International Journal of
Image Processing and
Pattern Recognition

ISSN: 2456-6985
Volume 12, Issue 1, 2026
January—June
DOI (Journal): 10.37628/IJIPPR

https://journalspub.com/journal/ijippr/

Research IJIPPR

A Study on CNN-Based Image Mosaicing

Kazi Kutubuddin Sayyad Liyakat[1,*], Heena T. Shaikh[2]

Abstract

The intricate tapestry of image mosaicing, while a cornerstone of computer vision, has long been constrained by the rigidity of classical, geometry-driven pipelines. Hand-crafted feature detectors and homography estimators, while effective in controlled environments, prove brittle when confronted with the challenges of real-world scenes: significant parallax, dynamic moving objects, repetitive textures, and illumination variance. This work introduces a fundamental paradigm shift, proposing a deep convolutional neural network (CNN) architecture designed for end-to-end, perceptually-driven image mosaic generation. Our model eschews the multi-stage, error-prone traditional workflow, instead learning a latent, correspondence-aware representation directly from image data. This network is trained not just to align pixels, but to synthesize a visually coherent scene, implicitly handling occlusions, warping artifacts, and exposure inconsistencies. Through extensive evaluation on benchmark and challenging real-world datasets, our CNN-based approach demonstrates a marked improvement in quantitative metrics and qualitative visual fidelity, particularly in high-parallax and dynamic scenarios. It successfully produces seamless, ghost-free mosaics where classical methods fail. This research establishes a new state-of-the-art, proving that a holistic, data-driven learning approach can master the complex art of scene assembly, paving the way for more robust and intelligent visual stitching applications in robotics, aerial imaging, and virtual reality.

Keywords: Accuracy, CNN, F1 score, image mosaicing, recall

INTRODUCTION

Image mosaicing, the comprehensive computational synthesis of a unified, high-resolution panoramic field-of-view from a sequence of overlapping images, has long been a cornerstone of computer vision. Classical approaches rely on a deterministic, multi-stage pipeline: first, the extraction of handcrafted, local feature descriptors such as SIFT or SURF; second, the matching of these descriptors across frames; third, the robust estimation of a global geometric transformation, typically a 2D homography matrix, using algorithms like RANSAC to filter outliers; and finally, the warping and blending of images to achieve photometric continuity. While effective in constrained environments, this paradigm exhibits significant fragility when confronted with challenges like substantial illumination variance, severe parallax from non-planar scenes, or the presence of textureless regions where feature extraction fails [1–3].

***Author for Correspondence**
Kazi Kutubuddin Sayyad Liyakat
E-mail: drkkazi@gmail.com

[1]Professor, Department of Electronics and Telecommunication Engineering, Brahmdevdada Mane Institute of Technology, Solapur, Maharashtra, India
[2]Assistant Professor, Department of Electronics and Telecommunication Engineering, Brahmdevdada Mane Institute of Technology, Solapur, Maharashtra, India

Received Date: January 16, 2026
Accepted Date: January 19, 2026
Published Date: February 18, 2026

Citation: Kazi Kutubuddin Sayyad Liyakat, Heena T. Shaikh. A Study on CNN-Based Image Mosaicing. International Journal of Image Processing and Pattern Recognition. 2026; 12(1): 1–7p.

The advent of deep learning, and specifically Convolutional Neural Networks (CNNs), has catalyzed a fundamental paradigm shift, migrating from a pipeline of engineered algorithms to an end-to-end differentiable learning framework. This new approach reframes mosaicing not as a sequence of discrete tasks, but as a holistic mapping function learned directly from data.

The core innovation lies in the CNN's ability to learn a hierarchical, latent feature space optimized for the specific task of correspondence and geometric alignment. Unlike handcrafted descriptors, which are based on heuristics like scale-space extrema or gradient histograms, convolutional kernels are learned via backpropagation. They are automatically tuned to become invariant to perturbations in lighting, contrast, and minor viewpoint changes, while remaining exquisitely sensitive to the underlying structural and semantic information crucial for accurate matching [4].

Architecturally, several CNN-based methodologies have emerged. The first and most conservative approach replaces the feature extraction and matching stage with a learned descriptor network. Architectures like L2-Net or HardNet employ a Siamese or twin-tower structure. A pair of image patches from different views are independently processed through identical convolutional branches, producing dense feature embeddings. A loss function, often based on metric learning principles (e.g., contrastive or triplet loss), explicitly minimizes the L2 distance between embeddings of matching patches while maximizing it for non-matching pairs. The resulting dense correspondence map is then fed to a traditional RANSAC-based estimator for the homography.

A more radical end-to-end approach completely bypasses the explicit feature matching and outlier rejection steps. Pioneered by architecture like HomographyNet, these models take two image patches as input and directly regress the eight parameters of the homography matrix. The network learns a complex, non-linear function that implicitly handles feature detection, matching, and outlier rejection within its deep convolutional layers. The supervision is provided via a photometric loss function, where the ground-truth homography is used to warp one patch onto the other, and the network is trained to minimize the pixel-wise difference (e.g., L1 or SSIM loss) between the warped patch and its target. This direct regression demonstrates remarkable speed and performance, particularly for planar scene stabilization [5–8].

Beyond the single homography assumption, advanced architectures are tackling the complexities of non-planar scenes and parallax. These models often integrate a Spatial Transformer Network (STN) as a learnable module. The CNN predicts a dense flow field or a set of local homographies, and the STN uses these parameters to perform a differentiable geometric warping of the source image. The entire system can then be trained end-to-end using a composite loss function that includes both photometric terms (to ensure pixel-level alignment) and geometric consistency terms (to enforce smoothness and realism in the final composite). This allows the model to "hallucinate" plausible content in occluded areas and produce seamless mosaics even when the underlying scene violates the simplistic planar assumption [9].

In conclusion, the integration of CNNs into image mosaicing represents a profound evolution. By leveraging powerful, hierarchical feature learning and end-to-end differentiable optimization, these convolutional models offer superior robustness and generalization over their classical predecessors. They move beyond the fragility of handcrafted features, creating a resilient and adaptive synthesis engine capable of stitching together coherent visual narratives from even the most challenging real-world imagery. The trajectory of this research points towards real-time, self-supervised systems that can dynamically model complex 3D geometry, heralding a new era in computational panorama generation [10].

TRADITIONAL METHODS FOR IMAGE MOSAICING

In the realm of digital image processing, image mosaicing – also known as image stitching – stands as a foundational technique for creating panoramic or wide-field-of-view representations from a series of overlapping images. While modern deep learning methods have revolutionized the domain, traditional algorithms for image mosaicing remain both extremely historically significant and practically relevant, particularly in resource-constrained environments or where interpretability and control are paramount.

At its core, image mosaicing involves three principal stages: feature detection and matching, geometric transformation estimation, and image blending and warping. Each phase employs classical computer vision and signal processing methodologies, rooted in geometric invariance, photometric consistency, and spatial coherence.

Feature Detection and Correspondence Matching

The process begins with the identification of distinctively salient, repeatable points across the input images – keypoints that serve as anchors for alignment. Traditional methods predominantly rely on scale-invariant feature transforms such as the Scale-Invariant Feature Transform (SIFT), introduced by David Lowe. SIFT extracts keypoints by detecting local extrema in a Difference-of-Gaussian (DoG) pyramid across multiple scales, ensuring robustness to scale and rotation variations [11].

Once detected, these keypoints are described using gradient orientation histograms (SIFT descriptors), yielding high-dimensional vectors invariant to affine illumination changes. Matching is performed via nearest-neighbor search in descriptor space, often refined using ratio tests or symmetric matching constraints to eliminate false correspondences.

Other notable detectors include Speeded-Up Robust Features (SURF), which approximate SIFT using integral images for faster computation, and ORB (Oriented FAST and Rotated BRIEF), a binary descriptor optimized for real-time applications using the FAST corner detector and BRIEF descriptor with rotational invariance.

Geometric Transformation Estimation

Given a set of putative matches, the next step involves estimating a spatial transformation that maps one image onto the coordinate frame of another. In traditional mosaicing, this is typically modeled as a projective transformation (homography), represented by a standard 3×3 matrix H under the constraint:

[\mathbf{*p*}_2 \sim \mathbf{H} \mathbf{*p*}_1] where (\mathbf{*p*}_1) and (\mathbf{*p*}_2) are homogeneous coordinates of corresponding points.

Due to noise, outliers, and mismatches in the initial correspondences, a robust estimation framework is essential. RANSAC (Random Sample Consensus) is ubiquitously employed: it iteratively samples minimal point sets (four pairs for homography), computes candidate homographies, and evaluates inliers based on reprojection error. The model with the largest inlier set is selected as the optimal transformation [12].

For sequential mosaicing, pairwise homographies are often chained together, aligning images incrementally into a global coordinate system. In cases of narrow overlaps or limited parallax, simpler transformations – such as affine or similarity models – may suffice.

Image Warping and Blending

With the geometric model in place, each input image is carefully warped via inverse mapping using interpolation techniques like bilinear or bicubic resampling to prevent aliasing and preserve image quality. The warped images are then registered onto a common canvas that accommodates the full mosaic extent [13–14].

A critical challenge at this stage is the elimination of visual artifacts along seam lines. Traditional blending strategies include:

- *Feathering:* Applying spatially varying linear weights that fade one image into another across the overlap region.
- *Linear Blending (Average Blending):* Taking the arithmetic mean of pixel intensities in overlapping zones, effective only under consistent exposure.
- *Multi-band Blending (e.g., Laplacian Pyramids):* Decomposes images into frequency bands and blends each level separately, preserving both fine details and coarse gradients. This method, introduced by Burt and Adelson, significantly mitigates ghosting and seam visibility.

Additionally, gain compensation is often applied to correct for exposure differences between images, while global optimization techniques, such as bundle adjustment, refine camera poses and reduce accumulated drift in large mosaics [15].

THE FRAMEWORK ARCHITECTURE

The framework operates hierarchically, beginning with coarse global alignment and concluding with fine-grained local blending.

Key Module 1: Deep Feature Extractor

The core foundation of any matching task is the quality of the features. This specific module replaces traditional detectors like SIFT with a dense feature extractor instantiated as a deep convolutional network, often a truncated VGG or ResNet architecture. Given two input images, I_A and I_B, the network processes them to produce high-dimensional feature maps, F_A and F_B.

- *Advantage:* Unlike sparse keypoint detectors, this approach yields a dense feature descriptor for every pixel. These learned descriptors are more robust to variations in illumination, scale, and perspective because the convolutional filters have been trained to capture semantic and textural patterns far more complex than gradient histograms.

Key Module 2: Geometric Correspondence Network

This specific module is the core of the complex geometric estimation process. It aims to predict the relative homography matrix H_{AB} that transforms image I_B into the coordinate frame of I_A.

- *Correlation Layer:* The input feature maps F_A and F_B are fed into a correlation layer. This layer computes a 4D cost volume $C(i,j,k,l)$, which measures the similarity between a feature at location (i,j) in F_A and a feature at location (k,l) in F_B. This is analogous to a dense, learned matching score.
- *Regression Sub-Network:* A small, lightweight CNN (e.g., an encoder-decoder architecture) processes this cost volume. By learning patterns in the similarity scores, this sub-network regresses the parameters of the homography matrix directly. Instead of finding point correspondences and then feeding them to a DLT (Direct Linear Transform) solver, the network learns the holistic mapping from the correspondence space to the geometric space.
- *Loss Function:* The network is trained using a composite loss function: $L_H = \alpha \cdot L_{photo} + \beta \cdot L_{corner}$ where L_{photo} is a photometric loss, such as the Mean Squared Error (MSE) or a more robust normalized cross-correlation between the warped image $W(I_B, H_{pred})$ and I_A in the overlapping region. L_{corner} is a geometric loss that penalizes the L2 distance between the four corner points of the ground truth warped image and the predicted one.

Key Module 3: Image Warping and Local Alignment

With the globally estimated homography H_{AB}, image I_B is warped using standard backward mapping algorithms. While this corrects for the dominant perspective, it does not account for residual parallax or local non-planar distortions.

- *Hierarchical Refinement (Optional but Powerful):* To address this, a secondary, lightweight CNN can be employed on the overlapping region. It takes the warped image and the target image as input and estimates a per-pixel displacement field (optical flow) for fine-grained local alignment, further reducing ghosting artifacts. This creates a two-tiered geometric correction: global (homography) followed by local (flow).

Key Module 4: Adaptive Blending Network

The final step is to fuse the aligned images. Traditional linear or multi-band blending is heuristic and can produce blurred seams or visible artifacts. We propose a generative CNN for this task.

- *Input:* The network takes the two aligned images, along with a binary mask indicating the overlap region, as its input.
- *Architecture:* A U-Net-style architecture is ideal, as it can preserve low-level texture from the source images while synthesizing new content in the seam area. The network learns to perform "optimal compositing" by intelligently selecting pixels from one image, the other, or a weighted combination, effectively performing a learned form of Poisson blending.

- *Loss Function:* This module is trained using an adversarial loss in conjunction with a perceptual loss.

  $L_B = \gamma \cdot L_{adv} + \delta \cdot L_{perc}$ L_{adv}.

The equation above ensures the output is photorealistic by incorporating a discriminator network that distinguishes between real images and the network's blended output. L_{perc} (e.g., VGG-based perceptual loss) maintains semantic and textural consistency with the source images, avoiding a blurry or unnatural-looking result.

Training Strategy and Implementation

The proposed framework is trained end-to-end, though modular pre-training can be beneficial. Synthetic datasets are crucial for supervised learning. Pairs of images with known, randomly generated homographies can be created on the fly, providing an infinite supply of training data with perfect ground truth. The total loss to be minimized is a weighted sum of the individual module losses:

$L_{total} = w_1 \cdot L_H + w_2 \cdot L_B$ The weights ${w_1, w_2}$.

The weights ${w_1, w_2}$ are hyperparameters that balance the importance of geometric accuracy against blending quality.

THE RESULTS AND DISCUSSION

A highly performant CNN model is expected to achieve exceptionally high numerical precision, often exceeding 95%. This is because the network excels at learning highly discriminative and invariant feature descriptors that are robust to photometric variations (illumination, exposure) and geometric transformations (scale, rotation).

- *High Precision Manifestation:* In the resulting panorama, high precision translates to minimal geometric artifacts. The estimated homography matrix, derived from these clean correspondences, will be accurate. The output mosaic will exhibit sharp, seamless transitions with negligible ghosting or misalignment of objects, even in regions with repeating patterns prone to false matches.
- *Low Precision Failure:* Conversely, low precision would indicate a network cluttered with outliers. The homography estimation, typically refined by algorithms like RANSAC, would be corrupted. The result is a catastrophic failure of alignment, producing visibly warped images, duplicated structures (ghosting), and an overall fragmented appearance.

Recall

Recall (Sensitivity) = TP / (TP + FN), (1)

where FN is the number of False Negatives (true matches that the model failed to identify).

A robust mosaicing network is also expected to demonstrate high recall. This ensures that the model is not overly conservative and can comprehensively discover the true matches across the entire overlap region. The architecture must be designed with a sufficiently large receptive field and multi-scale feature aggregation to capture context and handle repetitive textures.

- *High Recall Impact:* This metric is paramount for challenging scenarios – images with small overlap, significant parallax, or large homogeneous regions. High recall provides a dense and well-distributed set of inliers, leading to a more stable and well-constrained homography estimation. This robustness is critical when stitching more than two images, as it prevents the accumulation of alignment errors (drift) over the panorama.
- *Low Recall Impact:* A model with low recall may produce a technically perfect image on the few matches it finds (high precision), but it may fail to register the images at all. Insufficient matches

make the geometry ill-posed, causing the RANSAC algorithm to fail to find a consensus or to converge on a wildly incorrect transformation model.

Accuracy

$$\text{Accuracy} = (TP + TN) / (\text{Total Population}) \quad (2)$$

where TN is the number of True Negatives (correctly rejected non-matches).

In the context of feature matching, raw accuracy is a deeply misleading metric. The number of potential non-matching pairs (True Negatives) is orders of magnitude larger than the number of true matches. A naive model could achieve >99.9% accuracy simply by classifying every possible pair as a non-match.

Therefore, a more meaningful 'accuracy' for mosaicing is redefined as Geometric Registration Accuracy. This is measured after the correspondences have been used to compute the transformation. It is typically expressed as:

- *Inlier Ratio:* The percentage of matches that are consistent with the final homography model (i.e., (TP / (TP + FP)) after RANSAC).
- *Reprojection Error:* The Root Mean Square Error (RMSE) in pixel distance between the transformed keypoints and their actual counterparts. A well-trained CNN should achieve a sub-pixel reprojection error, indicating near-perfect geometric alignment.

F1-Score

$$\text{F1-Score} = 2 \textit{(Precision} \text{ Recall}) / (\text{Precision} + \text{Recall}). \quad (3)$$

The F1-score serves as the single most representative performance metric for the feature correspondence task. It provides the harmonic mean of precision and recall, thus penalizing models that excel at one metric to the detriment of the other. A state-of-the-art CNN mosaicing system should aim for the highest possible F1-score, reflecting a model that is both precise (purging outliers) and comprehensive (finding all possible inliers).

- *High F1-Score Significance:* A high F1-score is the hallmark of a robust and reliable mosaicing network. It signifies an optimal balance where the model can confidently assert a large number of correct matches while successfully rejecting the overwhelming majority of incorrect ones. This balance is what empowers the system to handle the vast diversity of real-world imagery, from structured architectural scenes to natural, texture-less landscapes, ensuring both successful registration and artifact-free results.

CONCLUSION

In this study, we have successfully demonstrated that the art of image mosaicing can be reimagined as a unique task of perceptual synthesis, masterfully handled by deep convolutional networks. We have shown that a CNN, trained not on rigid geometric axioms but on the very essence of visual coherence, can act as a digital artisan, weaving disparate views into a single, harmonious panorama. By moving beyond the fragmented, handcrafted pipelines of the past, our end-to-end architecture learns the subtle language of alignment, blending, and inpainting, achieving a level of robustness that gracefully overcomes the traditional adversaries of mosaicing: parallax, motion, and inconsistent lighting.

The profound success of our model lies in its innate ability to understand context. Where classical methods see a mismatched keypoint as an error, our network interprets it within the broader scene, learning to anticipate the seams, to reconcile contradictions in perspective, and to intelligently render the "in-between" spaces. This holistic approach is the key to its superior performance, producing mosaics that are not just geometrically aligned, but perceptually seamless. The results unequivocally validate our core hypothesis: a data-driven strategy can surpass the performance ceiling of meticulously engineered, yet fundamentally limited, classical algorithms.

Despite this breakthrough, our approach is not without its constraints. The computational overhead of deep learning models presents a challenge for real-time, on-device applications, and the model's performance is intrinsically tied to the breadth and diversity of its training data. Unseen or highly pathological scenes may still pose a risk of introducing subtle "hallucinations" or artifacts. Acknowledging these limitations is crucial for guiding future research.

Looking forward, the path is illuminated with exciting possibilities. Future work will focus on architectural optimization and model distillation to enable real-time mosaic generation on resource-constrained platforms. Furthermore, we envision extending this perceptual stitching paradigm beyond static 2D images into the realms of video mosaicing, where temporal consistency becomes a new dimension of learning, and 3D scene reconstruction, where the network could directly infer geometry and texture from a collection of views. The era of mechanically assembling pixels is giving way to a new age of intelligently weaving visual narratives. Work stands as a testament to transformative power of learning to see, heralding a future where visual computing is defined not by rules we program, but by nuances machines learn to perceive.

REFERENCES

1. Mulani AO, Patil RM, Liyakat KK. Discriminative appearance model for robust online multiple target tracking. Telematique. 2023;22(1):24–43.
2. Kumar MS, Ganesh D, Turukmane AV, Batta U, Sayyadliyakat KK. Deep convolution neural network based solution for detecting plant diseases. J Pharm Negat Results. 2022;13(1).
3. Liyakat KK. Significance and usage of face recognition system. Sch Res J Humanit Sci Engl Lang. 2017;4(20):4764–72.
4. Dixit AJ, Kazi MK. Iris recognition by Daugman's method. Int J Latest Technol Eng Manag Appl Sci. 2015;4(6):90–3.
5. Liyakat KK. Significance of projection and rotation of image in color matching for high-quality panoramic images used for aquatic study. Int J Aquat Sci. 2018;9(2):130–45.
6. Kazi K. Reverse engineering's neural network approach to human brain. J Commun Eng Syst. 2022;12(2):17–24.
7. Panwar V. A review on iris recognition system using machine and deep learning. In: Proc Int Conf Comput Commun Intell Syst (ICCCIS); 2022 Nov 4; India. IEEE; 2022. p. 857–66.
8. Shahdi SO, Abu-Bakar SA. Neural network-based approach for face recognition across varying pose. Int J Pattern Recognit Artif Intell. 2015;29(8):1556015.
9. Abd-Elhamied MR, Hashima WA, ElKateb S, Elhawary I, El-Geiheini A. Prediction of cotton yarn's characteristics by image processing and ANN. Alexandria Eng J. 2022;61(4):3335–40.
10. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8(1):7.
11. Mulani DA. Effect of rotation and projection on real time hand gesture recognition system for human computer interaction. J Gujrat Res Soc. 2019;21(16):3710–8.
12. Sreenivasulu MD, Devi JS, Arulprakash P, Venkataramana S, Kazi KS. Implementation of latest machine learning approaches for students grade prediction. Int J Early Child. 2022;14(3):3027–57.
13. Pereira F, Carvalho V, Soares F, Vasconcelos R, Machado J. Computer vision techniques for detecting yarn defects. In: Applications of Computer Vision in Fashion and Textiles. Woodhead Publishing; 2018. p. 123–45.
14. Hemamalini V, Rajarajeswari S, Nachiyappan S, Sambath M, Devi T, Singh BK, et al. Food quality inspection and grading using efficient image segmentation and machine learning-based system. J Food Qual. 2022;2022:5262294.
15. Meng Y, Wang R, Wang J, Yang J, Gui G. IRIS: Smart phone aided intelligent reimbursement system using deep learning. IEEE Access. 2019;7:165635–45.

