

Supplementary Materials: Depth-Aware Stitching Framework for Omnidirectional Vision with Multiple Cameras

Anonymous Authors

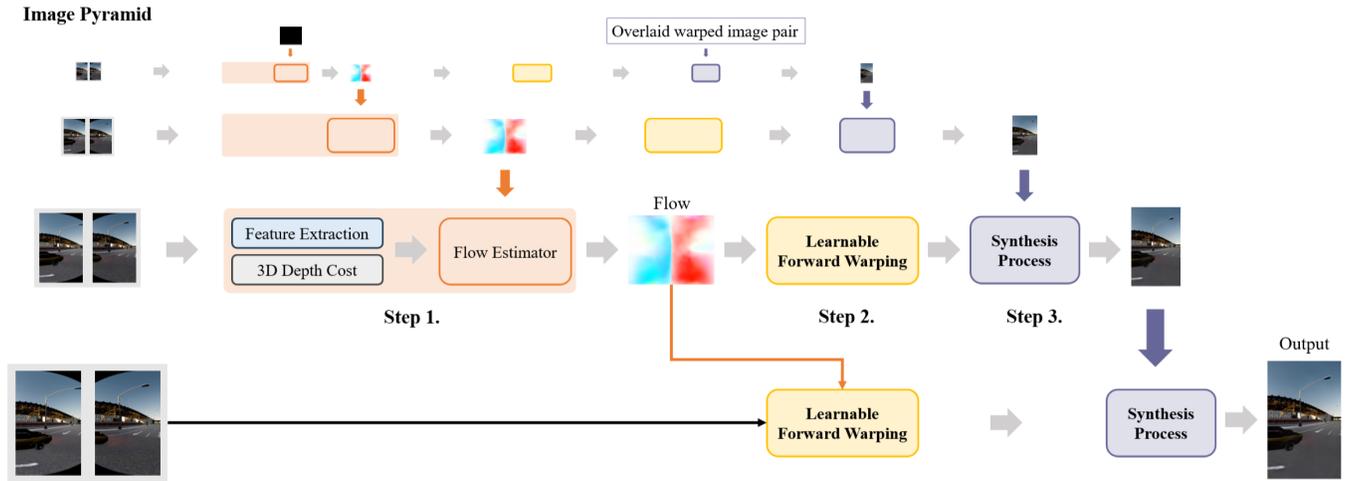


Figure 1: The overview of OmniStitch’s pyramid structure.

In this document, we provide the following supplementary context:

- Details of pyramid strategy in OmniStitch (Section A).
- Details of synthesis network architecture (Section B).
- Details of GV360 dataset (Section C).
- Qualitative results on GV360 dataset (Section D).

Regarding the network architecture, the precise channel count, number of layers, activation function, and other pertinent details of the OmniStitch network can be found in the code that will be provided.

A DETAILS OF PYRAMID STRATEGY IN OMNISTITCH

Recent developments in image stitching have predominantly employed a two-step warping approach. Initially, this involves globally warping the entire image to achieve global similarity, followed by locally warping the segmented image to enhance local similarity [1, 4, 5, 8]. While this method performs well in scenarios with small parallax, it tends to falter with wide parallax, as demonstrated by the qualitative assessments from both the GV360 and real-world datasets. This issue likely arises from using distinct objective functions for global and local warping without any integrative connection between the two stages. Additionally, these methods do not offer any refinement of the final output.

To overcome these limitations, we have adopted a pyramid structure commonly used in optical flow-based synthesis models [2, 3, 7]. This coarse-to-fine approach not only refines flow estimation and synthesis progressively using up-sampled outputs but also allows for the uniform application of the same network architecture across different pyramid levels, significantly reducing parameter count [2, 3]. OmniStitch leverages these benefits, yet it

introduces two principal distinctions in its structure, as shown in Figure 1.

Firstly, OmniStitch features a four-level pyramid designed to enhance stitching performance progressively. The flow estimation step is bypassed at the final pyramid level, corresponding to the original resolution. Instead, the refined flow is created by quadrupling the scale of the up-sampled flow. This modification has been experimentally proven to significantly boost the LPIPS metric significantly, mainly because estimating flow between images with significant parallax at full resolution can result in errors and blur artifacts.

Secondly, there is no provision for up-sampled flow or output at the highest pyramid level. Here, the top-level image pair is processed using a Learnable Forward Warping (LFW) network, the same type employed in step 2, although the LFW network is not trained during this phase. The warped image pair is overlaid and replaces the up-sampled output. Any additional up-sampled results are simply replaced with zeros of equivalent dimensions.

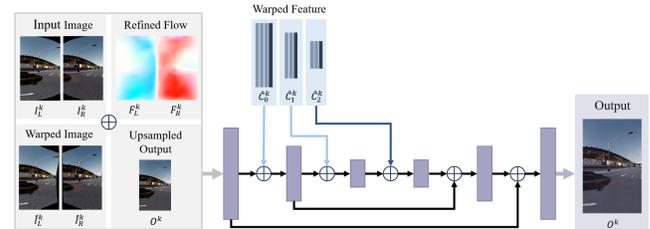


Figure 2: The detailed architecture of synthesis network of the OmniStitch.

B DETAILS OF SYNTHESIS NETWORK ARCHITECTURE

In this section, we describe the details of the synthesis network in Step 3: Synthesis Process (Section 3.3.3.). The precise structure is illustrated in the Figure 2. The synthesis network, detailed in Section 3.3.3 under Step 3: Synthesis Process, is pivotal in the image stitching framework. It harnesses the outputs of the preceding pyramid level (denoted as O^k) and integrates results from Steps 1 and 2 of the current level (denoted as $I_L^k, I_R^k, F_L^k, F_R^k, \tilde{I}_L^k$, and \tilde{I}_R^k).

This network uses an advanced encoder-decoder architecture with lateral connections, similar to the U-net configuration. Each encoder stage processes inputs comprising a series of warped outputs from the feature encoder at distinct stages (denoted as $\tilde{C}_0^k, \tilde{C}_1^k$, and \tilde{C}_2^k). These inputs undergo warping via Learnable Forward Warping (LFW) and are concatenated with the output of the preceding CNN layer, forming tailored inputs for each subsequent layer.

Notably, the contextual features \tilde{C}_0^k and \tilde{C}_1^k are derived by average splatting of C_0^k and C_1^k , while \tilde{C}_2^k is produced using softmax splatting [6]. This decision was based on empirical observations that showed minimal differences in the outcomes between these splatting methods for C_0^k and C_1^k , leading to the selection of average splatting to reduce parameterization.

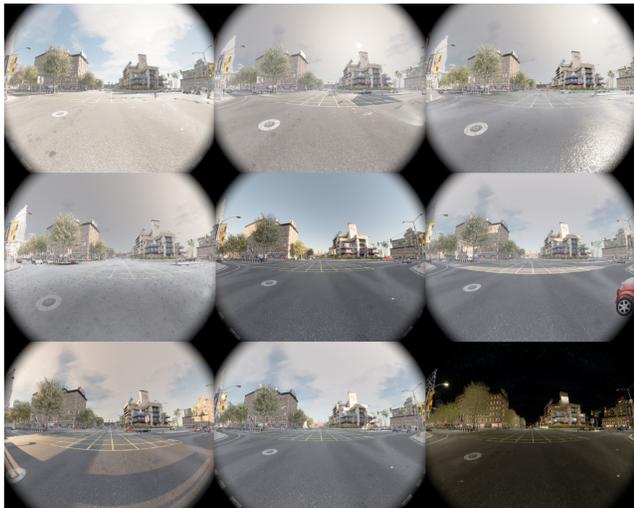


Figure 3: The configuration of the GV360’s weather and time settings.

C DETAILS OF GV360 DATASET

OmniStitch is a supervised model trained using the GV360 dataset to ensure robust performance across various environments. This dataset includes diverse settings for distance parallax, weather, time, map, and spawn points. Training data was collected using two maps and 18 spawn points 4, covering nine different weather and time conditions 3, with distance parallaxes ranging from 0.01m to 1.4m 5. Each setting was carefully configured to ensure a uniform distribution, providing a comprehensive range of scenarios.

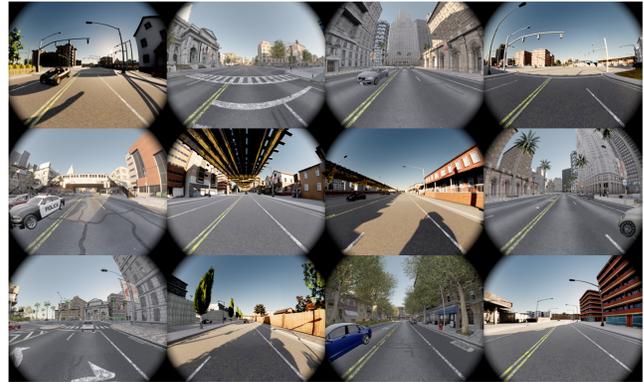


Figure 4: The configuration of the GV360’s map and spawn point settings.

During the testing phase, data was collected from three different maps using 9 spawn points not previously used in the training. The test data utilized four distance parallax : 0.01m, 0.5m, 0.8m, and 1.4m. Notably, each collection session was conducted with the vehicle being driven autonomously, ensuring that the test conditions closely simulated real-world driving scenarios.

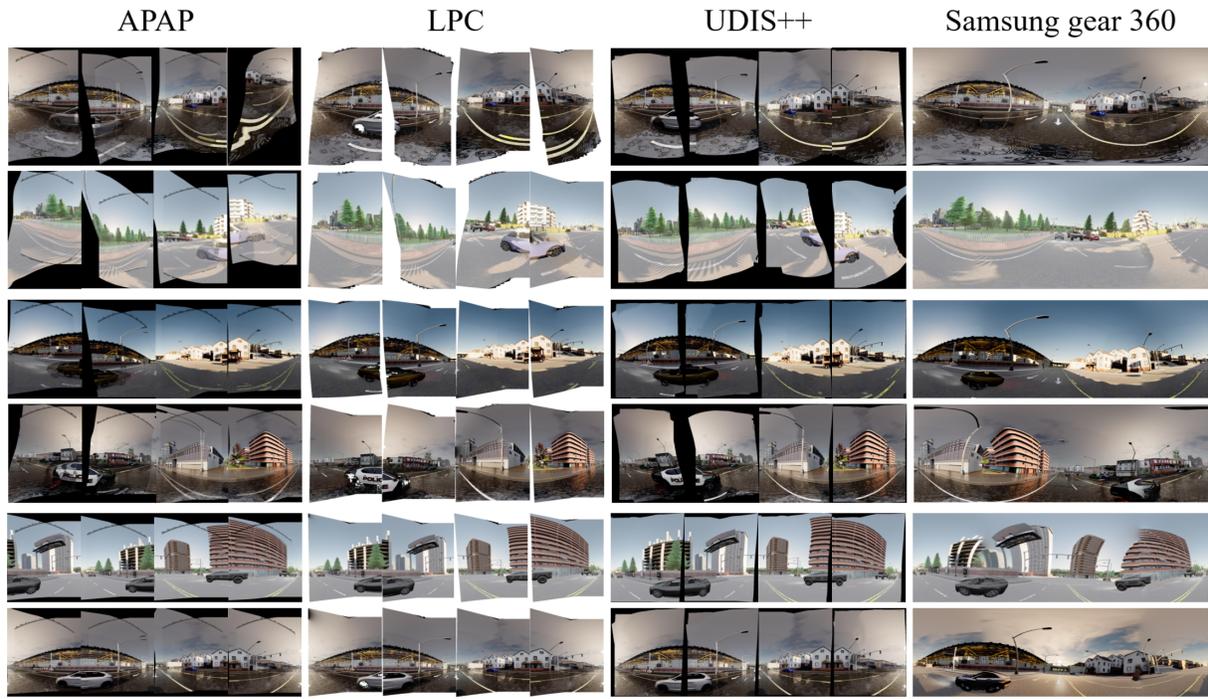
D QUALITATIVE RESULTS ON GV360 DATASET

This section provides a comprehensive overview of the qualitative results from various image stitching models, as shown in Figure 6 and Figure 7. We have focused our detailed analysis on the more advanced models—PTGui, VSLA-like, and OmniStitch—where meaningful comparisons are feasible. For clarity, the comparison setup is organized by Distance parallax: the first and second columns feature a distance of 1.4 meters, the third and fourth columns a distance of 0.8 meters, and the fifth and sixth columns a distance of 0.01 meters. To facilitate accurate comparisons, the output from each model has been scaled to match the size of the ground truth.



Figure 5: The configuration of the GV360’s inter camera distance settings. Distance parallax – 1.4 m (1 row), 0.8 m (2 row), 0.01 m (3 row).

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290



291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

Figure 6: Qualitative results with APAP, UDIS++, Samsung gear 360 on GV360 dataset. Distance parallax – 1.4 m (1,2 row), 0.8 m (3,4 row), 0.01 m (5,6 row).

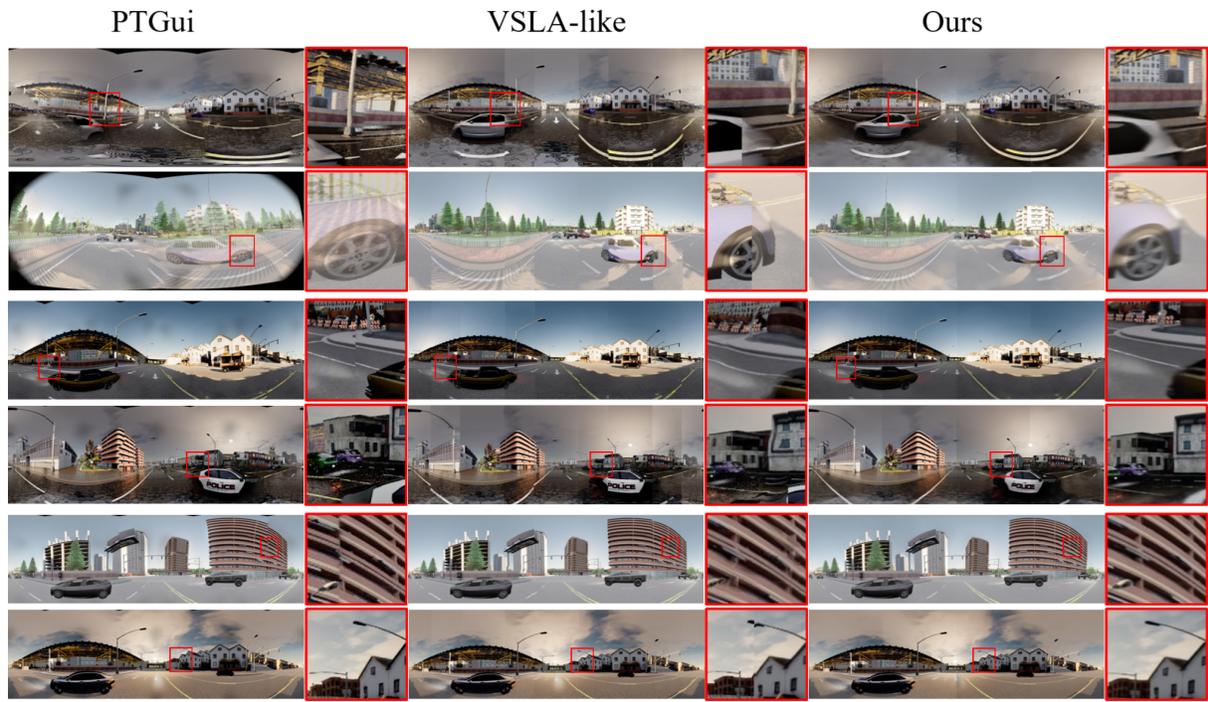


Figure 7: Qualitative results with PTGui, VSLA-like, OmniStitch on GV360 dataset. Distance parallax – 1.4 m (1,2 row), 0.8 m (3,4 row), 0.01 m (5,6 row).

REFERENCES

- [1] Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, and Jiaxin Wang. 2022. Geometric structure preserving warp for natural image stitching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3688–3696.
- [2] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. 2023. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1578–1587.
- [3] Xin Jin, Longhai Wu, Guotao Shen, Youxin Chen, Jie Chen, Jayoon Koo, and Cheul-hee Hahm. 2023. Enhanced bi-directional motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5049–5057.
- [4] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. 2017. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on multimedia* 20, 7 (2017), 1672–1687.
- [5] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. 2023. Parallax-tolerant unsupervised deep image stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7399–7408.
- [6] Simon Niklaus and Feng Liu. 2020. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5437–5446.
- [7] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.
- [8] Tian-Zhu Xiang, Gui-Song Xia, Xiang Bai, and Liangpei Zhang. 2018. Image stitching by line-guided local warping with global similarity constraint. *Pattern recognition* 83 (2018), 481–497.

Unpublished working draft.
Not for distribution.