SAGS: STRUCTURE-AWARE GEOMETRY-GUIDED STITCHING FOR COST-EFFECTIVE GLASS CURTAIN WALL INSPECTION WITH CONSUMER UAVS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image stitching for glass curtain walls is critical for unmanned aerial vehicle (UAV)-based infrastructure inspection, yet remains challenging due to specular reflections, transmitted background interference, and repetitive window structures. These factors introduce unreliable appearance cues and periodic correspondence ambiguity, degrading existing feature-based and learning-based matching methods. In this paper, we identify periodic matching ambiguity as a key geometric challenge and derive a sufficient condition for reliable correspondence under bounded displacement uncertainty. Based on this principle, we propose Structure-Aware Geometry-Guided Stitching (SAGS), which exploits facade structures as geometric priors. SAGS employs a fine-tuned Segment Anything Model 3 (SAM 3) to extract facade structural masks, dense-frame geometric calibration to estimate the scale factor and geometric priors, and mask-based cross-correlation for geometry-guided alignment. Experiments on a real-world glass curtain wall dataset demonstrate that SAGS achieves reliable facade panorama stitching using consumer-grade UAVs without real-time kinematic (RTK) positioning or LiDAR sensors by leveraging velocity telemetry as motion priors.

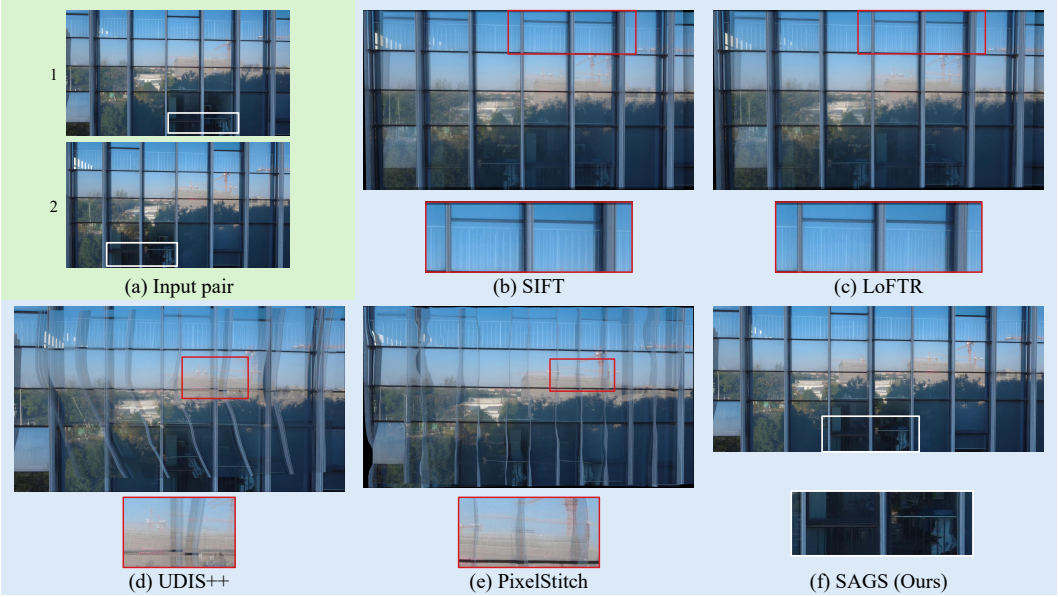


Figure 1: Qualitative comparison of stitching results on a glass curtain wall with repetitive structures and reflection interference. (a) Input image pair; (b–f) results of SIFT, LoFTR, UDIS++, PixelStitch, and SAGS, respectively. White boxes mark manually identified corresponding physical regions in the input images. Red boxes indicate the regions enlarged below each stitching result.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Unmanned aerial vehicle (UAV) inspection captures partial facade views across overlapping images. Image stitching organizes these observations into a continuous 2D panorama for visual review and annotation across window units, requiring consistent correspondence and geometry of physical facade structures. Conventional pipelines estimate geometric transformations from local correspondences using handcrafted features such as SIFT, SURF, and ORB Lowe (2004); Bay et al. (2006); Rublee et al. (2011). Learned matchers such as LoFTR improve correspondence estimation under weak-texture conditions Sun et al. (2021).

Glass curtain wall stitching presents two coupled challenges. Reflections and transmitted backgrounds introduce visual cues that do not consistently follow the physical facade geometry. Meanwhile, repetitive window layouts create multiple plausible correspondences, making the correct window unit difficult to identify. Figure 1 illustrates these difficulties: SIFT and LoFTR exhibit window-frame misalignment consistent with periodic ambiguity in (b,c), while UDIS++ and PixelStitch produce structural distortions in (d,e), potentially influenced by reflected content.

Existing approaches improve alignment through flexible warps and seam-aware processing Zaragoza et al. (2013); Li & Wang (2016), but these mechanisms alone do not explicitly resolve periodic correspondence ambiguity. Extracting physical window structures reduces appearance interference, yet their repetition still leaves matching ambiguous. This motivates combining structural references with UAV motion priors that constrain the displacement range and distinguish similar window units.

We propose Structure-Aware Geometry-Guided Stitching (SAGS) for glass curtain wall inspection using consumer-grade UAVs. Under a one-dimensional periodic model, we formulate a sufficient condition for excluding neighboring periodic correspondences under bounded displacement uncertainty. This provides a basis for using imperfect motion priors without requiring RTK positioning or LiDAR measurements.

SAGS uses a fine-tuned SAM 3 model to extract structural masks for both alignment and seam placement. Dense-frame calibration estimates the mapping from physical motion to image displacement, which is reused for subsequent sparse-frame stitching without processing the intermediate images of each sparse pair. Together, structural masks and calibrated motion priors reduce appearance interference and periodic ambiguity, while structure-aware seams reduce visible transitions.

The key contributions are summarized as follows:

- We propose SAGS, a structure-aware stitching framework for glass curtain walls using consumer-grade UAV imagery and velocity telemetry. SAGS combines dense-to-sparse geometric calibration with structural masks for motion-constrained alignment and structure-aware seam placement. An analysis of periodic correspondence ambiguity provides a sufficient condition under which the search interval includes the true displacement while excluding neighboring periodic alternatives.
- We construct a real-world glass curtain wall dataset containing 2,307 image pairs from 21 buildings with synchronized UAV telemetry. Comparisons with matching-based pipelines and deep stitching methods demonstrate improved alignment reliability, while ablation studies support the complementary roles of structural masks and motion priors.

2 RELATED WORK

2.1 IMAGE STITCHING AND GEOMETRIC ALIGNMENT

Feature-based stitching pipelines estimate geometric transformations from local correspondences. A representative example is AutoStitch, which combines invariant local features with homography estimation (Brown & Lowe, 2007). Handcrafted features such as SIFT, SURF, and ORB are commonly paired with robust estimators such as RANSAC to reduce the influence of outliers (Lowe, 2004; Bay et al., 2006; Rublee et al., 2011; Fischler & Bolles, 1981). To accommodate viewpoint changes and parallax, dual-homography warping, smoothly varying affine transformations, and APAP introduce more flexible alignment models (Gao et al., 2011; Lin et al., 2011; Zaragoza et al., 2013). Robust elastic warping further uses local deformation to improve alignment under parallax (Li et al., 2018).

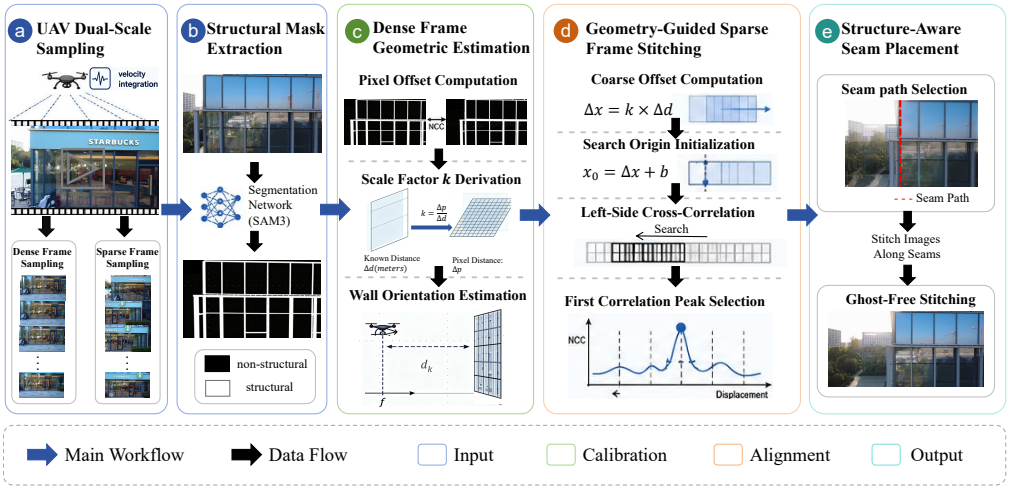


Figure 2: Overview of SAGS, comprising structural mask extraction, dense-frame calibration, geometry-guided sparse-frame alignment, and structure-aware seam placement.

Learned matchers, including SuperGlue, LoFTR, and LightGlue, improve correspondence estimation using learned feature interactions and matching strategies (Sarlin et al., 2020; Sun et al., 2021; Lindenberger et al., 2023). These methods provide correspondences for downstream geometric estimation rather than complete stitching pipelines. Deep stitching methods also learn alignment and image composition. UDIS explores unsupervised stitching through feature reconstruction (Nie et al., 2021), while UDIS++ combines global homography and local thin-plate spline deformation with learned composition masks (Nie et al., 2023). More recently, PixelStitch uses bidirectional global projection and pixel-wise residual warps to improve alignment while preserving image structure (Jin et al., 2025).

These advances improve correspondence estimation and deformation modeling, but identifying the correct physical correspondence remains challenging when reflections, transmitted backgrounds, and repetitive window layouts coexist. SAGS addresses this issue by combining facade structural masks with motion priors to constrain periodic correspondence ambiguity.

2.2 Structure-Guided Alignment and Seam Placement

Structural cues help preserve geometric regularity in image stitching. Line-point consistency constraints and structure-preserving warps reduce geometric distortion (Jia et al., 2021; Du et al., 2022), while seam-driven stitching and seam-guided local alignment incorporate seam quality into the stitching process (Li & Wang, 2016; Lin et al., 2016). For glass facades, however, extracted edges may originate from reflected or transmitted objects, complicating the selection of physical structures for alignment.

Promptable segmentation offers an alternative source of region information. SAM supports promptable object segmentation (Kirillov et al., 2023), and SAM 3 extends this capability to concept prompts (Carion et al., 2026). Our work uses facade structural masks jointly for motion-constrained alignment and structure-guided seam placement.

2.3 UAV-Based Facade Inspection

UAV-based reconstruction and inspection support infrastructure monitoring and mapping Colomina & Molina (2014). RTK positioning and LiDAR sensing can improve geometric accuracy but introduce additional hardware requirements. RGB-based UAV stitching has also been explored Cai et al. (2023).

For glass facades, Pandey et al. (2025) investigate drone-based imaging and automated classification for condition monitoring. Mao et al. (2023) study glass facade segmentation and repair in aerial



Figure 3: Qualitative comparison of stitching results in the presence of background buildings and reflections. From left to right: input image pair and results of SIFT, LoFTR, UDIS++, PixelStitch, and SAGS. Red boxes highlight visible stitching artifacts.

photogrammetric 3D models using geometric and multi-view constraints. Xu et al. (2025) provide a high-resolution facade parsing dataset with structural annotations and perspective-corrected images. Building on this research context, SAGS focuses on periodic correspondence ambiguity in glass facade stitching by combining structural masks with UAV motion priors.

3 METHOD

3.1 OVERVIEW

Image stitching on glass curtain walls faces two coupled challenges: reflections and transmitted backgrounds introduce appearance cues that are inconsistent with the physical facade geometry, while repetitive window layouts create ambiguity in structural correspondence. SAGS addresses these challenges by combining facade structural masks with UAV motion priors to constrain geometric alignment.

As illustrated in Fig. 2, SAGS first uses a fine-tuned SAM 3 model to extract facade structural masks. Dense-frame calibration then estimates the mapping from physical UAV motion to image displacement using structural mask alignment and velocity telemetry. This mapping guides sparse-frame alignment within a prescribed search tolerance, with the periodic uniqueness condition characterizing when neighboring periodic alternatives can be excluded.

Finally, the extracted structural masks guide seam placement along facade structures to reduce visible transitions in the panorama. We consider locally planar facades and predominantly translational motion along a known direction, with approximately stable camera orientation and camera-facade distance within each calibration segment.

3.2 STRUCTURAL MASK EXTRACTION

Glass curtain wall images contain visual cues from physical facade structures, reflections, and transmitted backgrounds. Reflected and transmitted content can appear within the facade region itself, introducing correspondences that do not consistently follow the physical facade geometry.

To reduce this appearance interference, SAGS employs a fine-tuned SAM 3 model Carion et al. (2026) to extract facade structural masks, as illustrated in Fig. 2(b). By representing the extracted structures as masks, the alignment process relies on their spatial arrangement rather than the underlying RGB appearance. Figure 3 illustrates the motivation for structural masking. SIFT and LoFTR exhibit facade misalignment consistent with matching background buildings, while UDIS++ and PixelStitch show structural distortions in reflection-rich regions. SAGS better preserves facade geometry by using structural masks as alignment references.

The resulting masks serve as geometric references for subsequent displacement estimation and guide structure-aware seam placement. However, their repetitive patterns can still produce multiple plausible correspondences, motivating the periodic ambiguity analysis in the following subsection.

3.3 PERIODIC STRUCTURE AMBIGUITY AND UNIQUENESS CONDITION

Structural masks reduce appearance interference, but repetitive window layouts still produce multiple plausible displacement candidates. We analyze this ambiguity using a one-dimensional periodic model with pixel period W.

Let d_{gt} and $\hat{d}$ denote the true and predicted displacements, respectively, with $|d_{\mathrm{gt}} - \hat{d}| \leq E$, where E bounds the prediction error. Correlation is restricted to $[\hat{d} - B, \hat{d} + B]$, where B is the search radius. A sufficient condition for including the true displacement while excluding neighboring periodic alternatives is

$$E \leq B < \frac{W}{2}. \tag{1}$$

The true displacement lies within the interval because $B \geq E$. Since the interval width $2B$ is smaller than W, it contains at most one candidate from the periodic family $\{d_{\mathrm{gt}} + kW : k \in \mathbb{Z}\}$. This ambiguity is analogous to aliasing in sampling theory (Shannon, 1949): repetitive structures admit multiple displacement interpretations. Here, motion priors resolve the ambiguity by restricting displacement uncertainty. In practice, SAGS uses a prescribed search tolerance without estimating W or E online. When Eq. 1 holds, bounded prediction errors can be tolerated, providing a basis for using consumer-grade UAV telemetry without requiring RTK positioning or LiDAR measurements. The condition excludes neighboring periodic alternatives, but does not guarantee correct correlation peak selection under mask errors or occlusion.

3.4 Dense Frame Geometric Estimation

Dense sampling keeps inter-frame displacement below one structural period. We select the first global correlation maximum rightward from zero, without a calibrated scale factor:

$$\Delta p_i = \min \left(\arg\max_{d \in \mathcal{D}_i} C(M_i, M_{i+1}, d) \right), \tag{2}$$

where C is normalized structural mask correlation and $\mathcal{D}_i$ is the rightward search range from zero. The corresponding physical displacement is obtained from velocity telemetry:

$$\Delta d_i = \int_{t_i}^{t_{i+1}} v(t)\, dt, \tag{3}$$

where $v(t)$ is the velocity component along the inspection direction.

To reduce sensitivity to pixel localization errors in short-baseline pairs, we estimate K over a longer calibration segment whose endpoint images retain sufficient overlap. Consecutive displacements initialize the first-to-last alignment:

$$\Delta P = \sum_{i=1}^{n-1} \Delta p_i. \tag{4}$$

To correct small accumulated errors, we compute the first-to-last mask correlation and select the peak nearest to ΔP:

$$\Delta P^* = \arg\min_{d \in \mathcal{P}_{1,n}} |d - \Delta P|, \tag{5}$$

where $\mathcal{P}_{1,n}$ denotes the displacement locations of local maxima of $C(M_1, M_n, d)$. Under the ideal periodic model, the correct periodic peak is nearest when the accumulated displacement error is less than $W/2$.

The scale factor is then computed as

$$K = \frac{\Delta P^*}{\Delta D}, \qquad \Delta D = \int_{t_1}^{t_n} v(t)\, dt, \tag{6}$$

where ΔD is the physical displacement between the first and last calibration frames along the inspection direction. For comparable pixel localization errors, the larger displacement reduces their relative contribution to the scale estimate. The calibrated K is reused for subsequent sparse-frame stitching.

3.5 Geometry-Guided Sparse Frame Stitching

After dense-frame calibration, SAGS reuses the scale factor K for geometry-guided sparse-frame stitching without processing the intermediate images of each sparse pair. Sparse sampling reduces computation, while the calibrated motion prior constrains periodic correspondence ambiguity.

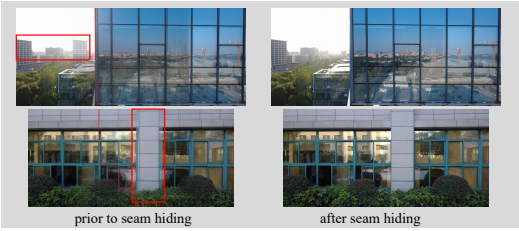


Figure 4: Comparison before (left) and after (right) structure-aware seam placement. In the left images, thin red lines indicate seam locations, while thick red boxes highlight visible artifacts.



Figure 5: Representative facade conditions: ordinary appearance, reflection, transmission, and repetitive structures. These challenges may coexist.

For a long-range frame pair, the UAV displacement Δs is converted into an estimated image displacement using the calibrated scale factor:

$$\hat{x} = K\Delta s. \tag{7}$$

With accurate calibration and motion estimates, $\hat{x}$ would directly determine alignment under the assumed geometry. In practice, the local search corrects residual prediction errors rather than searching over the full displacement range. Thus, even displacements spanning multiple window units can be disambiguated when the residual uncertainty satisfies $E \leq B < W/2$. The local search interval is:

$$x \in [\hat{x} - B, \hat{x} + B]. \tag{8}$$

The final displacement is obtained by maximizing structural mask correlation within this interval:

$$d^* = \arg \max_{d \in [\hat{x}-B, \hat{x}+B]} C(M_i, M_j, d). \tag{9}$$

When the periodic uniqueness condition in Eq. 1 holds, the constrained search interval includes the true displacement and excludes neighboring periodic alternatives.

3.6 Structure-Aware Seam Placement

After displacement estimation, aligned images are merged into a facade panorama. Seam placement helps reduce visible artifacts caused by inconsistent reflections and transmitted backgrounds Li & Wang (2016); Lin et al. (2016).

Within the overlap region, we define the seam cost as

$$c(x, y) = D(x, y) + \lambda \left(1 - \frac{M_1(x, y) + M_2(x, y)}{2}\right), \tag{10}$$

where D is the mean absolute difference across the three color channels, divided by 255 to normalize it to $[0, 1]$. The aligned binary masks M_1, M_2 label window frames as 1 and glass as 0. We use a fixed weight $\lambda = 3.0$ to favor seams within structural regions.

The seam minimizes the cumulative path cost:

$$s^* = \arg\min_s \sum_y c(s_y, y), \qquad |s_y - s_{y-1}| \leq 1, \tag{11}$$

where s_y denotes the seam column at row y. Dynamic programming followed by backtracking yields the optimal path. Pixels on opposite sides of the seam are taken from the respective aligned images. Figure 4 compares results before and after structure-aware seam placement.

4 Experiments

4.1 Experimental Setup

Hardware Platform. All data are collected using a consumer DJI UAV without RTK or LiDAR. Videos are recorded at 3840×2160 resolution with 83° horizontal field of view. Flight logs, including velocity, GPS position, and altitude, are recorded at 10 Hz for auxiliary geometric analysis.

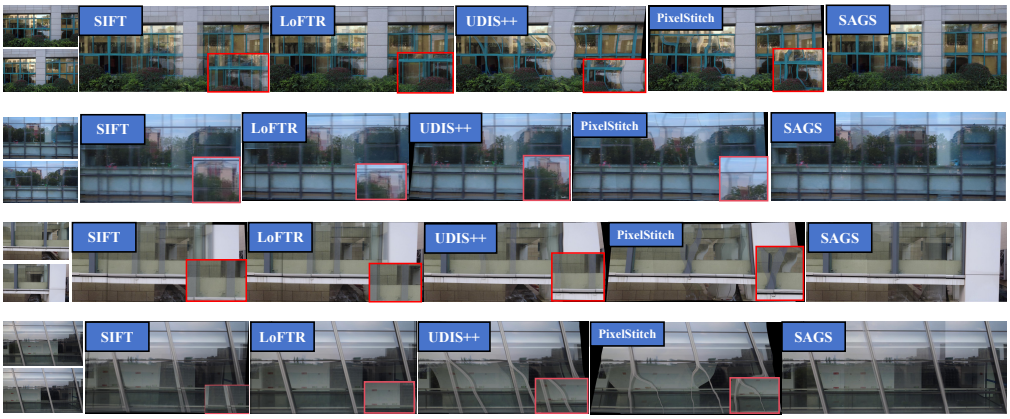


Figure 6: Qualitative comparison across four representative facade conditions. From top to bottom: normal facades, strong reflections, Strong Transmission, and Severe Repetition. Each row shows the input image pair followed by stitching results from SIFT, LoFTR, UDIS++, PixelStitch, and SAGS. Red boxes highlight visible stitching artifacts. The rows illustrate the selected challenges, which may coexist within the same image pair.

Table 1: Comparison with baseline methods under different glass curtain wall conditions using ASR (%).

Method	Overall ($n = 2133$)	Normal Facade ($n = 704$)	Strong Reflection ($n = 876$)	Strong Transmission ($n = 455$)	Severe Repetition ($n = 1155$)
SIFT	57.81	76.13	39.95	72.75	52.99
LoFTR	61.98	86.65	42.58	58.46	54.63
UDIS++	61.13	75.43	43.84	54.73	59.13
PixelStitch	59.68	62.22	47.94	63.08	57.32
SAGS	**90.76**	**87.64**	**91.21**	**90.32**	**93.59**

All experiments are performed offline on a workstation equipped with Intel i9-13900K CPU and NVIDIA RTX 4090 GPU.

4.2 Dataset

To study glass curtain wall stitching with UAV motion priors, we construct a real-world dataset using consumer UAVs. It contains 2,307 image pairs from 21 buildings, with 174 pairs allocated to SAM 3 fine-tuning and 2,133 pairs used for evaluation. An example sequence is shown in Appendix A.1.

Challenge labels are non-exclusive: reflection denotes visible reflected objects or textures in the overlap region; transmission denotes objects visible through the glass; repetition denotes multiple similar-looking window units. Ordinary pairs exhibit none of these challenges prominently. Figure 5 illustrates these conditions using manually aligned versions of the four image pairs shown in Fig. 6. Subset sizes are reported in Table 1, with each evaluation pair counted once in the overall results.

Alongside RGB images, the dataset provides timestamp-aligned flight logs containing velocity, GPS position, and altitude. Velocity telemetry supplies the motion priors used by SAGS. Only the SAM 3 structural extractor is fine-tuned; subsequent calibration and stitching require no additional training. Buildings used for evaluation are excluded from fine-tuning to avoid scene-level leakage.

4.3 Evaluation Metric

We evaluate stitching reliability using Alignment Success Rate (ASR), combining structural correspondence verification with a blinded visual assessment. Without ground-truth panoramas, overlap-based metrics such as PSNR and SSIM can be confounded by view-dependent reflections and transmitted content, even when physical facade structures are correctly aligned.

Corresponding points on physical window frames and panel boundaries are annotated in each input pair and shared across all evaluated methods. After mapping these points to a common stitching coordinate system, a pair passes the structural check if every annotated correspondence has a Euclidean alignment error of at most 30 pixels, measured at a common evaluation scale. This tolerance is chosen to be smaller than the window repetition periods in the evaluated images, providing a screen for mismatches between neighboring window units.

Results passing the structural check undergo visual assessment with method identities hidden to identify conspicuous geometric distortions, such as bent or discontinuous window frames. A pair is successful only if it passes both checks. Appearance differences caused solely by reflections or transmission are not counted as alignment failures. Missing outputs and results with unmappable annotated points are counted as failures.

4.4 Baselines

We compare SAGS with four baselines covering feature matching and learning-based image stitching.

Matching-based pipelines. SIFT uses FLANN matching, while LoFTR uses pretrained weights for correspondence estimation. Both pipelines use RANSAC to estimate a homography from the matched correspondences.

Learning-based stitching methods. UDIS++ and PixelStitch use their publicly available implementations and pretrained models for geometric alignment.

For alignment evaluation, all methods, including SAGS, use the same weighted-average blending procedure without seam optimization. All methods are evaluated on the same image pairs using the same ASR protocol. SAGS additionally uses UAV velocity telemetry, dense-frame calibration, and a facade-specific fine-tuned SAM 3 extractor.

4.5 Comparison with Baselines

Table 1 shows that SAGS provides modest gains on ordinary facades, but substantial improvements under reflection, transmission, and repetitive-structure challenges. Qualitative results are presented in Fig. 6.

Under strong reflection and transmission, baseline results exhibit facade misalignment consistent with correspondences following reflected or transmitted content rather than physical window structures. Under repetitive layouts, alignment may shift to neighboring window units. These challenges often coexist within an image pair, producing conflicting alignment cues. UDIS++ and PixelStitch exhibit combined alignment artifacts and pronounced structural distortions in the illustrated cases.

SAGS reduces reliance on misleading appearance cues through structural masks and constrains periodic correspondence using calibrated motion priors. Its results better preserve facade alignment and structural regularity in these examples. The contributions of these components are examined separately in the ablation study.

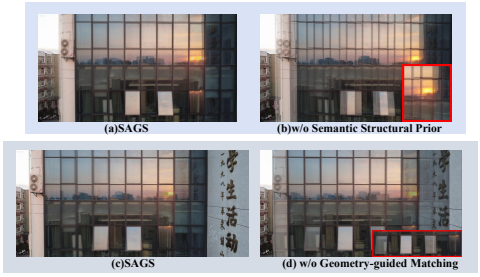


Figure 7: Qualitative ablation results of SAGS.

Table 2: Ablation studies of SAGS.

Variant	ASR (%) ↑
SAGS (Full)	**90.76**
w/o Semantic Structural Prior	69.15
w/o Geometry-guided Matching	62.82

4.6 Ablation Studies and Discussions

Structural Masks. We remove the structural masks and perform appearance-based normalized cross-correlation within the same motion-guided search interval. As shown in Table 2, ASR decreases from 90.76% to 69.15%, supporting the role of structural masks in reducing reflection and transmission interference during alignment.

Motion Priors. We retain the structural masks but remove the motion-guided search restriction, estimating displacement through global mask cross-correlation. ASR decreases from 90.76% to 62.82%, demonstrating the benefit of motion priors in constraining ambiguous structural correspondence.

Figure 7 illustrates appearance-induced misalignment without structural masks and periodic misalignment without motion priors. Together, these results support the complementary roles of structural representation and motion-guided search.

4.7 Validation of the Periodic Uniqueness Condition

We conduct a controlled simulation using synthetic periodic facade masks to examine the relationship between structural periodicity, displacement uncertainty, and matching success.

Synthetic curtain wall patterns are generated with controllable window periods, and motion-prior errors are simulated. We evaluate 14×14 parameter combinations, with $W/(2B)$ ranging from 0.50 to 3.00 and E/B ranging from 0.10 to 2.00. Each combination is evaluated over 150 trials, yielding 29,400 correspondence evaluations. A trial is successful only when the estimated displacement exactly equals the ground-truth displacement.

Figure 8 presents the exact-displacement recovery rates. The lower-right region satisfies $E \leq B < W/2$ and exhibits near-perfect recovery, consistent with the sufficient condition. Outside this region, the true displacement may fall outside the search interval when $E > B$, while neighboring periodic alternatives may enter the interval when $W \leq 2B$. Neither violation necessarily causes failure.

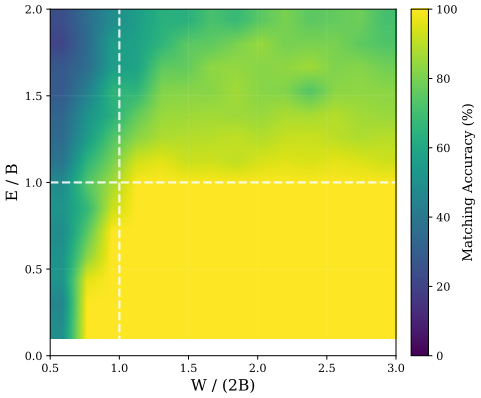


Figure 8: Heatmap validation of the periodic uniqueness condition under different window periodicity and displacement uncertainty settings.

4.8 Limitations and Future Work.

SAGS assumes approximately stable camera orientation and camera-facade distance when reusing the calibrated scale factor. Substantial changes may reduce displacement prediction accuracy. Curved or irregular facades may also exhibit residual misalignment that cannot be corrected by the translation model. Although periodic structures also occur on stone and other facades, the full framework has not been validated on these materials. Future work will investigate adaptive calibration, more flexible alignment models, and evaluation across facade materials and geometries.

5 Conclusion

We present SAGS for glass curtain wall stitching using consumer-grade UAV imagery and velocity telemetry. SAGS combines structural masks with motion-guided correlation and uses a periodic ambiguity exclusion condition to constrain alignment. Dense-frame calibration supports subsequent sparse-frame stitching, while structure-aware seam placement reduces visible transitions. Experiments demonstrate improved alignment reliability over representative matching-based and deep stitching methods under reflection, transmission, and repetitive-structure challenges, without requiring RTK positioning or LiDAR measurements.

REFERENCES

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pp. 404–417, 2006.

Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74:59–73, 2007.

W. Cai, S. Du, and W. Yang. Uav image stitching by estimating orthograph with rgb cameras. *Journal of Visual Communication and Image Representation*, 94:103835, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Ismael Colomina and Pere Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:79–97, 2014.

Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, and Jiaxin Wang. Geometric structure preserving warp for natural image stitching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3688–3696, 2022.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Jun Gao, Seon Joo Kim, and Michael S. Brown. Constructing image panoramas using dual-homography warping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49–56, 2011.

Qi Jia, Zhen Li, Xue Fan, Hao Zhao, et al. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *CVPR*, pp. 12181–12190, 2021.

Hengzhe Jin, Lang Nie, Chunyu Lin, Xiaomei Feng, and Yao Zhao. PixelStitch: Structure-preserving pixel-wise bidirectional warps for unsupervised image stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28125–28134, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. *IEEE International Conference on Computer Vision*, pp. 4015–4026, 2023.

Jing Li and Zhengming Wang. Seam-driven image stitching. *IEEE Transactions on Image Processing*, 25(3):1105–1117, 2016.

Jing Li, Zhengming Wang, Shihui Lai, Yuwei Zhai, and Ming Zhang. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on Multimedia*, 20:1672–1687, 2018.

K. Lin, N. Jiang, L. F. Cheong, M. N. Do, and J. Lu. Seagull: Seam-guided local alignment for parallax-tolerant image stitching. In *European Conference on Computer Vision*, pp. 370–385, 2016.

Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. Smoothly varying affine stitching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 345–352, 2011.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17627–17638, 2023.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

Zhu Mao, Xianfeng Huang, Hanyu Xiang, Yiping Gong, Fan Zhang, and Jie Tang. Glass façade segmentation and repair for aerial photogrammetric 3D building models with multiple constraints. *International Journal of Applied Earth Observation and Geoinformation*, 118:103242, 2023. doi: 10.1016/j.jag.2023.103242.

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing*, 30:6184–6197, 2021.

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yong Zhao. Parallax-tolerant unsupervised deep image stitching. In *ICCV*, pp. 7365–7374, 2023.

N. Pandey, K. Bingi, and R. Ibrahim. Drone-based imaging and automated classification for glass facade condition monitoring. In *2025 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp. 333–338, 2025. doi: 10.1109/I2CACIS65476.2025.11101021.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Claude E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949. doi: 10.1109/JRPROC.1949.232969.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Shuchang Xu, Fan Wu, Junjie Cheng, and Wenzhen Yang. High resolution building facade dataset for facade elements parsing. *Data in Brief*, 62:111950, 2025. doi: 10.1016/j.dib.2025.111950.

Julio Zaragoza, Tat-Jun Chin, Michael S. Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2339–2346, 2013.

A Appendix

A.1 Example Image Sequence

Figure 9 illustrates the dense calibration segment and subsequent sparse-frame stitching inputs.



Figure 9: Example UAV image sequence. The red box marks the densely sampled calibration segment used to estimate the scale factor K, which is reused for subsequent sparse-frame stitching.

A.2 Implementation Settings

Table 3 summarizes the stitching and SAM 3 fine-tuning settings.

Table 3: Implementation settings of SAGS.

Setting	Value
Stitching resolution	3840×2160
Dense-frame sampling interval	Typically 0.30 m
Sparse-frame sampling interval	Typically 2.5 m
Calibration frequency	Once per sequence
Physical search tolerance	0.45 m (search radius)
Correlation measure	Normalized cross-correlation
Correlation region	Valid overlap only
SAM 3 inference prompt	`window frame`
Fine-tuning preprocessing	Random resizing; padding to 1008^2
Frozen module	Text encoder
Optimizer	AdamW; weight decay 0.1
Learning rates	Vision backbone: 6.25×10^{-6}
	Other trainable parameters: 2×10^{-5}
Training epochs	50
Batch size / accumulation steps	1 / 4
Validation frequency	Every 2 epochs