



International Journal of Emerging Technologies and Advanced Applications

# Support-Gated Dense Hypothesis Selection for Three-View Aerial Video Stitching

Xuebing Wang<sup>1</sup>

<sup>1</sup>School of New Energy, Inner Mongolia University of Technology, Ordos 017010, China  
zxbx240130@sina.com

**Abstract**—Three-view aerial video stitching requires spatial alignment and control of historical warp drift. We evaluate a target-training-free dense hypothesis-selection system using released LoFTR and PixelStitch components, observed third-view support, and a bounded historical proposal. Published synthetic depth and real laser geometry evaluate the actual backward sampler independently of the matcher. A locally frozen study uses six held-out SynDrone towns, three held-out ETH3D rig scenes and five estimator seeds. The guarded model improves synthetic success within 3 pixels from 70.576% to 71.033%, a 0.456-percentage-point difference with a descriptive town-bootstrap 95% interval of [0.187, 0.747] and an exact paired sign-randomization  $p = 0.03125$ . Real-rig success changes from 92.435% to 92.604%, but an adapted official PixelStitch control achieves 93.973%. Ablations attribute nearly all same-backbone gain to current spatial verification; the extra historical mechanism contributes little and does not consistently beat the spatial-only model. Complete fresh-matching inference and common rendering take 6.785 seconds per triplet, versus 4.530 for the pairwise control on the measured configuration. Two public multi-UAV clips demonstrate wide-angle output without independent dense accuracy truth. The evidence supports a modest spatial-selection improvement, not a general advantage from temporal memory, real-time operation or universal superiority.

**Index Terms**—Video stitching, Three-view geometry, Dense warps, Temporal transport, Independent evaluation

## I. INTRODUCTION

Three overlapping aerial streams can extend spatial coverage, but their fusion requires more than concatenating images. Depth-dependent parallax breaks a single-plane model, changing overlap weakens geometric evidence, and camera or object motion can make a temporally smooth warp inaccurate. The intended output is one continuous wide-angle video whose observed content remains traceable to its three inputs.

Classical panorama construction combines robust registration and compositing [1], [2]. Shape-preserving and parallax-tolerant warps address the limitations of a global homography [3], [4]; learned methods provide increasingly flexible correspondence and deformation models [5]–[7]. These developments motivate a practical question: when several plausible dense warps are available, which current and historical hypotheses should a three-view system trust?

We examine a target-training-free hypothesis-selection system built on released pretrained components. Its specific addition is an eligibility and selection rule: use indirect third-view evidence only when observed outer-view support exists, and admit transported history only within a bounded neighborhood of the current spatial solution. A heuristic propagated-risk penalty further discourages uncertain history. These components are tested separately. Neither three-view input nor the combination of established modules is claimed as a new fundamental paradigm.

An equally important issue is how to evaluate a rendered result. A correspondence error computed from an internal forward field need not describe the output of a different backward sampler. Likewise, a reference estimated by the evaluated matcher cannot independently establish accuracy. We therefore derive reference correspondences from published synthetic depth and real laser geometry, and score the actual rendering maps. Genuine multi-UAV clips provide application evidence with explicitly different limits.

The study contributes a concrete support-gated candidate-selection implementation, a controlled assessment against matching and compute-related alternatives, and a renderer-consistent evaluation with retained negative development findings. Improvements, if supported by the held-out results, must be attributed to the relevant component rather than to stronger initialization, additional candidate evaluations or seam fusion. The protocol was locally frozen before independent model scoring; this does not constitute public preregistration.

## II. RELATED WORK

### A. Spatial models and learned stitching

Automatic panorama construction and multiresolution blending are established foundations [2], [8]. Shape-preserving half-projective and parallax-tolerant warps handle non-planar alignment while constraining distortion [3], [4]. Mixture-based deformation models and segmentation-guided multiple homographies also precede this work [9], [10]. Consequently, using more than one geometric candidate is not by itself an originality claim.

Unsupervised deep stitching learns alignment without a complete target panorama [5]. UDIS++ addresses parallax-tolerant alignment [6], and PixelStitch supplies the released pixel-wise bidirectional refinement used here [7]. Remote-sensing methods investigate large geometric and radiometric differences [11], [12]; optimal-plane work offers another recent direction [13]. We distinguish these publications from the particular author-code adaptations actually executed in our experiments.

### B. Matching, consistency and video

LoFTR, LightGlue, RoMa and Efficient LoFTR represent different modern matching approaches [14]–[17]. Robust estimators such as RANSAC and MAGSAC++ address outlier correspondences [1], [18]. Improving this initialization can improve an entire stitching pipeline, so the main controls share one frozen LoFTR input cache. Cycle-consistency fusion has also been investigated for image matching [19]; the presence of a cycle is not an independent novelty claim or a guarantee of correctness.

Spatiotemporal content-preserving video warping predates recent neural models [20]. StabStitch addresses stitching-induced shake, and StabStitch++ combines spatial, temporal and smoothing networks with a released three-video example [21], [22]. Recent temporal-propagation and dynamic-memory work further limits any claim that historical warp memory itself is new [23]. DIS provides the current implementation’s optical flow [24]; learned iterative flow such as RAFT is related but is not substituted into the frozen temporal module [25].

Our scope is a particular image-only candidate rule and its measured tradeoffs, not a new optical-flow backbone, a new blending algorithm or a comprehensive reproduction of all recent systems. The experiment explicitly checks whether history adds value beyond current spatial evidence.

## III. METHOD

### A. Overall architecture

Figure 1 summarizes the implemented system. Its design is a candidate-selection pipeline around frozen pretrained components: it uses the current three-view evidence to constrain whether a previous warp may compete, rather than directly averaging frames or learning a new end-to-end stitching network. The current branch constructs multiple bidirectional dense fields from LoFTR matches, sequential homographies and PixelStitch refinement. An initial third-view verification supplies a current spatial reference. In the historical branch, bidirectional DIS flow transports the previous selected field; history is admitted only where flow checks pass, a current candidate is valid, and the proposal lies within 3 pixels of that spatial reference. A texture-dependent propagated-risk penalty increases its cost. The final pool contains the current candidates and eligible history, undergoes one third-view verification pass, and selects the minimum-cost field per pixel. Selected fields and risk scores form the next-frame state. Reference-to-source sampling and common seam/blending operations then produce

the panorama. The first frame and each reset boundary use current evidence only. Independent depth/calibration references enter evaluation, never these inference branches.

### B. Task and candidate construction

At time  $t$ , three overlapping RGB images  $I_t^i$ ,  $i \in \{0, 1, 2\}$ , are mapped to a canvas referenced to view 1. We seek a continuous wide-angle video, not a metric orthophoto or arbitrary novel view. Inference uses RGB only. Depth, poses and visibility labels belong exclusively to evaluation. Frozen pretrained networks make the method target-training-free, but not free of supervised external pretraining.

LoFTR correspondences are shared by every homography-dependent control [14]. Sequential RANSAC fits up to four homographies per edge, removes the current inliers at a 3-pixel threshold, and repeats on the remaining matches [1]. Every estimator seed explicitly permutes the same correspondences. The three edges are  $(0, 1)$ ,  $(2, 1)$  and  $(0, 2)$ . The released PixelStitch network refines each homography bidirectionally [7]. Our reference-plane adapter differs from the author’s intermediate-plane pipeline; both the original-plane adaptation and a single-candidate version of our adapter are therefore evaluated separately.

Let  $F_{ij}^k(x)$  denote candidate  $k$ . Its unary cost is

$$u_{ij}^k(x) = \bar{e}_{\text{photo}}(x) + 0.01 \min(e_{\text{cycle}}(x), 20), \quad (1)$$

where  $\bar{e}_{\text{photo}}$  is the  $5 \times 5$  average of channel-mean absolute differences between normalized images. Cycle error is measured in the source image. Invalid candidates receive infinite cost. If all candidates are invalid, the first current homography is used; an unsuccessful fit falls back to the previous homography or identity. Such fallbacks remain eligible for failure in evaluation.

### C. Observed third-view support

Initial unary selections define the indirect path  $P_{ij} = F_{kj} \circ F_{ik}$ , where  $k \neq i, j$ . Let  $v_{ij} = u_{ik} + u_{kj} \circ F_{ik}$ . The reliability  $a_{ij} = \max(0, 1 - v_{ij}/0.15)$  is zero outside the valid path domain. Candidate selection adds

$$0.05 a_{ij}(x) \frac{\min(\|F_{ij}^k(x) - P_{ij}(x)\|_2, 20)}{20}. \quad (2)$$

We make one verification pass to avoid iterative reinforcement of correlated errors. The term is disabled if the outer-view homography has fewer than 12 inliers or an inlier ratio below 0.1. These image-only checks do not certify correctness: paths share images, features and pretrained weights. They merely determine whether indirect evidence is usable.

### D. Transport with a bounded historical proposal

DIS bidirectional optical flow supplies the inter-frame maps [24]. A previous warp becomes the current historical proposal

$$\hat{F}_{i,j,t} = T_{j,t-1 \rightarrow t} \circ F_{i,j,t-1} \circ T_{i,t \rightarrow t-1}. \quad (3)$$

Both endpoint flow checks must have forward–backward residual below 1.5 pixels, and a current candidate must already

## THREE-VIEW STITCHING WITH SUPPORT-GATED HISTORY

Frozen pretrained networks | No target-scene fitting

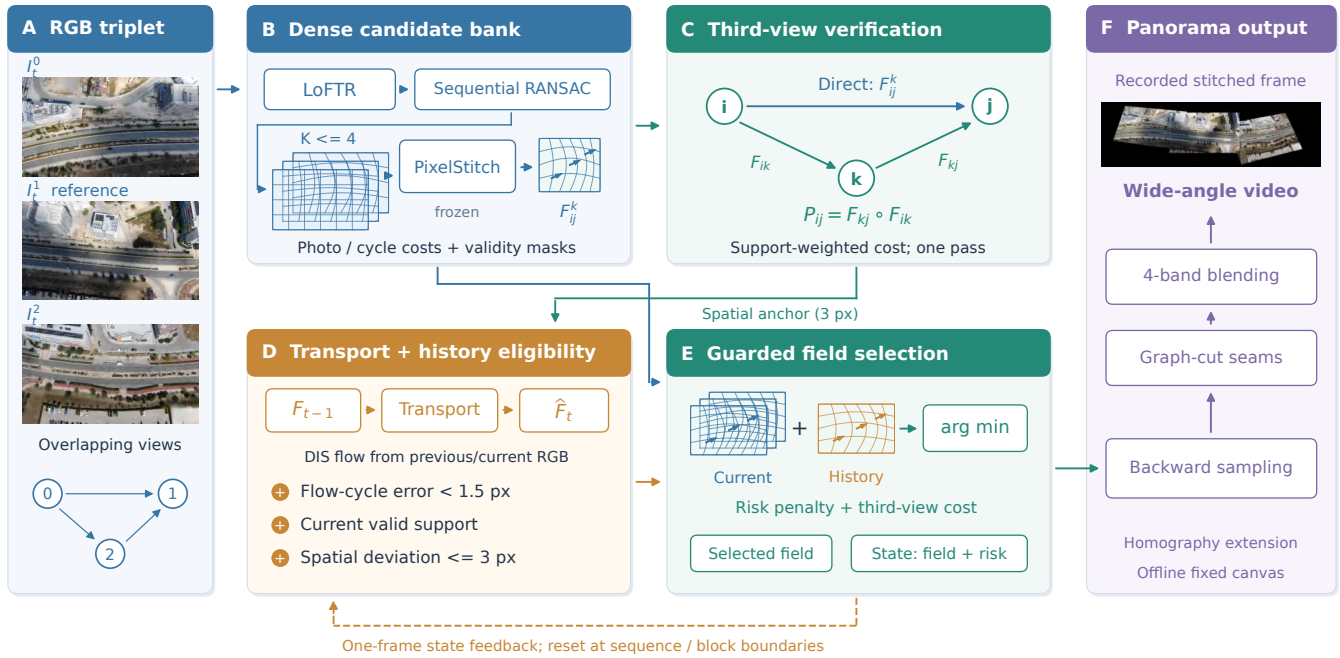


Fig. 1: Technical architecture of the guarded three-view stitching system. (A) Three overlapping RGB views. (B) Frozen matching/refinement with multiple homography hypotheses. (C) Direct and indirect geometric paths. (D) Historical transport and eligibility conditions. (E) Current/history competition and state update. (F) Common backward rendering and fusion. Internal grid/field glyphs are schematic; input and output thumbnails use recorded traffic frame 0 from the publisher-declared Apache-2.0 resource [26]. Dashed arrows indicate next-frame state. Component benefits are assessed separately by ablation.

support the location. Historical transport cannot expand this support domain.

For the minimum structure-tensor eigenvalue  $\lambda_{\min,i}$ , define

$$\sigma_i = \text{clip} \left( \frac{0.01}{\sqrt{\max(\lambda_{\min,i}, 0)} + 0.001}, 0.05, 10 \right). \quad (4)$$

With the transported previous score  $r_{t-1}$ , a Jacobian magnitude  $g$  clipped to  $[0.1, 10]$ , and endpoint flow-cycle residuals  $e_i, e_j$ , we use

$$r_t = \sqrt{r_{t-1}^2 + (g\sigma_i)^2 + \sigma_j^2 + e_i^2 + e_j^2}. \quad (5)$$

All quantities are sampled at their corresponding transported positions. History receives the additional cost  $0.01 \min(r_t, 20)$ ; selection of a current candidate resets the score to zero. This is an error-propagation-inspired heuristic, not a calibrated variance or probability.

Development experiments showed that this penalty alone can retain harmful history. The frozen candidate additionally requires

$$\|\hat{F}_{ij,t}(x) - F_{ij,t}^{\text{spatial}}(x)\|_2 \leq 3. \quad (6)$$

Here  $F^{\text{spatial}}$  is the current frame's spatially verified field. This restricts the historical proposal; it does not prove a bound on the final inverse-rendered endpoint, nor ensure that every

selected current hypothesis matches the spatial-only result. A mean-matched control replaces the spatially varying risk penalty by its mean over eligible locations for each frame and directed edge. It separates spatial variation from average penalty strength.

### E. Common rendering

The panorama directly samples each source through its reference-to-source field. Inside the reference rectangle, a 16-pixel border transition approaches the inverse homography; outside, the analytic inverse homography supplies the extension. Common graph-cut seams and four-band blending use established methods [8], [27]. Their visual benefit is not attributed to the proposed geometry. Feathered outputs expose residual ghosting.

For full application clips, an image-only homography prepass sets one fixed canvas from the union of all frame footprints. This viewport procedure is offline. It must not be confused with an online, constant-memory system. No global bijection, absence of duplicated content, or complete recovery of occluded regions is guaranteed.

#### IV. EXPERIMENTAL DESIGN

##### A. Data and evidence boundaries

SynDrone provides synchronized synthetic aerial RGB, depth and camera information [28]. Its three cameras share horizontal position but use different heights and pitches; they are not a real three-UAV formation. We use the released split: Town01 and Town02 training frames for development, and the official test blocks in the six remaining towns for held-out study evaluation. Each town contributes five disjoint 20-frame blocks sampled at 5 fps, yielding 600 held-out triplets. Temporal state resets between blocks. RGB is resized to  $960 \times 540$  using consistent pixel-center coordinates.

ETH3D supplies real synchronized camera-rig videos with laser-derived geometry [29]. The development scenes are delivery\_area and electro; forest, playground and terrains are held out, with 257, 240 and 165 frames respectively. This is a study split within the publicly labelled training release, not the official hidden test benchmark. Cameras 4, 5 and 6 are fixed, with camera 5 as reference. The rig is short-baseline and ground-based; it does not establish aerial-formation performance.

The public traffic and MP-ZEBRA clips provide application demonstrations [26], [30]. Neither supplies independent dense stitching truth in the downloaded material. MP-ZEBRA's timestamp pairing is not perfect synchronization; the recorded median and maximum offsets are 12 and 97 ms. These previously inspected clips are not held-out geometry evidence. Raw data remain on the experiment server.

Archive extraction uses file CRC checks; downloads retain source and transfer records. A decoded-image audit found no exact duplicates between the selected 504 development RGB images and 1,800 held-out SynDrone RGB images; no cross-split 64-bit perceptual-hash candidates met the preselected Hamming-distance threshold of four. This check neither removes shared simulator assets nor audits all external pretraining data.

##### B. Independent reference geometry

SynDrone reference correspondences are derived from published optical-axis depth and relative camera geometry, with stable-depth and mutual-visibility checks. The unknown common terrain-height offset cancels for same-frame relative geometry; it does not justify synthetic world-coordinate temporal truth. ETH3D references use calibrated projection, raw-to-undistorted coordinate conversion and laser-depth visibility checks. Source and target depth consistency, masks, strides and thresholds are fixed in the reference-generation code.

A source reference point is scored at the coordinate recovered by inverting the actual backward sampler. Interior roots use image-predicted initialization and a deterministic multiseed procedure; no target truth selects a root. If no interior root is found, a finite exterior homography root is accepted only outside the reference rectangle. Missing roots are failures. Published TPS baselines are similarly evaluated through their sampling maps and finite canvas. Local folding and missing roots are reported rather than silently excluded.

##### C. Endpoints and uncertainty

The primary endpoint is the fraction of independent landmarks aligned within 3 reference-image pixels. Average landmarks within each frame and edge, then edges, frames within block, blocks within town and towns equally. Report 1/5/10-pixel success, missing predictions, median error and separately labelled error capped at 50 pixels. An unconditional raw mean is infinite when a prediction is missing; a finite-only mean is not substituted.

The five estimator seeds are 17, 29, 43, 71 and 101. They permute the same correspondence sets and are repeated algorithm estimates, not independent scenes or few-shot training supports. Their mean and sample standard deviation are reported separately from scene uncertainty. The sole primary contrast is the guarded model minus its four-candidate pairwise control. A two-sided exact sign randomization uses the six seed-averaged town effects; a paired 10,000-draw town bootstrap provides a descriptive 95% interval. Six simulator towns support limited inference. Secondary contrasts do not receive uncorrected discovery claims.

ETH3D temporal evaluation compares alignment-residual changes at the same laser-derived world landmark in adjacent frames, re-anchoring every 20 frames. Missing predictions remain failures and contribute the declared cap. Absolute motion of the reference camera is not called stitching jitter. With only three held-out real scenes, per-scene effects are essential and broad statistical claims are unwarranted.

##### D. Controls and implementation

Controls share image preprocessing, correspondence order and reference points: a single homography; reverse meshes with 10- and 80-pixel residual screening; a single dense hypothesis; four- and six-candidate pairwise models; official PixelStitch intermediate-plane inference with the shared LoFTR initialization; and native and homography-prealigned StabStitch++ author networks [7], [22]. The last two are explicitly identified as three-view adaptations, and the pre-aligned variant further changes input preprocessing. Original-paper numbers are never inserted into our measured tables. Deterministic native StabStitch++ runs once; homography-dependent controls use all five seeds.

Ablations remove historical transport, its penalty, third-view verification, the witness gate or the new trust region. The mean-matched penalty control isolates spatial adaptivity. A six-candidate two-edge control matches the maximum 12 backbone calls of the four-candidate three-edge model. Shared-bank ablation runs compute extra candidates and are not used to claim single-method speed. A standalone pairwise implementation has separately checked numerical equivalence.

Experiments use an RTX 5070 Laptop GPU with 12 GB memory, driver 595.91.07 and PyTorch 2.11.0 with CUDA 12.8. Actual GPU arithmetic was checked after server recovery. The protocol, code, weights and seeds were locally frozen before held-out scoring. This record is not public preregistration. Development alternatives, incomplete attempts and negative findings remain archived. Efficiency requires

TABLE I: Frozen held-out geometry. Success is in percent;  $\pm$  is estimator seed SD, not a scene confidence interval. EPE is explicitly capped at 50 pixels. Native StabStitch++ is deterministic and has no estimated seed SD.

| Method                 | Synthetic success@3 | Synthetic capped EPE | Real-rig success@3 | Real-rig capped EPE |
|------------------------|---------------------|----------------------|--------------------|---------------------|
| Homography             | 66.07 $\pm$ 0.10    | 10.28                | 76.03 $\pm$ 0.06   | 3.63                |
| Mesh (10 px)           | 68.42 $\pm$ 0.08    | 9.51                 | 79.52 $\pm$ 0.05   | 3.03                |
| Mesh (80 px)           | 62.27 $\pm$ 0.13    | 8.26                 | 81.75 $\pm$ 0.01   | 2.10                |
| Single dense           | 70.62 $\pm$ 0.05    | 8.34                 | 92.34 $\pm$ 0.02   | 1.21                |
| PixelStitch H-adapted  | 69.13 $\pm$ 0.04    | 8.15                 | 93.97 $\pm$ 0.01   | 1.03                |
| StabStitch++ native    | 0.07                | 47.07                | 14.28              | 26.51               |
| StabStitch++ H-adapted | 0.04 $\pm$ 0.00     | 48.15                | 23.25 $\pm$ 0.06   | 17.85               |
| Pairwise K4            | 70.58 $\pm$ 0.05    | 9.03                 | 92.44 $\pm$ 0.04   | 1.51                |
| Pairwise K6            | 68.60 $\pm$ 0.07    | 10.29                | 90.29 $\pm$ 0.06   | 2.53                |
| Spatial only           | 71.02 $\pm$ 0.06    | 8.82                 | 92.61 $\pm$ 0.03   | 1.47                |
| Guarded history        | 71.03 $\pm$ 0.06    | 8.83                 | 92.60 $\pm$ 0.03   | 1.48                |

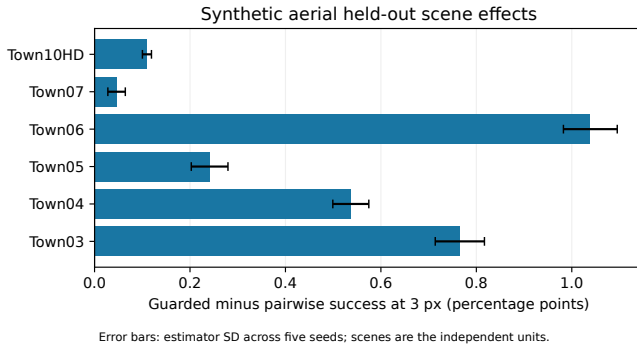


Fig. 2: Synthetic held-out effects. Error bars show estimator-seed SD, not scene confidence intervals.

isolated measurements that separate matching, fitting, refinement, temporal processing, transfer, rendering, decoding and encoding.

## V. RESULTS AND ANALYSIS

### A. Frozen primary comparison

All 45 shared-bank model jobs, 135 geometric/dense baseline jobs and 54 StabStitch++ jobs completed. Each scene and seed retained per-frame predictions. Table I reports equal-scene aggregates rather than pooled-pixel estimates. The guarded model reaches 71.0329% synthetic success, versus 70.5765% for the four-candidate pairwise control. The difference is 0.4564 percentage points, with a descriptive town-bootstrap 95% interval [0.1872, 0.7473] and the prespecified exact two-sided  $p = 0.03125$ . All six town effects are positive, but range from only 0.0462 to 1.0389 percentage points. Statistical consistency across these towns does not make the practical gain large.

The real-rig aggregate increases by 0.1689 percentage points, from 92.4351% to 92.6040%. Its three scene effects are positive, but three scenes provide weak population evidence; a positive bootstrap interval from such a small set must not be interpreted as robust cross-environment proof. The official PixelStitch adaptation is stronger on this data, with 93.9727% success and 1.0261-pixel capped error, compared with 1.4840 for the guarded model.

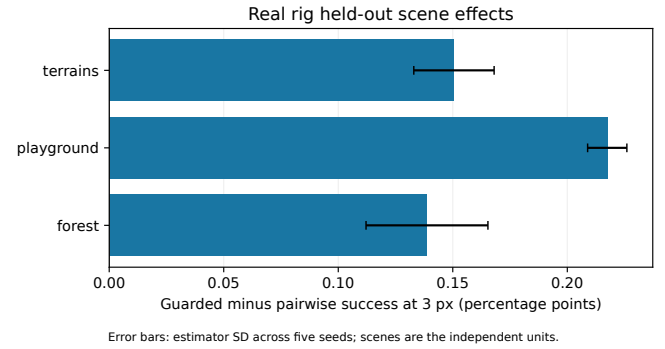


Fig. 3: Real-rig held-out effects. The three scenes are the outer experimental units.

### B. Stronger controls and error tradeoffs

A single dense candidate already achieves 70.6209% synthetic success and a lower capped error of 8.3419 pixels, compared with the guarded model's 8.8335. Thus multiple hypotheses do not improve every endpoint, and the comparison is not evidence that more candidates are universally beneficial. The six-candidate pairwise control is worse on both datasets. Its failure rules out a simple additional-compute explanation, but does not exhaust all possible compute-matched designs.

Final execution auditing found GPU out-of-memory errors in 86 edge records across three concurrent PixelStitch runs. These complete sequences were rerun with identical frozen code, weights and seeds on an exclusive GPU; original attempts are retained, and replacements are determined by execution failure rather than score. Corrected synthetic success is 69.1280% with 0.0421-point seed SD. Its capped error of 8.1462 pixels is lower than the guarded model's 8.8335, reinforcing the accuracy-tail-error tradeoff. Corrected values are reported in Table I. The shared LoFTR initialization does not remove differences in intermediate-plane geometry and adaptive warping. The very low synthetic StabStitch++ scores persist with homography prealignment. We report them as severe viewpoint/domain-shift outcomes of these particular author-code adaptations, not as a broad refutation of the published method or as the central evidence for our contribution.

Missing inverse predictions are failures, not omitted observations. The uncapped mean EPE is infinite whenever an

TABLE II: Ablation success at 3 pixels (%). Every row uses the frozen protocol.

| Variant                    | Synthetic | Real rig |
|----------------------------|-----------|----------|
| Pairwise K4                | 70.577    | 92.435   |
| Spatial only               | 71.018    | 92.608   |
| History only               | 70.332    | 91.661   |
| Unpenalized history + path | 70.778    | 91.945   |
| Constant penalty           | 70.981    | 92.148   |
| Texture penalty            | 71.183    | 92.409   |
| Mean-matched penalty       | 71.157    | 92.382   |
| Texture, no path           | 70.747    | 92.211   |
| No witness gate            | 71.272    | 92.409   |
| Guarded history            | 71.033    | 92.604   |

evaluated query has no prediction; such rows occur in every synthetic guarded frame–edge record. Finite 50-pixel-capped averages in Table I therefore must not be read as finite uncapped errors. The companion per-scene/seed tables retain invalid fractions, infinity flags and success at 1, 5 and 10 pixels.

### C. Ablation and temporal behavior

Spatial-only verification achieves 71.0179% synthetic success. Adding the final historical mechanism changes this by just 0.0150 percentage points; on the real rig, it decreases success by 0.0045 points. The data therefore do not establish a practically meaningful independent accuracy benefit from the added history rule.

Unrestricted historical propagation reduces same-backbone performance. The texture score and its mean-matched constant are close: 71.1833% versus 71.1572% on the synthetic set, and 92.4088% versus 92.3818% on the rig. The evidence for spatial adaptivity itself is small. Removing the witness gate improves the synthetic aggregate to 71.2721%; the frozen guarded model is retained rather than retrospectively replaced by this better test configuration.

The rig’s capped temporal residual change is 1.4712 pixels/frame for pairwise, 1.4211 for guarded and 1.4062 for spatial-only. Official PixelStitch achieves 0.6463, and the single dense candidate 0.7792. Historical restrictions mitigate a development failure but do not produce the best temporal endpoint. Geometry and temporal behavior must both be considered; a stable inaccurate homography is not a successful stitch.

## VI. EFFICIENCY AND QUALITATIVE EVALUATION

Fresh matching, image decoding/resizing, geometry, host/device transfer and common graph-cut/multiband rendering are included in Table III. Raw-video extraction, output-file writing and encoding are excluded. Eighty measured triplets per method come from four balanced-order timing sessions; these repeated frames are not independent scientific samples. The guarded pipeline takes 6.785 seconds/triplet, about 50% more than pairwise K4 and 27% more than pairwise K6. It is not real-time. Encoded application videos at 5 fps do not imply 5-fps processing.

The released matcher has 11,561,456 parameters and the dense backbone 5,650,312; their checkpoints occupy approximately 65.83 MiB together. The eligibility rules add no learned parameters or target-scene fitting stage. Per-frame robust fitting remains part of inference. Model loading and retained video-state arrays are recorded separately; allocated CUDA peaks are not disk size or parameter count.

Both clips contain 300 decoded output frames per method. The final all-frame fixed viewport has no base-homography corner outside its bounds. This addresses the detected canvas truncation, not missing geometry or occlusion. Irregular black borders are retained; no generative filling is used. Building-edge deformation, local ghosting and moving-content inconsistencies can remain despite seam fusion.

The deterministic worst/median selection in Fig. 5 exposes failure rather than showing only successful cases. Reference masks emphasize stable covisible geometry; image boundaries and dynamic-object quality require additional scrutiny.

## VII. LIMITATIONS AND REPRODUCIBILITY

Synthetic aerial depth permits controlled geometry evaluation but does not reproduce every flight, camera or scene condition. ETH3D supplies real geometry for a short-baseline ground rig, not a three-UAV formation. The genuine multi-UAV application clips lack independent dense geometry truth and include imperfect synchronization. These evidence sources cannot jointly be described as a fully validated operational aerial system.

The reference masks emphasize depth-stable, mutually visible regions. Their scores do not describe every boundary pixel or independently validate moving-object alignment.

The selected rules use several fixed heuristic thresholds. The propagated score is not a calibrated uncertainty model; current spatial agreement can reject a correct historical alternative or preserve a shared error. Repeated textures, moving objects, large occlusions and nearly disjoint outer views remain difficult. The backward field may fold or duplicate content, and deterministic inverse-root selection does not establish global invertibility. Graph-cut fusion can conceal an overlap error while leaving the geometry wrong.

Pretrained data may overlap unknown external sources. The selected-image duplicate audit addresses only the explicit development/held-out material and does not make an exhaustive pretraining-independence claim. Five estimator seeds measure algorithm variability, not five independent acquisitions. Six simulated towns and three real-rig scenes limit population-level conclusions.

Author baselines retain their native internal resizing and normalization after receiving the common input images. Their network architectures, training distributions and preprocessing differ. Both native and homography-prealigned StabStitch++ can be affected by domain shift; weak performance here cannot establish universal superiority. Recent related methods that were not executed remain a limitation of comparison coverage.

TABLE III: Isolated development-input efficiency, mean seconds per triplet over four sessions of 20 measured frames, each after five warm-up frames. Matching is fresh. GPU peak is maximum allocated memory, in GiB. Encoding is excluded.

| Method          | Read/resize | Matching | Geometry | Rendering | Total | GPU peak |
|-----------------|-------------|----------|----------|-----------|-------|----------|
| Single dense    | 0.015       | 0.400    | 0.474    | 2.371     | 3.260 | 1.733    |
| Pairwise K4     | 0.015       | 0.400    | 1.711    | 2.405     | 4.530 | 1.733    |
| Pairwise K6     | 0.015       | 0.400    | 2.553    | 2.359     | 5.326 | 1.733    |
| Guarded history | 0.015       | 0.587    | 3.877    | 2.306     | 6.785 | 1.733    |

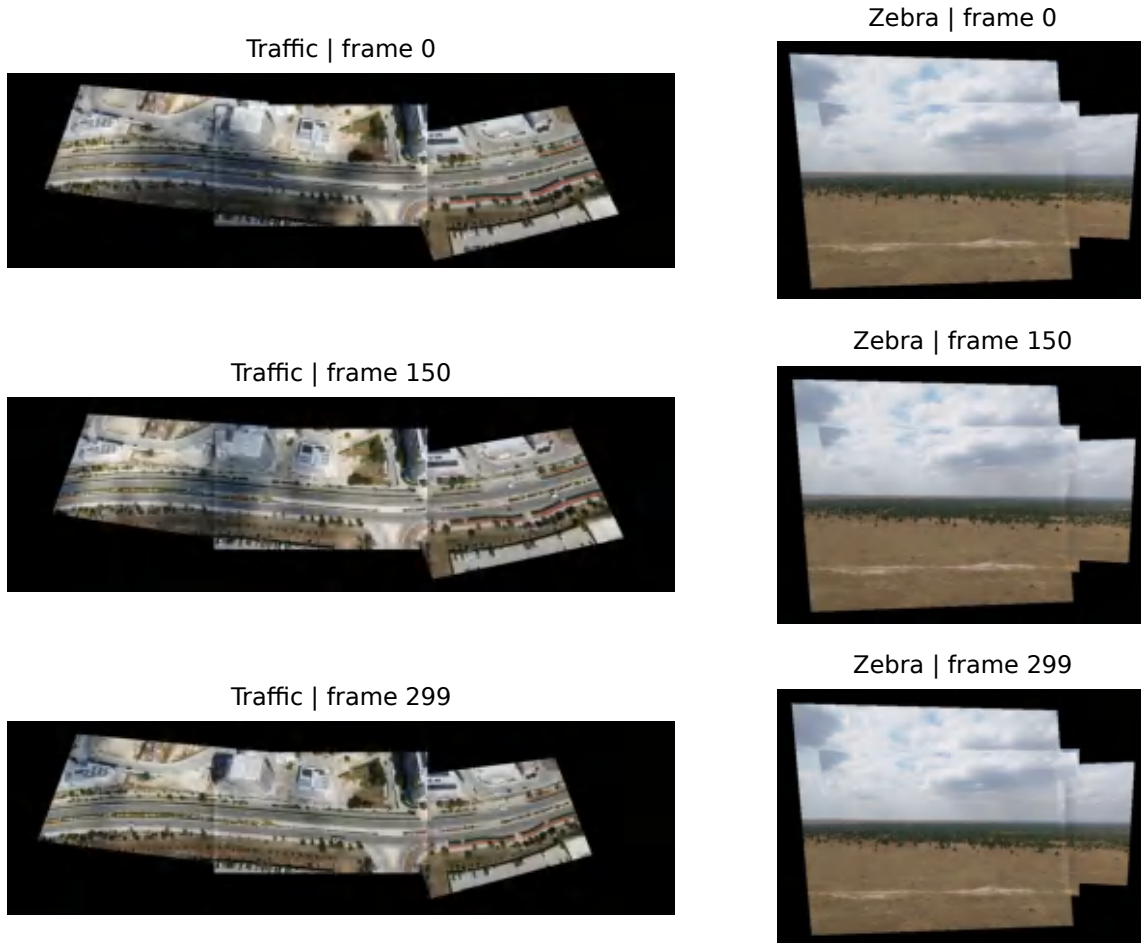


Fig. 4: Public traffic and MP-ZEBRA application outputs at fixed first, middle and last frames (0, 150, 299). No independent dense accuracy claim follows from these images. Source releases: [26], [30]; respective publisher-declared Apache 2.0 and CC BY 4.0 terms.

Source, configurations, frozen hashes, runtime versions, per-frame predictions and unsuccessful development alternatives are retained. Data redistribution follows the source-specific terms: SynDrone reports a CC-BY data license without a version asserted here; ETH3D uses CC BY-NC-SA 4.0; the traffic release declares Apache 2.0 and MP-ZEBRA declares CC BY 4.0. Repository licenses and pretrained-weight provenance are recorded separately. Raw datasets are not included in the anonymous manuscript source package.

#### DECLARATIONS

No external funding and no competing interests were reported by the author. The study uses public data and reports no new human-participant or animal intervention; no institutional ethics approval is asserted. AI assisted research planning, code development, experiment orchestration, literature metadata checking, drafting, translation and simulated review. This disclosure does not assert completed human verification. Author identity and contribution confirmation are supplied in a separate file.

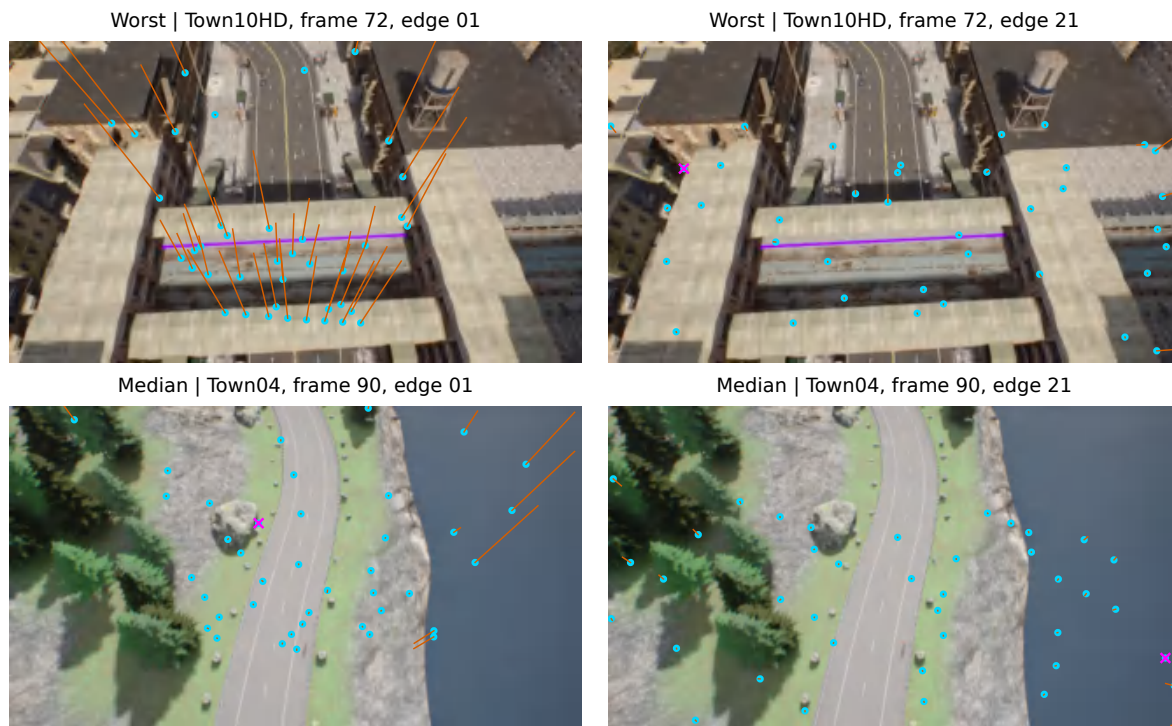


Fig. 5: Worst and upper-median held-out synthetic frames by the guarded model's seed-17 frame success, selected over all six towns. Cyan marks denote reference endpoints, orange segments point to predictions, and magenta crosses denote missing predictions. Forty uniformly indexed reference points per edge are displayed; all eligible points enter scoring. Source imagery: SynDrone [28], CC-BY as stated by its authors.

### VIII. CONCLUSION

The frozen study supports a small, consistent synthetic same-backbone improvement from three-view spatial hypothesis verification, with a smaller real-rig gain. It does not establish a substantial added benefit from historical risk modeling or superiority to the stronger real-rig PixelStitch control. The complete pipeline is slower and remains vulnerable to distortion, visibility changes and synchronization errors. Its useful outcome is a reproducible candidate-selection implementation and an independently grounded account of where its components help or fail. Real formation recordings with independent geometry, broader recent baselines and a demonstrably stronger temporal mechanism remain future work requiring new validation; the completed held-out set must not be reused as an untouched test after further tuning.

### REFERENCES

- [1] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981, DOI: 10.1145/358669.358692.
- [2] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007, DOI: 10.1007/s11263-006-0002-3.
- [3] C.-H. Chang, Y. Sato, and Y.-Y. Chuang, "Shape-Preserving Half-Projective Warps for Image Stitching," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3254–3261, DOI: 10.1109/cvpr.2014.422.
- [4] F. Zhang and F. Liu, "Parallax-Tolerant Image Stitching," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3262–3269, DOI: 10.1109/cvpr.2014.423.
- [5] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images," *IEEE Transactions on Image Processing*, vol. 30, pp. 6184–6197, 2021, DOI: 10.1109/tip.2021.3092828.
- [6] —, "Parallax-Tolerant Unsupervised Deep Image Stitching," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 7365–7374, DOI: 10.1109/iccv51070.2023.00680.
- [7] H. Jin, L. Nie, C. Lin, X. Feng, and Y. Zhao, "PixelStitch: Structure-Preserving Pixel-Wise Bidirectional Warps for Unsupervised Image Stitching," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 28 125–28 134, DOI: 10.1109/iccv51701.2025.02611.
- [8] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Transactions on Graphics*, vol. 2, no. 4, pp. 217–236, 1983, DOI: 10.1145/245.247.
- [9] Z. Zhang, X. Yu, and X. Yang, "Image alignment using mixture models for discontinuous deformations," *Signal Processing*, vol. 195, p. 108467, 2022, DOI: 10.1016/j.sigpro.2022.108467.
- [10] T. Liao, C. Wang, L. Li, G. Liu, and N. Li, "Parallax-tolerant image stitching via segmentation-guided multi-homography warping," *Signal Processing*, vol. 230, p. 109860, 2025, DOI: 10.1016/j.sigpro.2024.109860.
- [11] L. Qiu, C. Liu, G. Li, X. Dong, F. Xie, and Z. Shi, "Unsupervised Deep Image Stitching for Remote Sensing: Aligning Features Across Large-Scale Geometric Distortions," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 25 305–25 324, 2025, DOI: 10.1109/jstars.2025.3609808.

- [12] L. Qiu, F. Xie, C. Liu, X. Che, and Z. Shi, "Radiation-Tolerant Unsupervised Deep Image Stitching for Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–21, 2025, DOI: 10.1109/tgrs.2025.3605270.
- [13] L. Nie, Y. Mei, K. Liao, Y. Xu, C. Lin, and B. Xiao, "Robust Image Stitching With Optimal Plane," *IEEE Transactions on Visualization and Computer Graphics*, vol. 32, no. 7, pp. 5156–5166, 2026, DOI: 10.1109/tvcg.2026.3663425.
- [14] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-Free Local Feature Matching with Transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8918–8927, DOI: 10.1109/cvpr46437.2021.00881.
- [15] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 581–17 592, DOI: 10.1109/iccv51070.2023.01616.
- [16] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "RoMa: Robust Dense Feature Matching," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19 790–19 800, DOI: 10.1109/cvpr52733.2024.01871.
- [17] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21 666–21 675, DOI: 10.1109/cvpr52733.2024.02047.
- [18] D. Barath, J. Nuskova, M. Ivashechkin, and J. Matas, "MAGSAC++, a Fast, Reliable and Accurate Robust Estimator," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1301–1309, DOI: 10.1109/cvpr42600.2020.00138.
- [19] K. Sun, J. Yu, W. Tao, X. Li, C. Tang, and Y. Qian, "A unified feature-spatial cycle consistency fusion framework for robust image matching," *Information Fusion*, vol. 97, p. 101810, 2023, DOI: 10.1016/j.inffus.2023.101810.
- [20] W. Jiang and J. Gu, "Video stitching with spatial-temporal content-preserving warping," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 42–48, DOI: 10.1109/cvprw.2015.7301374.
- [21] L. Nie, C. Lin, K. Liao, Y. Zhang, S. Liu, R. Ai, and Y. Zhao, "Eliminating Warping Shakes for Unsupervised Online Video Stitching," in *Computer Vision – ECCV 2024*, 2025, pp. 390–407, DOI: 10.1007/978-3-031-73235-5\_22.
- [22] L. Nie, C. Lin, K. Liao, Y. Zhang, S. Liu, and Y. Zhao, "StabStitch++: Unsupervised Online Video Stitching With Spatiotemporal Bidirectional Warps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 7443–7456, 2025, DOI: 10.1109/tpami.2025.3568829.
- [23] X. Yun, F. Yu, K. Dong, and Y. Sun, "Unsupervised video stitching via spatiotemporal coupling with temporal propagation and dynamic memory," *Knowledge-Based Systems*, vol. 344, p. 116048, 2026, DOI: 10.1016/j.knosys.2026.116048.
- [24] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool, "Fast Optical Flow Using Dense Inverse Search," *Lecture Notes in Computer Science*, pp. 471–488, 2016, DOI: 10.1007/978-3-319-46493-0\_29.
- [25] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," *Lecture Notes in Computer Science*, pp. 402–419, 2020, DOI: 10.1007/978-3-030-58536-5\_24.
- [26] JYe9, "Multi-Camera Multi-Vehicle Tracking System (UAV Dataset)," n.d., Accessed: Sep. 21, 2026. [Online]. Available: <https://huggingface.co/datasets/jye9/Multi-Camera-Multi-Vehicle-Tracking-System>
- [27] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 294–302, 2004, DOI: 10.1145/1015706.1015718.
- [28] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "SynDrone – Multi-modal UAV Dataset for Urban Scenarios," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 2202–2212, DOI: 10.1109/iccvw60793.2023.00235.
- [29] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2538–2547, DOI: 10.1109/cvpr.2017.272.
- [30] S. Afridi, S. Bullock, A. Jarabo Penas, and L. Laporte Devylder, "MP-ZEBRA: Multi-Perspective Dataset of Plains Zebras," n.d.,

Accessed: Sep. 21, 2026. [Online]. Available: <https://huggingface.co/datasets/edouard-rolland/multi-perspective-dataset-plain-zebras>