Supplementary Material for: RetailGlue: Semantic Product-Level Image Stitching for Retail Shelf Panoramas

1. Overview

This supplementary document provides extended qualitative results to support the findings presented in the main manuscript. Specifically, we present comprehensive visual comparisons between our proposed semantic pipeline (**RetailGlue**) and the baseline local feature matcher (**DISK + LightGlue**).

Due to the extreme visual repetition inherent to retail environments—such as identical product packaging, uniform price tags, and parallel shelving structures—local feature-based pipelines frequently establish false correspondences. These erroneous matches cause sequential homography estimations to accumulate catastrophic drift, ultimately leading to severe geometric distortion or total pipeline failure. The following sections illustrate how our product-level approach robustly resolves these ambiguities across various challenging shelf sequences.

2. Additional Qualitative Results

2.1. Success Cases in Dense Retail Scenes

In Fig. 1, we present three distinct retail shelf sequences. For each sequence, the left column displays the panorama stitched by the DISK + LightGlue baseline, while the right column shows the result produced by RetailGlue.

As observed in the baseline results, dense matchers struggle to differentiate between adjacent identical products, leading to misaligned shelves, overlapping products, and "ghosting" artifacts. In contrast, RetailGlue anchors the stitching process to semantic product identities defined by robust DINOv3 embeddings. This effectively eliminates the search space of ambiguous background textures, yielding geometrically consistent and perfectly aligned full-shelf panoramas.

2.2. Failure Cases in Large-Scale Scenarios

To better understand the operational boundaries of our method, we provide examples of failure cases in Fig. 2. As discussed in the main manuscript, our product-level formulation relies entirely on detected bounding boxes to establish geometric constraints.

When a retail scene is dominated by massive items (e.g., bulk diaper packages or large toilet paper rolls) and captured with narrow overlap, the number of fully visible, complete products in the overlapping region may drop below the geometric requirements for RANSAC. In these specific edge cases, RetailGlue fails to estimate a reliable homography, resulting in disconnected panoramas. Conversely, pipelines relying on local feature extractors like DISK can successfully leverage the abundant textural details printed on a single large package to anchor the stitch, as seen in the left column.

Figure 1. **Qualitative Comparisons (Success Cases).** Side-by-side evaluation in highly repetitive retail environments. The baseline local feature matcher (left column) frequently suffers from geometric drift and structural misalignment. Our semantic approach (right column) successfully resolves these visual ambiguities, ensuring accurate alignment.

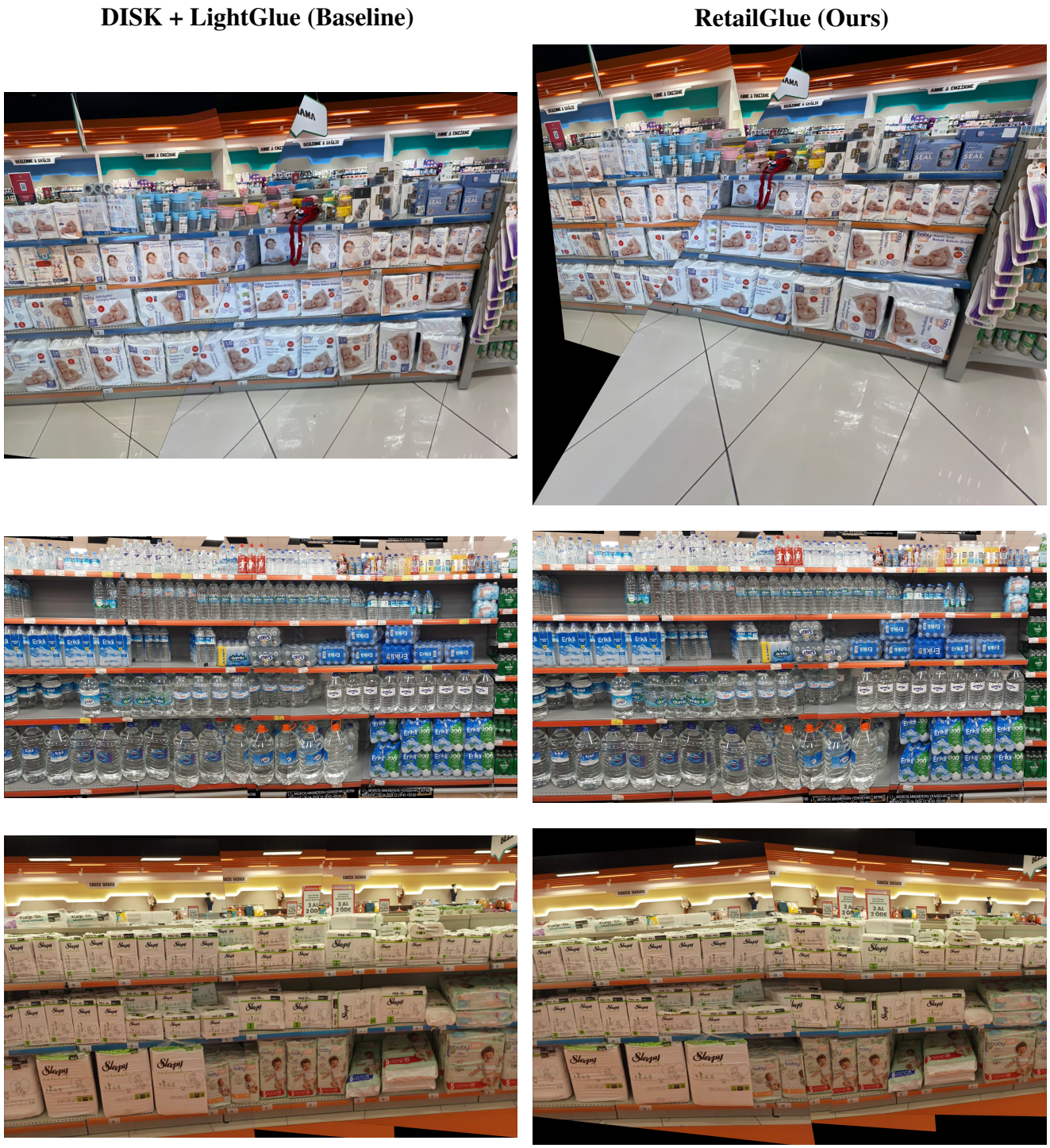

Figure 2. **Failure Cases.** Edge cases in scenes dominated by large-scale packaging with narrow capture overlap. RetailGlue (right column) struggles due to an insufficient number of semantic product keypoints to compute a homography. In contrast, the baseline pipeline (left column) successfully extracts abundant textural features from the large packaging to align the images.