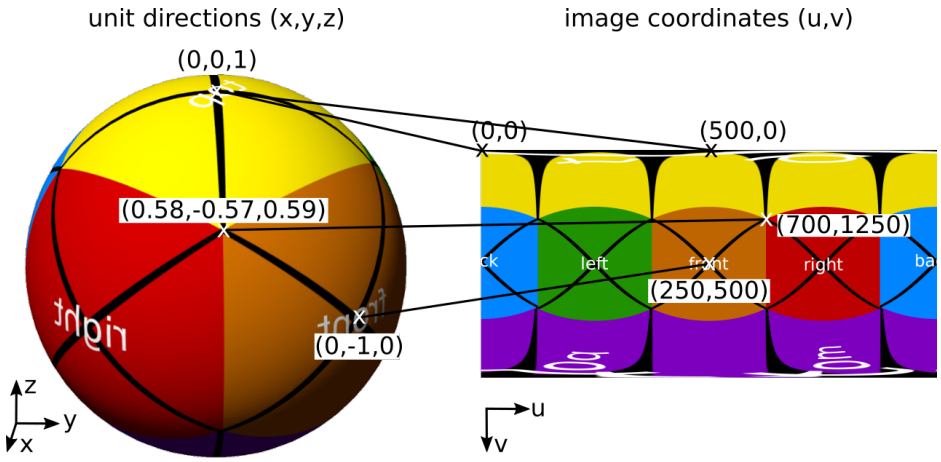


Figure 2.2.: Example of uv mapping for spherical geometry

leads to directional discontinuities at the many seams. This type of mapping is often used to simulate complex environments in 3D scenes (e.g. for game or animation graphics), as it is easy to use and reduces render time significantly compared to a 3D model of the same environment.

[RWPD05, p. 540]

Ideal Mirrored Sphere The ideal mirrored sphere is a mapping to a circle within a square. It represents how the surroundings would be reflected if one placed a small sphere with a perfectly reflective surface ("mirrored" sphere) into a scene and then photographed it using an orthographic camera. This mapping, like all the mappings presented here, shows the complete surroundings, albeit very distorted toward the edges. Figure 2.3cb shows where each direction is mapped and the extent of the distortion. It is clear that the farther away from the "front" area, the more distorted the mirrored sphere mapping is. The ideal mirrored sphere mapping can be used for calculating average illumination color for high dynamic range calculations, however, the type of distortion at the edges can cause problems with sampling, which is why the angular map mapping tends to be preferred. [RWPD05, p. 535]

Angular Map At first sight, the angular map seems very similar to the ideal mirrored sphere. It also maps to a circle within a square, however it samples the input in such a way that the back of the image is allotted more space and is less distorted than the mirrored sphere (see Figure 2.3e). [RWPD05, p. 537]

Equirectangular Map The equirectangular, or latitude-longitude (latlong) mapping is a common type of mapping in cartography. The data is mapped to a rectangular image space, in which the width is twice the height. The azimuth (around the circumference) of the unit directions is mapped to the map's horizontal coordinate and the elevation to the vertical coordinate. The main problem of this representation is well known in cartography: The

[3]Figure 2.3b does not perfectly represent the distortion in cube maps. It was chosen as a baseline because cube maps have relatively small distortion compared to other mappings and it visualizes clearly which parts of the image are mapped where and how they are distorted in other mappings.

distortion increases significantly towards the poles, as can be seen in Figure 2.3h. Otherwise, this mapping is convenient as it has very few seams and all pixels are valid (i.e. no "black" areas). It is used as a storage format for 360° images. [RWPD05, p. 538]

All of these projections are static, showing the entirety of the 360° image at once. It is also possible to view 360° images interactively. In this case the field of view tends to be limited, so only a certain part of the image needs to be projected: the part of the image the viewer is "facing" virtually. Once the viewing direction has been determined, the projection can be calculated such that the center of the image has minimal distortion. Theoretically, any of the above projections could be used for this.

2.1.2. Optical Flow

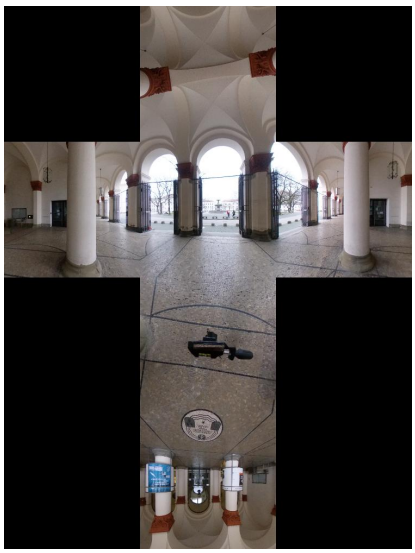
Optical flow describes the displacement of specific points between two images. It is generally used on consecutive frames of video sequences, for example for semantic segmentation, structure-from-motion, data compression or other applications where information about movement between images is required. To illustrate, Figure 2.4 shows two consecutive frames of a video sequence. On a high level, an optical flow algorithm should recognize that the pixels representing the bicyclist are moving towards the bottom left of the image, and the pixels representing the background are moving to the right (because the camera is panning slightly to the left).

There are two types of optical flow: Sparse optical flow and dense optical flow. Sparse optical flow algorithms calculate the motion of several select points that can be either chosen manually, or by some kind of automatic selection (e.g. based on features). This type of optical flow can be used to track only specific objects in a scene (e.g. the direction and relative velocity of a certain car in traffic).

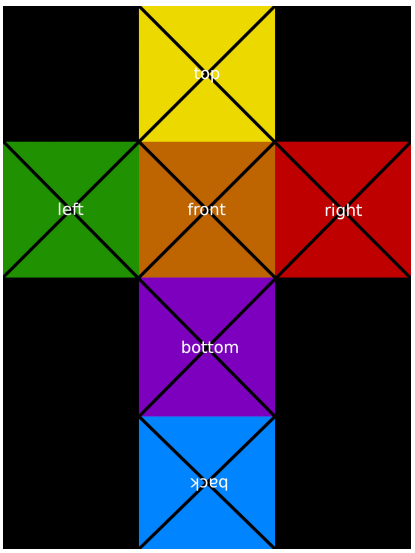
Dense optical flow algorithms compute the motion of *each pixel* between two images, instead of single points. This can be used for more general object tracking (e.g. direction and relative velocity of complete surroundings in traffic), to estimate 3D geometry (in structure-from-motion algorithms), or to identify static sections of the image for video compression [FBK15]. Dense optical flow can also be used for image synthesis, such as in Richardt et al's Megastereo [RPZSH13] described in Section 2.2.2, which is also the basis of the flow-based interpolation presented in this thesis.

There are a number of optical flow algorithms, ranging from methods using parametrization, or regularization [FBK15], to methods relying on Deep Learning [SET20]. Although these algorithms differ greatly in approach, they have in common the type of result, which is a vector field. For dense optical flow, this vector field contains a vector for each pixel, describing the displacement of this pixel between the input images. Sparse optical flow only contains a vector for each pre-chosen point, not for every pixel.

Figure 2.5 shows two different visualizations of the vector field calculated by the dense optical flow algorithm by Farnebäck [Far03] between the frames in Figure 2.4. Figure 2.5a is a color-based visualization: the hue encodes the vector direction, the saturation encodes the vector length for each pixel. Using this visualization, it is possible to roughly distinguish two separately moving areas of the image, which could be used for semantic segmentation. Figure 2.5b shows the pixel displacements with vectors: the origin of the vector is shown by a point and the direction and length of the vector are represented by a line.

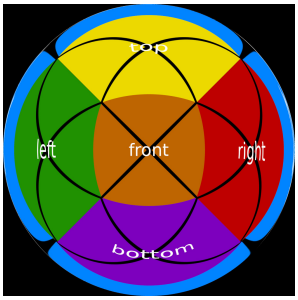


(a) Cube map example



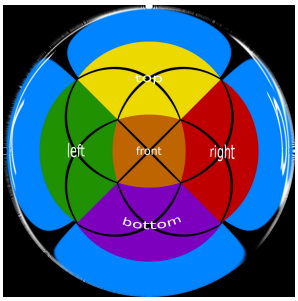
(b) Cube map distortion visualization

(c) Mirrored sphere mapping example



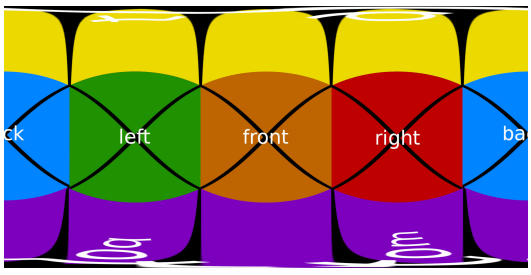
(d) Mirrored sphere distortion visualization

(e) Angular mapping example



(f) Angular mapping distortion visualization

(g) Equirectangular (latlong) mapping example

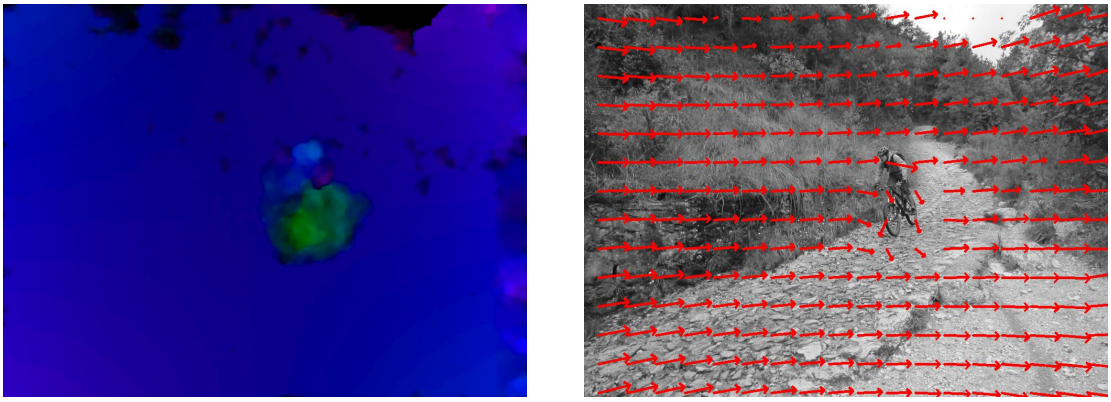


(h) Equirectangular distortion visualization

Figure 2.3.: Common mappings for 360° images

(a) Frame 1 (b) Frame 2

Figure 2.4.: Example frames that optical flow is calculated on



(a) Visualization of optical flow with color: the hue encodes the vector direction, the saturation encodes the vector length for each pixel.

(b) Visualization of optical flow with arrows: The vectors are represented by arrows (only a limited number of vectors are shown)

Figure 2.5.: Optical flow visualizations

These visualizations can help in understanding if and how well an optical flow algorithm is working. Although there are a large number of different algorithms, most of them still struggle with common issues such as occlusions, too-large displacements and intensity changes [FBK15]. Occlusions are problematic, since the displacements between two images may reveal or cover image areas that, as a result, have no correspondence in the previous image. This problem is exacerbated when displacements are very large (e.g. due to fast-moving objects). Large displacements are also problematic by and of themselves, as most algorithms are not designed to handle them. How these limitations affect the use of optical flow for image synthesis will be explored in Chapters 3 and 4.

2.2. Related Work

Image-based Rendering (IBR) and viewpoint interpolation[4] started gaining interest with the advent of virtual walkthroughs, for example for Apple's QuickTime® VR, in order

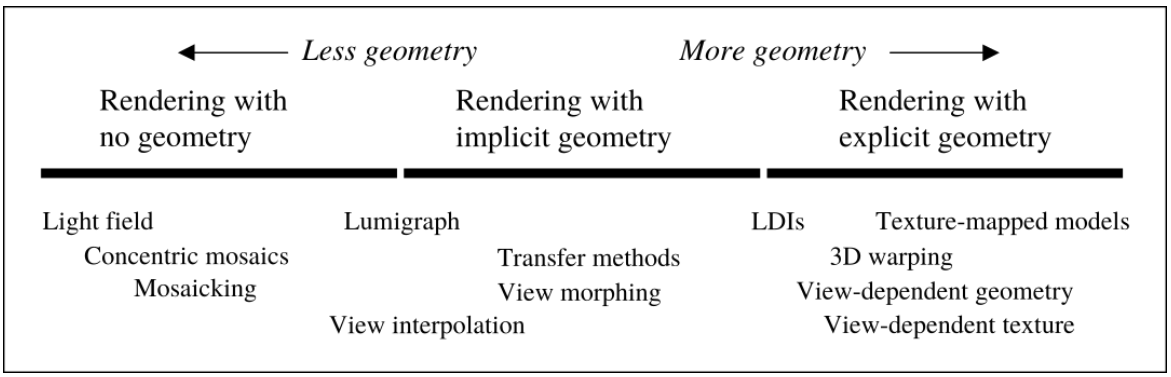


Figure 2.6.: Categorization of IBR techniques with representative members, taken from Kang and Shums "A Review of Image-based Rendering Techniques"[SK00]

to save render time by generating views from images instead of using complex 3D scenes including textures, lights and complex geometric models [Che95]. Chen and Zhang, in their survey on image-based rendering [ZC04], use the terms *source description* and *appearance description* to compare these basic rendering techniques: "Traditional" rendering techniques use *source description*, i.e. the scene is described by the objects within it, their positions and properties. On the other hand, IBR techniques try to achieve the same goal through *appearance description*. Vision itself has little to do with 3D geometry; it is the processing of a dense set of light rays by the brain which are "captured" by the eye. This process can also be performed by a capture device like a camera. So, instead of trying to describe the scene through the objects it contains, IBR rendering techniques try to model the *light rays* that reach the viewer.

While the differentiation between source description and appearance description is helpful in understanding the basic differences between image-based and traditional rendering, many image-based rendering methods also utilize some form of source description, predominantly in the form of 3D geometry. In different survey on image-based rendering techniques, Kang and Shum [SK00] classify IBR rendering techniques on a continuum, which ranges from techniques using no geometry, to techniques using implicit geometry (or more generally, feature correspondences), to techniques using explicit geometry (see Figure 2.6).

The algorithm presented in this thesis is a combination of two different approaches: the first step uses no geometry whatsoever, and the second step uses feature correspondences to correct some problems that arise in the first step. This differentiation guides the related work presented here: The first section presents synthesis approaches using no geometry, and the second section presents approaches using implicit geometry or feature correspondences.

2.2.1. Image Synthesis without Image Correspondences

A theoretical model for image synthesis with no geometry or other image features, i.e. "pure" appearance description was developed by Adelson et al. [AB91]: the *plenoptic function* (Equation 2.1). The plenoptic function is a 7D function that describes the observable light

[4]There seems to be no explicit difference between the terms "image-based rendering", "viewpoint interpolation" and "viewpoint synthesis" in literature, so they will be used interchangeably in this thesis, although "interpolation" is favored for simpler blending techniques, and "synthesis" is used to describe more complex algorithms.

at every point in space V_x, V_y, V_z, from every direction θ, φ, at every wavelength λ, at every possible point in time t.

$$P = P(\theta, \varphi, \lambda, t, V_x, V_y, V_z) \tag{2.1}$$

In practice, it is unfeasible, if not impossible, to cover all dimensions of this function, as this would require a capture device at every location, at every point in time, capturing light rays coming from every direction. However, by making different assumptions, IBR techniques try to reconstruct simplified versions of the plenoptic function. Common assumptions are the reduction of wavelengths to RGB, the removal of the temporal dimension, and the assumption that the light is constant along a ray in empty space (i.e. does not change its wavelength over distance) [ZC04]. The methods for resampling the plenoptic function, or a lower-dimensional version are diverse, from ... [LH96] ...

finish, and maybe mention other approaches, such as neural network approaches

"A Simple Method for Light Field Resampling" [Kaw17]

Kawai [Kaw17] approaches the problem of synthesizing new images with two degrees of freedom without using 3D geometry. Their basic setup is to capture four 360° images at each corner of a rectangular area and use resampling to synthesize a new image anywhere within this area.

The resampling is done by inserting a virtual sphere centered at the synthesized viewpoint representing a projection screen on which to project rays from the captured viewpoints. The locations of the captured viewpoints are known, so the outbound rays of these viewpoints can be calculated by using the image-to-world coordinate conversion from the equirectangular representation. The intersections of these captured rays with the virtual sphere are calculated and the corresponding pixel values used. The projection screen where no rays have intersected are approximated by repeating the reprojection with different resolutions.

In cases where several rays share an intersection, they propose several methods. The first is to take an average of the rays. As an alternative, they suggest a rating based on the inner product of the ray direction and the viewing direction and use the ray with the smallest score. A final option is to prioritize one specific captured viewpoint over the others to completely avoid ghosting artefacts.

To evaluate their method they capture four viewpoints in a scene (the distances between the viewpoints are not mentioned and calculate an image along a diagonal. They then compare the results of the different ray combination methods by describing visual artefacts.

"On the Use of Ray-tracing for Viewpoint Interpolation in Panoramic Imagery" [SLDL09]

Shi et al. [SLDL09] examine how ray tracing can be used to calculate arbitrary new viewpoints based on knowledge of relative positions between the viewpoints which are stored as cube maps. For every pixel in the target image, a ray is cast into the scene. In order to find the correct value of that point, they use a color consistency constraint, which determines whether the pixel values of the reference images are similar. The assumption is that if the colors differ, the rays must not be intersecting the same point.

In order to calculate an intersection with the scene, they propose two different methods: A brute-force depth search using no scene geometry which searches along all of the captured

rays until the pixel values are similar enough to fulfill their color constraint requirements, or a guided depth search using sparse 3D reconstruction.

To evaluate their method, they use a set of five captured input images with a maximum distance of one meter, from which they remove one to use as ground truth. They evaluate the algorithm by comparing the brute-force to the guided depth search based on the artefacts in the results and the computation time.

"Unconstrained Segue Navigation for an Immersive Virtual Reality Experience" [HDR+17]

Herath et al. [HDR+17] propose a system that enables casual users to capture their surroundings with a smartphone in a grid, and then navigate that environment with two degrees of freedom. In order to interpolate between two captured 360° images (1 DoF), they differentiate between faces that are parallel to the axis of movement and faces that are perpendicular to the axis of movement. For faces that are parallel, they stitch the faces of two adjacent viewpoints together and interpolate by using a sliding window. For faces that are perpendicular, they calculate a homography between the faces of two adjacent viewpoints and morph the image accordingly. To interpolate any image within a rectangular area bounded by four captured viewpoints (2 DoF), they recursively interpolate intermediary viewpoints until they reach the desired position.

The focus of this work is on the whole process of capture, navigation, and viewing and the interpolation step is not explicitly evaluated.

2.2.2. Image Synthesis using Image Correspondences

Leveraging image correspondences for synthesis has been a popular method almost since the beginning of viewpoint synthesis. Chen and Williams [CW93] were one of the first to use "the morphing method" that simultaneously blends the shape and texture of two images using image correspondences. A comparable method, based on optical flow, is used by Richardt et al. [RPZSH13] for planar images.

Adaping planar algorithms (e.g. optical flow, structure-from-motion) for 360° images is a common challenge in 360° image synthesis. Kolhatkar et al. [KL10] and Huang et al. [HCCJ17] solve this problem by extending the faces of the cube maps to account for pixels moving across borders. [KL10] then use optical flow for interpolation between two images, whereas [HCCJ17] estimates the scene geometry with an SfM algorithm to extend monoscopic 360° videos to stereo. Zhao et al. [ZWF+13] propose a method for adapting sparse correspondence matching for the spherical domain, circumventing the need to use an extended cube map.

Morphing the images to create a new viewpoint can be done with pixel-based blending [RPZSH13], [KL10] or by triangulating the image and calculating homographies between the triangles [HCCJ17], [ZWF+13].

The flow-based blending approach in this thesis builds on the approaches of [RPZSH13] and [KL10], which are presented in more detail in the following sections.

"Megastereo: Constructing High-Resolution Stereo Panoramas" [RPZSH13]

Richardt et al. [RPZSH13] present an approach to combine planar images captured casually on a radius to create a panoramic image that is viewable in stereo in high resolution. In

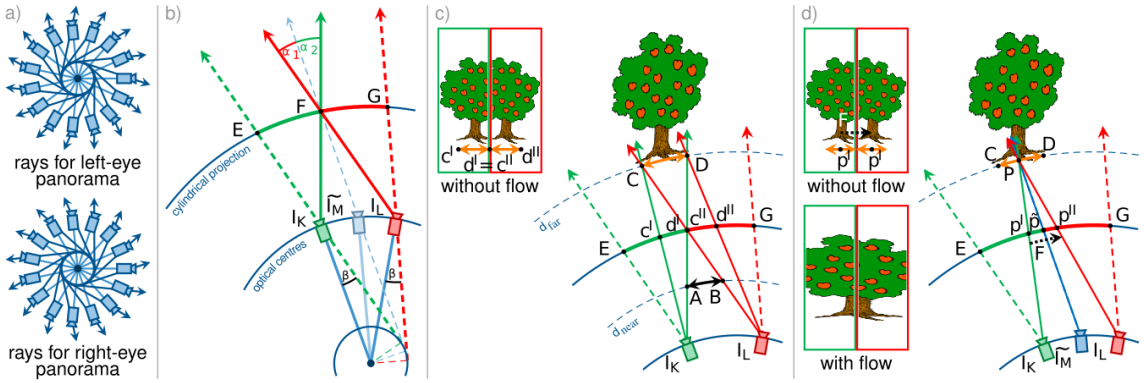


Figure 2.7.: (a) Illustration of rays required for creating a stereoscopic panorama and (b) deviation angles α. (c) Duplication and truncation artefacts caused by the aliasing. (d) Flow-based upsampling to synthesize required rays. *Adapted from [RPZSH13]*

place of scene geometry, they use a cylindrical imaging surface that is concentric to the capture center. For each eye at a given viewing orientation, they project a ray into the scene (Figure 2.7a) and calculate the deviation angle α between the desired ray and the nearest captured rays (Figure 2.7b).

Linearly blending the rays of the two closest captures would lead to artefacts due to the difference between the real geometry and the cylindrical surface (Figure 2.7c), and using a nearest-neighbor technique results in discontinuities. Instead, they propose a "flow-based blending" technique: For each ray of the final image that is not a captured ray (deviation angle 0), a new ray is synthesized using optical flow. The vertical image strip captured by the synthesized ray is interpolated by taking the two closest viewpoints I_K and I_L and interpolating $\widetilde{I_M}$ (Figure 2.7d) using the optical flow vectors $F_{k\to l}$ and $F_{l\to k}$. The corresponding strip is then taken from this new viewpoint which contains the matching ray.

The interpolated image $\widetilde{I_M}$ at point η between the images I_K and I_L is calculated by shifting I_K by $\eta \cdot F_{k\to l}$ and by shifting I_L by $(1-\eta) \cdot F_{l\to k}$. The two shifted images are then blended linearly, using η as the weight. Instead of calculating the entire image for each ray, only the necessary image areas are extracted and the interpolation is calculated pixel-wise.

To evaluate their method, they leverage datasets used by other approaches, as well as capturing their own images. They visually compare the results, noting improvements based on visible artefacts.

"Real-Time Virtual Viewpoint Generation on the GPU for Scene Navigation" [KL10]

Kolhatkar and Laganière [KL10] propose a method for smoothly interpolating between pairs of 360° images. Their approach leverages the optical flow calculated between Their approach is similar to the approach in Megastereo, where optical flow between the images is used to incrementally morph the two images. Since they use 360° images, they extend the cube map representation to account for points moving across edges, which is the method that is used in this thesis, as well. To reduce artefacts in the obtained optical flow, they perform a matching and smoothing step. They then implement their algorithm on the GPU, which

	method			evaluation				
	input type	**DoF**	**extracted features**	**defined parameter space?**	**visual eval.**	**error metrics**	**computational cost**	**comparison to other approaches**
[Kaw17]	360° images	2	none	✗	✓	✗	(✓)	✗
[SLDL09]	360° images	2	none/ dense geo	(✓)	✓	✗	(✓)	✗
[HDR+17]	360° images	2	none	✗	✗	✗	✓	✗
[RPZSH13]	planar images	1	dense flow	✗	✓	✗	(✓)	✓
[KL10]	360° images	1	dense flow	✗	✓	✓	✓	✗
[HCCJ17]	360° video	3*	dense geo	✗	✓	✗	✓	✓
[ZWF+13]	360° video	1	sparse feature	(✓)	✓	✓	✓	✓

*on a constrained path

Table 2.1.: Comparing the methods and evaluations of different approaches

allows them to interpolate between images in real-time.

They evaluate their method by capturing scenes at "reasonable" distances and removing every other image in order to obtain ground truth data. The computation time of the optical flow calculation of the extended cubes is measured, as well as the actual interpolation time. Furthermore, they compare the interpolated results with ground truth images, both visually and by using a per-pixel metric.

2.2.3. Discussion

The overview of the presented approaches (Table 2.1) show that most approaches using explain table

3. Pixel-based 2 DoF Synthesis of 360° Viewpoints with Flow-based Blending

The approach presented in this chapter combines two types of methods presented in the previous chapter: At its base, the basic pixel-based synthesis uses no geometry or correspondences whatsoever. It is similar to Kawai's "A Simple Method for Light Field Resampling" [Kaw17] in that it also uses a sphere to resample the surroundings. However, where Kawai's work suggests using only one viewpoint for resampling to avoid ghosting artefacts (i.e. doubled edges), the pixel-based synthesis presented in this chapter uses flow-based blending to try to remove ghosting artefacts. The flow-based blending is based on the method presented in Richardt et. al's Megastereo [RPZSH13].

First, the general approach is presented (Section 3.1), and then the details of the implementation are described (Section 3.2).

3.1. Approach

This section presents the assumptions and simplifications made for the pixel-based synthesis of 360° viewpoints (Section 3.1.1), the basic pixel-based approach using a proxy geometry (Section 3.1.2) and an improvement to the basic approach based on flow-based blending from Richardt et al.'s Megastereo [RPZSH13] (Section 3.1.3).

3.1.1. Assumptions

In order to simplify the process of synthesis, some assumptions are made based on the scene and the viewpoints in the scene:

- the scene is static
- all images are captured on a plane parallel to the floor (viewpoint plane)
- all synthesized viewpoints are located inside the scene boundaries and are also located on the viewpoint plane
- the positions and orientations of the captures are known
- the scale (max radius) of the scene is known

Furthermore, for the moment, it is assumed that the optical flow algorithm used in the flow-based blending calculates a decent result between any pair of viewoints.

3.1.2. Basic 2 DoF Synthesis

With these assumptions and using a basic proxy geometry with approximately the same scale as the captured scene, it is already possible to synthesize new viewpoints with varying accuracy, depending on the scene. The process presented here for basic 2 DoF synthesis is a combination of texture lookup through raytracing, and mosaicking by using a constraint based on the ray deviation angle.

Raytracing-based Texture Lookup

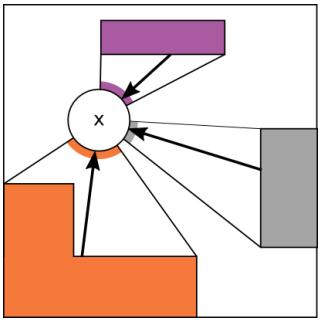
The first step is to map the texture (i.e. pixel values) of an existing viewpoint to a new viewpoint according to its position in the scene. Theoretically, any 360° viewpoint can be mapped to any other, since each 360° image captures each point in the scene. This is only theoretically the case, since image resolution and occlusions in the scene will conceal some areas for some viewpoints whereas they are visible for others. However, at this point, this will be ignored and it will be assumed that each viewpoint image contains all the points of the scene albeit at different image coordinates and different sampling rates[1].

Additionally, 3D geometry of the scene is needed for raytracing. However, since the approach in this thesis does not capture or infer any real geometry, a proxy geometry is used that has approximately the same scale as the scene that was captured. The proxy geometry is a sphere, as this is a simple, very general geometry to represent a variety of different scenes. The radius of the sphere is chosen so that the sphere contains all possible points in the scene, for which the scale of the scene needs to be known. Under these assumptions, it is possible to map the image at one viewpoint to a new position by combining raytracing and texture lookup.

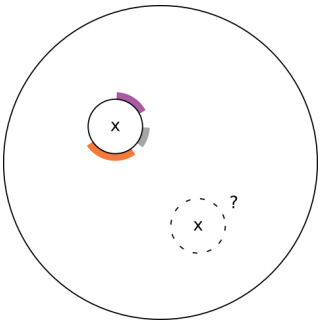
In order to do this, several steps of raytracing are necessary, which are visualized in Figure 3.1. Figure 3.1a shows how a camera at a specific viewpoint captures the light rays reflected from the objects in the scene. The captured pixel values are visualized on a circle around the center of projection of the camera (for simplicity's sake, only one row of pixels is shown). Once the viewoints have been captured (there is only one viewpoint in this example), a new viewpoint is ready to be synthesized. The proxy geometry is visualized as a circle[2] in Figure 3.1b, with the new viewpoint to be synthesized represented by a dotted circle around a center of projection. For each pixel of the synthesized image, a ray is projected into the scene (Figure 3.1c) and its intersection with the scene is calculated. Then, the ray from the center of projection of the captured viewpoint to the scene intersection is calculated, which, when normalized, is equivalent to a unit direction vector of the unit sphere. Using the unit direction and the image mapping function, the pixel value at that position is retrieved (Figure 3.1e) and copied back to the new viewpoint (Figure 3.1f). This way, the pixel values (i.e. texture) of a captured viewpoint are mapped to the new viewpoint (Figure 3.1g). Figure 3.1h compares the mapped values to the actual scene. It is immediately visible that most points have the value they would have, had the viewpoint been captured instead of synthesized (ground truth value), whereas some are incorrect. This is due to the disparity between the proxy geometry and the real scene geometry.

[1] By design, areas closer to the camera are captured with a higher sampling rate per point than areas farther away.

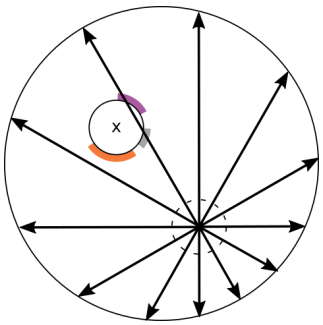
[2] In this example, the sphere does not surround the complete scene (the corners of the scene are outside of the circle). This is only for visualization purposes, normally the sphere would contain the complete scene, including the corners.



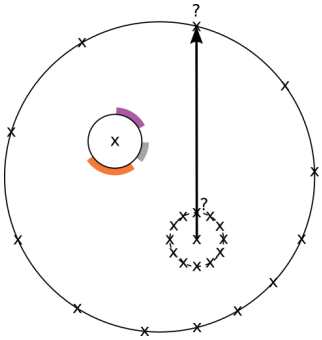
(a) The real scene viewed from above: The camera captures light rays reflecting from objects



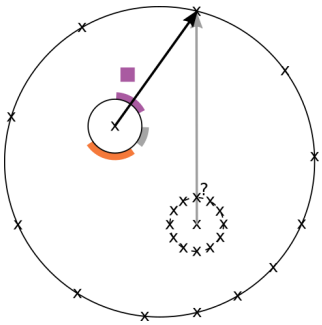
(b) A new viewpoint to be synthesized using the proxy geometry (sphere)



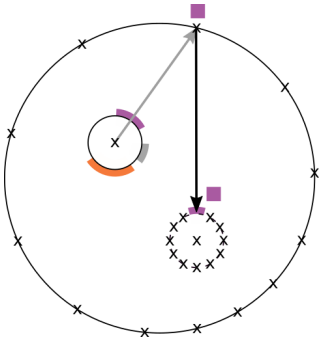
(c) A ray for each pixel in the synthesized viewpoint is traced to its intersection with the proxy geometry



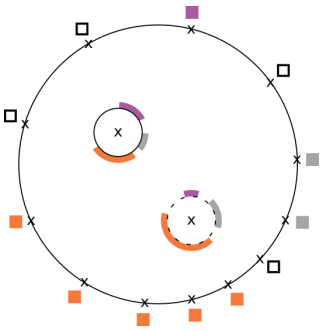
(d) For each ray, a texture lookup is performed



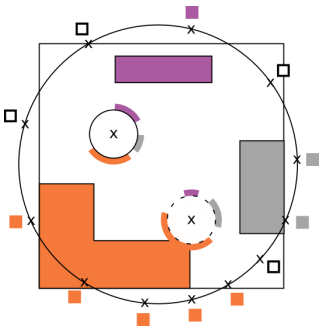
(e) From the ray-proxy geometry intersection, a second ray is traced to the center of projection of the captured viewpoint and the pixel value is looked up



(f) The pixel value is copied back to the new viewpoint



(g) This process is repeated for all pixels



(h) The resulting texture mapping from the captured viewpoint to the proxy geometry in comparison with the original scene

Figure 3.1.: Process of texture lookup through raytracing

Deviation-angle-based Mosaicking

The example in Figure 3.1 contains only one captured viewpoint, so the choice of which viewpoint to use for texture lookup is trivial. In cases where several viewpoints are available, a choice must be made as to which viewpoint should be used. In the case where the real scene has the same geometry as the proxy sphere, this is practically irrelevant, since the raytracing is always accurate. However, this is unrealistic, since the number of spherical rooms containing no objects is negligible. As a result, as soon as the real scene differs from the proxy sphere, some viewpoints yield better results than others. Figure 3.2a shows how a discrepancy between the real scene and the proxy sphere can lead to inaccurate results.

When comparing rays from different viewpoints, two metrics can be examined: the euclidean distance of the captured viewpoint from the synthesized viewpoint, and the deviation angle between the rays. Figure 3.2b visualizes the two metrics and in the example, the $vp2$ with the smaller deviation angle is a better match. In fact, assuming that there is no obscuring element in the air such as fog, and disregarding diffusion and scattering over distance, the same light ray is captured by any viewpoint located on the ray in question. This means the closer the deviation angle is to zero, the more accurate the result will be, no matter the distance of the viewpoints with the best result being a deviation angle of zero (Figure 3.2c). However, sampling rates and resolution also have an effect on the sampled point, so the euclidean distance cannot be completely ignored.

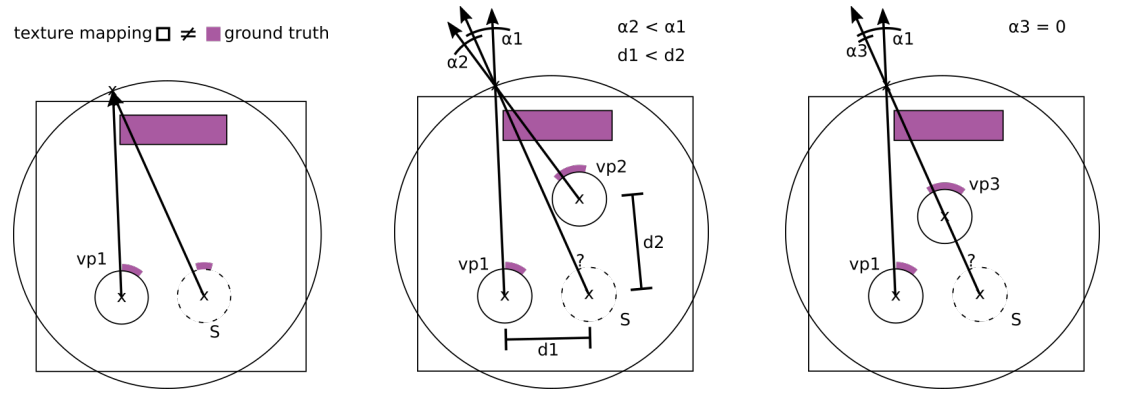
Instead of combining these metrics in a function that might have to be weighted differently depending on the scene size, the distribution viewpoints, and more, the choice of which viewpoints to use is divided into two different steps: The choice of input viewpoints from all captured viewpoints for the synthesis of a specific point, and the choice of which of the input viewpoints to use for each ray of the synthesized point.

The pre-selection of input viewpoints from all captured viewpoints is based on the assumption that the closer the captured viewpoints are to the synthesized viewpoint, the more accurate the sampling of the surroundings will be (i.e. the relative size of the objects). As a result, all viewpoints further than a certain distance are discarded from the input.

After selecting the most appropriate captured viewpoints, the synthesized image is created by comparing the deviation angles of these viewpoints for each ray (i.e. pixel). The two closest viewpoints are then blended together on a per-pixel basis, so that there are no abrupt edges between mosaic areas. The blending function is presented in more detail in Section 3.2.

3.1.3. 2 DoF Synthesis using Flow-based Blending

Using basic 2 DoF Synthesis works fairly well as long as the real scene geometry corresponds roughly to the proxy sphere geometry. The basic shape of many rooms can be approximated by a sphere, however the objects within these rooms can diverge greatly from the proxy geometry. In these cases, ghosting and doubling artefacts become visible, such as areas appearing twice, not at all, or two areas overlapping inconsistently. This problem is exacerbated when the synthesized viewpoint is very close to an object, as is visualized in Figure 3.3: In this example the synthesized viewpoint is very close to a detailed object whose geometry diverges significantly from the proxy sphere geometry. The values of the points captured by the two viewpoints $vp1$ and $vp2$ differ (orange and purple), and neither of them is the desired ground truth value (gray) (Figure 3.3a). In order to improve the result, an adapted variation of the flow-based blending method from Richardt et al.'s al.'s Megastereo



(a) Differences in the real scene from the proxy geometry can lead to inaccurate texture mappings

(b) The deviation angle α of the rays can be compared, as well as the euclidean distance d

(c) The optimal case (disregarding sample resolution) is when the deviation angle is 0, as for viewpoint 3 in this example

Figure 3.2.: Choosing the appropriate viewpoint to improve the result

[RPZSH13] is introduced. This method allows interpolation between two 360° viewpoints using optical flow. Figure 3.3b shows how the interpolation can be used to achieve a more accurate result: A new viewpoint $vp1\text{-}2$ is interpolated between $vp1$ and $vp2$ such that the interpolated viewpoint is located on the ray in question[3] This new viewpoint is then used for the texture lookup to create the synthesized image with the goal of improving the accuracy of the mapped point.

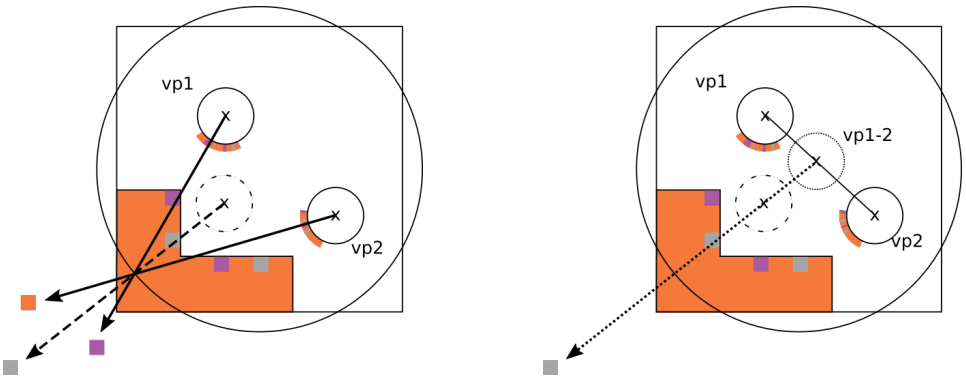
Adapting Flow-based Blending in Megastereo for 1 DoF Interpolation of 360° Images

Megastereo [RPZSH13] aims to generate high-resolution stereo panoramas by combining images captured on a circle. Their approach is to combine corresponding strips of the captured images and to create a view for each eye (see Section 2.2.2). In order to mitigate artefacts such as ghosting, they use "flow-based blending" to combine two images A and B. This consists of using the optical flow vectors $F_{A\to B}$ and their inverse $F_{B\to A}$. To get the interpolated image at position δ between image A and B, first, image A is shifted by $\delta \cdot F_{A\to B}$ and image B is shifted by $(1-\delta) \cdot F_{A\to B}$, yielding I_A and I_B, respectively. Then, I_A is multiplied by $(1-\delta)$ and I_B by δ and these pixel values are added together to give the resulting interpolation. This is described by the following function, in which each pixel at position x of the synthesized image S is defined by:

$$\begin{aligned} S(x) = (1-\delta) \cdot A(x + \delta \cdot F_{A\to B}(x)) \\ + \delta \cdot B(x + (1-\delta) \cdot F_{B\to A}(x)) \end{aligned} \tag{3.1}$$

The flow-based blending in Megastereo operates on planar images. In order to use it for 360° synthesis, it is necessary to adapt the method for 360° images.

[3]Positioning the interpolated viewpoint directly on the ray is only possible for rays that are on the 2D plane containing all the viewpoints. All other cases must be approximated.



(a) Detailed areas where real geometry is very different from the proxy geometry are problematic for basic 2 DoF Synthesis

(b) Interpolating between input viewpoints using optical flow may improve results

Figure 3.3.: Introducing flow-based blending to improve accuracy

A 360° image can be projected in several ways, as described in Section 2.1.1. The output of these projections is a planar image, meaning that it would be possible to apply flow-based blending directly. However, optical flow algorithms are generally designed to handle planar images without seams or distortions and would most likely produce unexpected results if used naively on planar projections of 360° images. As a result, the 360° images must first be projected and adapted in such a way that optical flow can be calculated accurately on them.

Of the projections presented in Section 2.1.1, only the cube map representation is applicable. Spherical representations are impractical, as aligning seams is not feasible and the distortion towards the edges is extreme. The equirectangular representation has only four seams to handle, but also distorts the image greatly around the poles. The cube map representation contains a number of seams but does not distort the image more than a planar image would be. The challenge presented by the many seams of the cube map is to be able to track points that move across the seams created by the six faces. Figure 3.4 shows an example of different points moving across seams, illustrating why calculating optical flow on each face separately would not be enough. Figure 3.4 also shows the linear discontinuities at the seams (e.g. at the upper edge of the carpet): since each cube face was captured by a different virtual camera, angles are not consistent across seams. As a result it is not possible to use the cube map as it is, since the optical flow assumes linear movement[4].

To solve this problem, an *extended* cube map is introduced, which was also used by Huang et al. [HCCJ17] and Kolhatkar et al. [KL10] to adapt 360° images for structure-from-motion and optical flow algorithms. Instead of projecting a field of view of 90° for each camera, which covers exactly 360° of the image, the extended cube map uses a larger field of view for each camera (in this case, 150°). As a result, some areas of the scene are represented several times: the areas of the image that are near a seam are represented on each face that is adjacent to the seam. This way, when calculating optical flow on each face separately, points that move across where the seam would be in a regular cube map remain on the face with the corresponding projection. Figure 3.5 shows the front, top, and left faces of the

[4]Optical flow uses vectors to describe movement, which are inherently linear

(a) Viewpoint A in cube map representation (90° FoV per face)

(b) Viewpoint B in cube map representation (90° FoV per face)

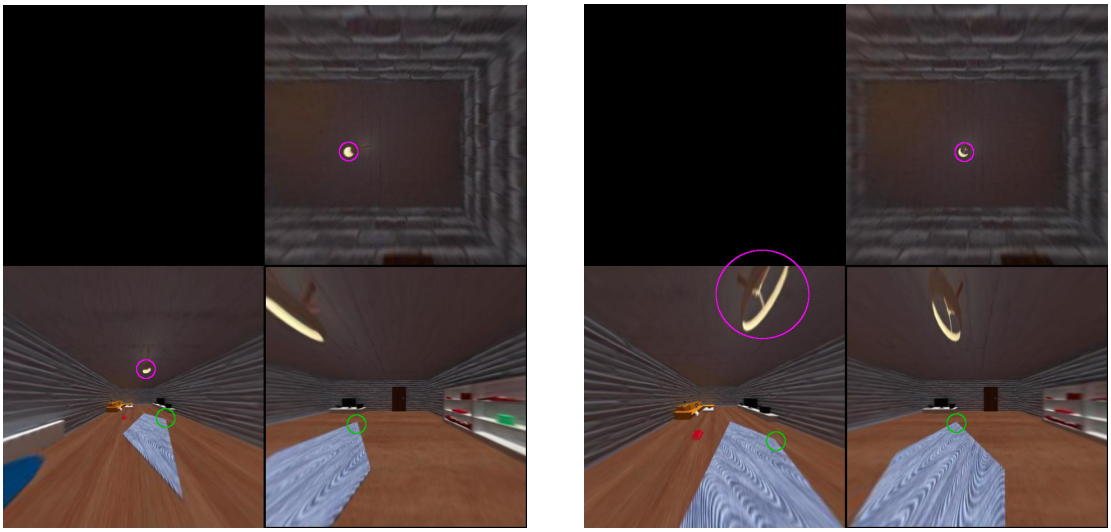
Figure 3.4.: Points in the scene moving across seam edges need to be tracked by optical flow (the back face of the cube map was omitted for simplicity's sake)

extended cube map representation of the cube map shown in Figure 3.4. In the extended cube map, the tracked points do not traverse a seam on any of the faces, meaning optical flow can be calculated on the entire original face.

Nonetheless, this method is still limited by the field of view used by the virtual cameras. If the maximum displacement is larger than the face extension, the extended cube map will not be sufficient, as the points can also traverse the extended seams. Also, the larger the field of view, the more the image will be distorted towards the edges of a face, which may lead to distorted optical flow results. Both problems are visible in Figure 3.5: The lamp in the left face has such a large displacement between a and b that it is partly cut off in b. On top of being cut off, it is also greatly distorted, which will result in distorted optical flow values for that area.

In general, this means that displacement between two images is limited. However, the displacement that is trackable by optical flow algorithms is also limited. The effect of these limitations will be explored in Chapter 4.

Despite these limitations, using the extended cube map makes it possible to calculate optical flow on each face separately, meaning that Megastereo's flow-based blending method can be applied to two 360° viewpoints A and B: First, the extended cube map projections A_{ext} and B_{ext} are created from the image data. From this point, each set of faces A^i_{ext} and $B^i_{ext}, i \in [top, left, front, right, bottom, back]$ is handled separately. Optical flow $F^i_{A \to B}$ and inverse optical flow $F^i_{B \to A}$ are calculated for A^i_{ext} and B^i_{ext}. Then, the shifted image is calculated using Equation 3.1. Finally, for each face, the extended parts of the extended cube map are clipped so that each face once again has a field of view of 90°, resulting in the blended 360° image at position δ between viewpoints A and B. Since the flow-based blending method is applied to two complete images instead of image strips like in Megastereo, it is equivalent to interpolation with one degree of freedom (i.e. on a line) between viewpoints A and B. This interpolated viewpoint can then be used for texture lookup just like a captured



(a) Viewpoint A in extended cube map representation (150° FoV per face)

(b) Viewpoint B in extended cube map representation (150° FoV per face)

Figure 3.5.: Points that traversed a seam in the regular cube map can be tracked across the original seams in the extended cube map

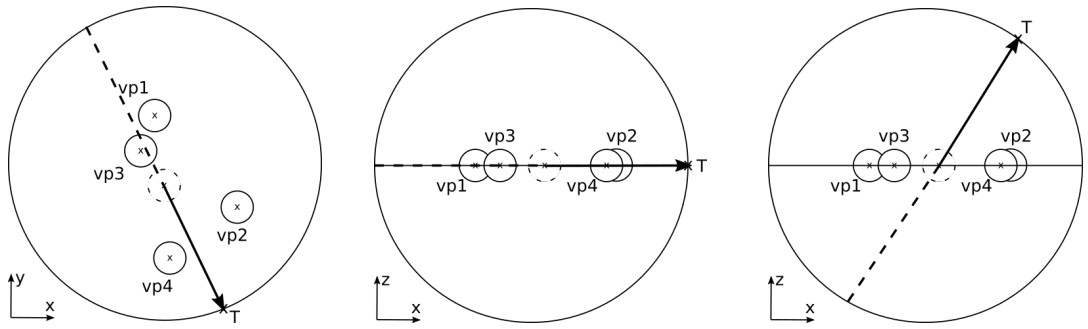
viewpoint.

2 DoF Synthesis with Flow-based Blending

Adapting Megastereo's flow-based blending for 1 DoF interpolation of 360° images allows the creation of a new viewpoint between A and B that is closer to the actual ray (Figure 3.3b). In order to leverage this to improve the basic 2 DoF synthesis, the 1 DoF interpolation needs to be integrated in the 2 DoF synthesis algorithm. For each pixel of the synthesized image, a set of input viewoints A and B needs to be chosen for use in the 1 DoF interpolation. Then, based on the positions of A and B, and the ray in question, an interpolation distance δ must be calculated that defines the position of the 1 DoF interpolation between A and B.

For these steps, an approximation needs to be made due to the 2 DoF restriction: Figure 3.6a and b show the ideal case, where the target point T is on the same horizontal plane as the captured points and the synthesized point (the viewpoint plane). In this case there are a number of different positions directly on the ray that are also on the plane. Depending on the input viewpoints, the most convenient can be chosen and an interpolated viewpoint calculated at that position. However, this is only the case for all target points *on this plane*. All other target points in the scene lie either above or below the viewpoint plane, for example in Figure 3.6c, where T is above the plane. In these cases, the only intersection of the ray and the viewpoint plane is at the synthesized viewpoint. Finding the set of viewpoints A and B that allow the closest 1 DoF interpolation to this point is not trivial, as it would require comparing the minimum distance of the point to all vectors between all the possible sets of viewpoints. In order to simplify this problem, all the rays that are not on the viewpoint plane are approximated.

To approximate a ray pointing at target point T at the spherical coordinates (r, θ, φ), the elevation θ is reduced to 0, assigning the spherical coordinates to $(r, 0, \varphi)$, which is equivalent

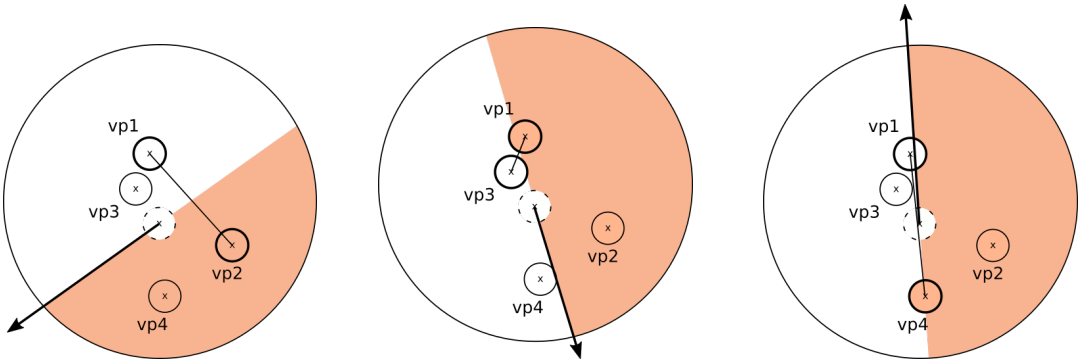


(a) Scene view from above: The target point T is on the viewpoint plane

(b) Scene view from the side: The target point T is on the viewpoint plane

(c) Scene view from the side: The target point T can also be above (or below) the viewpoint plane

Figure 3.6.: Example of different target points in the scene



(a) A: vp1, B: vp2

(b) A: vp1, B: vp3

(c) A: vp1, B: vp4

Figure 3.7.: Examples of the choice of viewpoints A and B for 1 DoF interpolation based on deviation angle and position on either side of the ray. The two sides of the ray are color coded in white and orange.

to moving T to the viewpoint plane by the shortest path. This will yield less accurate results as the actual ray moves towards the poles, since the deviation angle between the actual ray and the approximated ray increases towards the poles. However, this approximation is computationally and mathematically much simpler and should yield acceptable results.

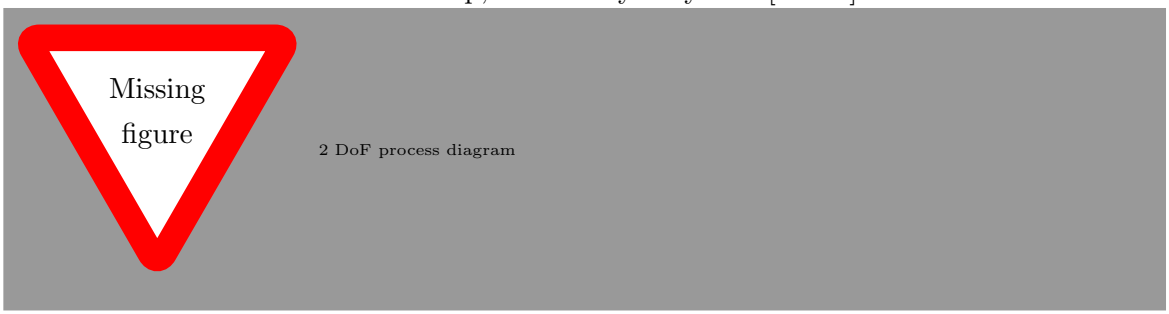
With this approximation, the viewpoints A and B used for 1 DoF interpolation can be chosen. As in basic 2 DoF synthesis, the metric for choosing A and B is the deviation angle. The actual rays, not the approximated rays, are used for the calculation and comparison of the deviation angles, since there is no need to use the approximated rays at this point. However, for the choice of A and B, an additional constraint is included: The two viewpoints chosen must be on either side of the approximated ray, so that there is an intersection between the vector connecting the viewpoints A and B and the approximated ray (see Figure 3.7).

Once the viewpoints A and B are chosen for each target point T, the interpolation distance $\delta \in [0, 1]$ is calculated, which is the the point on the vector $\overrightarrow{AB}$ that intersects the approximated ray. The calculation of δ is a simple line intersection calculation, explained in Section 3.2.

Using the chosen viewpoints A and B, and the calculated interpolation distance δ, a 1 DoF interpolation is calculated for each approximated ray. Using this interpolated viewpoint, a texture lookup is then performed for each pixel associated with that ray. This results in a mosaicked image where each image area (vertical strip in equirectangular representation) is an interpolated, reprojected viewpoint. The 1 DoF interpolation step should improve some of the artefacts caused by the use of the proxy sphere instead of the actual scene geometry. Its effectiveness and limitations are explored in Chapter 4.

3.2. Implementation Details

This section presents the technical and mathematical details on which the basic 2 DoF synthesis and the 2 DoF synthesis with flow-based blending are based. The basic 2 DoF synthesis and the 2 DoF synthesis with flow-based blending are implemented in Python3 [Pyt18], using the libraries NumPy [The19b], OpenCV [The19a], SciPy [The20], and scikit-image [vdWSN+14]. For the conversion between different 360° projections, as well as the calculation of the extended cube map, the library "skylibs" [Hol20] is used.



3.2.1. Preprocessing

The input data consists of a set of captured viewpoint images in equirectangular representation, a text file containing the metadata (positions and orientations) of these viewpoints, and the approximate scene radius. In order to easily and intuitively access the locations and image data of the captured viewpoints, the data is encapsulated in the CaptureSet class. The CaptureSet class first parses the metadata, then, with this information, rotates all the images so that they have the same orientation, and shifts the viewpoint cloud so that it is centered around the origin (0,0,0). This is done under the assumption that images were captured in a regular distribution throughout the room. The proxy sphere representing the scene is also centered at the origin. Instead of storing the image data directly in the CaptureSet, the file paths are stored so that the images can be dynamically loaded when needed. The CaptureSet can then be used by the ImgSynthesizer to synthesize a new image either using regular blending or flow-based blending at any given location within the scene[5].

3.2.2. Basic 2 DoF Synthesis

For the basic 2 DoF Synthesis, the steps described in detail in this section are the selction of input viewpoints, the calculation of the ray-scene intersection and the texture lookup, as well as the deviation angle calculation and blending function introduced in Section 3.1.2.

[5]Given that it is within the convex hull of the captured viewpoints

Selecting Appropriate Input Viewpoints

Before calculating ray-scene intersections and performing texture lookup, the most appropriate input viewpoints are selected from the complete set of input viewpoints. In reality, resolution does play a role in the resampling, meaning that captured viewpoints that have a large euclidean distance from the synthesized viewpoint will not necessarily yield the "correct" value. Since the euclidean distance is not factored in to the weighting function (see 22), the captured viewpoints are filtered before synthesis. All viewpoints outside a certain radius (approx. 1m) are discarded. If the set of viewpoints within the radius is empty, all viewpoints are used, regardless of distance.

Calculating Ray-sphere Intersection

The ray-sphere intersection is used in the raytracing-based texture lookup to find the scene points captured by the synthesized viewpoint (see 3.1c). This raytracing process is a basic raytracing technique used in computer graphics. The vectors representing the rays of a viewpoint can be easily derived from the unit directions of the 360° image (see Section 2.1.1): Each unit direction is a vector on the unit sphere, representing the location of an image value (pixel). These coordinates are in *model space*, meaning that they are centered around zero. Translating them into *world space* moves the rays to their respective location in the scene, where the vectors represent the rays cast into the scene, where they will intersect with the proxy geometry.

The intersections of these rays with the proxy sphere geometry can be calculated analytically: The proxy sphere, which is centered at the origin, can be represented implicitly by Equation 3.2. The set of points P defined by this equation make up the surface of the sphere (Equation 3.3). The equation describing any point on the ray can be expressed by Equation 3.4, where O is the origin of the ray, which is the center of projection of the new viewpoint, t is the length of the ray and D is a unit vector describing the direction.

$$x^2 + y^2 + z^2 - R^2 = 0 \tag{3.2}$$

$$P^2 - R^2 = 0 \tag{3.3}$$

$$P = O + tD \tag{3.4}$$

The point P in Equation 3.3 can be substituted with the equation of the any point on the ray which yields Equation 3.5. This equation can be developed into Equation 3.6, which is a quadratic function with $a = D^2$, $b = 2OD$, $c = O^2 - R^2$ (Equation 3.7).

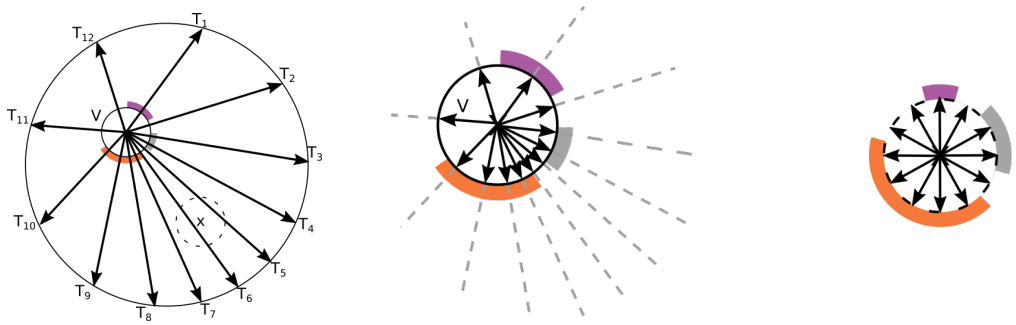
$$|O + tD|^2 - R^2 = 0 \tag{3.5}$$

$$D^2t^2 + 2ODt + O^2 - R^2 = 0 \tag{3.6}$$

$$a = D^2, b = 2OD, c = O^2 - R^2$$

$$f(t) = at^2 + bt + c \tag{3.7}$$

$$t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3.8}$$



(a) The rays $T_i - V$ in world space

(b) The normalized rays $T_i - V$ in model space

(c) The result of resampling based on uv coordinates

Figure 3.8.: Texture lookup by uv remapping

This equation can then be solved for t. Since the radius of the sphere is chosen so that it contains the complete scene and no viewpoints are synthesized outside of the scene, the quadratic function will always have two solutions (i.e. two intersections): one for which the vector length t is negative, and one for which the vector length is positive. Since the original ray used for the calculation is unidirectional (i.e. it cannot invert its direction), it needs to be extended by a positive value. The original ray, being a unit ray of length 1, can then be multiplied by the positive t, which yields the intersection point.

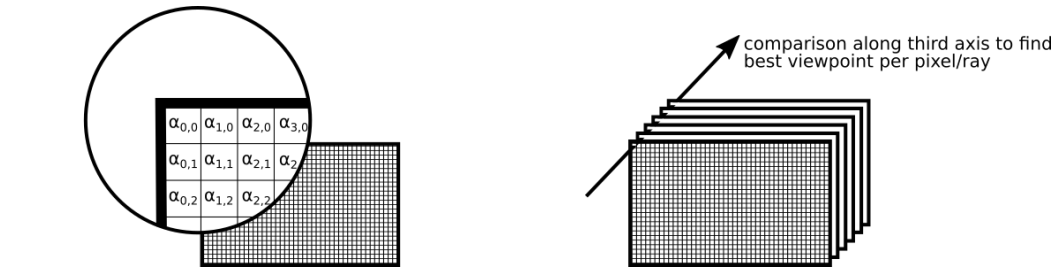
The vectors and intersection points are each calculated and stored in latlong representation (i.e. matrix of vectors of the same shape as the latlong image), which means that they can handily be associated with the unit directions, as well as the uv coordinates (and thus, pixel values) using the latlong mapping function. By storing the values in this representation (i.e. 3D matrix), Numpy's vectorization can be used, which greatly facilitates implementation.

Texture Lookup

The texture lookup shown in Figure 3.1d-f is performed by resampling using uv coordinates. Given a captured viewpoint at the location V and the ray-scene intersections from the synthesized viewpoint T_i, where i denotes which ray is being examined, the rays from the V to the intersections can easily be calculated by $T_i - V$ (see Figure 3.8a). These rays are then normalized to have length 1 and returned to model space (see Figure 3.8b). The normalized rays in model space are in the same format as the unit directions and can therefore be transformed to image coordinates using the latlong mapping function (see Section 2.1.1), implemented in the library Skylibs [Hol20]. These image coordinates (i.e. uv coordinates) can be used to resample the data, which is equivalent to actually performing a ray-by-ray texture lookup, but much faster. The result of the resampling along with the "new" rays is shown in Figure 3.8c. This resampling based on uv coordinates is also implemented in Skylibs and utilizes Scipy's function *scipy.ndimage.map_coordinates*.

Deviation Angle Calculation and Knn Blending

The deviation angle calculation is a simple angle calculation between vectors. This calculation is performed for each ray of the synthesized point and the corresponding rays of all of the input viewpoints. Again, the deviation angles per viewpoint are stored in latlong



(a) The deviation angles of one input viewpoint in latlong format

(b) The stacked deviation angles of all input viewpoints

Figure 3.9.: Visualization of deviation angle storage

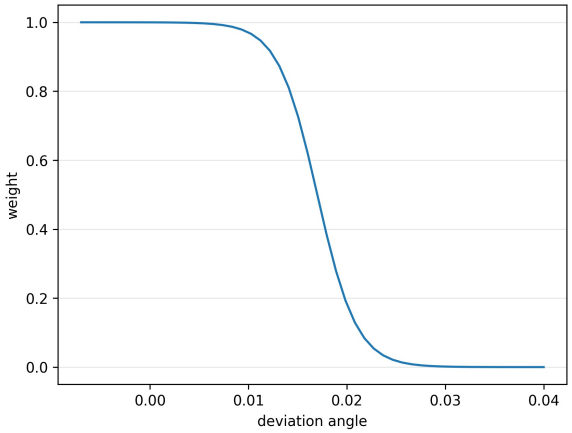


Figure 3.10.: The inverse sigmoid function used for weighting

representation (Figure 3.9a). The deviation angles for all viewpoints are stacked in a three-dimensional matrix, which allows comparison of the deviation angles per pixel/ray of all the viewpoints (Figure 3.9b). This makes the selection of the best viewpoint per pixel very straightforward, since the id of the k best input viewpoints can easily be extracted and the corresponding pixel value retrieved. With this information, the "regular", k-nearest-neighbor (knn) blending is performed.

The regular blending function combines the values of the rays of the k closest deviation angles α. The idea behind it is to weight deviation angles of 0 very highly and all larger deviation angles with exponentially low values. This is done with an inverse sigmoid function (Function 3.9, visualized in Figure 3.10). The parameters of the function were found by trial and error.

$$w(\alpha) = \frac{1}{(1 + e^{500 \cdot (\alpha - 0.017)})} \tag{3.9}$$

Then, the per-pixel weights w are normalized so that their sum for each pixel is one, so that the final image is not oversaturated. Finally, the pixel values of the different remapped viewpoints are multiplied by the normalized weights associated with that viewpoint and the weighted latlong images are added up to give the final synthesized image. Blending the

(a) $k = 1$ (b) $k = 2$

Figure 3.11.: K-nearest-neighbor blending with different values for k. The images are clipped from the latlong representation of an image synthesized using regular blending.

two best viewpoints per pixel results in smoother transitions between the mosaicked areas. Figure 3.11 shows the difference between using $k = 1$ and $k = 2$: It is clearly visible that the border between two mosaicked areas (e.g. on the rug in the center of the room) is very abrupt in Figure 3.11a, becaulse the two areas do not align perfectly. This border is much smoother in Figure 3.11b, since the two areas are gradually blended into one another. Using $k > 2$ does not have much impact, since most deviation angles where $k > 3$ are too large to have an effect.

3.2.3. Flow-based Blending

The 2 DoF synthesis with flow-based blending utilizes some steps from the basic 2 DoF synthesis, such as the deviation angle calculation and the texture mapping. The details of the input viewpoint selection, the choice of viewpoints A and B for the 1 DoF interpolation, the calculation of the interpolation distance δ, as well as the 1 DoF intepolation are explained in this section.

Selecting Appropriate Input Viewpoints

The adaptation of optical flow algorithms for 360° images, as well as most optical flow algorithms themselves, are limited to a maximumum displacement. This is not considered in the algorithm itself, so it is handled in the input viewpoint selection. The goal of the input selection is to find a minimal convex hull around the point to be synthesized from the set of captured viewpoints. The solution is if there is at least one captured viewpoint in each quadrant around the synthesized viewpoint. If there is no point in one of the quadrants, the points in the quadrants adjacent to the empty quadrant are selected so that they are as close to the empty quadrant as possible. Since the synthesized point is in the convex hull of all the captured points, there must exist a solution for the minimal convex hull. In the worst case, the minimal convex hull is so large that the optical flow algorithm fails.

Choosing the Viewpoints A and B

The choice of the input viewpoints A and B for the 1 DoF interpolation is important in that two viewpoints need to be chosen that are on either side of the ray in question (see Figure 3.7), so that the line on which the images can be interpolated actually intersects the

approximated ray. Since all the viewpoints are on a plane, the "side" a viewpoint is on is defined by whether the deviation angle is between 0 and 180° (one side) or between 180 and 360° (other side). To get the best deviation angles on either side, the angles are sorted and the angles closest to 0 and closest to 360 are chosen.

The only problem with this process would arise if there was no viewpoint on the other side of the axis. However, because of the restriction that all synthesized points must be within the convex hull of the captured points, this can never happen. In the worst case scenario, there is only one viewpoint on the other side of the axis with a very large deviation angle. This is acceptable, since this viewpoint is not directly used, instead the 1 DoF interpolation creates a new viewpoint with a ideally very small deviation angle.

Determining the 1 DoF Interpolation distance δ

Given the two viewpoints A and B, it is now possible to calculate the intersection of $\overrightarrow{AB}$ and the approximated ray (elevation $\theta = 0$) between the synthesized point S and the target point P in the scene (Figure 3.12). The calculation of the intersection between two lines is a fairly simple method adapted from [Wei]. Using these four points A,B,P and S, two infinite lines can be defined (Equation 3.10). The intersection point at $t * \overrightarrow{AB}$ is given by Equation 3.11.

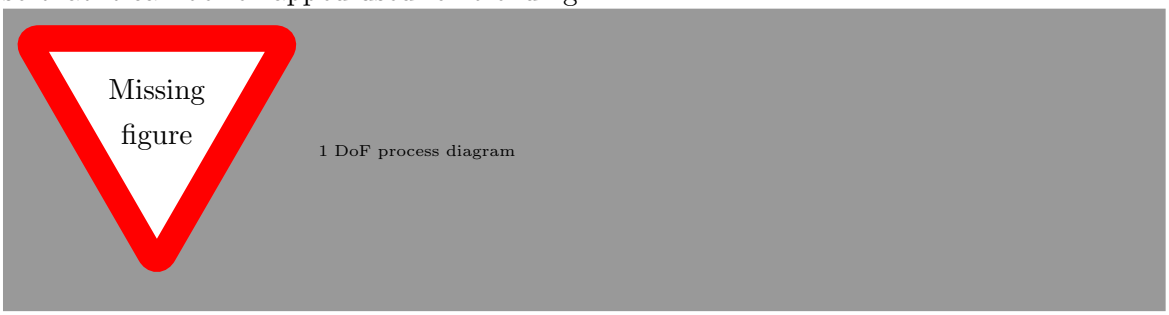
$$\overrightarrow{AB} = A + t * (B - A) \quad \overrightarrow{SP} = S + u * (P - S) \tag{3.10}$$

$$t = \frac{(x_A - x_S)(y_S - y_P) - (y_A - y_S)(x_S - x_P)}{(x_A - x_B)(y_S - y_P) - (y_A - y_B)(x_S - x_P)} \tag{3.11}$$

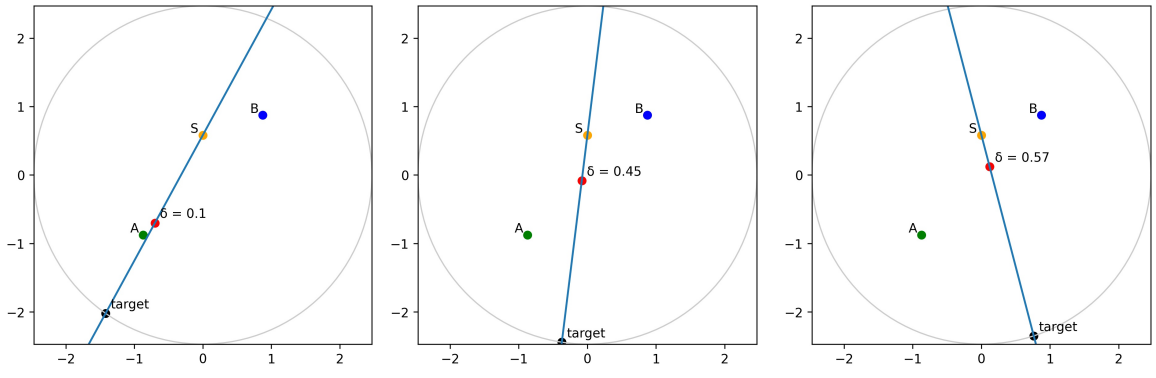
Since the synthesized viewpoint S is within the convex hull, there is always an intersection $t \in [0, 1]$. This value can then directly used as δ. The only case that merits an exception is if the synthesized viewpoint is directly on the border of the convex hull, i.e. directly on the line between two captured viewpoints. In this case, the ray that is parallel to vector $\overrightarrow{AB}$ is equal to the line AB. As a result t is not defined and δ must be found by dividing the distance $|\overrightarrow{AS}|$ by $|\overrightarrow{AB}|$.

1 DoF Interpolation

Given the two viewpoints A and B and the interpolation distance δ, the interpolated image at δ between A and B can be calculated. The different steps of the process are extending the cube map, calculating optical flow on the extended cube map, shifting the images by the optical flow, and transforming the shifted, extended cube map back into latlong representation so that it can be remapped used for blending.



1 DoF process diagram



(a) (b) (c)

Figure 3.12.: Depending on the scene point, a different δ is calculated based on the intersection of $\overrightarrow{AB}$ and $\overrightarrow{S, target}$

Extended Cube Map The class *ExtendedCubeMap* uses the 360° image data and the virtual camera provided by skylibs [Hol20] to "extend" the cube map by capturing a virtual image for each face of the cube with a 150° field of view. This is done for both viewpoints A and B. The *ExtendedCubeMap* class handles the extended faces and can perform different functions on them, for example the optical flow calculation, which is required for the next step.

Optical Flow Calculation and Image Shifting The optical flow algorithm used in the implementation is Farnebäck's algorithm implemented by OpenCV (*cv2.calcOpticalFlowFarneback*) [The19a]. For calculating optical flow between two *ExtendedCubeMaps* A and B, the *ExtendedCubeMap* class provides a function *optical_flow*, which takes as arguments the optical flow algorithm, and a second *ExtendedCubeMap* to use for optical flow calculation. Passing the optical flow function dynamically greatly simplifies exchanging the optical flow algorithm for future work.

The *ExtendedCubeMap* then calculates the optical flow *separately* between each corresponding pair of extended faces. On top of the optical flow from *ExtendedCubeMaps* A to B, the inverse optical flow from *ExtendedCubeMaps* B to A is required as well. The inversion of optical flow is nontrivial, since inverting the optical flow vectors is not enough: The vectors must be shifted by their identity and then reversed. Alternatively, the optical flow can just be calculated on the *ExtendedCubeMaps* in reverse, i.e. from B to A. This option is used in the proof-of-concept implementation in order to avoid bugs, and since the impact on performance is not very high for images with small resolution, which are used in Chapter 4.

maybe illustrate

Using the optical flow, the inverted optical flow, and the interpolation distance δ, the shifted images I_A and I_B are calculated seperately for each face by again using Scipy's *scipy.ndimage.map_coordinates*, with image coordinates shifted by the optical flow vectors and δ, and the inverse optical flow vectors and $(1 - \delta)$, respectively. The shifted *ExtendedCubeMap* images are then combined by multiplying them by $(1 - \delta)$ and δ, respectively, and adding the pixel values. The result is an interpolated *ExtendedCubeMap* at interpolation distance δ.

Before passing this on to the blending step of the 2 DoF synthesis, the *ExtendedCubeMap* is transformed back into a regular cube map by clipping each face back to its original size

and mapping the cube map back to a latlong map, since all other processing steps use the latlong format.

Flow-based Blending

The flow-based blending step combines all of the previous steps: first, for each ray, the viewpoints A and B are selected and the interpolation distance δ is calculated. Then, a mask is created for each set (A, B, δ), masking all pixels that do not belong to the set, and the interpolated image for that set is calculated. This is repeated for all sets.

The complexity of the flow-based blending is directly bound to the precision of δ. Given a precision of two decimal points results in 101 different interpolated images ($\delta \in [0.00, 0.01, 0.02 \ldots 0.99, 1.00]$), whereas precision of one decimal point results in only 11 calculated images. A precision of two is used in the implementation, since this is an acceptable compromise between complexity and a result with smoother transitions.

Finally, the masked interpolated latlong images are added together to give the final result.

3.2.4. Performance

As this is a first attempt at 2 DoF synthesis, performance was not deemed important, and no parallelizations were incorporated. Since the algorithms operates pixel-wise, the computation time increases exponentially for larger image resolutions ($O(n^2)$ coplexity). The performance of the flow-based blending depends on the location of the synthesized point in relation to the captured points. In the worst case, one interpolated image must be calculated per pixel, at best one interpolated image must be calculated for the whole image (if a synthesized viewpoint is directly on a line between two captured viewpoints). The non-optimized interpolation of a whole image for pixel regions (and in the worst case, single pixels), makes the flow-based blending significantly slower than the regular blending.

The synthesis of an image of 1000x500 pixels with no optimization using regular blending with 4 input viewpoints takes approximately 4s on a single Intel Xeon (Skylake) processor, whereas flow-based blending with the same input takes between 10 and 30 minutes, depending on the constellation of the viewpoints. Fortunately, many of the operations are "embarrassingly parallel", meaning they can be very easily parallelized, which will be discussed in Section **??**.

3.2.5. Implementation-related Problems

Unfortunately, there are some implementation-related problems that are unrelated to the algorithm, but nonetheless have an effect on the result. The extension of the regular cube map to the *ExtendedCubeMap* with the skylibs library slightly distorts the image for an as-of-yet unknown reason, so that when the map is clipped back down to its regular size, it is not identical to the original. Although the difference not extreme, it is nonetheless noticeable.

The other, more visible problem, also from the skylibs library, concerns the conversion from the cubemap back to the latlong representation. Due to a bug in this operation, a black line sometimes appears at on or more borders of a face. Since the synthesized images are made up of mosaicked, reprojected areas, it is possible that in some cases, these lines shift into the middle of faces.

Although these errors seem relatively small, it is necesssary to keep them in mind during the evaluation, since they only apply to the flow-based blending and may skew the results.

4. Evaluation and Results

Unfortunately, there are no publicly available benchmarks for 360° image synthesis with two degrees of freedom without 3D geometry, as not many methods exist that try to achieve this. Since the approach presented in Chapter 3 is a first, basic attempt at solving this problem, this chapter presents a basic evaluation of the algorithm, based on mathematically calculable error metrics. These metrics measure the accuracy of a synthesized image compared to the ground truth and are used to assess the effect of a limited number of parameters.

First, possible parameters are discussed (Section 4.1), followed by the presentation of the evaluation methodology (Section 4.2). Then a limit evaluation is performed on virtually generated scenes to explore the limits of the algorithm in a controlled environment (Section 4.3). Based on the knowledge gained in the limit evaluation, a proof-of-concept evaluation is performed on real scenes (Section 4.4). Finally, the aggregate evaluation findings are discussed (Section **??**).

4.1. Parameters

Before defining the parameters to test in the limitation and proof-of-concept evaluations, this section gives an overview of possible parameters in the context of the 2 DoF algorithm presented in Chapter 3. The 2 DoF algorithm already makes a few assumptions, for example the constraint to the viewpoint plane, the fact that the scene is static, and more (stated in Section 3.1 on page 19). These assumptions are upheld in the evaluation, as they are prerequisite to the function of the algorithm.

define performance as accuracy of results based on error metrics

The remaining parameters (that are not constrained by the assumptions) can be divided into two categories:

Internal parameters i.e. based within the algorithm, such as the blending type and the selection of input viewpoints

External parameters i.e. based on the properties of the captured scene, such as the viewpoint density, or the geometry of the scene.

The internal parameters can be modified after the scene has been captured, the external ones cannot. The most prominent internal parameters based on the implementation from Chapter 3 are:

- the location of synthesized points within the scene (near walls, objects, etc)
- the location of synthesized points relative to the captured points
- the blending type, i.e. flow-based blending or deviation-angle-based knn blending ("regular blending")
- the optical flow algorithm used for flow-based blending

There are more internal parameters that could theoretically be modified, such as the knn blending function (the inverse sigmoid function on page 31), or the method of approximation for 2 DoF in flow-based blending (page 26), but these will be assumed immutable for this evaluation, as the variation of these parameters would require developing new functions, which would be outside the scope of this thesis.

As for the possible external parameters, the number of different possible scenes is infinite, but the assumed key parameters are:

- type of scene (outdoor, indoor, etc) → size and shape of scene
- objects within the scene
- density of captured viewpoints
- distribution of captured viewpoints

External parameters such as the camera settings (e.g. aperture, shutter speed, white balance) and the lighting throughout the scene are not considered; it is assumed that all the captures have the same camera settings and the white balance is comparable throughout the scene. Furthermore, the evaluation is restricted to indoor scenes of approximately $25m^2$. This reduces the parameter space significantly, since the fact that indoor scenes tend to be enclosed by walls enforces a maximal distance of objects to the camera.

The evaluation presented in this thesis aims to examine the effects of a few select internal and external parameters, instead of exhaustively examining all of them. In order to do this, a *scenario* is designed for each selected parameter that attempts to demonstrate the effect of this parameter on the accuracy of the result. Limiting the evaluation to specific scenarios reduces the testing space but might also lead to missing some interactions between parameters. This is acceptable, since the evaluation does not aim to be comprehensive.

4.2. Evaluation Methodology

The evaluation is divided into two distinct phases: an evaluation of limits using virtual scenes, and a proof-of-concept evaluation using real scenes. Both evaluations follow the methodology depicted in Figure 4.1, and consist of four steps: *scenario definition*, where a scenario is designed to exemplify the parameter to test, *synthesis*, where the synthesized images are calculated using the 2 DoF synthesis presented in Chapter 3, *error calculation*, where the accuracy of the synthesized images is measured, and *result analysis*, where the the cause and effect of the parameters is examined. The details of these steps are described in the following sections.

Scenario Definition

A scenario is defined by the parameter that it tests, the static parameters that are used, and the scene where the data was captured. Although a scenario is designed to test a specific parameter, which then is "dynamic" (i.e. will be modified throughout the scenario), there might also be more dynamic parameters. For example, in a scenario for exploring viewpoint density, the blending type might also be modified to see what effect the viewpoint density has on regular and flow-based blending. The static parameters include for example the location of the synthesized viewpoints, or any other parameter that remains unchanged throughout

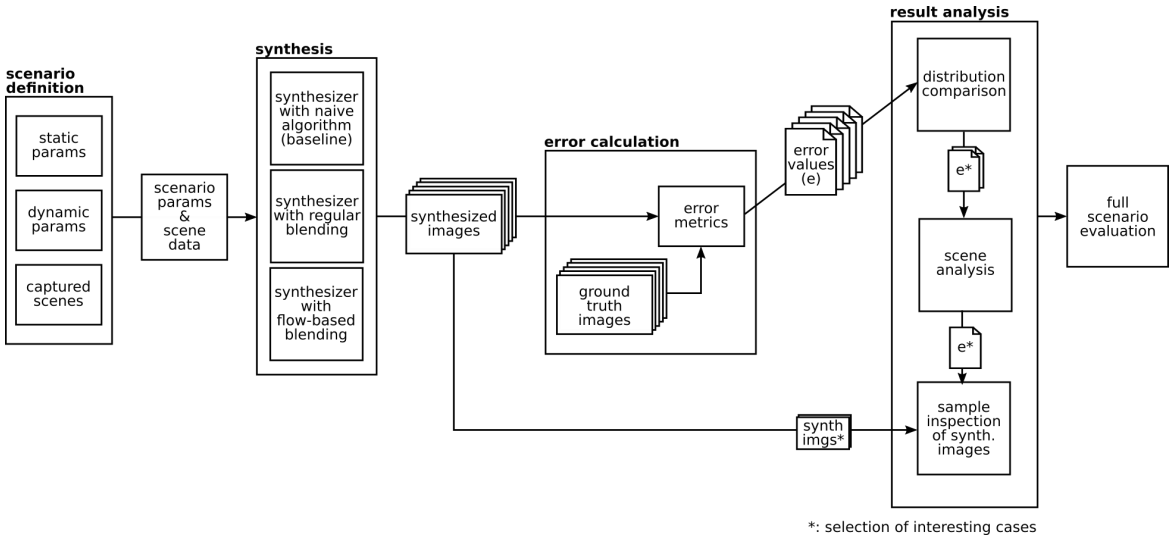


Figure 4.1.: Methodology for the evaluation of a scenario

the scenario. One other defining factor of a scenario is the scene data that the scenario is tested on. Although the scene is part of the parameters (the external parameters, to be exact), it merits particular mention, as it contains the actual image data that is used for synthesis. This data, along with the other parameters defined in the scenario, are then passed on to the synthesis step.

Synthesis

The synthesis step consists of two parts: the 2 DoF synthesis presented in Chapter 3 using either flow-based blending or regular blending depending on the scenario parameters, and a baseline synthesis using a naïve algorithm. The naïve algorithm consists of simply selecting the nearest neighbor viewpoint based on euclidean distance. The input parameters are the same for both algorithms and both results are passed on to error calculation. The results of the naïve algorithm serve as a baseline comparison to verify whether the developed 2 DoF algorithm is an improvement to a naïve approach. If either the regular or the flow-based blending generally performed worse than the naïve algorithm, this would be an indication of a substantial flaw in the approach.

Error Calculation

There are many properties that a synthesis algorithm can be evaluated for, for example performance, visual acceptability (based on user studies), number of artefacts, or distance from the ground truth. In this evaluation, mathematical error metrics are used to compare each result to its ground truth image. Two different metrics are chosen based on different image features so that potential limitations of each metric can be compensated for by the other. However, it must be taken into account that neither are perfect for the task of measuring accuracy on 360° synthesized images, so their validity should always be taken with a grain of salt.

(a) (b) (c) L1 error visualization of (a) and (b)

Figure 4.2.: Example visualization of L1 RGB error. The RGB error values have been intensified so that they are more visible.

L1 error on RGB The first metric is the L1 error calculated on the ground truth and result images in RGB color space. This is a simple error metric that calculates the mean absolute difference of the RGB values and therefore indicates the mean accuracy of each pixel of the image. The RGB errors of each pixel are added together for the complete image and then divided by the number of pixels in the image. This results in an error value $e \in [0, 3]$, since the maximum error per pixel is 3 for floating point RGB color values $\in [0, 1]$.

The L1 error can also be visualized by calculating the absolute difference per pixel without averaging the values. Figure 4.2 shows an example visualization of the L1 error between two images. The visualization encodes areas of the image where there is a very large difference with a value closer to white and areas where there is no difference as black, which clearly highlights areas of the image that are problematic.

The L1 error is useful because it gives a rough estimation of how accurately each pixel is synthesized. The visualization indicates in which areas the synthesized image is inaccurate, which is helpful for classifying problems. However, a drawback of the L1 error is that it relies on color values, which means that images with large differences in pixel values will generally produce a higher error value than images with smaller differences in pixel values, even though the distortion and displacement may be the same.

As in the case of optical flow calculation, some adjustment must be made to adapt this metric for 360° images. Since the equirectangular projection is not equal-area, the areas towards the poles would intrinsically have higher weighting, since RGB L1 is calculated per pixel. To avoid this problem, the cube map projection is used, since it does not significantly distort the image. The average value is then calculated using the six faces of the cube, omitting the black background.

SSIM error on Grayscale The metric to complement the RGB L1 error is a variation of the the structural similarity index (SSIM) [ZBSS04], which measures the *structural similarity* between two images. Instead of comparing the images pixel by pixel, the SSIM uses the luminance, contrast and structure of the images for comparison. It compares these locally, i.e. it operates on smaller areas instead of the image as a whole. As a result, it is possible that the SSIM does not register small displacements in the scene if the objects are not distorted. However, the additional comparison with the RGB L1 error should mitigate this potential

problem.

The SSIM metric in general, and the implementation used in the evaluation[1] return a value $\in [-1, 1]$ with 1 signifying an extremely similar image and -1 signifying a very different image. In order to more easily compare it with the RGB L1 error, the SSIM value is converted to an error value $e \in [0, 1]$, with 0 signifying an identical image (0 error) and 1 signifying a very different image.

The SSIM error is calculated on the grayscale image in cubemap representation. There is no need to use an RGB image, since it does not use the color values of an image. To avoid possible problems with distortion, the cubemap representation is again used.

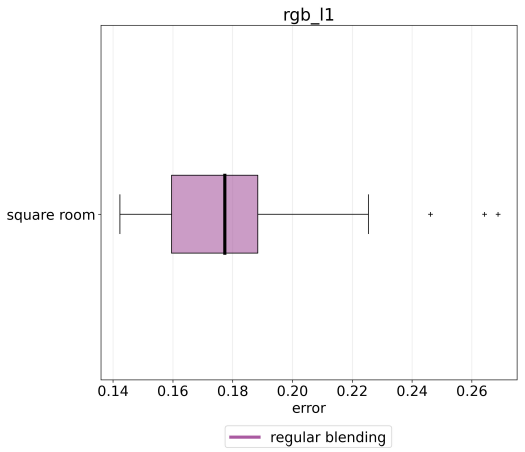
Result Analysis

For each scenario evaluation, the number of results is made up of the dynamic parameters, the number of synthesized viewpoints, plus the baseline results. For each image, there are two error metrics to be considered. In order to analyze these results effectively, it is necessary to break them down by creating different visualizations that highlight different attributes of the results.At first, an overview is created, from which interesting cases are selected and examined in more detail.

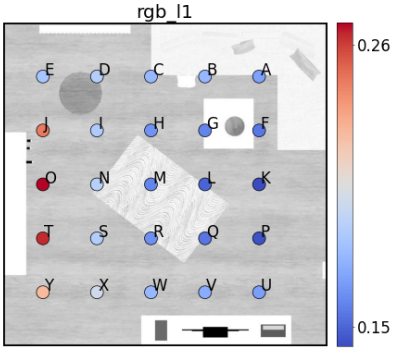
Distribution Comparison The first step of the analysis is a comparison of error value distribution. In order to compare all the error values of a scenario, they are plotted using a boxplot (Figure 4.3a). The different parameter combinations of the scenario are plotted on the y axis (e.g. viewpoint densities a, b, c) and the error distribution (i.e. the error values of all the synthesized viewpoints) is plotted on the x axis. The box plot shows the distribution of these values: the thick black line in the colored box is the median value (approx. 0.184 in Figure 4.3a), the colored ranges to the left and right of the median describe the "interquartile range" (IQR), the range of the closest half of the data (25% on each side). The "whiskers" of the plot extend to the minimum and maximum of the values. The minimum and maximum are defined as $1.5 \cdot IQR$. Any data outside of the the range between the minimum and the maximum, are the "outliers", depicted as small crosses. The boxplot gives a general overview of how the error values of the specific scene are distrubute. The distribution of the error values of the results in Figure 4.3a, for example, shows that the first three quartiles of the results are fairly close to the median (between approx. 0.14 and 0.19), whereas the fourth quartile extends, over almost the same range (0.19 to 0.225) and there are some extreme outliers. This indicates that there are some viewpoints that performed significantly worse than most of the others.

Scene Analysis Based on the insights gained in the distribution comparison, several interesting cases are selected for closer analysis. These cases are examined by color coding the error values and assigning the colors to the positions in the scene. This puts the error values in context with the scene surroundings. Figure 4.3b shows the synthesized points in the context of the scene, color coded by their error values. The maximum and minimum values of the points (also clearly visible in Figure 4.3a) are coded as light green and dark blue, respectively. This visualization gives a more detailed overview over the values of the different points. In Figure 4.3b, for example, the synthesized points near the right wall of

[1] skimage.metrics.structural_similarity [vdWSN+14]



(a) Boxplot example for distribution comparison of the results of regular blending in the "square room"



(b) Visualization for scene analysis: The error values from the boxplot in (a) are mapped to colors at the positions of the viewpoints in the scene. Light green represents the worst value (approx. 0.27) and dark blue the best value (approx. 0.145).

Figure 4.3.: Different types of result visualizations for L1 error values ("rgb_l1") for example results of regular blending in a scene

the room have much better values than the row on the left side of the room. The four light green values are clearly the outliers that were visible in Figure 4.3a. Using this information, it is possible to draw som conclusions about the effect of the position of the synthesized viewpoint relative to the objects in the scene, and select a few synthesized images that merit closer examination.

replace figures (colors and error metric names

Sample Inspection In order to further understand the effects of the parameters on specific positions, some of the synthesized viewpoints from the scene analysis are examined manually by comparing the synthesized image to the ground truth image. The visual examination may also reveal information that the error metrics are unable to extract, such as specific types of artefacts. For example, by using the information presented in Figure 4.3b, it is possible to choose one of the best results, for example synthesized point "K" near the right wall in the middle. The close inspection of the synthesized images is shown in Figure A.1 (page 82 in Appendix A[2]): In the left column from top to bottom are the ground truth image, the synthesized image using regular blending, and the synthesized image using flow-based blending. In the right column are the L1 difference images to the ground truth image. They can help with understanding the error values. For example, the rug in the bottom face (green ring) is improved in the flow result compared to the regular result in Figure A.1. However, the flow result also has a fairly large artefact at the top of the door in the left face (magenta ring), which is not the case in the regular result. The green rings show improvements, and the magenta rings show artefacts or other problems.

color coding of marks?

[2]The synthesized images are shown in Appendix A, in order to avoid interrupting the flow of the text, since they take up a fairly large amount of space.

4.3. Evaluation using Virtual Scenes

In the first part of the evaluation, virtual scenes are used to evaluate the effect of a chosen set of internal and external parameters on the accuracy of the results of the 2 DoF synthesis. Virtual scenes allow full control over the external parameters, for example the scene geometry, and the positions of the captured viewpoints, as well as being less error prone than manual capture, where positional and rotational inaccuracies may be introduced. This also As a result, virtual scenes make it possible to test setups that would be unfeasible for eek real scenes, for example using a large number of captured viewpoints at varying locations in different scenes. Additionally, since the evaluation is based on comparing the synthesized images to the ground truth, it is necessary to capture the ground truth images, along with the input viewpoints. As a result, the scenarios can be designed independent of the constraints of the manual capture of real scenes. The parameters that will be explored in the scenarios are:

- proxy-scene difference (how dissimilar is the scene geometry from the proxy sphere geometry, including the objects within the scene)
- density of the captured viewpoints
- position of the synthesized points relative to the captured points

A scenario is designed for each of the three parameters, containing several different setups that are meant to test the limitations of the algorithm developed in Section 3.

4.3.1. Data Acquisition and Featured Scenes

Three different scenes were modeled for testing, using the animation software Blender [Ble20]: the *checkersphere*, the *square room* and the *oblong room*.

Checkersphere The *checkersphere* (Figure 4.4) is a perfect sphere with a diameter of approximately 2m. Its surface is covered with a checkerboard pattern with alternating colors of dark blue, dark red, and dark green. It represents a scene where the geometry of the scene is exactly the same as the proxy geometry. Although this kind of room is not likely to exist in reality, it serves as a good baseline, since the result of the reprojection should be very close to the ground truth. The checkerboard pattern was chosen so that distortions or inaccuracies are more visible.

Square Room The *square room* (Figure 4.5) is a room whose basic shape is a perfect cube with a side length of 3.5m. It contains an assortment of furniture[3]: In one corner, there is an orange, L-shaped couch with dark blue and white cushions and a white marble coffee table with a dark blue bowl on it, and several small, simple black and white pictures on the wall behind the couch. There is a blue and white rug in the middle of the room, and to the left of the couch are a round blue table, as well as a white radiator. On the wall next to the blue table is a white marble bookshelf containing several red books, as well as three wine bottles, and two decorational objects in green and purple. Across from the couch is a low marble cabinet with a black monitor, a black laptop and a black speaker on it. To the

left of the cabinet is a wooden door with a gray handle. The walls are brick, painted a dark purple and there is a lamp with a white lampshade hanging from the middle of the ceiling.

The intention of using the square room is to allow approximate accuracy for the reprojection step, while at the same time offering a more realistic indoor setting.

Oblong room The *oblong room* (Figure 4.6) has a room size of approximately 5.5m x 3.5m and contains the same basic elements as the square room. It has the exact same furniture layout as the square room, except that some objects

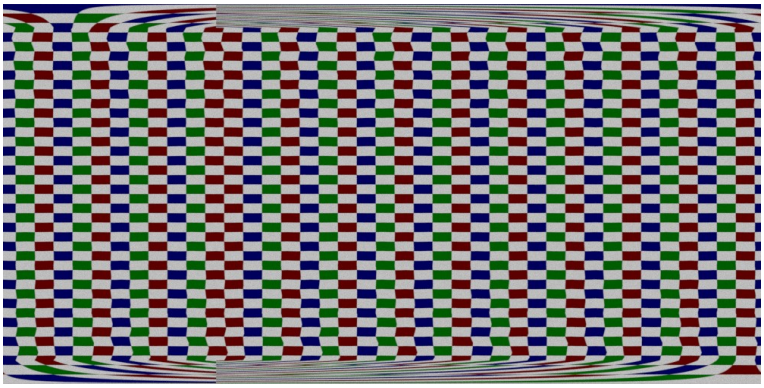
Given the different scenes, it is necessary to choose comparable viewpoints to capture that will be used as input for the synthesis. Since the scenes all have similar scale, the viewpoint layout was chosen to be identical in all of the scenes. This means that all scenes contain 36 captured viewpoints, aligned in a regular grid of 6x6 viewpoints, with 60cm spacing. The grid of viewpoints is centered within the scene. This means that in the checkersphere (Figure 4.7a) and the square room (Figure 4.7b), the viewpoints cover approximately the complete scene, and in the oblong room, there is about a 1m area on each side of the grid that is not captured (Figure 4.7c and Figure 4.7d). It is necessary to take this difference in viewpoint coverage into account for the evaluation, since the viewpoints in the oblong room have a larger distance to two of the walls, which may have an effect on the accuracy of the results.

Since Blender is designed for creating animated movies, and not the capture of static viewpoints, some adjustments had to be made to be able to capture the chosen viewpoints (as well as the ground truth images): After choosing the positions of the input viewpoints for the scenes, the position of the camera for each captured viewpoint was stored as a keyframe, so that the batch of viewpoints could be rendered like an animation. In order to automatically assign the viewpoints and the ground truth points to the keyframes, and write out the metadata, a blender script was implemented. This way the locations of the viewpoints would always be perfectly accurate, and the "capture" of the viewpoints required no manual effort. The images were rendered with a resolution of 1000x500 for all of the scenes, in order to reduce the computation time for the image synthesis in the tests.

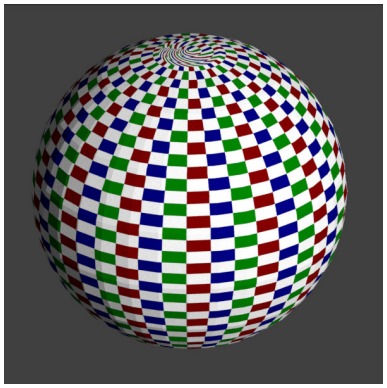
4.3.2. Synthesizing Ground Truth Optical Flow

Using Blender to create virtual scenes not only facilitates capture, but also offers an alternative to calculating optical flow. As mentioned in Section 2.1.2, most optical flow algorithms struggle with large displacements. The flow-based blending step in the 2 DoF synthesis algorithm, on the other hand, assumes decent optical flow and there is no attempt to judge whether the optical flow calculation is feasible between two selected viewpoints. As a result, given the wrong circumstances (two viewpoints A and B that are far apart), the optical flow algorithm may fail, leading the flow-based blending to output undesirable results. The

[3]The furniture models used in the square room and the oblong room are adapted from `https://www.cgtrader.com/free-3d-models/interior/living-room/low-poly-interior-57731178-c955-4625-9e44-109c8eea5ee2`, by user "miha29076", and the textures are adapted from `https://www.poliigon.com/texture/plaster-17`, `https://www.poliigon.com/texture/fabric-denim-003`, `https://www.poliigon.com/texture/wood-fine-dark-004`, and `https://www.poliigon.com/texture/interior-design-rug-starry-night-001`. *All accessed last on September 23, 2020*

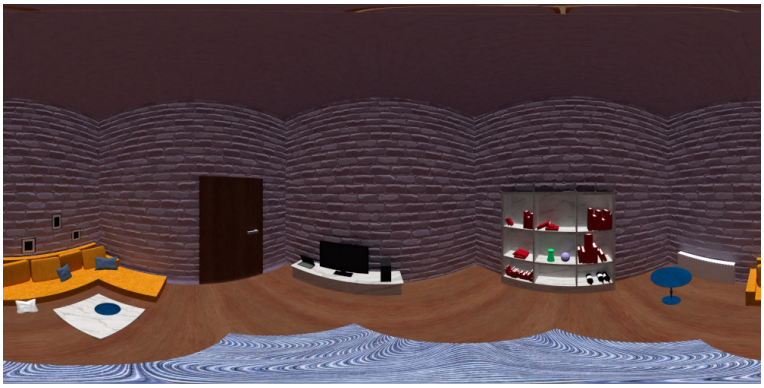


(a) Latlong image from the center of the sphere



(b) View from the outside for reference

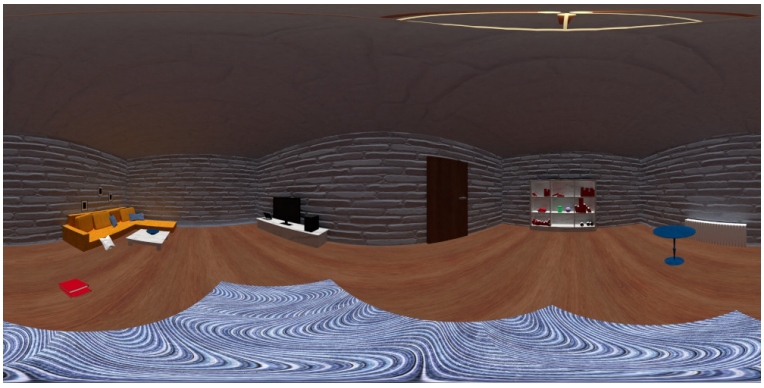
Figure 4.4.: Overview of the “checkersphere”



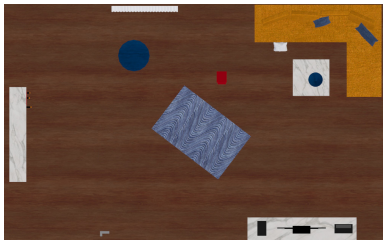
(a) Latlong image from the center of the room

(b) Top view

Figure 4.5.: Overview of the “square room”

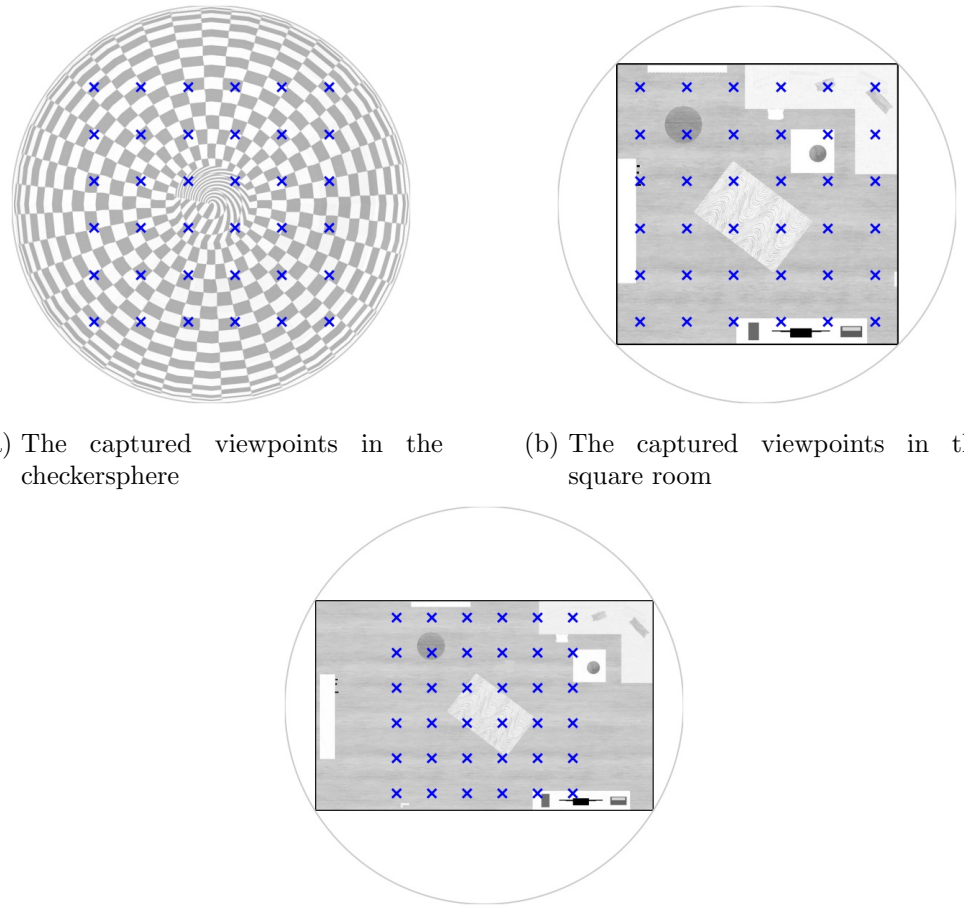


(a) Latlong image from the center of the room



(b) Top view

Figure 4.6.: Overview of the “oblong room”



(a) The captured viewpoints in the checkersphere

(b) The captured viewpoints in the square room

(c) The captured viewpoints in the oblong room

Figure 4.7.: The grid of captured viewpoints in each scene, including the proxy geometry (not to scale)

success of the optical flow algorithm is a prerequisite for the success of the 2 DoF algorithm with flow-based blending.

However, the focus of this evaluation is not the accuracy of an arbitrary optical flow algorithm. In the best case, it would be possible to emulate "perfect" optical flow, thus decoupling the success of the optical flow from the success of the flow-based blending. While this is practically impossible for real scenes, virtual scenes theoretically contain all necessary information for retrieving "ground truth" optical flow. Blender, for example, is capable of "rendering" motion vectors using its vector speed render pass, which calculates the movement between points from one frame to the next in pixel space. The result is a motion vector field, which corresponds to the result of a "classic" optical flow algorithm. This, in the best case, "ground truth" optical flow was first presented by Butler et al. [BWSB12] as a benchmark for optical flow algorithms.

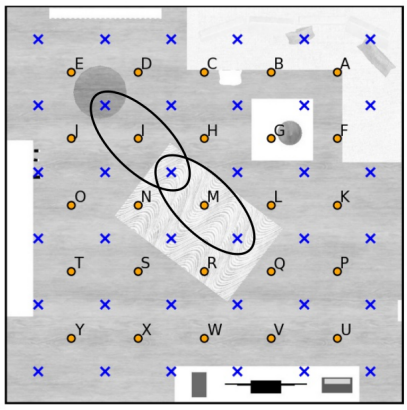
In order to demonstrate the improvement of Blender optical flow compared to Farnebäck optical flow, which is the optical flow algorithm used in the implementation, Figure 4.8a shows a scene setup in which 25 viewpoints (A-Y) are synthesized using Farnebäck and Blender optical flow at the interpolation distance $\delta = 0.5$ between captured points. Figure 4.8b shows the improvement of the results using Blender optical flow: In all but one case for the L1 values, and in the majority of the SSIM values the Blender optical flow improved the error values. The visual difference of the 1 DoF interpolation is clear: Figure 4.8c shows the viewpoint "I" interpolated using Farnebäck optical flow, and Figure 4.8d the same viewpoint using Blender optical flow. There are distincly fewer artefacts with Blender optical flow, for example the rug and the couch both have a much more distinct outline, and the bookshelf is also clearer. Figure 4.8e and Figure 4.8f show the same for interpolated viewpoint "M". In this case, the TV cabinet and the rug are much clearer in the Blender optical flow version (Figure 4.8f).

It is notable that using Blender's optical flow tends to improve the results, compared to Farnebäck's algorithm, but that does not mean that the resulting optical flow is necessarily accurate. For example, the bookshelf in the right and middle faces in Figure 4.8d still shows warping and doubling effects, indicating that there are still some inaccuracies. The same is true for the coffee table and the blue round table in the bottom face. There are several possible reasons for this, mostly based on the fact that the process in Blender, like most optical flow algorithms, is designed for frame-to-frame use, and has in this case been "misused" for distances that are infeasible for an actual animation. Nevertheless, no definitive explanation can be made at this point, since this would require in-depth understanding of Blender's vector speed render pass, which is outside of the scope of this thesis. Based on the results shown in Figure 4.8, and experience gained from testing both variants, the Blender optical flow is used for the limit evaluation, because, even though it is not perfect, it still seems to mostly yield better results than Scipy's implementation of Farnebäck's optical flow algorithm and as such decouples (to a degree) the success of the 2 DoF synthesis with flow-based blending from the success of the optical flow algorithm.

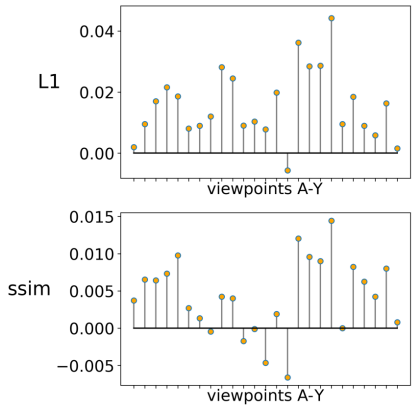
4.3.3. Scenarios and Results

Using the generated scenes and optical flow, as well as the parameters defined for this evaluation, it is now possible to design, test, and evaluate the following scenarios:

- "Different Scene Geometries"



(a) The viewpoints for testing optical flow

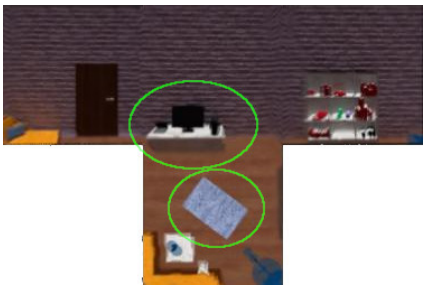


(b) The improvement of error values using Blender optical flow vs Farnebäck optical flow

(c) 1 DoF interpolated viewpoint "I" using Farnebäck optical flow

(d) 1 DoF interpolated viewpoint "I" using Blender optical flow

(e) 1 DoF interpolated viewpoint "M" using Farnebäck optical flow



(f) 1 DoF interpolated viewpoint "M" using Blender optical flow

Figure 4.8.: Comparing 1 DoF interpolation results using Farnebäck to results using Blender optical flow

- "Density of Captured Viewpoints"
- "Position of Synthesized Viewpoints Relative to Captured Viewpoints"

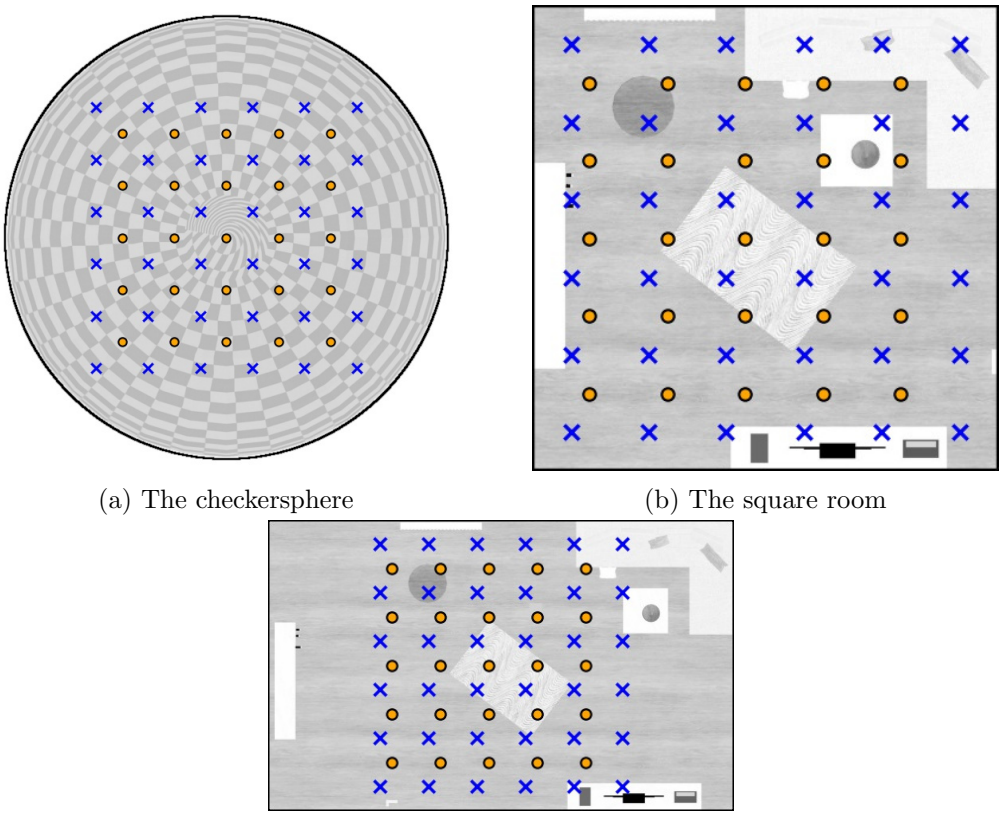
This section presents the scenes, viewpoint setups, and tested parameters used in each of these scenarios, as well as the results of the tests, which are evaluated using distribution comparison, scene analysis, and sample inspection, as described in Section 4.2. First the effect of the tested parameter on the regular blending results is examined, then the effect on the flow-based blending results is examined, and finally, the the results of the flow-based blending are compared to the results of the regular blending.

Different Scene Geometries

The first parameter to be examined is the effect of different scenes on the accuracy of the results. There are two attributes of a scene that may have an influence on the result: the basic shape of the scene (e.g. sphere, cube, rectangular prism, or arbitrary polygon), and the objects within the scene. All the scenes presented in 4.3.1 are used, as they differ both in their basic shape, as well as in the arrangement of the objects within, although the square room and the oblong room are more similar to each other than to the checkersphere.

The arrangement of captured viewpoints in all the scenes is identical (a 6x6 grid with a spacing of 60cm), and 25 viewpoints are synthesized in each scene (Figure 4.9). The synthesized points are near or in the center of each grid cell, since these are the areas where the deviation angles are the highest, and where the largest artefacts for the regular blending are expected to emerge. In the square and oblong rooms, the synthesized points are slightly offset from the center of each grid cell (Figure 4.9). The offset is important in order to test actual 2 DoF synthesis, instead of just 1 DoF interpolation, since synthesizing in the center of a grid cell could be done with only 1 DoF interpolation (e.g. by interpolating by 0.5 between the top right to the bottom left captured viewpoint, as was done in Section 4.3.2 for testing Blender optical flow). No offset was used in the checkersphere scene, since the checkersphere scene is expected to have excellent results for the regular blending (since the proxy geometry and scene geometry are identical). In this case it is more interesting to use one of the presumably best positions for the flow-based blending to see how well it holds up in comparison.

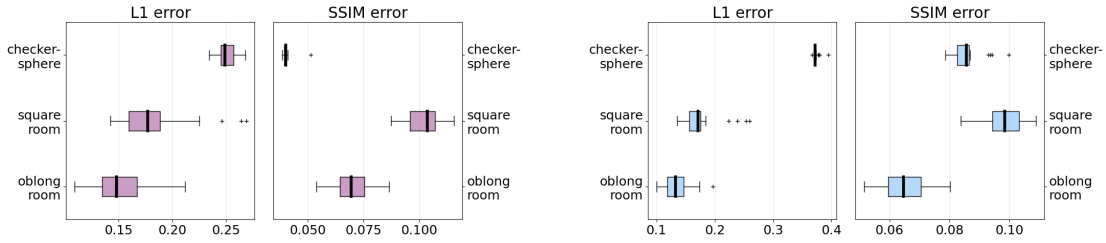
Regular Blending Results Figure 4.10a shows the distributions of the error values for the regular blending results in the three scenes. The most striking feature of this distribution is that the checkersphere results show the highest error values for the L1 error, whereas they show the lowest values for the SSIM error. At first, this seems surprising, both because the error metrics do not "agree", as well as because the results of the regular blending are expected to be very good since the scene geometry is identical to the proxy geometry. The reason for the ambiguous results is the sensitivity of the L1 metric to different color values. L1 errors on the RGB images of the checkersphere will produce a higher value in general because of the checkerboard texture: The difference between a dark blue pixel of a dark checkerboard field, and a white pixel on a white checkerboard field is close to 1, whereas the difference between a dark brown pixel and a dark purple pixel (e.g. between the door and the wall in one of the rooms) will produce a lower error value, although the distortion or displacement may be identical. The SSIM error metric, which does not take the color values of the pixels into account, produces a distribution that is much closer to the expected result:



(a) The checkersphere (b) The square room

(c) The oblong room

Figure 4.9.: The captured (blue) and synthesized (orange) viewpoints in the different scenes (scenes are not to scale)



(a) Regular blending (b) Flow-based blending

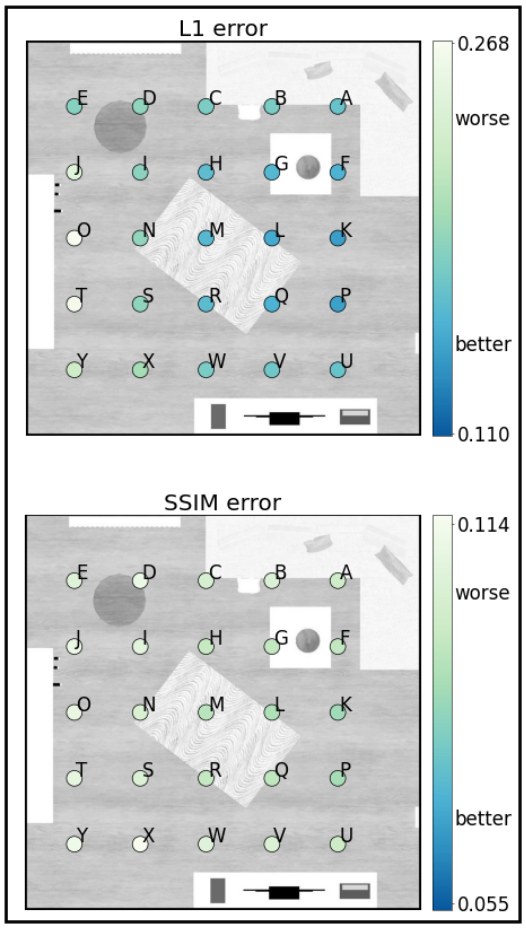
Figure 4.10.: Comparing the distributions of the results in different scenes

Since the scene geometry is identical to the proxy geometry, the result of the reprojection should be almost perfect. And in fact, when visually comparing the results of a synthesized viewpoint to the ground truth (Figure A.2, page 83), the only difference is a little bit of blurriness due to sampling differences. The L1 difference depicted in the right column of Figure A.2 shows how the blurriness caused the high L1 error.

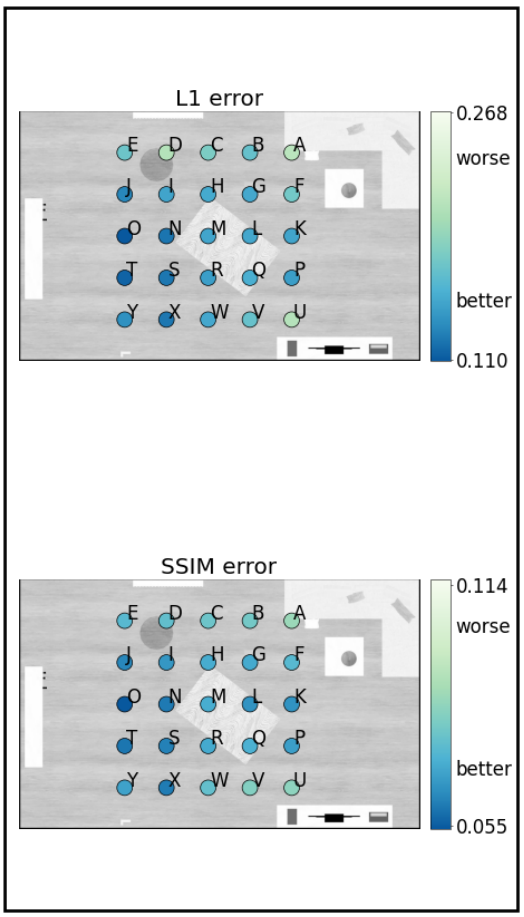
The trend of the error metrics of the other two rooms is consistent: Both the L1 and SSIM error values of the square room are generally higher than those of the oblong room. In this case, comparing the RGB color values is more reliant, since both rooms use the same textures. The results however, are surprising: Although the basic geometry of the square room is closer to the proxy geometry, the error values from the square room are generally higher than those from the oblong room. A closer scene analysis (Figure 4.11) shows the probable reason for this: The error values in the square room (Figure 4.11a) are particularly high near the bookshelf on the left side of the room. Comparing the synthesized images at location "O" (right in front of the bookshelf) for the two scenes (Figure A.3, page 84) confirms this impression: The bookshelf is so close to viewpoint "O" in the square room, that there are extreme inaccuracies in the synthesis due to large deviation angles. In the oblong room, the bookshelf is much further away, making the devation angles much smaller and the result more accurate. The larger net distance to the walls and some of the objects in the oblong room is the likely reason for the lower error results throughout the scene.

Flow-based Blending Results The distribution of error values of the synthesis using flow-based blending in the different scenes (Figure 4.10b) show similar tendencies as the error values of the regular blending, in that the L1 values of the checkersphere are very high, and the results of the oblong room generally have lower values than those of the square room. One exception is the SSIM error value of the checkersphere: Where the SSIM error in the checkersphere scene is significantly lower than the other two for the regular blending results, in the flow-based blending results, the SSIM error yields worse results than the oblong room, but still mostly better than the square room. The worse performance of the flow-based blending in the checkersphere is likely due to the fact that the flow-based blending introduces some inaccuracies, for example due to imperfect optical flow, or ray approximation. Figure A.2 (page 83) shows slightly higher inaccuracies for the flow result, as well as some artefacts and noise, that are the likeliest causes for the higher error values.

In the other cases, the results of the flow-based blending mirror those of the regular blending: The error values in the oblong room are generally lower than those in the square room. Looking at the example viewpoint "O" in Figure A.4 (page 85) shows that, like in the case of the regular blending, this is due to the relative postion of the walls and objects to the points: like the regular blending, the flow-based blending also performs a lot worse when there is a detailed object in close proximity. In the case of the regular blending, the reason for this are the large deviation angles leading to inaccurate reprojections. Generally, the flow-based blending should improve this problem, however, in this case it did not (or not considerably), which is mostly due to the optical flow calculation failing in proximity to the bookshelf, which was demonstrated in Section 4.3.2. If the optical flow algorithm does not produce accurate results, the flow-based blending will not, either. The result in the square room in Figure A.4 clearly shows the effect of the failed optical flow on the synthesized image. The normally straight lines of the books are warped (marked in magenta), but a comparison to the ground truth shows that they are warped in a way that bears no resemblance to the



(a) Results of regular blending in the square room



(b) Results of regular blending in the oblong room

Figure 4.11.: Scene analysis of regular blending results in the square and oblong rooms

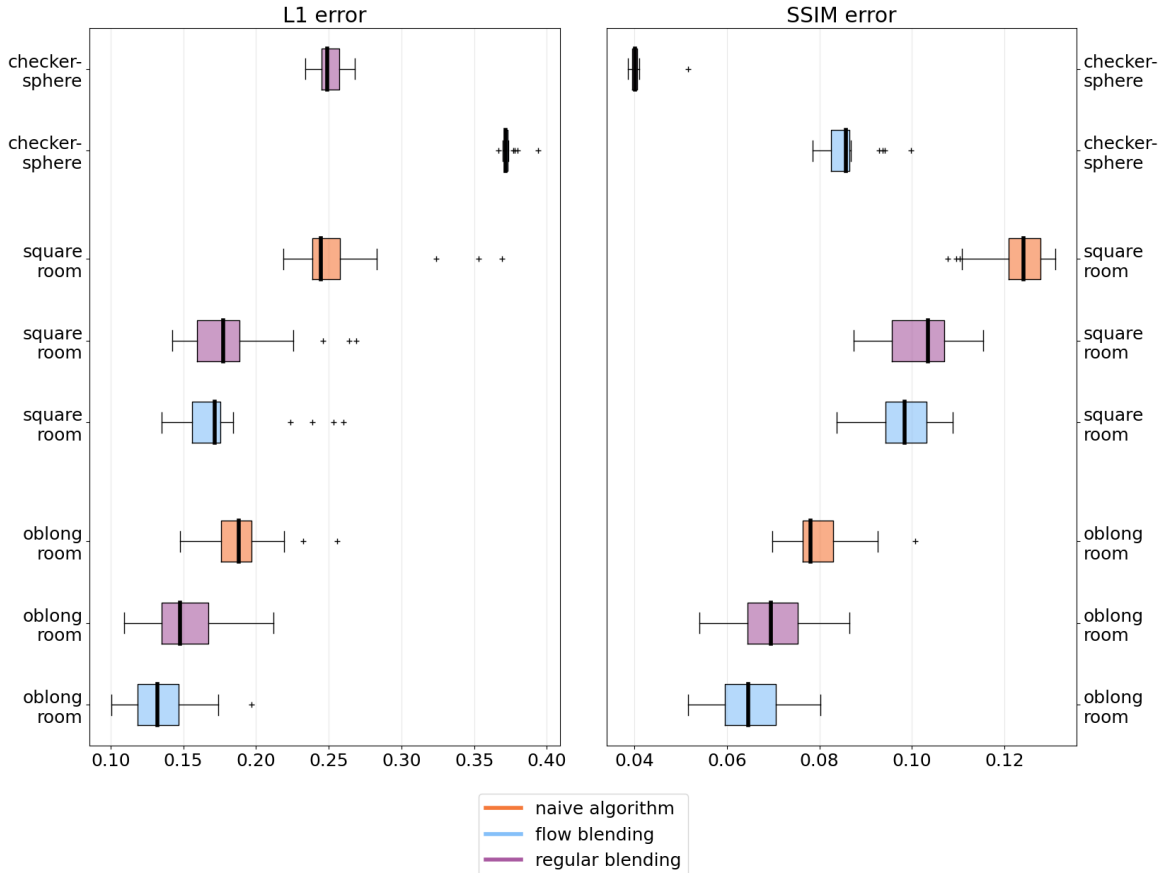


Figure 4.12.: Comparing the distribution of the results in different scenes for the naïve algorithm, regular blending, and flow-based blending

actual position or shape.

Comparing Regular Blending to Flow-based Blending Results The distributions of the error values of both the regular blending (purple) and the flow-based blending (blue) are shown in Figure 4.12, as well as the error value distribution of the naïve algorithm (orange) for comparison[4]. The graph shows that the error values of the results of the flow-based blending are generally slightly better than those of the regular blending (except in the case of the checkersphere) and that the error values of the results of the regular blending tend to be distinctly better than those of the naïve algorithm.

In the case of the checkersphere, where the scene geometry is identical to the proxy geometry, the regular blending distinctly outperforms the flow-based blending. However, this is to be expected. The goal of the flow-based blending approach aims to improve problems that arise due to the difference between the real and proxy geometries, and it introduces possible inaccuracies (viewpoint selection, optical flow, ray approximation) in order to do so. Therefore, it is not surprising that the results are less accurate than the results of the regular blending.

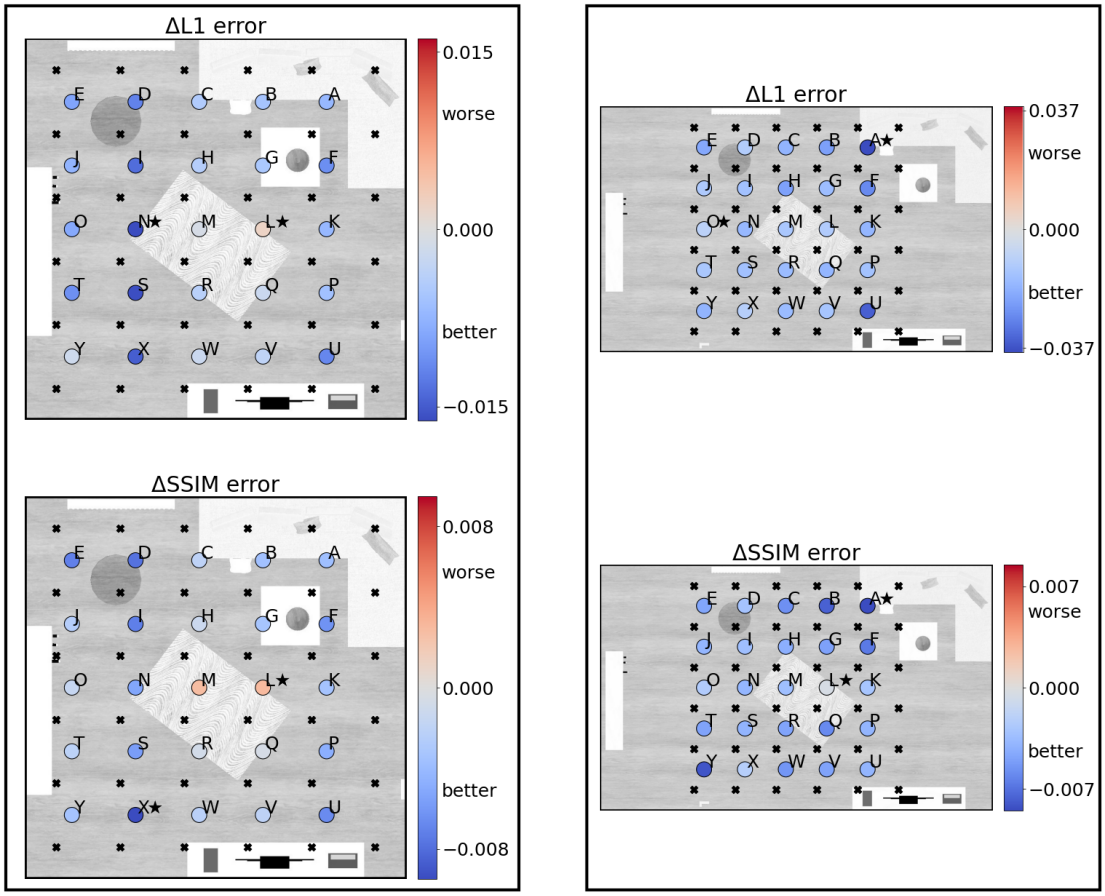
[4]The naïve algorithm error values of the checkersphere were omitted because they were so much higher than the other values that the scale of the plot became too small.

For the square room and the oblong room, the $\Delta L1$ and $\Delta SSIM$ of the flow-based blending versus the regular blending are shown in Figure 4.13. Positive $\Delta L1$ and $\Delta SSIM$ values (red) signify that the flow-based blending produced a *worse* result than the regular blending, and negative $\Delta L1$ and $\Delta SSIM$ (blue) signify that the flow-based blending produced a *better* result than the regular blending, with stars marking the highest and lowest values. At a glance, it is clear that in all cases in the oblong room, and in all but one or two cases in the square room (depending on the metric) the flow-based blending improves the result based on the L1 and SSIM error metrics. However, the L1 value range is larger in the oblong room than it is in the square room (max. improvement of 0.037, compared to 0.015 in the square room), which most likely indicates more visible improvements in the oblong room.

Figure A.5 (page 86) shows the viewpoint with the lowest (best improved) $\Delta L1$ difference in the square scene. The most visible difference is the rug on the bottom face: In the regular blending result, it shows some ghosting (doubled, offset edges), which has been mostly fixed in the flow-blending result. Other than that, the white coffee table with the blue bowl in the left face is slightly blurrier, but covers a more accurate area than the result of the regular blending, and the bottom part of the bookshelf is more accurate. Otherwise the two results are very similar visually. For the "worst improved" viewpoint L (Figure A.6, page 87) the regular and flow results are also visually very similar. The most visible differences are that the rug shows some ghosting artefacts in the regular blending result, which are gone in the flow-based result. However, the flow-based blending introduced some new artefacts, namely a sharp discontinuity and offset on the rug, and a distorted edge on the coffee table.

Figure A.7 (page 88) shows viewpoint "A" with the best improvement in the oblong scene. Here, the most visible difference is the couch in the left and bottom faces. Where the inner edge of the couch shows severe ghosting artefacts in the regular image, it is much cleaner in the flow-based result. The rug also covers a more accurate area of the image. However, the flow-based blending also introduces new artefacts, for example in the bottom face, where a part of orange couch appears detached from the rest of the couch. The "worst improved" viewpoint L (Figure A.8) shows even more of these artefacts: The rug in the bottom face has a few extreme discontinuities, as well as the coffee table in the left face (although the coffee table is positioned more accurately in the flow-based blending result). The severe discontinuities are caused by the selection of input viewpoints for flow-based blending: For each ray of the synthesized image, two input viewpoints are selected based on their deviation angle, and whether they are "on either side" of the ray in question (see Section 3.1.3). This means that while traversing the rays of synthesized viewpoint there comes a point where the selected input viewpoints for the 1 DoF interpolation A and B suddenly switch to B and C (this will be described in more detail in Section 4.3.4). In these cases, the less accurate the reprojection step is, the more extreme the discontinuity at that place in the image seems to be. Although these discontinuities are visually extremely irritating, it is possible that they do not have a large impact on the error metrics.

Scenario Synopsis The goal of this scenario was to evaluate the effects of the scene geometry on the result for regular and flow-based blending. This includes the details of the geometry given by the shape and placement of objects within the scene, as well as the general shape of the scene. Concerning the geometry details, i.e. the objects within the scene, the results of the scenario are fairly clear: both blending techniques perform better, the further away specific objects are from the synthesized point. The closer the synthesized point gets to



(a) Square room (b) Oblong room

Figure 4.13.: $\Delta L1$ and $\Delta SSIM$, "improvement" of flow-based blending over regular blending results in the square and oblong rooms. The stars denote the best and worst cases.

an object, the more ghosting and doubling artefacts and positional inaccuracies show up in the regular blending results. In some of these cases, the flow-based blending synthesizes a more accurate image, showing fewer positional inaccuracies and correctly synthsizing some of the object shapes. However, the flow-based blending also introduced some new artefacts, predominantly abrupt discontinuities. These discontinuities were not necessarily recognized by the error metrics and needed to be identified by visual inspection. In the case of extreme proximity to an object (e.g. in front of the bookshelf), both the regular blending and the flow-based blending produced highly inaccurate results. In the case of the flow-based blending, this was due to inaccurate optical flow. However, even though the optical flow was inaccurate, the flow-based blending result was not worse (in terms of the error metrics) than the regular blending result.

As for the impact of the general shape, the results are not very significant, due to the choice of the oblong and square rooms along with the chosen captured and synthesized viewpoints. The different distances of the objects to the synthesized viewpoints within the two scenes overpowered most effects that the different basic shape might have had. It was clear, however, that in the case where the scene geometry matched the proxy geometry, the flow-based blending performed worse than the regular blending.

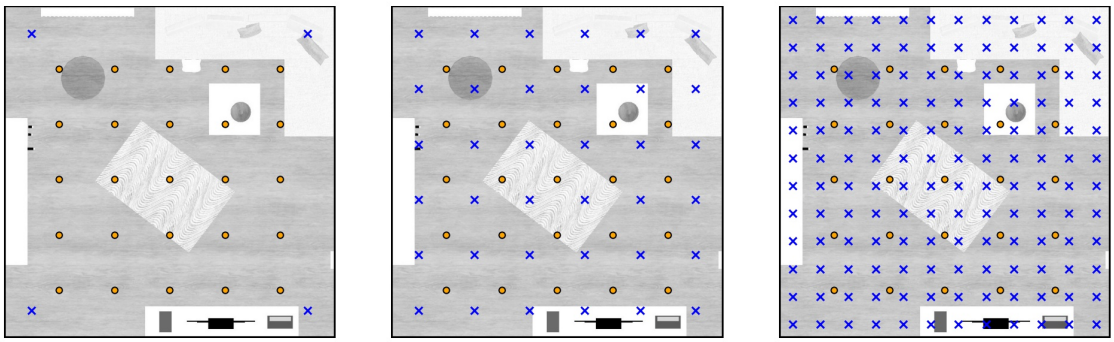
Density of the Captured Viewpoints

The previous scenario demonstrated that close proximity of a synthesized viewpoint to an object can have a strong adverse effect on the accuracy of the result. A possible way to alleviate this problem is to improve the density of the captured viewpoints near objects. The scenario presented in this section explores the impact of viewpoint density on the accuracy of the results.

To test the effect of viewpoint density, the square room is used, as the viewpoint grid covers the entire area of the room, which guarantees higher proximity to all the objects. Three different versions of the captured viewpoint grid are used: a 2x2 grid with a spacing of approximately 2.3m (Figure 4.14a), the 6x6 grid with a spacing of approximately 60cm, as was used in the previous scenario (Figure 4.14b), and a 12x12 grid with a spacing of approximately 30cm (Figure 4.14c). The 2x2 grid was chosen since it is the minimal grid to cover the entire room, and the 12x12 grid was chosen, since it halves the distance of the 6x6 grid and as such, retains the relative position of the captured viewpoints compared to the synthesized viewpoints. If a different grid was used, for example 10x10, some synthesized viewpoints would be closer to captured viewpoints than other synthesized viewpoints, which may have an effect on the overall results. Like in the previous scenario, 25 viewpoints are synthesized, located near the center of each grid cell. They are offset slightly from the exact center, like in the previous scenario, to demonstrate true 2 DoF synthesis.

Regular Blending Results Figure 4.15a shows the general distribution of the results of the regular blending with varying viewpoint densities. Unsurprisingly, the accuracy of the results improves, the higher the density of the input viewpoints is. In the 6x6 and the 12x12 setup, there are several outliers with a higher error value for the L1 error, which are likely due to the proximity of some points to the bookshelf. In the 2x2 setup, there are no significant outliers, possibly because the error values are more uniform throughout the scene.

A look at the scene visualization in Figure 4.16 confirms these assumptions: The outliers in the 6x6 and 12x12 setups are caused by the bookshelf on the left side of the room. In the

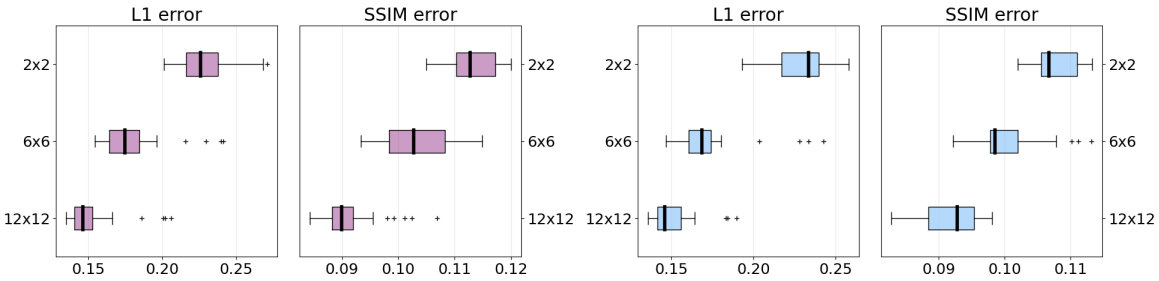


(a) A density of 4 viewpoints (2x2 grid)

(b) A density of 36 viewpoints (6x6 grid)

(c) A density of 144 viewpoints (12x12 grid)

Figure 4.14.: The different captured viewpoint densities (blue) in the square room with the synthesized viewpoints (orange)



(a) Regular blending

(b) Flow-based blending

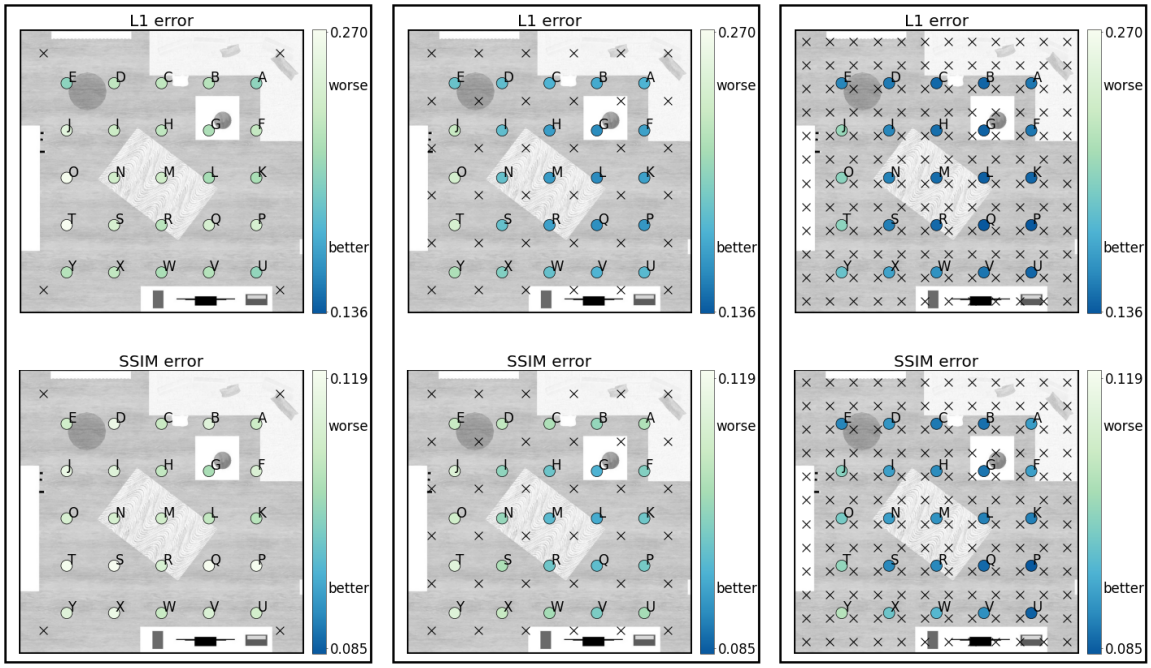
Figure 4.15.: Comparing the distributions of the results in the square room with different densities separately

case of the 2x2 setup, the error values are fairly high throughout, compared to the other two setups, which is the reason that the viewpoints near the bookshelf do not register as outliers. Other than the area around the bookshelf, both the 6x6 and the 12x12 scenes have a fairly uniform distribution for the L1 error. For the SSIM error, the areas close to the walls have a slightly higher error than those in the middle of the scene. This is in line with the results of the previous scenario, where the error values near objects or walls tended to be higher.

Although the error values near objects and walls are slightly higher in both the 6x6 and 12x12 setups, the improvement when going from the 6x6 to the 12x12 density is also higher in these areas, as can be seen in Figure 4.17a. This figure shows the improvement when using the 12x12 scene versus the 6x6 scene. In areas near walls (e.g. the top row for both metrics, viewpoints A-E), or objects (in front of the bookshelf for L1, viewpoints O and T, or in front of the TV screen for L1 and SSIM, viewpoints U,V,W), the improvement tends to be higher than near the center of the scene (e.g. viewpoints H, G, M, L).

In order to understand more about how the density affects ghosting and other artefacts in the images, one of the better and one of the worse results for all of the setups is examined visually. According to the general tendencies of the values in the different setups (Figure 4.16), synthesized viewpoint "T" near the bottom left of the room, next to the bookshelf has a comparatively high error value in both metrics for all three scenes and synthesized viewpoint "G", above the coffee table towards the top right corner has a comparatively low error value for all of the scenes. Figure A.9 (page 90) shows why T has a high error value: for the 2x2 scene, while most of the scene is moderately accurate (the objects are in approximately the right place), the bookshelf is shown from the wrong perspective, i.e. from the side versus from the front. The reason for this is that the perspective of the bookshelf of the viewpoint used for reprojection was from the side. Using only this information, it is impossible to reconstruct the front of the bookshelf. One other effect of the fairly large jump in perspective using only regular blending are the warped walls. This is due to using the sphere as proxy geometry. The warping problems are much less visible in the 6x6 and 12x12 results, since the reprojection jumps are much smaller due to the captured viewpoints being much closer together. While the L1 difference images on the right do show an improvement of the accuracy in the 12x12 image over the 6x6 image, it is not as visually obvious in the result image as might be expected from halving the distance between the viewpoints. The bookshelf still has many inaccuracies due to its proximity and high detail. However, while the difference between the 12x12 and the 6x6 image is not very evident in one of the worse results, Figure A.10 (page 91) better illustrates the effect of the denser grid. The 2x2 image excellently demostrates the problem with using input images with large deviation angles in conjunction with a proxy geometry that is not identical to the scene geometry: The rays used to synthesize this image captured the coffee table at different positions and the reprojection does not account for this, since it only approximates the scene geometry by using the proxy geometry. As a result, the coffe table appears four times. In this example, the reprojection errors decrease visibly as the density increases: in the 6x6 image, a slight doubling effect of the coffe table is still visible, whereas in the 12x12 image, the table only appears slightly blurry.

Flow-based Blending Results The error values of the flow-based blending results (Figure 4.15) show similar tendencies as the regular blending results. Like for the regular blending, the 6x6 and 12x12 setups have several outliers in the L1 metric, but otherwise a fairly



(a) Results of regular blending with a 2x2 grid

(b) Results of regular blending with a 6x6 grid

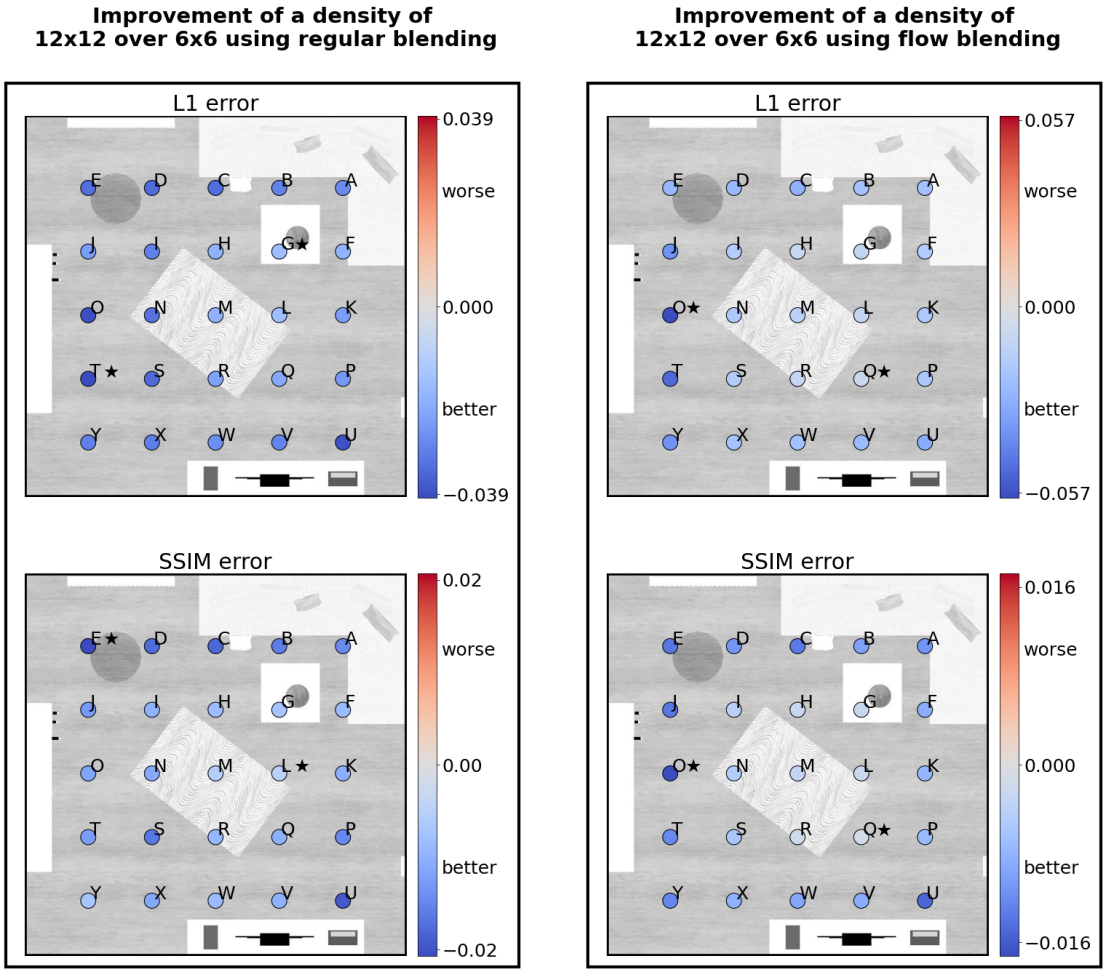
(c) Results of regular blending with a 12x12 grid

Figure 4.16.: Scene analysis of the regular blending results in the square room with different densities

close distribution. The scene visualization (Figure 4.18 also shows a similar pattern: In the 2x2 scene, the error values are generally high, whereas in the 6x6 and 12x12 scenes, the error values are high near the bookshelf, but generally similar everywhere else. Like in the regular blending scene, the improvement of the values in the 12x12 setup compared to the 6x6 setup is slightly higher near the walls. Since the tendencies are very similar as in the regular blending, it is more interesting to examine the comparison of the regular blending to the flow-based blending, instead of the flow-based blending by itself, as there do not seem to be any findings other than the improvement of the results with higher density, especially near the walls.

Comparing Regular Blending to Flow-based Blending Results The tendencies of the regular blending results compared to the flow-blending results shown in Figure 4.19 are not immediately apparent. Only the results of the 6x6 scene show a consistent improvement of the flow-based results compared to the regular results. For the 2x2 scene, the median L1 error is higher for the flow-based blending, while the error range is lower. The SSIM error for the 2x2 scene, however, is clearly lower for the flow-based blending. In the 12x12 scene, the L1 error shows an almost identical distribution, whereas for the SSIM result, the general distribution of the flow-based blending is wider, which implies that some results were improved while others were worsened.

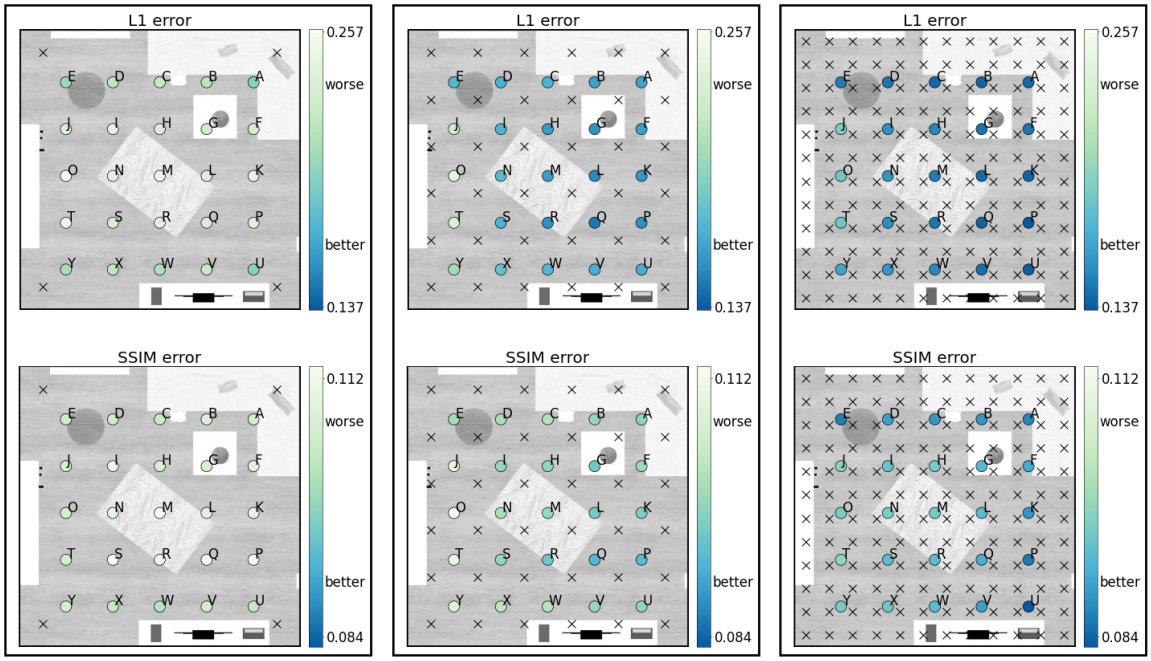
Figure 4.20 shows the improvements of the flow-based blending over the regular blending for the 2x2 and the 12x12 scenes. Both the 2x2 and the 12x12 scenes show improved values nearer to the walls, and a slightly worse values near the center of the room. However,



(a) Improvement for regular blending

(b) Improvement for flow-based blending

Figure 4.17.: Improvement of results using 12x12 density compared to 6x6 density



(a) Results of flow-based blending with a 2x2 grid

(b) Results of flow-based blending with a 6x6 grid

(c) Results of flow-based blending with a 12x12 grid

Figure 4.18.: Scene analysis of the flow-based blending results in the square room with different densities

although this overview shows similar tendencies, a look into the synthesized images shows some of the fallacies of the metrics.

For the 2x2 scene, the regular blending produced relatively inaccurate results, due to the extreme perspective changes. The flow-based blending relies on optical flow, which does not handle large displacements well (even with Blender optical flow). The results show that in the 2x2 scene, the optical flow algorithm did not produce very accurate results, which is visible in both the "best" and "worst improvement". Figure A.11 (page 92) The "best improvement" image (Figure A.11a) is the image that was one of the worst rated for the regular blending, since it did not correctly show the bookshelf. In the flow-based version, a blurry shadow of the bookshelf is visible, however, the whole image shows blurry distortions of all of the objects. So although the metrics show an improved value, visually it is much harder to distinguish the different objects. The same holds true for the "worst improvement" image, where both metrics measured a slight increase in error value from the regular to the flow-based result. Figure A.11b shows point "K", and while the regular blending result shows extreme artefacts due to doubling, the flow-based version, due to the failure of optical flow, exacerbates the problem: some objects, such as the rug, are reduced to a blurry spot, whereas others, such as the coffee table, suffer from even worse blurring. In summary, it is safe to say that the captured viewpoints in the 2x2 density are too far apart to produce any satisfying results.

As for the 12x12 scene, The error metrics show a slight decrease in error values near the walls, and a slight increase of error values near the middle of the scene. However, when looking at the best and worst improvement in Figure A.12 (page 93, very little difference

is actually visible. In the best improvement image (Figure A.12a), the flow-based blending slightly improves the accuracy of the bookshelf in the right and back faces, although some ghosting artefacts remain. It also does not display an artefact present in the front face of the regular blending result (a white blurry spot, which appears due to the use of an input viewpoint that had a deviation angle of zero at that position, but seems to have captured a part of the bookshelf with that ray). However, other than that, the two images are visually extremely similar. The same holds true for the "worst improvement" (Figure A.12b): Looking at the L1 difference images, it is hard to see any difference between the two, but when comparing the synthesized images, it is noticeable that the flow-based blending removed the ghosting artefacts on the rug and the coffee table. This change is so slight that it is not surprising that it did not register with the metrics. Visually however, the flow-based blending result seems more accurate, since it improves the ghosting artefacts. An explanation for the lower rating by the SSIM metric could be that some of the straight lines that have ghosting artefacts in the regular blending result (e.g. the right edge of the coffee table) are rated higher than the same lines in the flow-based blending result, that do not display ghosting artefacts, but are slightly warped. Since the SSIM metric uses structural elements, it is possible that it prefers the version with the ghosting artefacts.

The 6x6 scene is more straightforward. Figure 4.21 shows the improvement of the flow-based blending results over the regular blending results. As this is very similar to the improvement in the previous scenario for the square room (Figure 4.13a), where an improvement was achieved by the flow-based blending, the images are not examined in detail.

In summary it can be said that for the tested room (and presumably in most other cases, as well), increasing the density of the captured viewpoints also increases the accuracy of the results. In the tested cases, the improvement was more apparent near objects and walls.

In the extreme case of the 2x2 density, both the regular and the flow-based blending lead to extreme artefacts, due to the large jumps in reprojection and the failure of the optical flow algorithm. For the 6x6 density, the results are unambiguous: both metrics produce lower error values for the flow-based blending than for the regular blending. sing the 12x12 density results in fairly low errors and few artefacts for both regular and flow-based blending. In this case, the metrics are ambiguous, since the difference between the two versions is very small. A visual evaluation of some of the samples indicates that the flow-based blending improves ghosting artefacts, which visually improves the images, even though the error metric may be slightly higher.

Position of Synthesized Viewpoints Relative to Captured Viewpoints

In most of the previous results, the flow-based blending showed an improvement over the regular blending. However, the locations of the synthesized viewpoints were chosen explicitly so that the regular blending would most likely have the most difficulty: areas near the center of the grid cells with comparably high deviation angles. Naturally, this is not the only location that comes in question for synthesis. In fact, all locations within the convex hull can potentially be synthesized by the 2 DoF algorithm. In order to gain a more general understanding of the impact of the relative positons of the captured viewpoints on regular versus flow-based blending, this scenario synthesizes a dense grid of viewpoints, instead of choosing a few select locations in the scene, like in the previous scenarios.

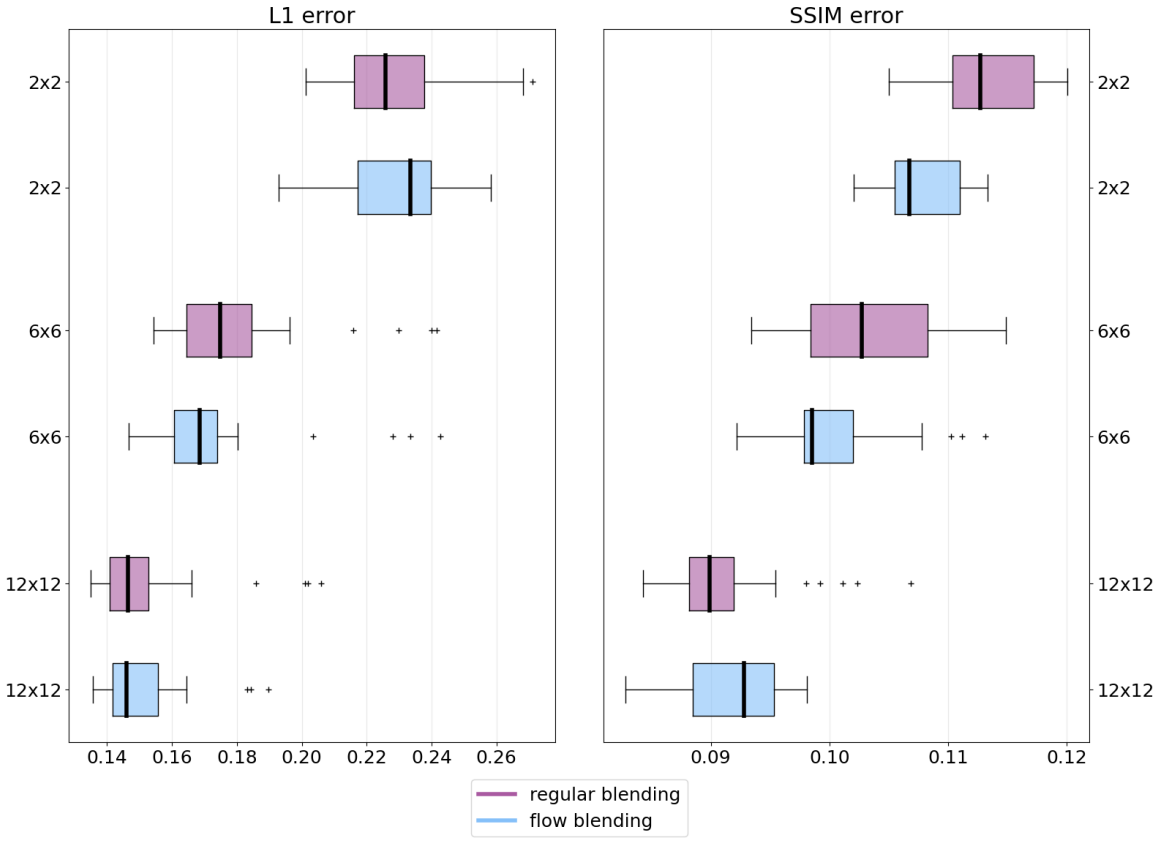
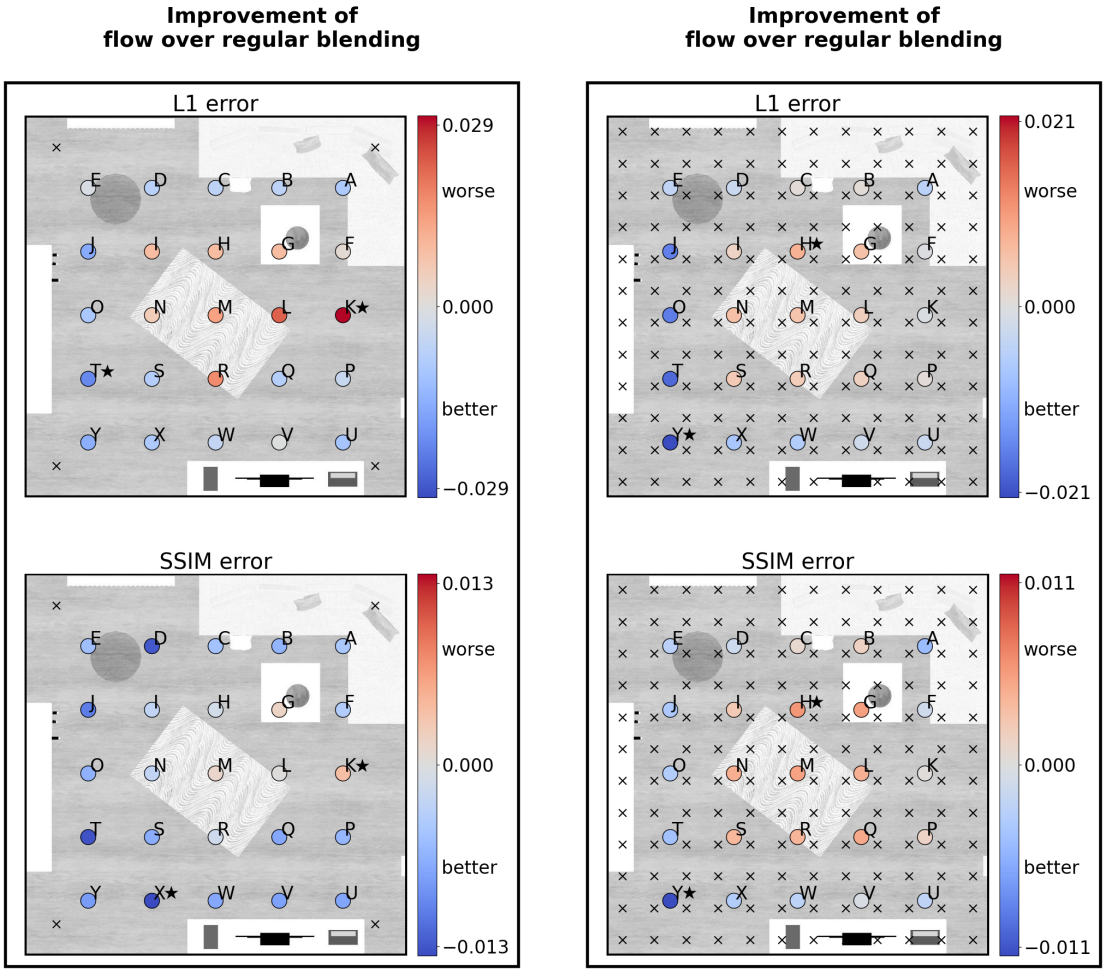


Figure 4.19.: Comparing the distributions of the results in the square room with different densities of captured viewpoints for the regular blending and flow-based blending



(a) Improvement in the 2x2 scene

(b) Improvement in the 12x12 scene

Figure 4.20.: Improvement of flow-based blending results over regular blending results in the 2x2 and 12x12 scenes

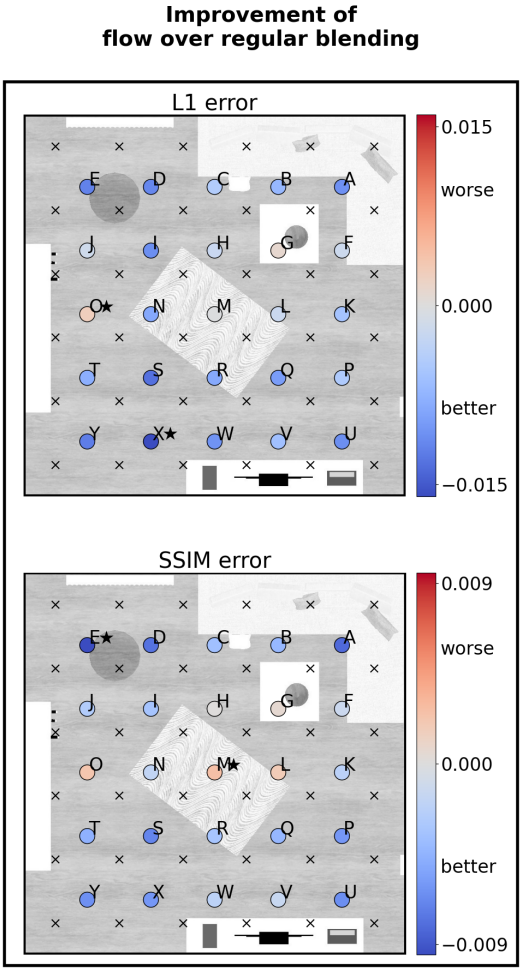


Figure 4.21.: Improvement of flow-based blending results over regular blending results in the 6x6 scene

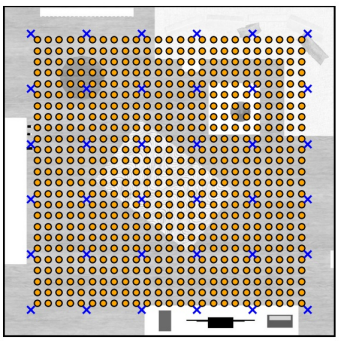


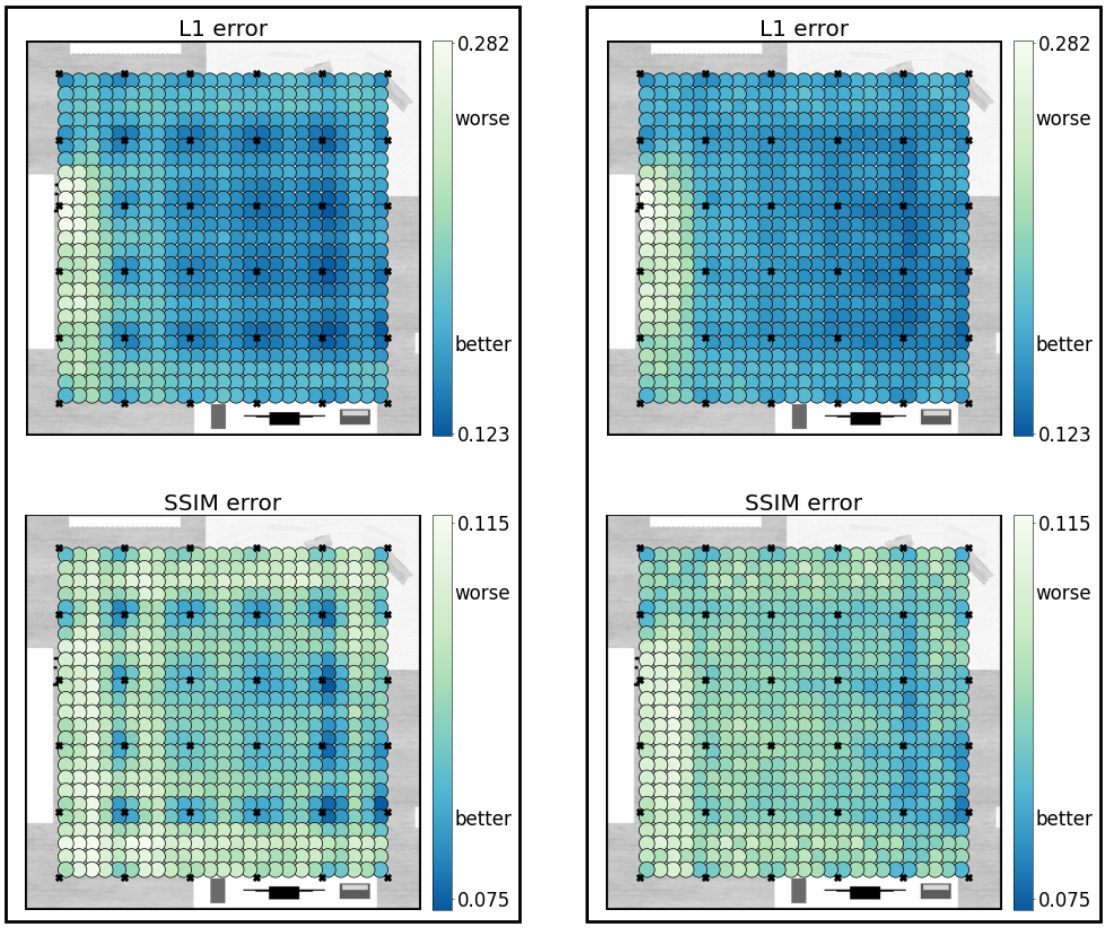
Figure 4.22.: The dense grid of synthesized viewpoints (orange) and the 6x6 grid of captured viewpoints (blue) in the square room

Based on the insights gained from the last two scenarios, the square room is used with a captured viewpoint density of 6x6. The square room gives the advantage of covering the whole possible space, and a density of 6x6 with a spacing of 60cm is a conceivable distance for extrapolation to real scenes. A grid of 25x25 viewpoints is synthesized, totaling 625 synthesized points (Figure 4.22).

Regular Blending Results In this scenario, only the position of the viewpoints within a single scene is examined, so the distribution can be inspected directly in the scene analysis visualization (Figure 4.23a). The dense coverage of synthesized viewpoints gives a fairly detailed picture of the effects of the location relative to the scene, and especially relative to the location of the captured viewpoints. It is immediately striking that the synthesized viewpoints in the close vicinity of a captured viewpoint have a distincly lower error value than viewpoints that are farther away. The exceptions to this are the synthesized viewpoints near the walls and near the bookshelf. The observation of higher error values near the bookshelf and walls are consistent with the observations in the previous scenarios. However, this scenario shows that the error values are higher near walls and objects, independent of whether the synthesized viewpoints are close to the captured viewpoints or not. This phenomenon is clearly visible in the L1 error visualization, but even more so in the SSIM error visualization, where the accuracy dropoff is even more extreme.

Flow-based Blending Results The error values of the flow-based blending results (Figure 4.23b) display a different pattern than those of the regular results: While the error values are lower in the vicinity of captured viewpoints, they are comparably low in the vertical and horizontal spaces between the captured viewpoints, as well. This implies that it does not make a big difference to the error values whether a synthesized viewpoint is very close to a captured viewpoint, or anywhere between a set of horizontally or vertically adjacent viewpoints. In general, the majority of the values is fairly similar, with clear outliers only visible near the bookshelf.

Comparing Regular Blending to Flow-based Blending Results Figure 4.24 shows the Δ L1 and SSIM error values when using the flow-based blending compared to the regular blending. Here, the results are clear: In general, the error values of synthesized points in



(a) Results of regular blending

(b) Results of flow-based blending

Figure 4.23.: Scene analysis of regular and flow-based blending results

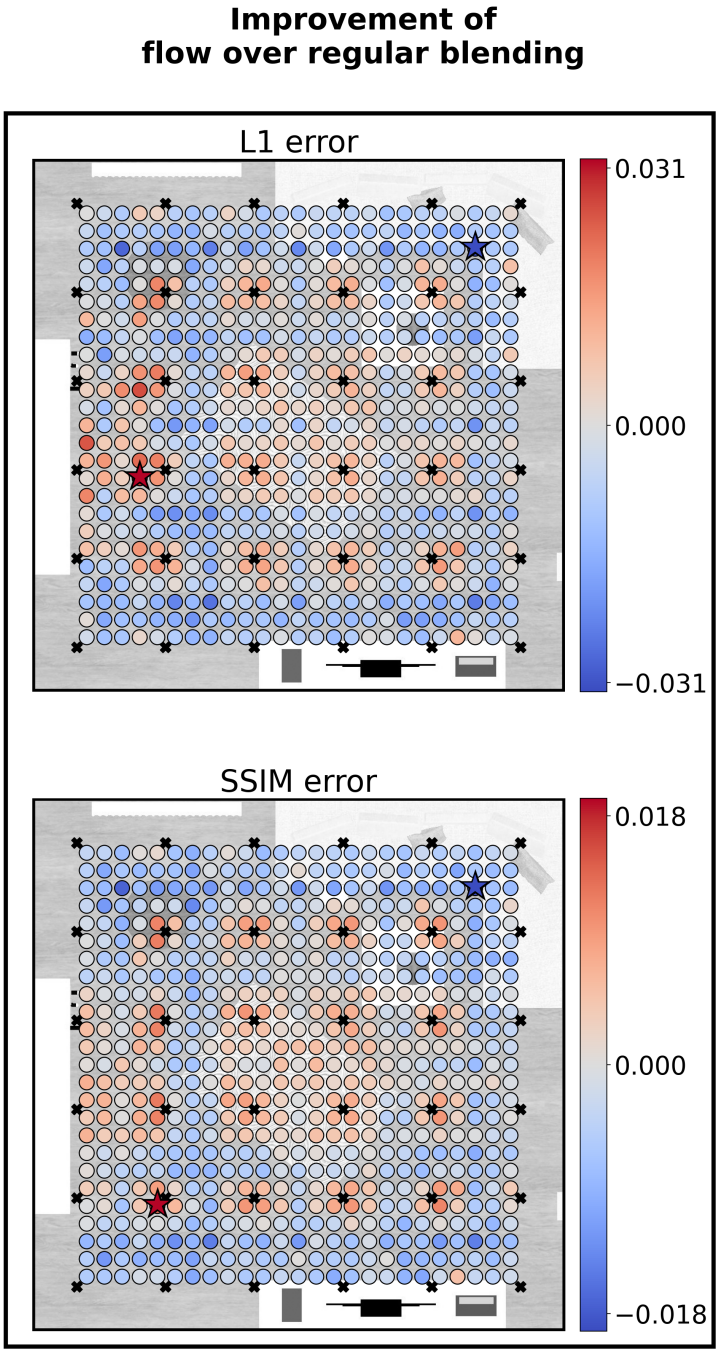


Figure 4.24.: Improvement of flow-based blending results over regular blending results for 625 synthesized images in the square room

close proximity to captured points go up (i.e. the accuracy gets worse) when using flow-based blending. An exception to this are the captured points near the walls of the room: In these cases, the flow-based blending improves the results in the majority of cases. Also, in the areas where the distance to the captured viewpoints is highest, the flow-based blending also performs better. An area where the results are ambiguous is the area near the bookshelf. Here, the improvement or degradation cannot be clearly attributed to the proximity to a captured viewpoint or the proximity of the bookshelf. It must be kept in mind, however, that in the case of the bookshelf, the Blender optical flow does not yield good, or even acceptable results, which has a direct, detrimental effect on the flow-based blending, so the results near the bookshelf should not be taken into strong consideration.

In many of the cases where the flow-based blending improves the result, it does so by shifting objects into more accurate positions. Figure A.13a shows one of these cases: The couch and the coffee table in the bottom face of the regular blending result are offset by a noticeable amount from the ground truth. The flow-based blending result shows almost no offset. It does introduce some new artefacts, for example blurriness and some discontinuities, which will be explained in Section 4.3.4. Figure A.13b shows a similar example, where the rug in the bottom face shows a ghosting effect in the regular result, as well as being slightly offset from the ground truth position. Again, the flow-based result improves these errors, in this case only introducing a slight visual discontinuity.

In the areas close to captured viewpoints, where the flow based blending performs worse than the regular blending, a look at the images themselves is required to attempt to explain this unexpected result. Figure A.14a shows an example where most of the image is very similar, except some distinctive artefacts in the bottom face (the blue table is misshapen) and in the left face (the coffee table displays a distinct edge and displacement). Figure A.14b shows a different example, where the difference between the flow-based and the regular results is barely visible: There is a slight distortion at the top of the door in the left face, and the coffee table has a slightly larger offset.

Some of the artefacts in the flow-based blending results are undoubtedly due to the approach (e.g. ray approximation). However it is also possible that part of the reason for the higher error values for the flow-based blending results are part be due to the implementation problems discussed in Section 3.2.5, since these only affect the flow-based blending and not the regular blending.

4.3.4. Discussion

The results of the 2 DoF synthesis with regular blending in the tested scenarios, were very much as expected. Areas where the proxy geometry deviated strongly from the actual geometry (e.g. near detailed objects) showed moderate to severe ghosting artefacts or displacements. Nonetheless, the results using regular blending generally had lower error values than the naïve algorithm, and in some cases, even than the flow-based blending.

In the majority of the tested cases, the flow-based blending produced lower error values than the regular blending. However, the evaluation also uncovered some factors that decreased the accuracy of the results, originating from the implementation details as well as the underlying approach:

- failure of the optical flow algorithm
- bugs in external libraries

- deviation-angle-based choice of input viewpoints with abrupt change
- ray approximation

The failure of the optical flow algorithm, notably in the cases where a viewpoint was in close proximity to the bookshelf, presumably degraded the results of the flow-based blending considerably. However, since it was not possible to test with perfect optical flow, it remains unclear whether even "perfect" optical flow could improve the accuracy of viewpoints that are extremely close to objects, as there would still be large occluded areas. However, for areas where there is little occlusion, different optical flow algorithms may

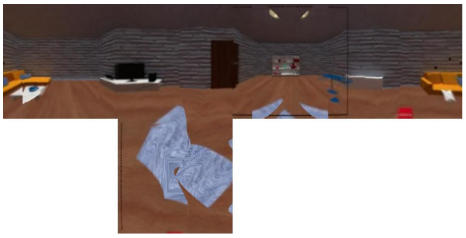
Furthermore, a factor that may have caused elevated error values is the due to reprojection problem in the external library Skylibs [Hol20]. This bug, which leads to slight pixel offsets and thin black lines in the flow-based blending results, exclusively (as the regular blending does not use the problematic operation), may have a noticeable impact on the Δ L1 and SSIM error values in cases where the results are very similar. However, this error is very difficult to quantify, since the flow-based blending result contains strips of a number of 1 DoF interpolated images, which display the problem to varying degrees. As a result, it can only be assumed that this bug may have a noticeable impact on the decrease of accuracy in the flow-based blending results.

The accuracy of the optical flow, and the bug in the external library are both implementation-related problems, as they are not part of the approach. However, there are also some problems that are intrinsic to the approach and thus need to be discussed in more detail.

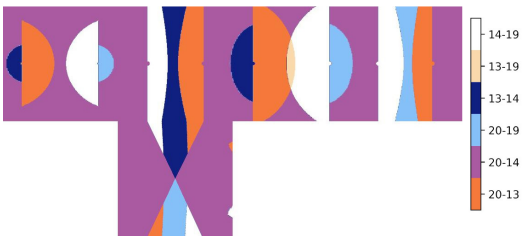
The most visible problem that is intrinsic to the approach is the choice of input viewpoints per pixel. This, in combination with the reprojection using the proxy geometry, is the reason for the abrupt jumps and discontinuities in the flow-based blending results. Figure 4.25 breaks down the cause of the problem step-by-step: Looking at an arbitrary synthesized viewpoint in the oblong room (Figure 4.25 that displays this problem, the abrupt discontinuities are clearly visible, for example in the rug in the bottom and right faces. In the synthesis process for each pixel of this image, two viewpoints A and B are chosen that are on either side of the ray corresponding to the pixel (Section 3.1.3). If more than one viewpoint is found on either side of the ray, the points with the smallest deviation angles are chosen. The interpolation distance between these points A and B is then calculated and the 1 DoF interpolation is performed. Finally the interpolated point is reprojected to the position of the synthesized viewpoint. This succession of operations is performed for each ray of the synthesized image. As a result, there are areas in the synthesized image, where, from one pixel to the next, a different set of input viewpoints is chosen. This leads to image areas ("panels") originating from a 1 DoF interpolation between different sets of input viewpoints (Figure 4.25b). When applying these panels to the synthesized image (Figure 4.25c), it becomes clear that these are the source of the abrupt continuities. Figure 4.25d shows a visualization of the scene with an approximation of the rays that are associated with each panel[5].

The switch between viewpoint sets within the image is not necessarily a problem in itself, given the 1 DoF interpolation worked well, and the reprojection was more or less accurate. However, in the cases where the proxy geometry deviates strongly from the actual geometry, the reprojection will introduce inaccuracies, which are exacerbated by the discontinuities. For example, the square room's geometry is closer to the proxy geometry. When comparing

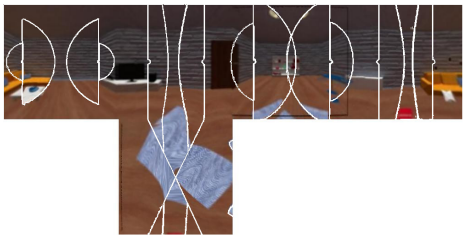
[5]The rays were inserted by hand, so they are not 100% accurate in terms of deviation angles



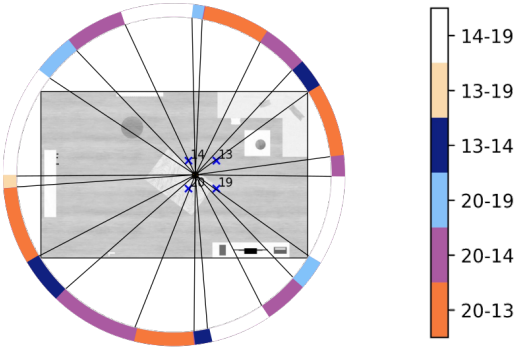
(a) An arbitrary viewpoint in the oblong room



(b) The panels of the image that have been assigned to a set of input viewpoints A, B $\in [13, 14, 19, 20]$



(c) The panels applied to the image



(d) The panels visualized in the scene (approximation)

Figure 4.25.: The input viewpoint choice problem in the oblong room (geometry further from proxy geometry)

a very similar viewpoint from the square room with the example from the oblong room, (Figure 4.26), some discontinuities are hardly visible, for example at the coffee table, or on the door. Other discontinuities are visible, but less apparent since the displacement is relatively small, for example on the rug, or on the TV cabinet.

Since the scenario evaluating the effect of the scene geometry focused mostly on objects within the scene instead of the basic scene geometry, it is difficult to judge whether this problem of abrupt change in viewpoint sets could affect the flow-based blending so much that it generally produced higher error values than the flow-based blending. However, if the general scene geometry deviated strongly from the proxy geometry, the regular blending algorithm would presumably also generate poor results.

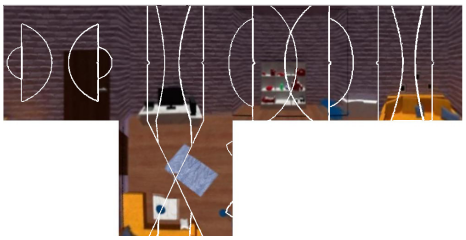
The other intrinsic problem of the approach is the ray approximation. Since the ray approximation is uniform for all of the flow-based images, it is difficult to judge how much of an impact it has on the results. Based on the analyzed images, it does not seem to create noticeable artefacts that can unambiguously be traced back to it.

4.4. Proof-of-Concept Evaluation using a Real Scene

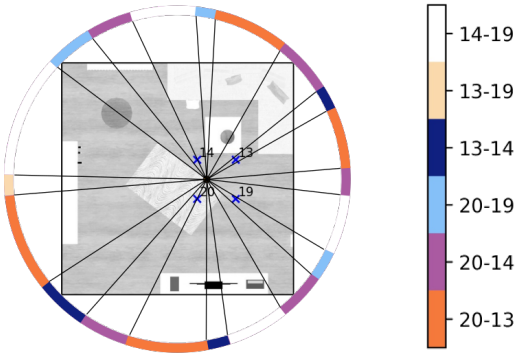
After the extensive evaluation of the virtual scenes, this section presents the results of the 2 DoF synthesis tested on a real scene. The goal of this “proof-of-concept” evaluation is to

(a) A similar viewpoint to Figure 4.25 in the square room

(b) The panels of the image that have been assigned to a set of input viewpoints A, B $\in [13, 14, 19, 20]$



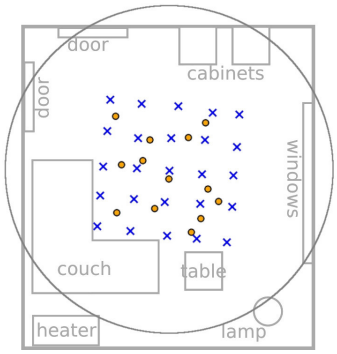
(c) The panels applied to the image



(d) The panels visualized in the scene (approximation)

Figure 4.26.: The input viewpoint choice problem in a the square room (geometry closer to proxy geometry)

(a) Latlong image from the center of the room



(b) The captured (blue) and synthesized (orange) viewpoints in the scene

Figure 4.27.: Overview of the real scene

determine to what extent the insights gained in the evaluation of the virtual scenes hold true when applied to data captured in the real world.

As for the internal and external parameters, only a single scene is tested, using a grid of captured viewpoints with a single density. Within this grid, several viewpoints are synthesized at random positions, and the regular blending results are compared to the flow-based blending results.

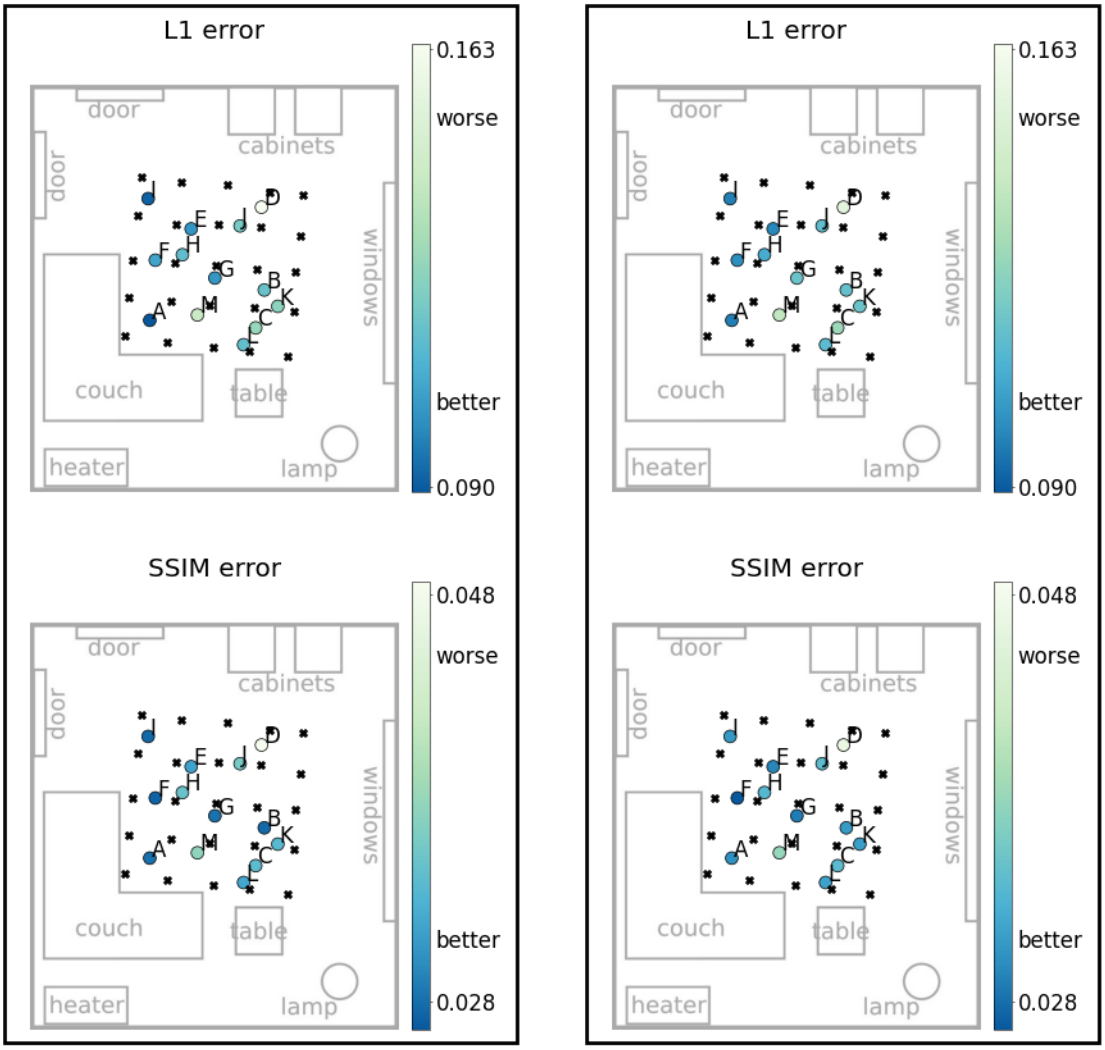
4.4.1. Data Acquisition

The images were captured using a Ricoh Theta Z1 360° camera with a tripod in an approximately 3m by 2.5m living room with a sloped roof (Figure 4.27a). For the locations of the captured viewpoints, a 5x5 grid was mapped out on the floor, so that the distances between the captured viewpoints would be approximately uniform. The spacing used for the grid was 40cm, as this was estimated to be a feasible distance for optical flow calculations. For the ground truth data, the tripod was randomly placed within the boundaries of the grid, capturing 13 different viewpoints to be used as ground truth data.

After capturing the data, the structure-from-motion library OpenSfM [Map] was used to automatically determine the positions and rotations of all of the images. The positions of the ground truth viewpoints were then used as positions for the 2 DoF synthesis. Also, instead of using a radius encompassing the complete area of the room, a slightly smaller radius was chosen so that the proxy geometry would be closer to the actual scene geometry. Figure 4.27b shows the acquired scene, including the proxy geometry as a gray circle. The proxy geometry is centered around the points, which were captured slightly offset from the center of the scene. This should be kept in mind, as this was not tested in the virtual scenes and may decrease reprojection accuracy.

4.4.2. Results

Figure 4.28 shows the scene analysis for the regular and flow-based blending. At a glance, they are very similar, although there does not seem to be a clear pattern to the error values. Since the results are so similar, first the possible reason for the general tendencies of the error



(a) Regular blending

(b) Flow-based blending

Figure 4.28.: Scene analysis of error values for regular and flow-based blending results

values is examined (e.g. why the error values are especially high for both blending types at a specific viewpoints), and then the results of the regular blending are compared to those of the flow-based blending.

use this title format for all of the scene analysis figures

General Tendencies

A closer look at Figure 4.28 shows that viewpoints “D” and “M” have particularly high error values for both metrics. A closer examination of “D” (Figure A.15a, page 96) shows that the high error values for both methods are due to the proximity of the cabinet elements on that side of the room. Unsurprisingly, the regular blending does not correctly reproject these, since they are not part of the proxy geometry. However, the flow-based blending also fails, presumably because the displacement between the input viewpoints was too large for the optical flow algorithm to handle. Figure A.15b shows the results of viewpoint “M”. In

this case, there is practically no difference between the two results. In fact, it seems like both are either rotated or positioned with a slight offset. It is possible that some positional or rotational error was introduced in the ground truth data. This is one of the difficulties of using real data instead of virtually generated data.

A look at some of the better rated points, viewpoints "I" and "F" (Figure A.16a and Figure A.16b, page 97a), show that both the regular blending, and the flow-based blending worked fairly well: Most of the objects are in the right positions, the largest error being on the tripod in the bottom face, which is not actually of interest, on the overhead light, and on the windows. The windows have a comparably high error, not necessarily because they were synthesized incorrectly, but mostly because the lighting changes between the captured viewpoints and the ground truth viewpoints. This may also be the reason for the SSIM error being comparatively lower for viewpoint "F", since the SSIM error metric is less sensitive to lighting changes.

Comparing Regular Blending to Flow-based Blending Results

improvement vs Delta

Although both the regular and the flow-based blending produced results with similar error values, the results are not identical. Figure 4.29 shows the Δ L1 and SSIM of the regular blending compared to the flow-based blending. For most of the viewpoints, the flow-based blending performed slightly better than the regular blending for both metrics. However, all of the values are very close together, and the difference between the highest and lowest Δ L1 and SSIM is generally very small. In this case, the error values alone are not sufficient to understand the differences between the results of the regular versus the flow-based blending.

A closer inspection of viewpoint "G" (Figure A.17, page 98) shows that most of the details (outlines of the couch, reading lamp, cabinets) are more accurate in the flow-based blending result. However, like in the virtual scenes, the flow-based blending also introduced some artefacts, such as noticeable displacements on the wall and door in the front face, as well as some extreme ghosting on the top face, where the 1 DoF interpolation was not able to correctly calculate optical flow of the ceiling lamp. The large RGB difference of the pixels of the ceiling lamp in the flow-based blending is most likely also the reason why the $\Delta L1$ error value is significantly higher than the $DeltaSSIM$ value for viewpoint "G".

Examining a result that has a relatively high $\Delta L1$ and $\Delta SSIM$, viewpoint "A" (Figure A.18a, page 99), also clearly demonstrates two of the problems of the flow-based blending: The red pillow in the front face is distorted due to failed optical flow, and the discontinuities are clearly visible, both in the front face on the back of the couch, and in the back face on the window.

Viewpoint "K" (Figure A.18b, where the flow-based blending result produced lower L1 and SSIM values than the regular blending result, shows a clearly visible correction in the left face, where the reading lamp shows a distinct artefact in the regular blending result, but clean lines in the flow-based result.

4.4.3. Discussion

In summary, the proof-of-concept evaluation showed that the insights gained in the evaluation of virtual scenes held true in the real scene: In most cases, the flow-based blending performed better than the regular blending, according to the L1 and SSIM metrics. It should also be taken into account that the flow-based blending introduced some interpolation-related

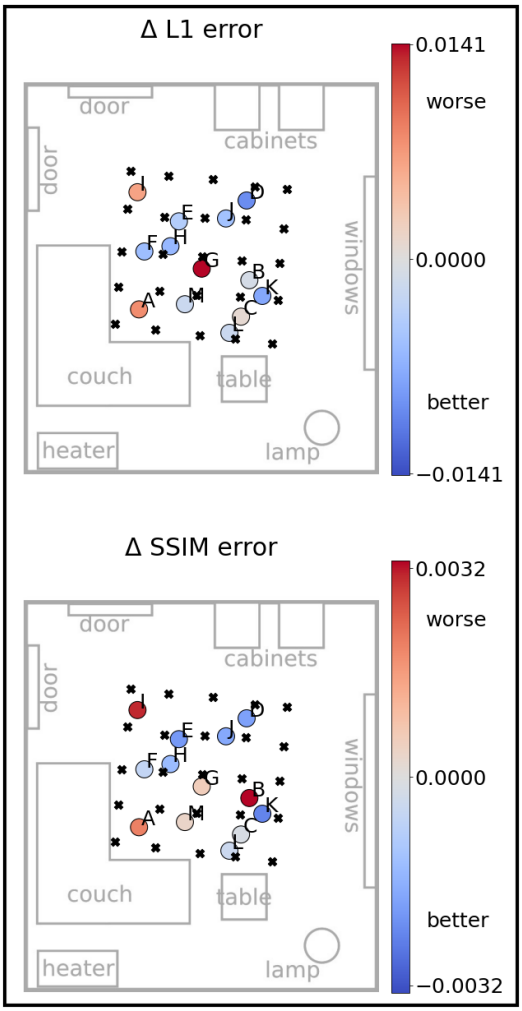


Figure 4.29.: Δ L1 and SSIM error values for regular blending compared to flow-based blending results

artefacts due to the Skylibs bug (predominantly visible as black lines), which will have had a detrimental impact on the error values of the flow-based blending.

Visually, the flow-based blending managed to improve some of the artefacts introduced by using the proxy geometry, however, the artefacts caused by the abrupt change in input viewpoints for 1 DoF interpolation were visually irritating. In order to improve this, one approach would be to use a proxy geometry that is closer to the scene geometry, that could for example be approximated using a structure-from-motion algorithm to calculate a sparse scene geometry. An alternative would be to change the selection of input viewpoints for the 1 DoF interpolation, or to introduce some kind of constraint (e.g. a color constraint) in order to blend and soften the abrupt edges.

Also, in the places where the optical flow algorithm failed, the results were blurred or warped, which is visually more noticeable than the regular blending result, where the objects were left undistorted but possibly viewed from an inaccurate position. It could be very advantageous to explore different optical flow algorithms, especially ones that focus on capturing large displacements. Furthermore, it could be possible to undistort the extended cube map before calculating optical flow, which could also improve the accuracy of the optical flow.

- "guess" an optimal radius without using viewpoint locations e.g. outside
- find a good weight function that balances deviation angle and distance appropriately
- undistort extended cubemap e.g. by using methods like [SLL19] which can undistort images up to 120°
- extend to 3D → input viewpoints could improve flow-based blending for areas towards the poles
- parallelization and offloading to gpu
- human perception evaluation with user study
- tradeoff between line consistency and perspective accuracy → would be interesting to introduce color constraints / optical flow smoothing, or other techniques to avoid discontinuities. this may come at the cost of accuracy

4.5. Limitations

4.5.1. Limitations of the Algorithm

4.5.2. Limitations of the Evaluation

The detailed evaluation of virtual scenes and the proof-of-concept evaluation of the real scene showed that flow-based interpolation can improve the results of basic pixel-based synthesis with proxy geometry. However, there are some limitations to this evaluation.

Firstly, only a limited number of external parameters were tested, for example leaving out more extreme room shapes that deviate strongly from the basic rectangular shape. It is possible that, given a strong deviation from the proxy geometry, the artefacts in the flow-based blending caused by the abrupt change in viewpoint would increase to a degree where the regular blending would be preferable. Also, only indoor scenes were considered. It is

possible that the behavior of the algorithms is is completely different for outdoor scenes with much larger distances between the viewpoint and the scene geometry. Furthermore, the parameters that were tested were not tested exhaustively, so it is possible that some interactions between parameters were overlooked.

Secondly, the metrics that were used (i.e. the L1 error and the SSIM error), can show an improvement between different results, but it is difficult to quantify this improvement. For example, it is not possible to explicitly compare the error values of different scenes (unless they contain very similar colors and patterns), since the L1 error metric is very dependent on pixel color values, and it is difficult to differentiate the exact cause-and-effect of the SSIM error metric. Futhermore, the L1 and SSIM error metrics give no indication on believability or visual preference, since they cannot measure the severity of artefacts for human perception. This means that it is possible that users may prefer the regular blending results over flow-based blending results, due to the smoother transitions of the regular blending between different areas. Although user acceptability was not a criterium for this evaluation, it is crucial for the intended area of appliation, namely Virtual Reality. Another factor that plays into the visual acceptability is temporal consistency, i.e. whether the parallax movement of objects in the scene is believable when navigating through the scene using synthesized views. This is also not tested in the evaluation.

5. Conclusion and Future Work

The approach and implementation presented in this thesis laid the groundwork for a pixel-based 2 DoF synthesis using 1 DoF interpolation in order to apply flow-based blending. The In order to improve this approach for use in actual Virtual Reality applications, the artefacts created by the flow-based approach could be improved by either adding a constraint to the viewpoint selection, or by increasing the accuracy of the proxy geometry. The implementation could then be optimized in a way that would enable real-time navigation of previously captured scenes with 2 DoF, which could be used to inhance a number of Virtual Reality applications.

A. Synthesized Images

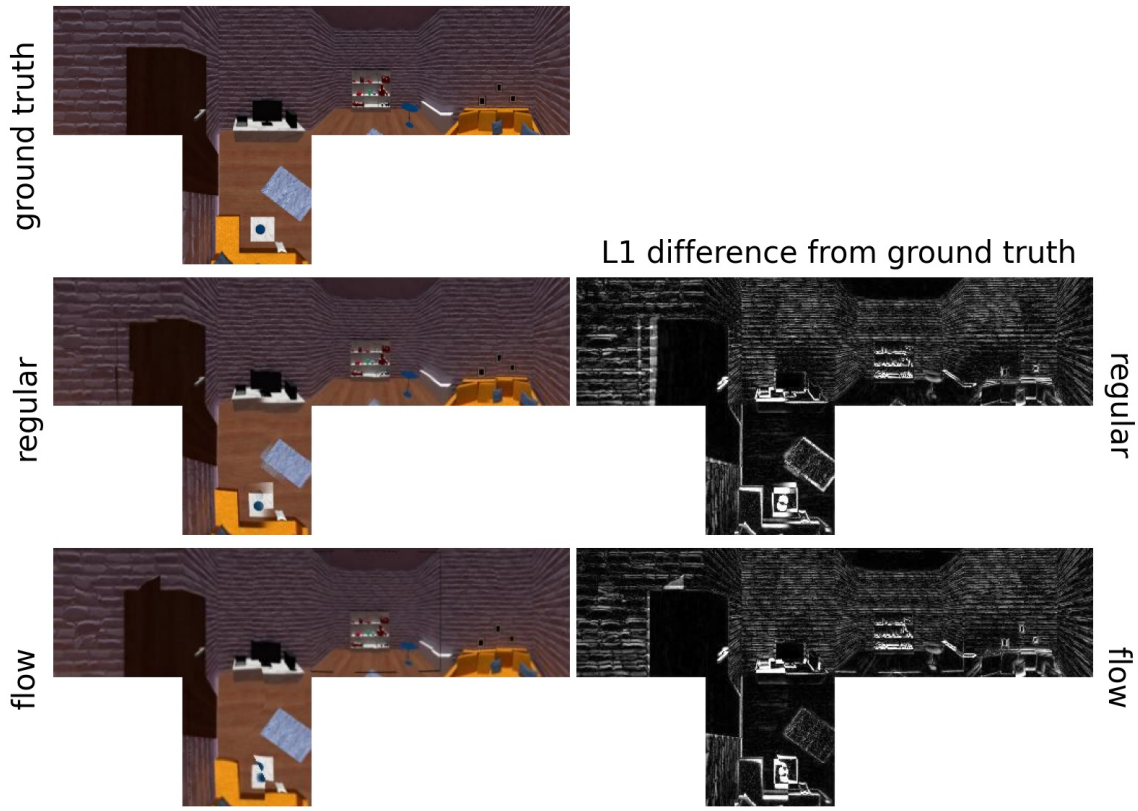


Figure A.1.: Sample inspection of example viewpoint "K": The images are in cube map representation, as this is tends to be more intuitive to understand than latlong representation. The top face is omitted for a more compact representation.

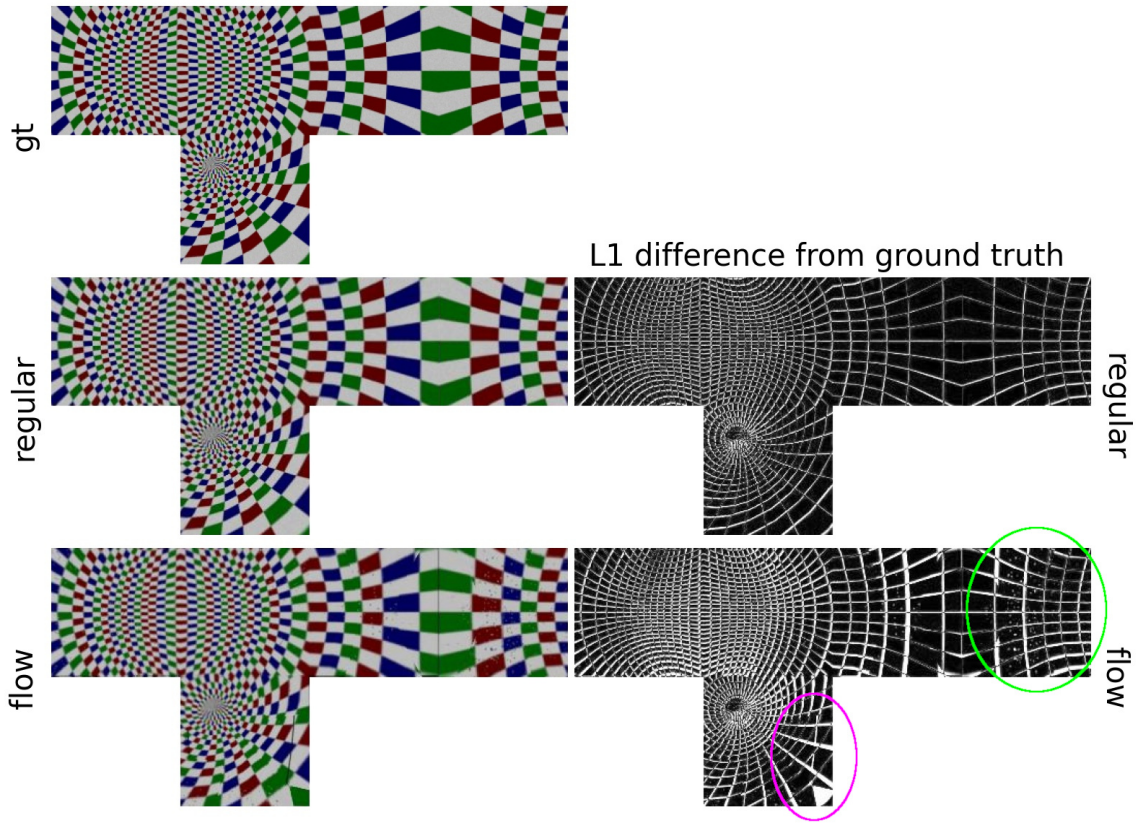


Figure A.2.: Results for synthesized viewpoint “Y” in the checkersphere scene: The regular blending result is very close to the ground truth, except for some blurriness. The flow-based blending result shows some inaccuracies (magenta) and noise (green)

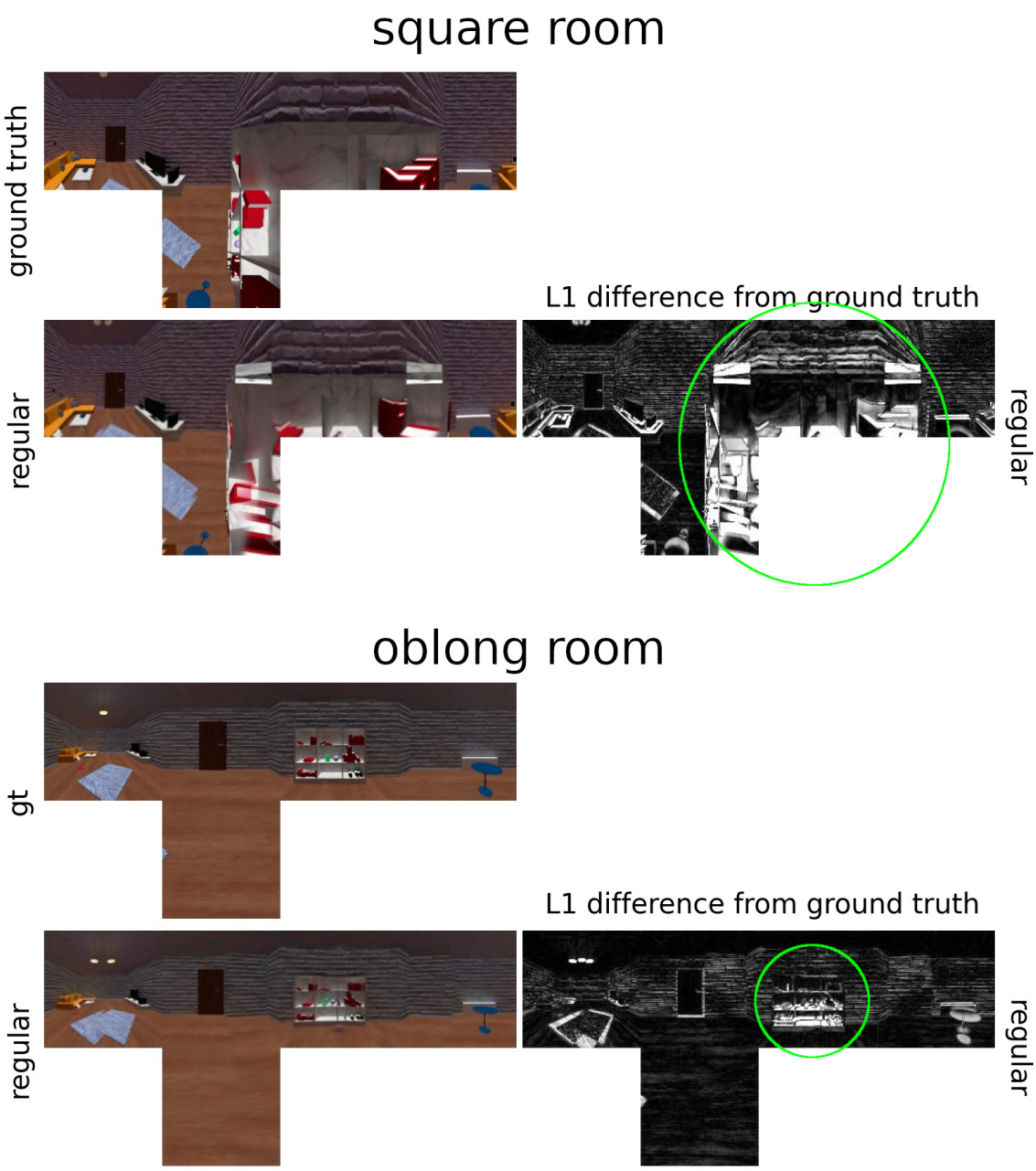


Figure A.3.: Regular blending result of viewpoint "O" in the square and oblong rooms: The bookshelf has a strong impact on the difference in error values (marked in green)

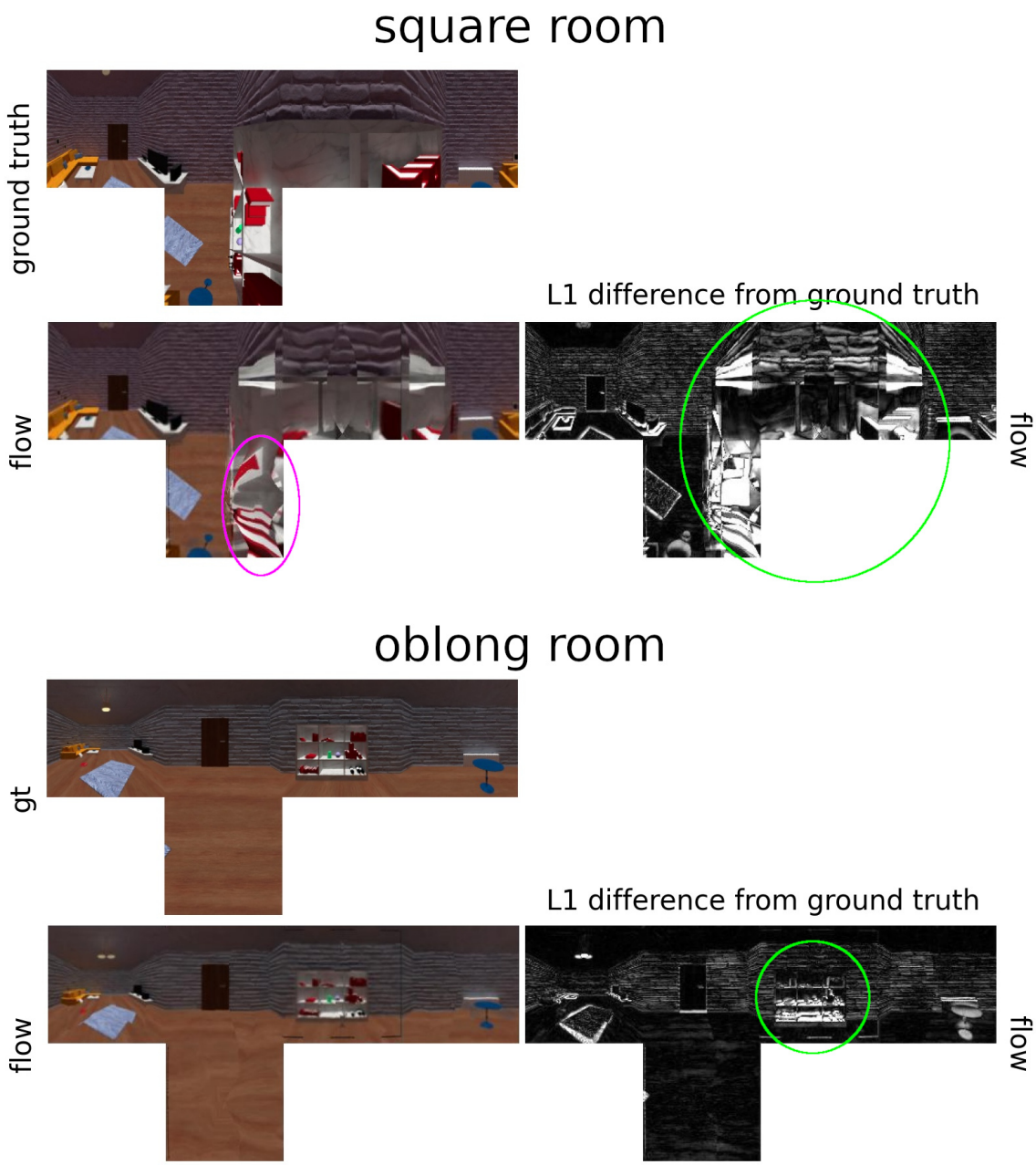


Figure A.4.: Flow-based blending result of viewpoint "O" in the square and oblong rooms: The bookshelf has a strong impact on the difference in error values (green) and the details of the bookshelf are warped due to inaccurate optical flow (magenta).

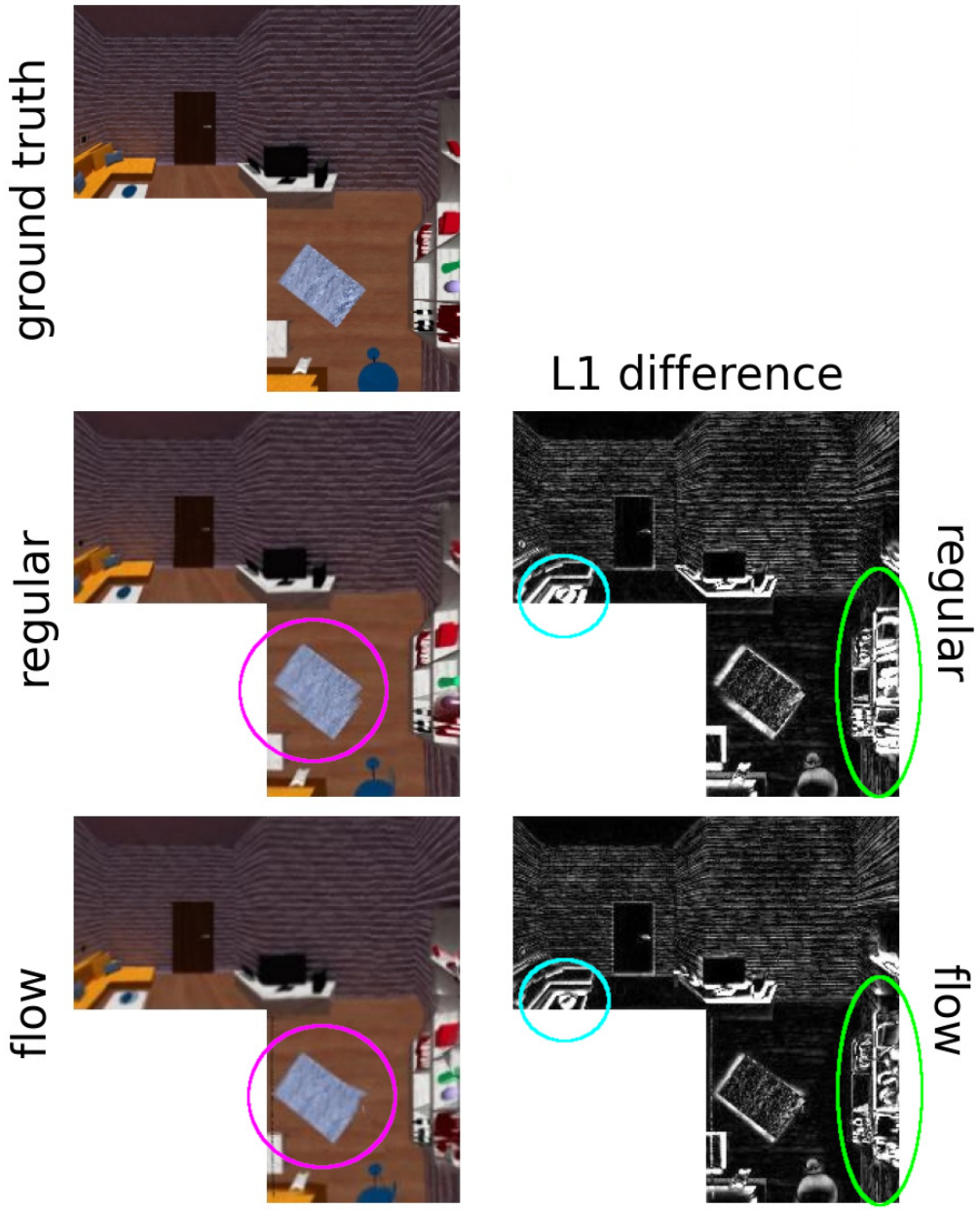


Figure A.5.: Synthesized point "N" in the square room (best improvement for L1, good for SSIM): The flow-based blending removed the ghosting artefacts on the rug (magenta) and improved the accuracy on the coffee table (cyan), and the lower part of the bookshelf (green). The rest of the scene is very similar for both results.

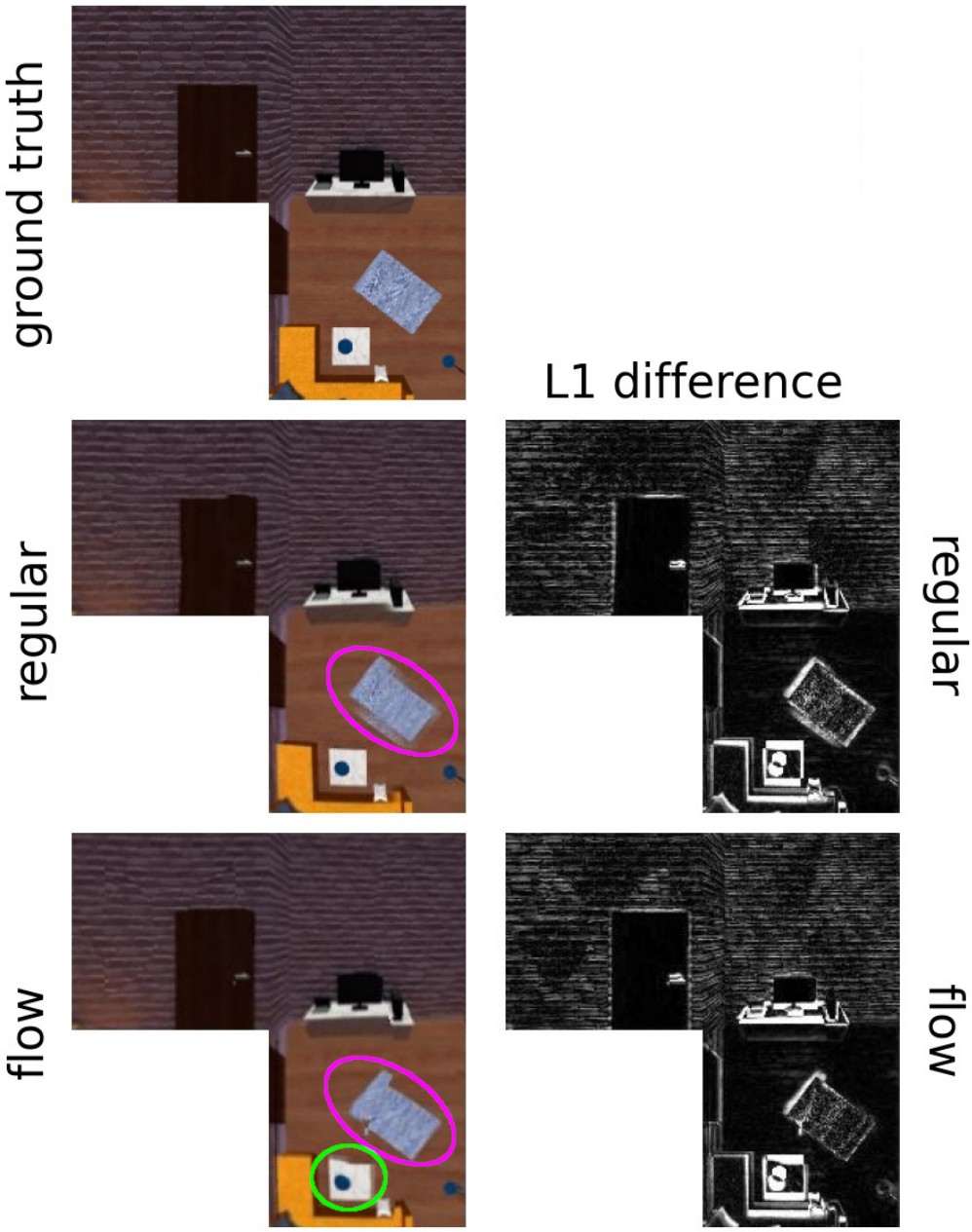


Figure A.6.: Synthesized point "L" in the square room (worst "improvement": slight increase of error for both metrics): The ghosting artefact on the rug in the regular blending result was replaced by a displacement artefact in the flow-based blending (magenta), which also introduced a new artefact, namely the warped top edge of the coffee table (green). Otherwise the scenes are very similar.

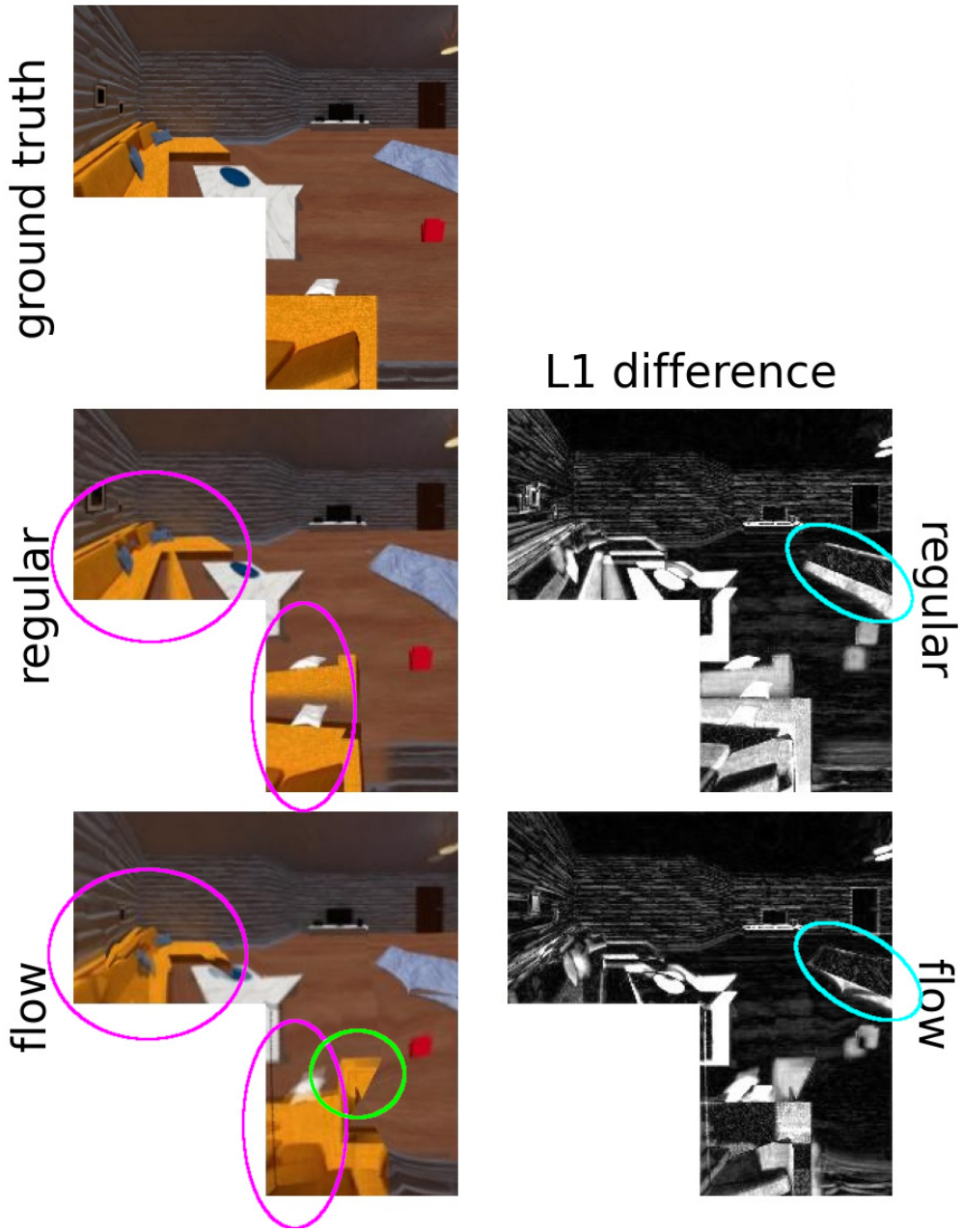


Figure A.7.: Synthesized point "A" in the oblong room (best improvement): The flow-based blending drastically improved ghosting artefacts on the couch (magenta) and the accuracy of the rug (cyan), but also introduced new artefacts (green).

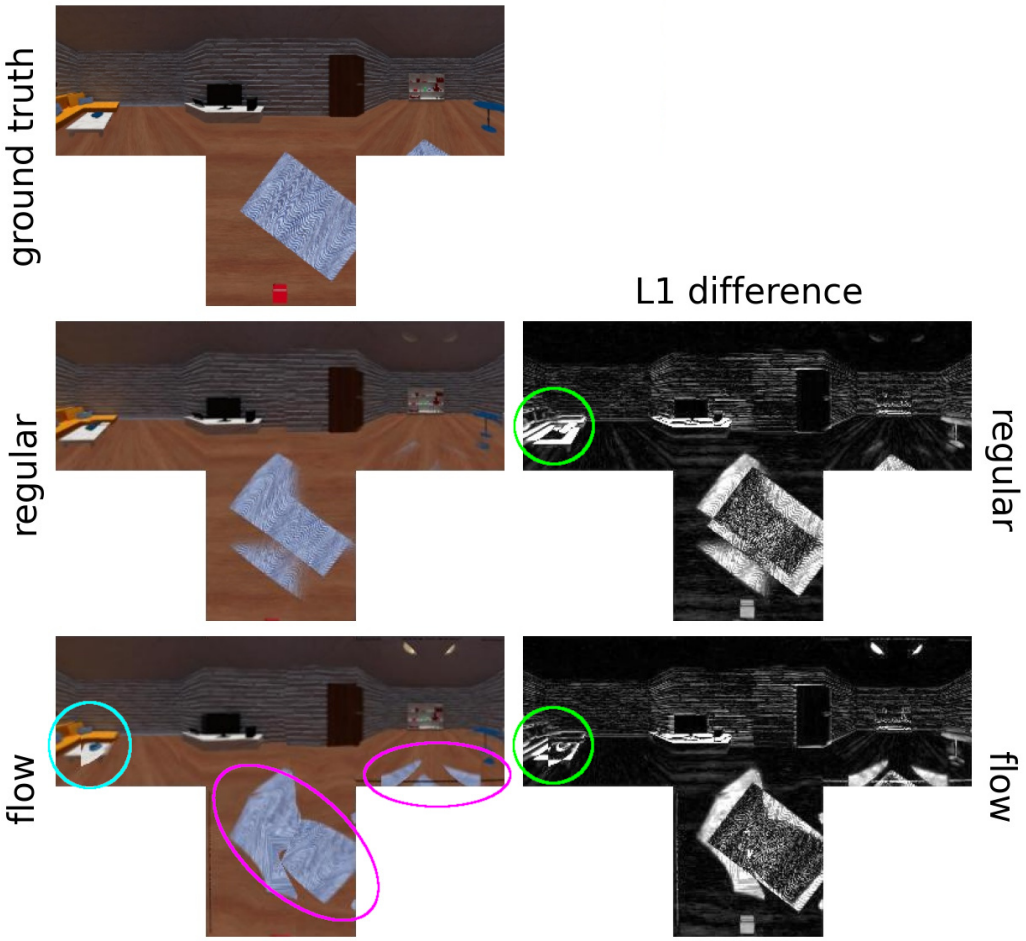
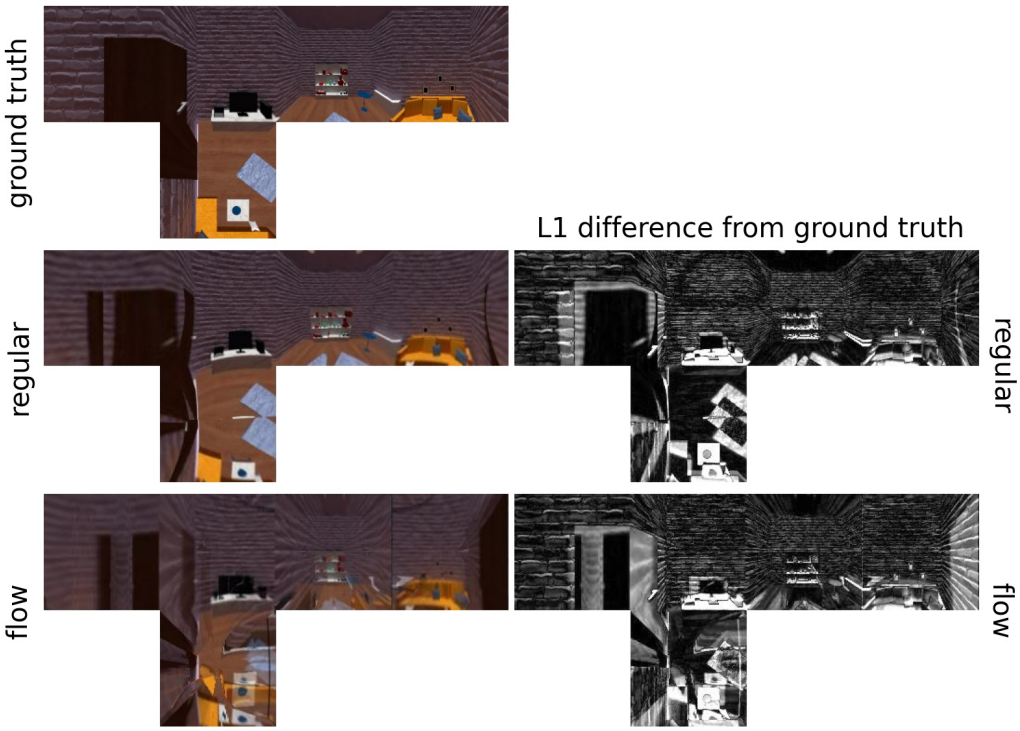


Figure A.8.: Synthesized point “L” in the oblong room (worst improvement): The flow-based blending result introduced some severe discontinuity artefacts on the rug (magenta) and on the coffee table (cyan), although the coffe table is positioned more accurately in the flow-based blending result (green)

Figure A.9.: The regular blending results for point "T" (one of the worse results) in the square room with a viewpoint density of 2x2, 6x6, and 12x12

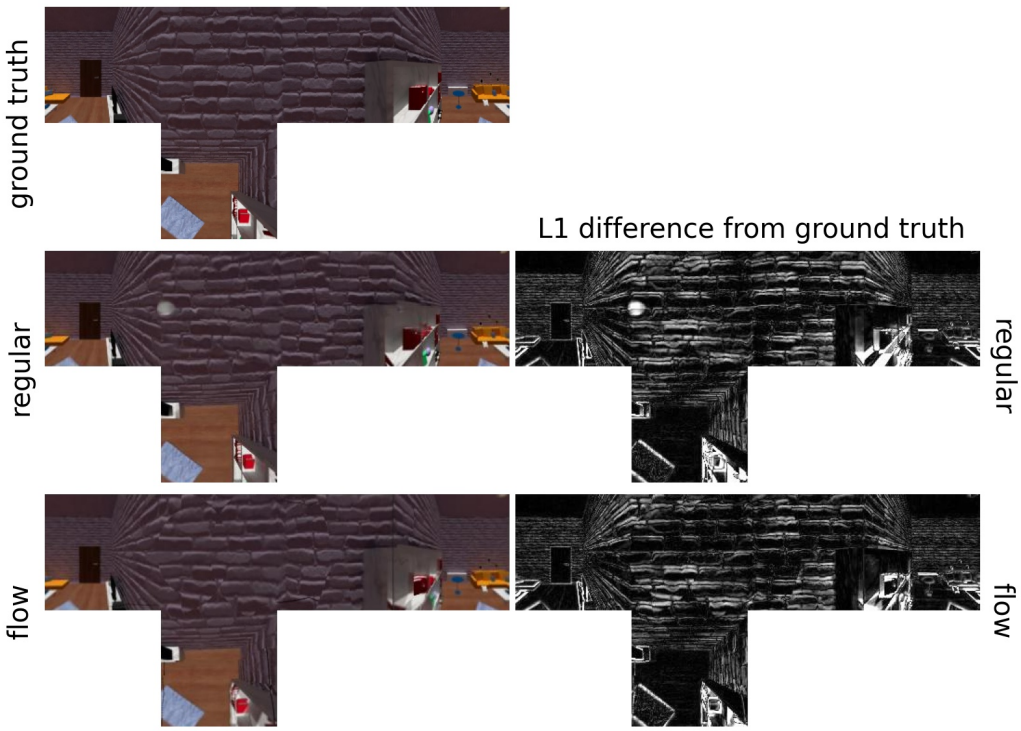
Figure A.10.: The regular blending results for point “G” (one of the better results) in the square room with a viewpoint density of 2x2, 6x6, and 12x12

(a) Synthesized point “T” (best improvement for L1, good for SSIM)

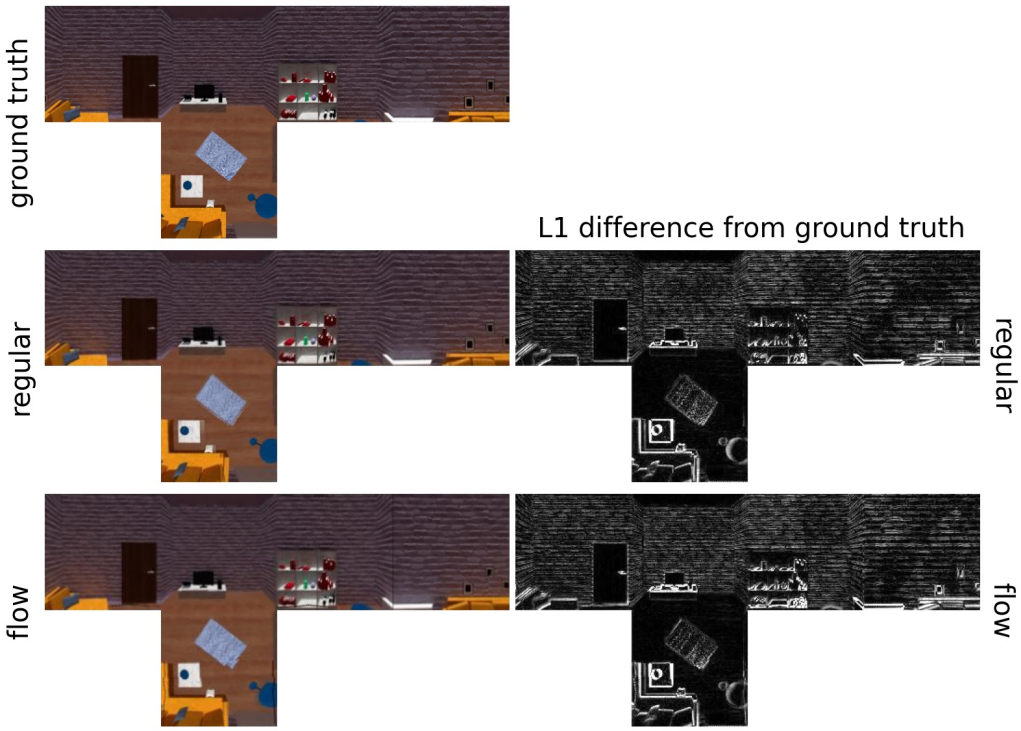


(b) Synthesized point “K” (worst “improvement”: slight increase of error)

Figure A.11.: Best (T) and worst (K) improvements of flow-based blending over regular blending with the 2x2 setup in the square room. The “worst” improvement is slightly worse than the regular blending result



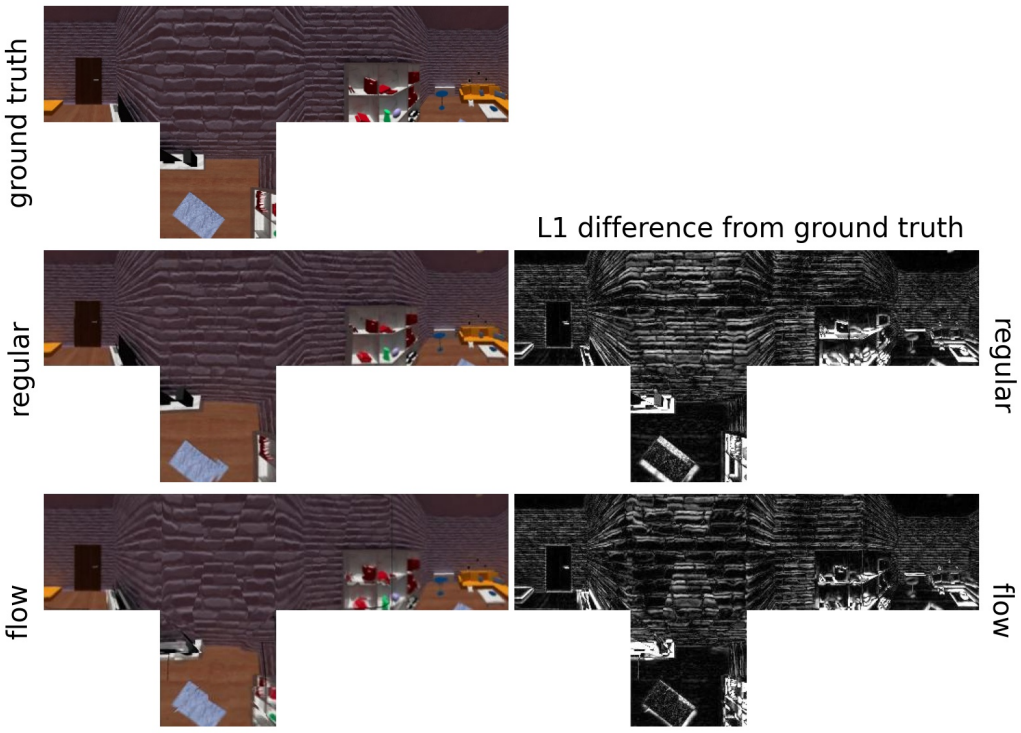
(a) Synthesized point “Y” (best improvement for L1 and SSIM)



(b) Synthesized point “H” (worst “improvement”: slight increase of error)

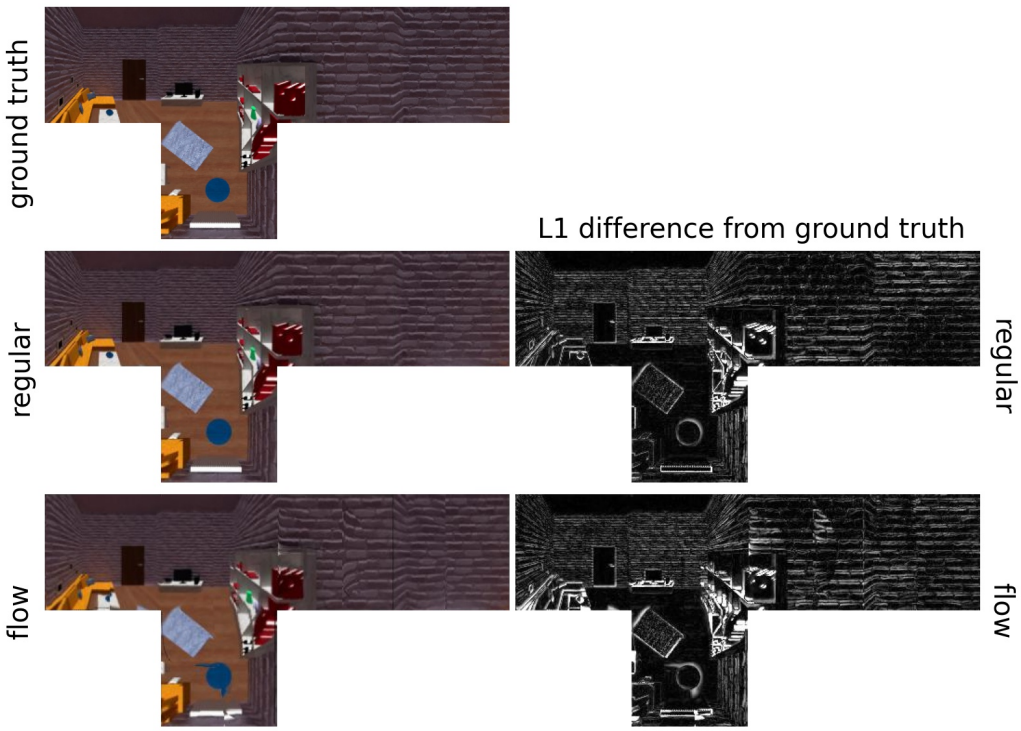
Figure A.12.: Best (Y) and worst (H) improvements of flow-based blending over regular blending in the 12x12 setup in the square room. The “worst” improvement is slightly worse than the regular blending result

(a) The coffee table is in a more accurate position, even though it shows some blurriness

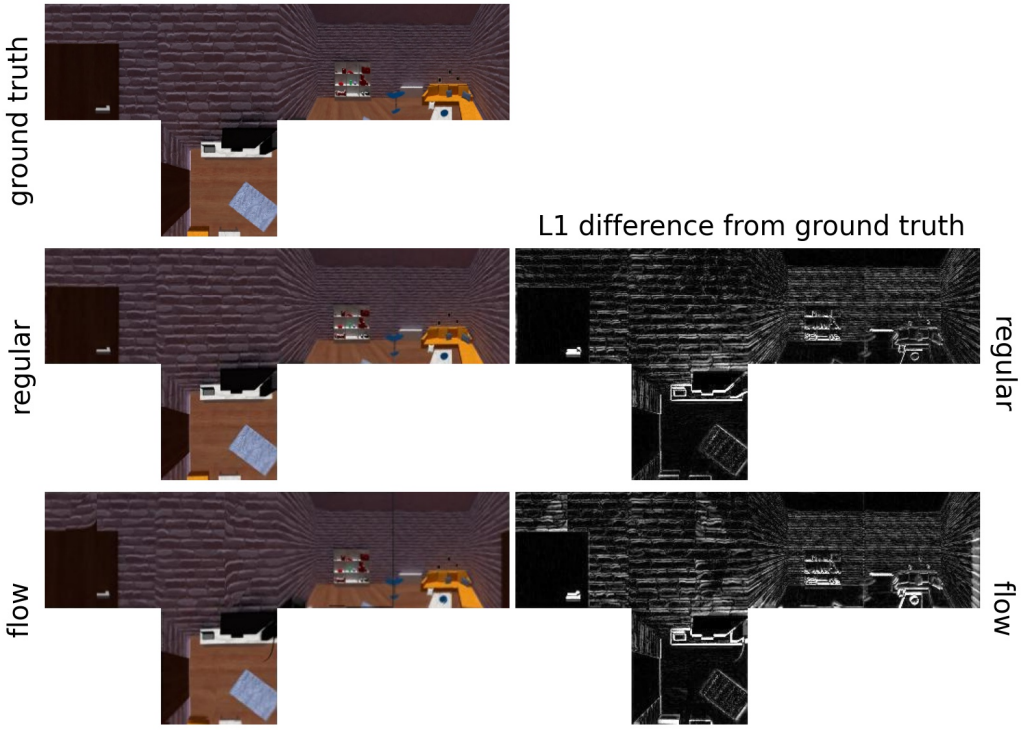


(b) The rug no longer has doubled edges, however, there are still some artefacts

Figure A.13.: Results for which the flow-based blending improved the accuracy

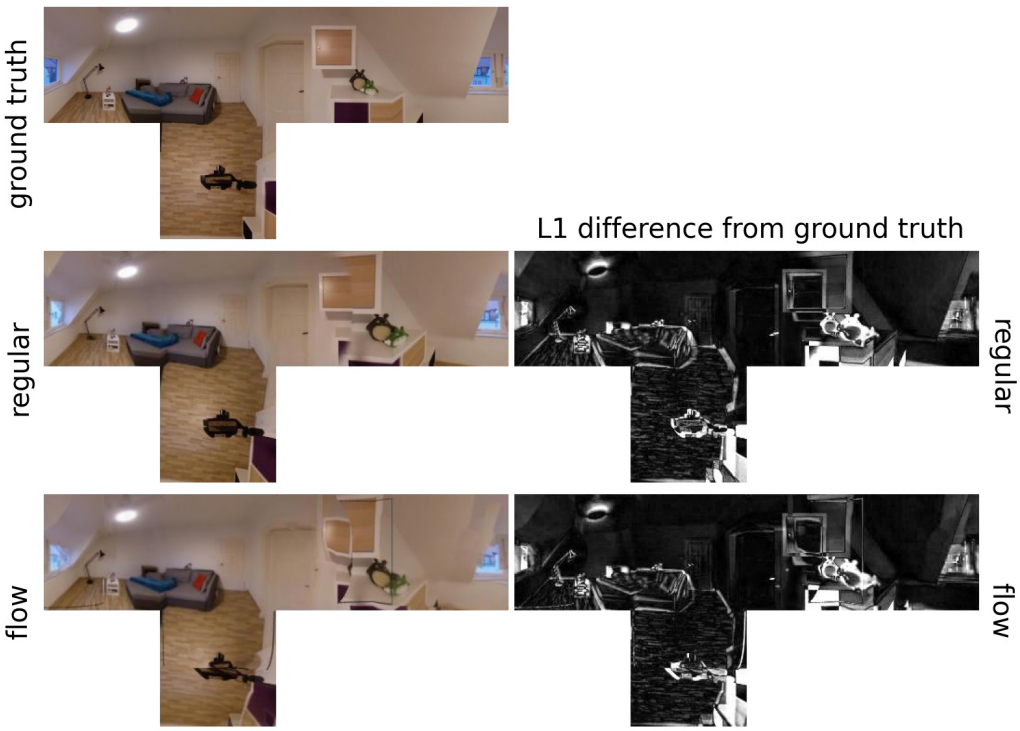


(a) The results are very similar, except on the blue table in the bottom face, and the white coffee table in the left face, where there are distinct artefacts in the flow-based result.

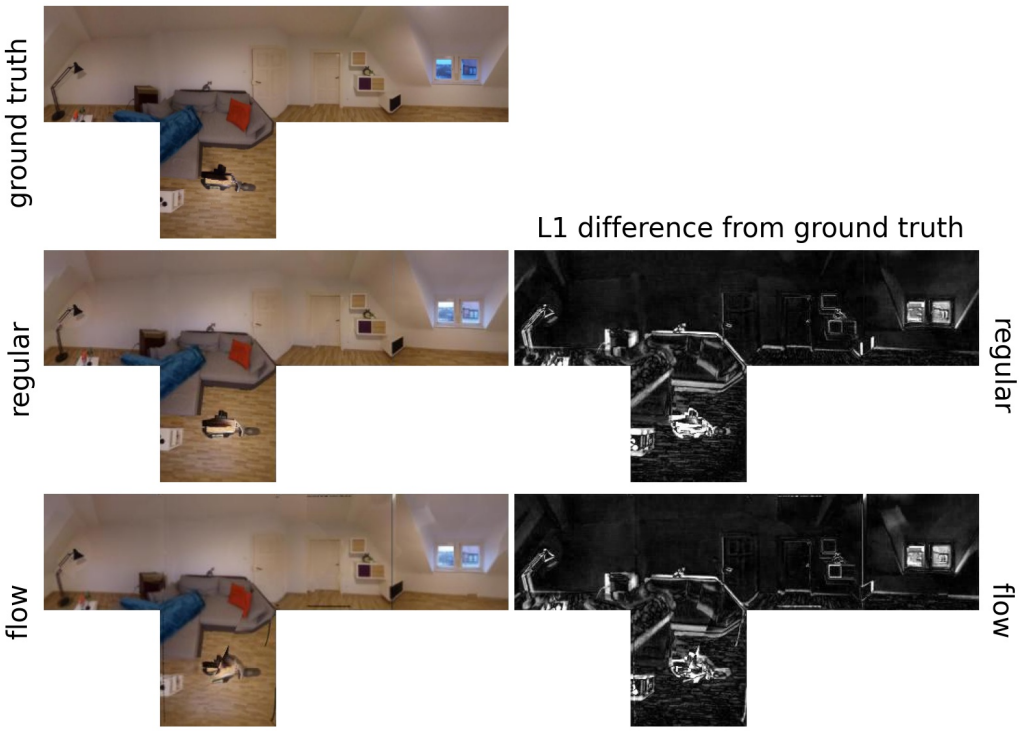


(b) In this case, there is hardly a visible difference between the two results

Figure A.14.: Results for which the flow-based blending decreased the accuracy (both examples are in the direct vicinity of a captured viewpoint

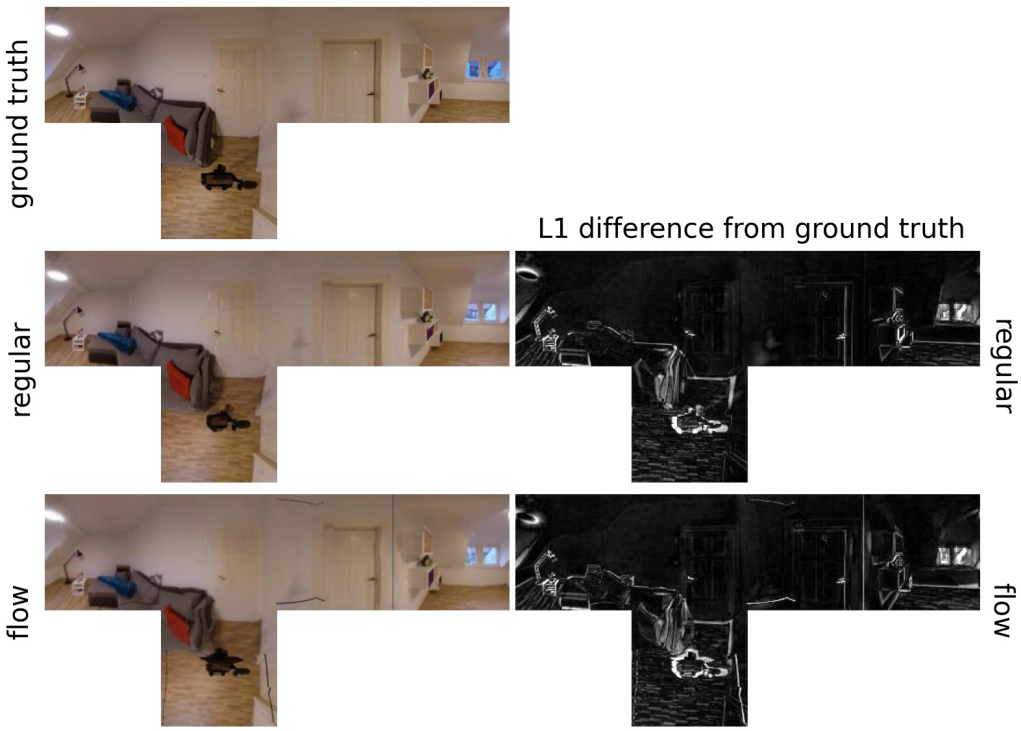


(a) Viewpoint “D”

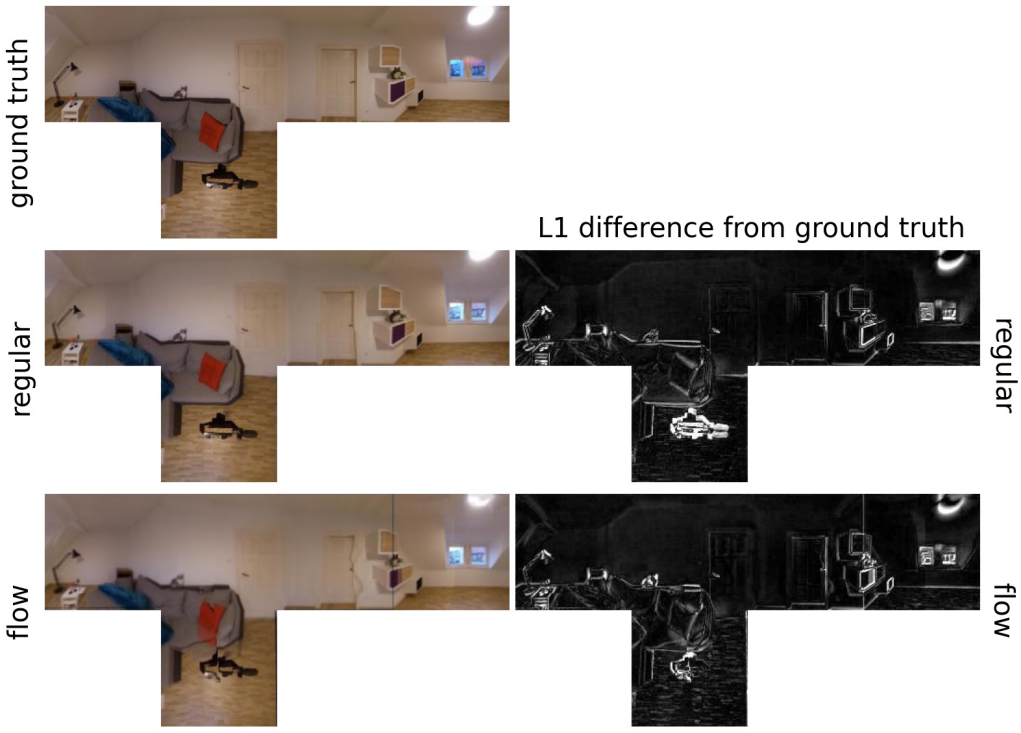


(b) Viewpoint “M”

Figure A.15.: Viewpoints in the real scene where both regular and flow-based blending produced results with high error values



(a) Viewpoint “I”



(b) Viewpoint “F”

Figure A.16.: Viewpoints in the real scene where both regular and flow-based blending produced results with low error values

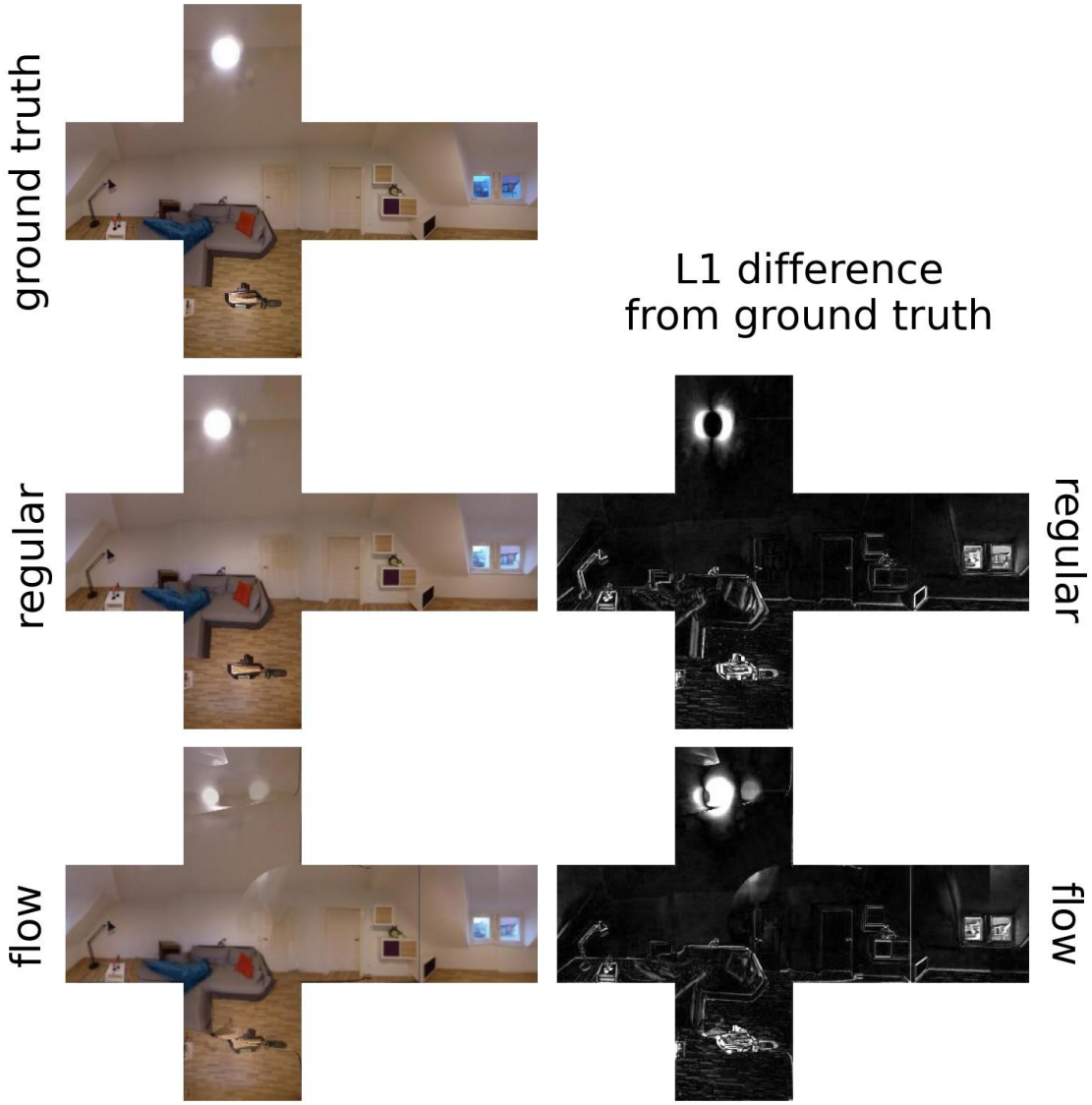
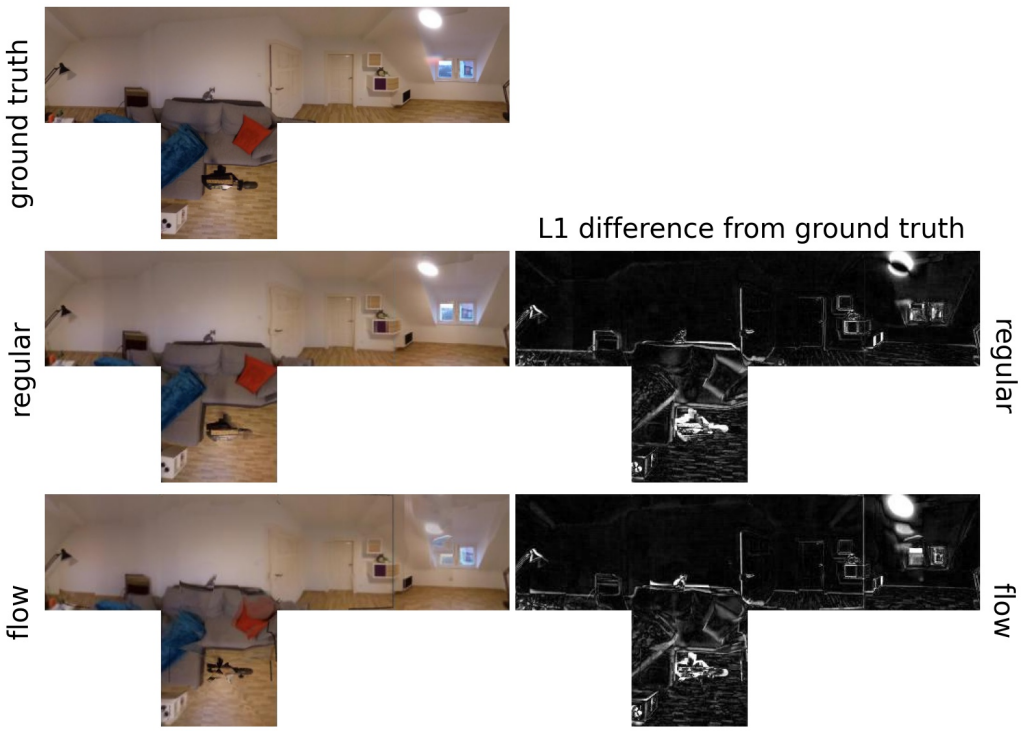
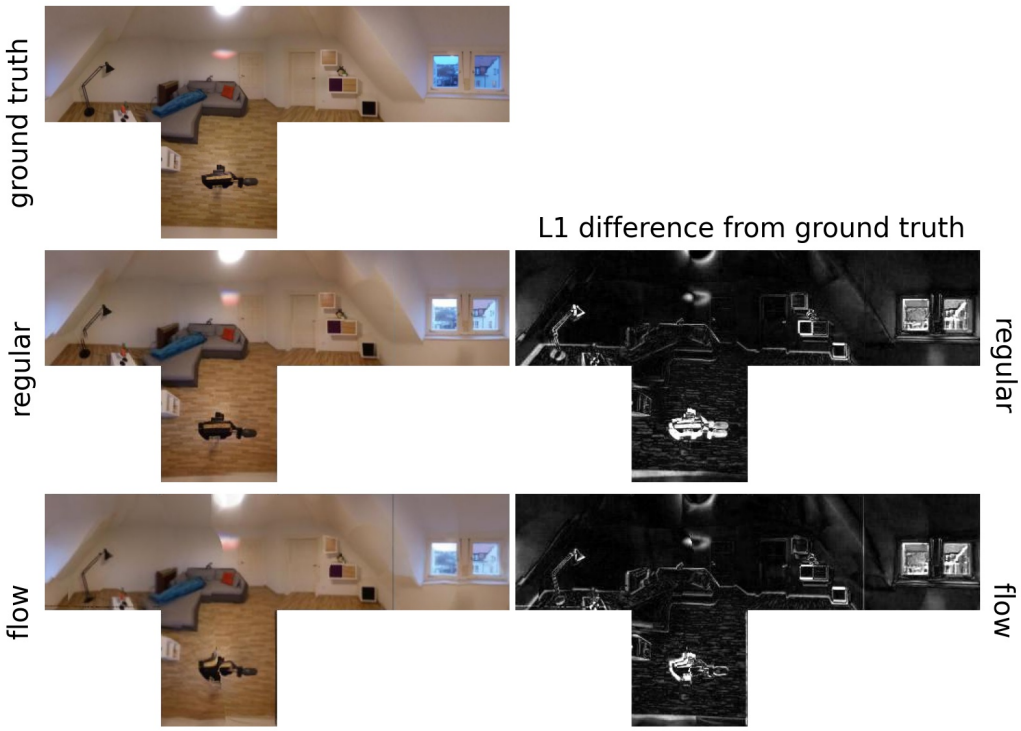


Figure A.17.: Viewpoint "G"



(a) Viewpoint “A”, where the flow-based blending produced worse results than the regular blending



(b) Viewpoint “K”, where the flow-based blending produced better results than the regular blending

Figure A.18.: Viewpoints in the real scene where there is a clear difference between the regular and the flow-based blending results

List of Figures

Bibliography

[AB91] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[Ble20] Blender Online Community. Blender - a 3d modelling and rendering package, 2020.

[BWSB12] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 611–625, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[Che95] Shenchang Eric Chen. QuickTime VR: An Image-Based Approach to Virtual Environment Navigation. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, page 29–38, New York, NY, USA, 1995. Association for Computing Machinery.

[CW93] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, page 279–288, New York, NY, USA, 1993. Association for Computing Machinery.

[Far03] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[FBK15] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1 – 21, 2015. Image Understanding for Real-world Distributed Video Networks.

[HCCJ17] J. Huang, Z. Chen, D. Ceylan, and H. Jin. 6-DOF VR videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR)*, pages 37–44. IEEE Computer Society, 2017.

[HDR+17] Sachini Herath, Vipula Dissanayake, Sanka Rasnayaka, Sachith Seneviratne, Rajith Vidanaarachchi, and Chandana Gamage. Unconstrained segue navigation for an immersive virtual reality experience. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, 50:13, 10 2017.

[Hol20] Hold, Yannick (Soravux). Skylibs. `https://github.com/soravux/skylibs`, 2020.

[Kaw17] Naoki Kawai. A simple method for light field resampling. In *ACM SIGGRAPH 2017 Posters*, SIGGRAPH '17, New York, NY, USA, 2017. Association for Computing Machinery.

[KL10] S. Kolhatkar and R. Laganière. Real-time virtual viewpoint generation on the gpu for scene navigation. In *2010 Canadian Conference on Computer and Robot Vision*, pages 55–62, 2010.

[LH96] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, page 31–42, New York, NY, USA, 1996. Association for Computing Machinery.

[Map] Mapillary. OpenSfM. `https://www.opensfm.org/docs/`.

[Pyt18] Python Software Foundation. Python 3.7.9 documentation. `https://docs.python.org/3.7/`, 2018.

[RPZSH13] Christian Richardt, Yael Pritch, Henning Zimmer, and Alexander Sorkine-Hornung. Megastereo: Constructing high-resolution stereo panoramas. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1256–1263. IEEE Computer Society, June 2013.

[RWPD05] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[SET20] Stefano Savian, Mehdi Elahi, and Tammam Tillo. *Optical Flow Estimation with Deep Learning, a Survey on Recent Advances*, chapter 12, pages 257–287. Springer International Publishing, 01 2020.

[SI14] Davide Scaramuzza and Katsushi Ikeuchi. Omnidirectional camera. *Computer Vision: A Reference Guide*, 2014.

[SK00] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In King N. Ngan, Thomas Sikora, and Ming-Ting Sun, editors, *Visual Communications and Image Processing 2000*, volume 4067, pages 2 – 13. International Society for Optics and Photonics, SPIE, 2000.

[SLDL09] F. Shi, R. Laganiere, E. Dubois, and F. Labrosse. On the use of ray-tracing for viewpoint interpolation in panoramic imagery. In *2009 Canadian Conference on Computer and Robot Vision*, pages 200–207. IEEE Computer Society, 2009.

[SLL19] YiChang Shih, Wei-Sheng Lai, and Chia-Kai Liang. Distortion-free wide-angle portraits on camera phones. *ACM Trans. Graph.*, 38(4), July 2019.

[The19a] The OpenCV team. OpenCV 4.2 documentation. `https://docs.opencv.org/4.2.0/`, 2019.

[The19b] The SciPy community. NumPy v1.16 Manual. `https://numpy.org/doc/1.16/`, 2019.

[The20] The SciPy community. SciPy v1.5.2 Reference Guide. `https://docs.scipy.org/doc/scipy-1.5.2/reference/`, 2020.

[vdWSN+14] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.

[Wei] Weisstein, Eric W. Line-Line Intersection. `https://mathworld.wolfram.com/Line-LineIntersection.html`.

[ZBSS04] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[ZC04] Cha Zhang and Tsuhan Chen. A survey on image-based rendering—representation, sampling and compression. *Signal Processing: Image Communication*, 19(1):1 – 28, 2004.

[ZWF+13] Q. Zhao, L. Wan, W. Feng, J. Zhang, and T. Wong. Cube2video: Navigate between cubic panoramas in real-time. *IEEE Transactions on Multimedia*, 15(8):1745–1754, 2013.