Depth-Supervised Fusion Network for Seamless-Free Image Stitching

Zhiying Jiang[1] **Ruhao Yan**[2] **Zengxi Zhang**[2] **Bowei Zhang**[2] **Jinyuan Liu**[2]*
[1] College of Information Science and Technology, Dalian Maritime University
[2] School of Software Technology, Dalian University of Technology
zyjiang0630@gmail.com atlantis918@hotmail.com

Abstract

Image stitching synthesizes images captured from multiple perspectives into a single image with a broader field of view. The significant variations in object depth often lead to large parallax, resulting in ghosting and misalignment in the stitched results. To address this, we propose a depth-consistency-constrained seamless-free image stitching method. First, to tackle the multi-view alignment difficulties caused by parallax, a multi-stage mechanism combined with global depth regularization constraints is developed to enhance the alignment accuracy of the same apparent target across different depth ranges. Second, during the multi-view image fusion process, an optimal stitching seam is determined through graph-based low-cost computation, and a soft-seam region is diffused to precisely locate transition areas, thereby effectively mitigating alignment errors induced by parallax and achieving natural and seamless stitching results. Furthermore, considering the computational overhead in the shift regression process, a reparameterization strategy is incorporated to optimize the structural design, significantly improving algorithm efficiency while maintaining optimal performance. Extensive experiments demonstrate the superior performance of the proposed method against the existing methods. Code is available at `https://github.com/DLUT-YRH/DSFN`.

1 Introduction

Image stitching is a fundamental task in computer vision, aiming to combine multiple images captured from different perspectives or positions into a single, high-resolution image with an extended field of view. This task plays a crucial role in various applications, such as panoramic photography [1, 2], remote sensing [3, 4], medical imaging [5, 6], and virtual reality [7, 8], where a comprehensive and seamless representation of a scene is required.

Conventional stitching methods predominantly follow a feature-driven paradigm, relying on hand-crafted local feature descriptors (e.g., SIFT [9], ORB [10]) for feature detection and robust matching algorithms (e.g., RANSAC [11]) to compute homography transformation matrices for reference and target images alignment. However, the planar scene assumption inherent to homography models fails to accommodate the complex geometric relationships arising from multi-depth layers in real-world scenarios, resulting in ghosting artifacts and structural misalignment in stitched results. To mitigate these issues, conventional methods adopt two primary optimization strategies: (1) enhancing local alignment accuracy through region-adaptive deformation techniques (e.g., mesh warping [12, 13, 14]), and (2) concealing residual artifacts by optimizing seam paths via energy functions [15, 16]. While effective in most scenarios, their performance is critically constrained by feature density and quality, leading to failures in low-texture, repetitive patterns, or large parallax scenarios.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-4339

Recent advancements in deep learning have introduced novel solutions for image stitching. Convolutional Neural Network (CNN)-based methods enable end-to-end learning of implicit geometric correlations between images, such as directly predicting transformation parameters through deep homography networks or modeling non-rigid deformations using deformable convolutions [17, 18]. Some works further integrate spatial attention mechanisms to enhance robustness against dynamic objects [19, 20] or employ unsupervised learning frameworks to address the scarcity of annotated real-world data [21]. Nevertheless, existing deep learning methods still face significant challenges: dependency on synthetic training data limits cross-domain generalization capabilities, while ensuring structural consistency in large-parallax scenarios remains challenging.

To address the aforementioned challenges, this paper proposes a depth-supervised image stitching that focuses on resolving co-planar alignment in large-parallax scenarios and ensuring seamless consistency transitions in multi-view overlapping regions. First, a two-stage depth-aware transformation estimation mechanism is introduced for large-parallax alignment. This mechanism leverages depth information to differentiate feature disparities of identical objects across varyi ng depth layers, while a recursive global-local deformation strategy integrates global homography estimation with localized adaptive warping, addressing the rigidity of conventional single-homography models in multi-plane scenes. During multi-view planar fusion, the optimal stitching seam is determined via graph-structured low-cost computation. A diffusion-based soft-seam propagation then generates pixel-wise confidence maps to define adaptive blending regions, effectively suppressing misalignment and ghosting artifacts caused by parallax. Additionally, we design a reparameterized strategy to optimize the shift regression model, ensuring the optimal effectiveness and the efficiency. The contributions are summarized as follows:

- We propose a depth-supervised image stitching, which focuses on addressing the alignment challenges caused by large parallax of significant depth differences, enabling the seamless fusion of multi-view images.
- The proposed method employs a depth-aware two-stage transformation estimation, coupled with a reparameterization strategy, which significantly enhances alignment performance in scenarios with large parallax.
- The determination of the soft-seam region enables a flexible adjustment for multi-view fusion, effectively avoiding issues such as misalignment and ghosting.
- Extensive experiments demonstrate that our method outperforms the state-of-the-arts, in terms of the large parallax alignment and seamless fusion.

2 Related Work

2.1 Image Stitching

Feature-Based Image Stitching. The core of manual feature-based image stitching lies in achieving accurate alignment through effective feature extraction and matching, which relies on sufficient geometric features in the scene. Brown *et al.* [22] pioneered this field by employing scale-invariant feature extraction combined with random sampling consistency to establish global rigid transformations. To address parallax issues, Li *et al.* [23] developed an analytical warp function based on point correspondences, enabling improved alignment through geometric constraints. Recognizing the limitations of single global transformations, Gao *et al.* [24] introduced dual-plane alignment by establishing separate warping models for distinct scene layers, though this approach faced challenges in complex environments with ambiguous planar divisions. Further advancing spatial adaptability, Zaragoza *et al.* proposed As Projective As Possible (APAP) [25], which localized mesh-based projective transformations, significantly increasing parameter flexibility while introducing artifacts at depth-discontinuous regions such as object boundaries.

The inherent alignment challenges in multi-view image stitching often manifest as ghosting artifacts within overlapping regions, necessitating sophisticated seam selection strategies. Zhang *et al.* [26] proposed a dual-scale alignment framework that preserves global structural consistency through optimal homography while enabling local seam-driven adjustments. Subsequent approaches focused on optimizing seam placement through energy minimization principles, with Kwatra *et al.* [27] introducing graph-based segmentation techniques to avoid object intersections. However, the computational intensity of these pixel-level optimization methods presents significant practical limitations

for real-world applications.

Deep Learning-Based Image Stitching. While contemporary feature descriptors [28, 29, 30] demonstrate potential for learned representations, their isolated application within traditional pipelines has limited practical adoption, driving research toward fully learned stitching frameworks. Deep learning approaches circumvent manual feature engineering by learning semantic representations through supervised (Lai *et al.* [19], Kweon *et al.* [20]), weakly supervised (Song *et al.* [31]), or unsupervised (Nie *et al.* [32]) paradigms, offering enhanced robustness in complex scenes. However, supervised methods' reliance on labeled data constrains their effectiveness in high-parallax scenarios. Nie *et al.* [21] pioneered unsupervised frameworks with improved cross-scene generalization and parallax tolerance, though persistent plane misalignments in extreme depth-varying scenes reveal fundamental limitations of current learning architectures.

2.2 Single Image Depth Estimation

Single image depth estimation aims to recover per-pixel depth from monocular visual data. Traditional approaches relied on geometric priors [33] or non-parametric depth transfer mechanisms [34], fundamentally constrained by color consistency assumptions. The advent of CNNs revolutionized this field through data-driven feature learning. Li *et al.* [35] pioneered multi-scale superpixel-to-pixel mapping via shallow CNNs with CRF refinement, though limited by local receptive fields. Liu *et al.* [36] advanced this by integrating CRF potentials within CNN frameworks, yet remained constrained by insufficient global context modeling. Eigen *et al.* [37] introduced a two-stage architecture that significantly enhanced spatial reasoning capabilities. Subsequent breakthroughs by Laina *et al.* [38] demonstrated the critical role of deep residual architectures in capturing holistic scene geometry through expanded receptive fields. The recent emergence of Depth Anything [39] marks a paradigm shift, establishing new state-of-the-art performance through unified representation learning.

3 The Proposed Method

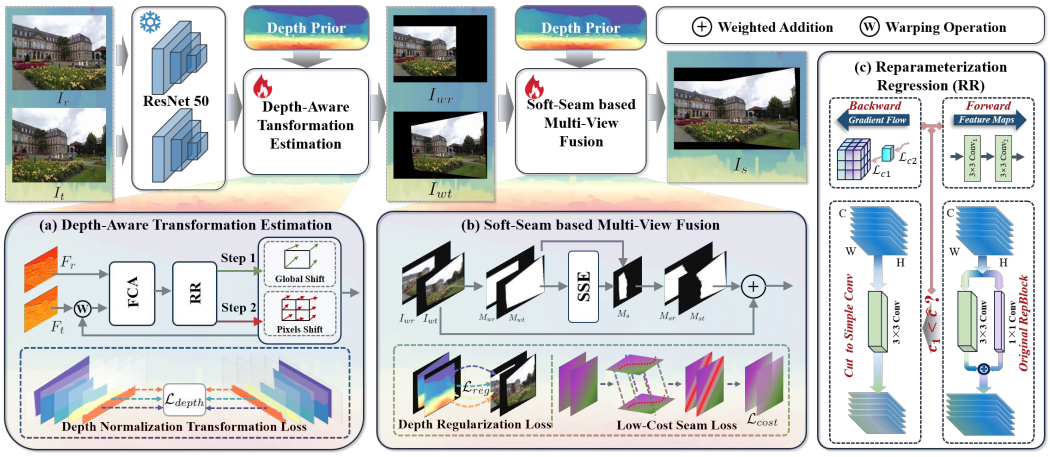


Figure 1: Workflow of the proposed method. It consists of two procedure: depth-aware transformation estimation and soft-seam based multi-view fusion. Besides, the transformation estimation process incorporates reparameterized regression to establish the optimal model.

As illustrated in Fig. 1, in the proposed method, we first feed the target image I_t and reference image I_r into the ResNet50 [38] for feature encoding. The extracted features from both views are then processed through a depth-aware transformation estimation module to obtain the warping matrices. To address alignment challenges in large parallax scenarios, a two-step recursive strategy is employed for the shift regression. Subsequently, the resulting transformations are applied to the observed images for alignment, and a soft-seam based multi-view fusion module is employed to blend the aligned images, producing a wide-field-of-view result I_s with natural transitions and no visible artifacts. The comprehensive description of each module is provided in the following.

https://doi.org/10.52202/085713-4339

3.1 Depth-Aware Transformation Estimation

We employ ResNet50 [38] to initiate the multi-scale feature extraction from both reference and target images, generating feature pairs at $1/16$ and $1/8$ resolutions, denoted as $\{F_r^{1/16}, F_t^{1/16}\}$ and $\{F_r^{1/8}, F_t^{1/8}\}$. Beginning with the coarser $1/16$-scale features, the Feature Correlation Aggregation (FCA) block [40] computes inter-view correspondences through:

$$C_{i,j} = FCA(F_r^{1/16}, F_t^{1/16}), \tag{1}$$

where $C_{i,j}$ is the correlation volume. A regression block then predicts quadrilateral vertex offsets $\Delta p \in \mathbb{R}^{4\times 2}$, from which a coarse homography matrix $H_C \in \mathbb{R}^{3\times 3}$ is derived via Direct Linear Transformation (DLT) [41]:

$$H_C = \arg\min_H \sum_{k=1}^{4} \|p'_k - H \cdot p_k\|_2^2. \tag{2}$$

p_k denotes the coordinate of the k-th point in the reference image, and p'_k means the corresponding point of the target image. This initial alignment warps the target feature to $\hat{F}_t^{1/8} = H_C(F_t^{1/8})$. Subsequently, a mesh-based refinement stage employs grid-wise offset estimation for sub-pixel precision. Let $\mathcal{M} = \{(x_i, y_j)\}$ define the mesh grid, with Radial Basis Function (RBF) interpolation generating the continuous deformation field:

$$\Delta(x, y) = \sum_{m=1}^{M} w_m \phi(\|(x, y) - (x_m, y_m)\|), \tag{3}$$

where $\phi(r) = -e^{-(\epsilon r)^2}$ denotes the Gaussian basis function with shape parameter ϵ. The final dense warping field $\mathcal{W}$ combines coarse homography and residual deformation:

$$\mathcal{W}(p) = H_C \cdot p + \Delta(p). \tag{4}$$

In the training process, we calculate the mean pixel error in the overlapping region after the coarse and residual transformations separately, which can be expressed as:

$$\begin{aligned} \mathcal{L}_{alignment} =& f_{alignment}(I_r, I_t, \lambda, \gamma, \eta) \\ =& \lambda \|I_r \cdot M_H - \mathcal{W}_H(I_t)\|_1 + \\ & \gamma \|I_t \cdot M_{H^{-1}} - \mathcal{W}_{H^{-1}}(I_r)\|_1 + \\ & \eta \|I_r \cdot M_N - \mathcal{W}_\Delta(I_t)\|_1, \end{aligned} \tag{5}$$

where M_H, $M_{H^{-1}}$ and M_N are homography transformation masks, inverse homography transformation masks and residual masks, which are obtained through homography H, inverse homography H^{-1}, and residual transformation Δ. λ, γ, η are balance weights. For ease of calculation, we choose to transform the mask, and since nonlinear transformations do not always support inverse operations, we do not design inverse losses.

To preserve structural consistency with the original scene, we impose a shape-preserving constraint for the mesh. We design the mesh loss from the point of the edge size of a single mesh and the offset of adjacent meshes. The number of control points is recorded as $U \times V$, $\vec{e}_w$ and $\vec{e}_h$ are the set of two adjacent edges in the mesh, and the edge loss of a single mesh can be described as:

$$\mathcal{L}_{edge} = \frac{1}{U \times (V-1)} \sum_{\vec{e}_w} \sigma(\langle \vec{e}, \vec{i} \rangle - 2W_{mesh}) + \frac{1}{(U-1) \times V} \sum_{\vec{e}_h} \sigma(\langle \vec{e}, \vec{j} \rangle - 2H_{mesh}), \tag{6}$$

where $\vec{i}$ and $\vec{j}$ are unit horizontal and vertical vectors, $\sigma(\cdot)$ is a non-linear activation function, H_{mesh} and W_{mesh} are the length and width of a single mesh. By calculating the loss in the mesh, the mesh stretching is limited and the distortion is reduced. We believe that the adjacent edges between the meshes in the non-overlapping region should be as parallel as possible, so we constrain the mesh angle as:

$$\mathcal{L}_{angle} = \frac{1}{a} \sum_{\vec{e}_{e1}, \vec{e}_{e2}} \delta(1 - cos\theta), \tag{7}$$

where a is the number of edge pairs, δ is the region label, and is denoted as 1 when the edge pair is in the non-overlapping region, 0 when the edge pair is in the overlapping region, and θ is the angle between the edge pairs. Considering the large parallax caused by significant depth variation, we incorporate the depth information as knowledge prior to supervise the learning of the transformation estimation. Specifically, we obtain the depth map through Depth Anything [39], characterizing the relative depth rather than the absolute depth. To this end, we perform normalization in the overlapping region of the reference and target images to reduce the relative error caused by the depth mutation of the non-overlapping region, expressed as:

$$\mathcal{L}_{depth} = f_{alignment}(I_{dr}, I_{dt}, \lambda', \gamma', \eta'). \tag{8}$$

where I_{dr}, I_{dt} mean the depth maps of I_r, I_t. The total loss for depth-aware transformation estimation can be expressed as:

$$\mathcal{L}^t = \mathcal{L}_{alignment} + \mu\mathcal{L}_{edge} + \zeta\mathcal{L}_{angle} + \xi\mathcal{L}_{depth}. \tag{9}$$

3.2 Soft-Seam based Multi-View Fusion

Based on the results inferenced from the transformation estimation, we obtain the aligned image pair I_{wr}, I_{wt}. However, precise alignment remains challenging in real-world parallax scenarios, and the multi-view image fusion process must additionally ensure authenticity and accurate reconstruction, such as preserving structures and achieving natural transitions between multi-view scenes. To address this, we relax the conventional definition of “seams" in the stitching, and suppose that any region requiring fusion within overlapping areas can be treated as a potential seam. We build upon the low-cost seam localization and establish a soft-seam region diffused from the distinct seam to serve as the adaptive fusion adjustment, aiming to resolve the ghosting and misalignment artifacts while enabling natural transitions.

Specifically, we calculate the corresponding region masks M_{wr}, M_{wt} based on the aligned images. These masks are fed into a Soft-Seam Estimation (SSE) to obtain soft-seam mask M_s within overlapping areas, serving as the candidate region for fusion. SSE module is built upon a UNet architecture [42, 43], in which 3×3 convolutions are replaced with dilated convolutions, with dilation rates are set as $1, 2, 3, 4$, and 5. At the four skip connections, the same-scale features from both input images are first upsampled using nearest-neighbor interpolation and then passed through a 1×1 convolution to reduce the number of channels. The difference map is first computed by subtracting the feature maps of the two images pixel by pixel. It is then concatenated with the upsampled features along the channel dimension, and the resulting representation is fed through two dilated convolution layers to further advance the decoding process.

M_s is then integrated with the original aligned image masks through a single filter and the sigmoid function, yielding two more flexible masks M_{sr}, M_{st} with pixel-level regional adaptability. These adaptive masks are subsequently applied to weighted fusion processing of the aligned images, enabling refined fusion tailored to local pixel characteristics.

In the training process, we first need to determine the terminal points of the seam, expressed as:

$$\mathcal{L}_{terminal} = \|(I_s - I_{wr}) \cdot (M_{wr} \odot \neg M_{wr})\|_1 + \|(I_s - I_{wt}) \cdot (M_{wt} \odot \neg M_{wt})\|_1. \tag{10}$$

It combines the inverted and original masks via element-wise multiplication to restrict the fusion mask boundary to the intersection area, controlling its endpoints. In which $\odot$ represents the pixel-by-pixel calculation of the two masks. $\neg M_{wr}$ and $\neg M_{wt}$ represent the inverted and expanded mask. In order to improve the sensitivity of the difference values, we choose the pixel square difference to construct the cost map and the cost loss is defined as:

$$\mathcal{L}_{cost} = \sum_{i,j} |M_s^{i,j} - M_s^{i+1,j}|(D^{i,j} + D^{i+1,j}) + \sum_{i,j} |M_s^{i,j} - M_s^{i,j+1}|(D^{i,j} + D^{i,j+1}), \tag{11}$$

where D is the squared difference between the warped image I_{wr} and I_{wt}. In order to constrain the smoothness of the fused image, the smoothness loss, which calculates the smoothness penalty by measuring the distance between adjacent pixels within the fusion region of the stitched image, is also adopted:

$$\mathcal{L}_{smooth} = \sum_{i,j} |M_s^{i,j} - M_s^{i+1,j}|(I_s^{i,j} - I_s^{i+1,j}) + \sum_{i,j} |M_s^{i,j} - M_s^{i,j+1}|(I_s^{i,j} - I_s^{i,j+1}). \tag{12}$$

 https://doi.org/10.52202/085713-4339

Furthermore, we also introduce the depth consistency to supervise the inference results. Specifically, the base depth maps are first aligned with the estimated transformation. Then, a secondary local regularization of the aligned depth images I_{wdr}, I_{wdt} is performed to further calibrate the local relative depth, expressed as:

$$\begin{aligned}\mathcal{L}_{reg} = &\sum_{i,j} |M_s^{i,j} - M_s^{i+1,j}|(\mathcal{F}(I_{wdr}, I_{wdt})^{i,j} - I_{wdt}^{i+1,j}) \\ &+ \sum_{i,j} |M_s^{i,j} - M_s^{i,j+1}|(\mathcal{F}(I_{wdr}, I_{wdt})^{i,j} - I_{wdt}^{i,j+1}).\end{aligned} \tag{13}$$

$\mathcal{F}$ denotes the fusion process. The total loss for the soft-seam based multi-view fusion can be expressed as:

$$\mathcal{L}^f = \rho\mathcal{L}_{terminal} + \tau\mathcal{L}_{cost} + \iota\mathcal{L}_{smooth} + \sigma\mathcal{L}_{reg}. \tag{14}$$

3.3 Reparameterization Regression

In the realm of learning-based image stitching methodologies, the adoption of fully connected architectures for shift regression often incurs significant computational costs while yielding suboptimal performance. To mitigate this issue, we leverage reparameterization techniques [44] to identify the optimal structural configuration during the parameter regression process.

Although reparameterization techniques enhance feature diversity and structural flexibility through the introduction of Reparameterization Blocks (RepBlocks), existing reparameterized architectures fail to achieve robust performance improvements due to inherent limitations in RepBlocks [45]. To address this, we propose a Reparameterization Block Adaption (RBA) algorithm, which dynamically adapts the model training process by selectively integrating either a RepBlock or a standard convolutional layer based on the specific requirements of the convolution layer.

Specifically, in the model forward process, different convolution structures provide distinct feature representations to extract diverse feature maps from input features. Therefore, in the initial model, following the research in [46], we replace the 3×3 convolution in our regression block located behind the FCA block of the proposed depth-aware transformation estimation to a RepBlock consists of a 1×1 convolution layer $Conv^1$ and a 3×3 convolution layer $Conv^3$. To evaluate the contribution of these two layers, we formulate a linear combination. Given a set of input feature maps $f_{in} \in \mathbb{R}^{C_{in}\times H\times W}$, where C_{in}, H and W are the input channel number, height and width, the output features f_{out} of a RepBlock are calculated as:

$$f_{out} = Re(\mathbf{w_1} \times Conv^1(f_{in}) + \mathbf{w_3} \times Conv^3(f_{in}) + \mathbf{b}), \tag{15}$$

where $\mathbf{w_1}$, $\mathbf{w_3}$ and $\mathbf{b} \in \mathbb{R}^{C_{out}\times1\times1}$ are the weights and bias. Re denotes the ReLU function. $f_{out} \in \mathbb{R}^{C_{out}\times H\times W}$, where C_{out} is the output channel number. In our formulation, $\mathbf{w_1}$ and $\mathbf{w_3}$ can be trained to evaluate the contribution of the two branches. The contribution c_1 of the $Conv^1$ in the RepBlock is calculated as:

$$c_1 = \frac{\frac{1}{C_{out}}(\sum \mathbf{w_1})}{\frac{1}{C_{out}}(\sum \mathbf{w_1}) + \frac{1}{C_{out}}(\sum \mathbf{w_3})}. \tag{16}$$

The effect of our RBA is to prevent the output features of the $Conv^1$ layer from playing a damaging role in feature extraction. Therefore, in the training process, if $c_1 < \hat{c}$, where $\hat{c}$ presents the hyperparameter of the minimum threshold, we cut the $Conv^1$ by coupling it to the $Conv^3$ layer, and the parameter weight $\mathbf{W}_3^{new}$ of the new 3×3 layer $Conv^{3,new}$ is calculated as:

$$\mathbf{W}_3^{new} = \mathbf{w_3} \cdot \mathbf{W_3} + \mathbf{w_1} \cdot pad(\mathbf{W_1}), \tag{17}$$

where $\mathbf{W_1}$ and $\mathbf{W_3}$ are the weights of $Conv^1$ and $Conv^3$, and the pad operation indicates adding value 0 around the $\mathbf{W_1}$ [44]. After cutting the 1×1 branch, we ensure the model can be adapted to the training task while avoiding the feature degradation caused by the additional branches.

4 Experiments

4.1 Implementation Details

Our method is implemented using the PyTorch framework and executed on an NVIDIA RTX 3090 GPU. For training both the depth-aware transformation estimation and soft-seam based multi-view

fusion models, we employ the Adam optimizer [47], and the learning rate decays exponentially, with an initial value of 10^{-4}. The transformation model is trained for 100 epochs, with the hyperparameters λ, γ, and η set to 3, 3, and 1, respectively. The values of λ', γ', η' are identical to those of λ, γ, and η. μ, ζ, ξ are set to 10, 10 and 0.3. For the multi-view fusion model, we initially train the model for 50 epochs on the training set, with the hyperparameters ρ, τ, ι, σ set to 10000, 1000, 1000, 10.

The UDIS-D training set [32] is employed as the training data. To enhance the reliability of the experimental results, we evaluate the model on the UDIS-D testing set and further validate it using real-world data from the IVSD dataset [40].

4.2 Performance Comparison

We compare our method with APAP [25], ELA [23], LPC [48], SPW [49], UDIS [32], UDIS++ [21], TRIS [50] and SRS [51], where each method adopts the pre-trained model and configuration parameters provided by the official.

4.2.1 Qualitative Evaluation

The qualitative comparison on the UDIS-D dataset is shown in Fig. 2. In the first example, our method achieves a clearer fusion result for the wall area, effectively avoiding issues such as blurring and ghosting artifacts. Additionally, within the region marked by the red framework, the proposed method successfully preserves the complete information of the bicycle without any loss of content. In the second example, while other comparative methods exhibit ghosting or content loss, our method delivers a stitching result that is both visually clear and realistic, demonstrating a better reconstruction quality. Furthermore, to compare the alignment accuracy of different methods, we visualize the alignment errors in the lower-right corner of the corresponding figures. It can be observed that the proposed method achieves significantly higher alignment precision compared to the others.

Visual results on the IVSD dataset are presented in Fig. 3. The proposed method demonstrates superior stitching performance across varying depths of the captured scene, with alignment errors further confirming its effectiveness. The consistent performance across both datasets robustly validates the efficacy of the proposed method.

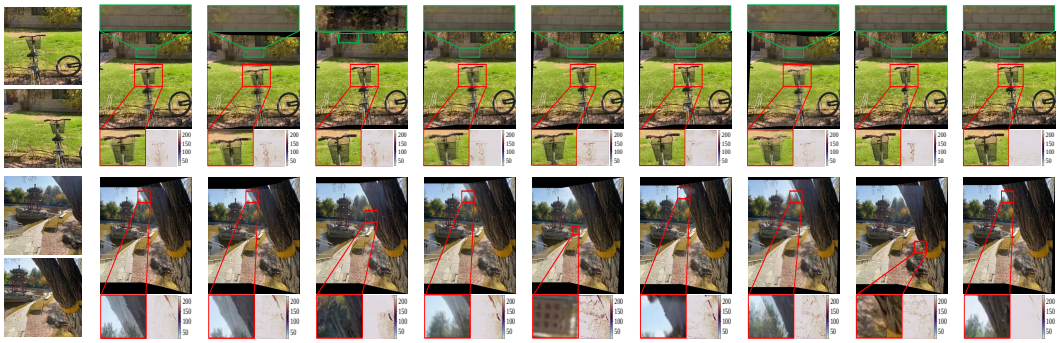


Figure 2: Visual comparison of stitched images from UDIS-D dataset. The alignment error is visualized in the lower right corner.

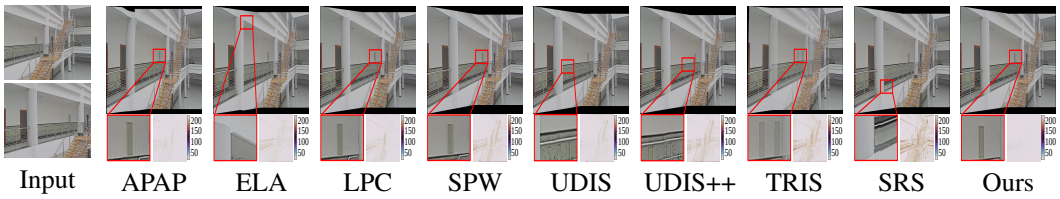


Figure 3: Visual comparison of stitched images from IVSD dataset. The alignment error is visualized in the lower right corner.

https://doi.org/10.52202/085713-4339

Table 1: Quantitative comparison on UDIS-D and IVSD datasets. The best and second results are marked in **red** and **blue**.

Method	UDIS-D				IVSD			
	PSNR(↑)	SSIM(↑)	SIQE(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	SIQE(↑)	LPIPS(↓)
APAP [25]	23.792	0.794	41.707	0.472	22.904	0.681	39.281	0.454
ELA [23]	24.012	0.808	41.781	0.470	23.452	0.701	37.186	**0.435**
LPC [48]	22.595	0.736	**43.616**	0.467	20.996	0.641	37.517	0.447
SPW [49]	21.606	0.687	41.060	**0.466**	18.868	0.575	36.156	0.449
UDIS [32]	21.171	0.648	42.186	0.475	23.535	0.743	40.474	0.451
UDIS++ [21]	**25.426**	**0.837**	43.184	0.469	**26.649**	**0.819**	**46.383**	0.439
TRIS [50]	24.476	0.821	41.621	0.476	24.187	0.753	40.873	0.448
SRS [51]	24.828	0.811	41.857	0.473	24.234	0.796	35.641	0.445
Ours	**25.467**	**0.839**	**43.732**	**0.462**	**26.778**	**0.820**	**46.568**	**0.436**

Table 2: Efficiency comparison against the state-of-the-art methods.

Methods	APAP [25]	ELA [23]	LPC [48]	SPW [49]	UDIS [32]	UDIS++ [21]	TRIS [50]	SRS [51]	Ours
Time (ms)	6683.14	8347.79	13435.47	11651.68	193.66	79.73	107.98	83.17	67.04

4.2.2 Quantitative Evaluation

We employ a set of evaluation metrics, including PSNR [52], SSIM [53], SIQE [54], and LPIPS [55], to conduct a comprehensive performance assessment. According the UDIS++ [21], the test sets of UDIS-D are categorized into three levels based on their complexity, and the corresponding quantitative results are summarized in Table 1. Furthermore, to rigorously evaluate the generalization capability of the proposed method, quantitative results on the IVSD dataset are also presented in Table 1. It is evident that the proposed method outperforms other methods across multiple metrics, substantiating its superiority further.

To further evaluate the efficiency of the proposed method, we present the processing time required by each comparative method for image stitching at a size of 512×512, with the results summarized in Table 2. It can be observed that, although our method involves deep estimation and inference processes, the overall running time is still better than that of the other methods, making it more suitable for practical applications.

4.2.3 User Study

To evaluate the subjective performance of our method, we conducted a user study to assess the visual quality of the stitched images. The input and output images were organized according to their scene categories, and participants were presented with a set of input images alongside the corresponding output images generated by each comparison method. Participants were asked to evaluate quality from multiple perspectives, including ghosting, misalignment, structural accuracy, and realistic scene restoration. They were allowed to zoom in or out for detailed observation and were instructed to rate each image on a scale of 1 to 5 based on visual quality. The study involved 50 participants, including 30 researchers or students with a background in computer vision and 20 individuals without specific expertise. The results of the user study, as illustrated in Fig. 4, demonstrate that our method consistently received higher ratings compared to other methods.

4.3 Ablation Studies

Loss Function Evaluation. We conduct a comprehensive ablation study to analyze the impact of different constraints in the depth-aware transformation estimation and soft-seam based multi-

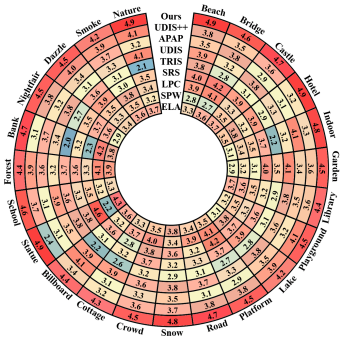


Figure 4: Visual quality survey.

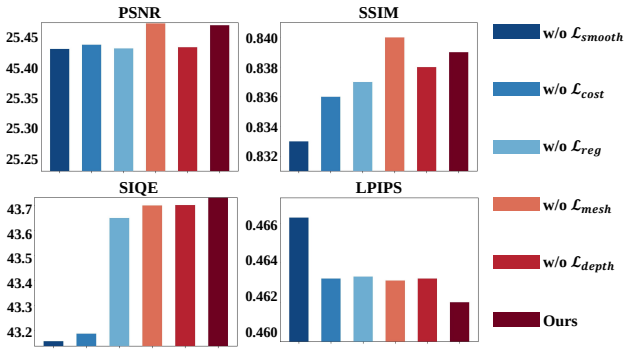


Figure 5: Ablation study on different loss components.

view fusion modules, respectively. During the experiments, we would like to clarify that $\mathcal{L}_{edge}$ and $\mathcal{L}_{angle}$ exhibit complementary effects, where retaining one constraint alone renders the other nearly ineffective. Therefore, we combine these constraints into a unified loss term, denoted as mesh loss $\mathcal{L}_{mesh}$ for ablation analysis. As shown in Fig. 6, the removal of mesh constraints in the transformation estimation leads to noticeable distortions in the deformed images, while the absence of fusion module constraints hinders the model's ability to compensate for stitching errors caused by insufficient local alignment. Furthermore, the lack of depth supervision not only increases the visual errors in both transformation and fusion but also degrades the overall accuracy of the method. The quantitative results across the entire UDIS-D dataset are presented in Fig. 5 and Table. 3. Although the ablation of mesh constraints marginally improves certain metrics by relaxing the image distortion limits, this improvement is not meaningful from a visual perspective.

Soft-Seam Fusion Evaluation. We also perform an ablation study of the fusion strategy compared with average fusion and seam cutting [27]. As shown in Fig. 7, the first row exhibits the gradient results from different fusion strategies and the second row visualizes the corresponding fusion regions. We can see that a larger fusion region increases the likelihood of ghosting artifacts. However, an excessively small fusion region may result in insufficient gradient smoothness. The proposed soft-seam fusion strategy adaptively preserves gradient continuity to the greatest extent, facilitating natural and seamless results.

Adaptive Mask Evaluation. The fusion mask in our method is derived from an adaptive weight matrix based on the soft-seam region. A comparison with the absolute weight derived from traditional seam-based methods is illustrated in Fig. 8. The result produced by our method exhibits a more visually appealing effect and smoother region.

Reparameterization Evaluation. We test 10 thresholds for hyperparameter selection, where $\hat{c} = 1$ represents the original model and $\hat{c} = 0$ represents the model without RBA. Results shown in Fig. 9 indicate that both the original model and the full reparameterization model cannot perform the best. The model with $\hat{c} = 0.25$ realizes the best performance, while maintaining efficient training.

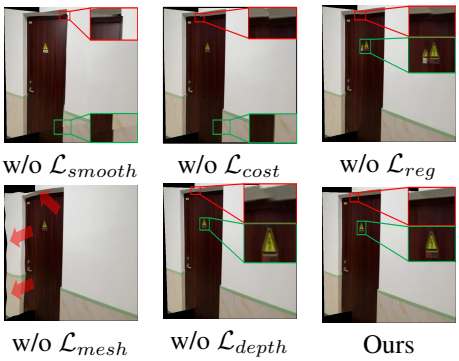
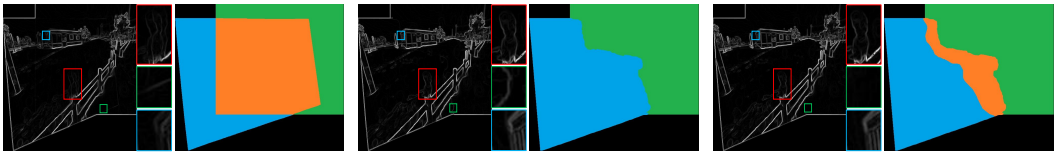


Figure 6: Ablation study on loss function.

Loss	PSNR(↑)	SSIM(↑)	SIQE(↑)	LPIPS(↓)
w/o $\mathcal{L}_{smooth}$	25.431	0.833	43.156	0.466
w/o $\mathcal{L}_{cost}$	25.438	0.836	43.186	0.463
w/o $\mathcal{L}_{reg}$	25.432	0.837	43.651	0.463
w/o $\mathcal{L}_{mesh}$	**25.473**	**0.840**	43.701	**0.463**
w/o $\mathcal{L}_{depth}$	25.434	0.838	**43.703**	0.463
Ours	**25.470**	**0.839**	**43.732**	**0.462**

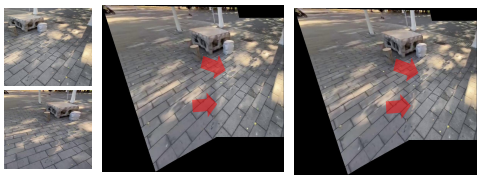
Table 3: Ablation study on loss components of transformation estimation and multi-view fusion models.

https://doi.org/10.52202/085713-4339



Average Fusion Seam Cutting Ours

Figure 7: Ablation study of the fusion strategy. The corresponding fusion regions are visualized on the right part.



Input Absolute Weight Ours

Figure 8: Visual comparison between the proposed adaptive mask and the absolute mask.

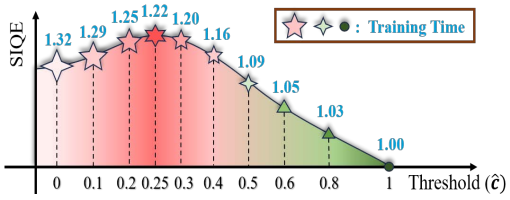


Figure 9: Ablation study on the hyperparameter $\hat{c}$ in the RBA.

5 Limitations

The proposed method is primarily designed for stitching two images and currently lacks full capability to address the challenges associated with multi-image panoramic stitching. Key limitations include difficulties in maintaining loop consistency and mitigating global error propagation in complex scenarios. These issues can compromise the geometric coherence of the final output, thereby constraining the method's robustness and broader applicability.

6 Conclusion

This paper proposed a depth-supervised image stitching method designed to address the alignment challenges in large parallax scenarios and achieve seamless wide field-of-view reconstruction. Firstly, a depth-aware two-stage transformation estimation is developed, which leverages depth-consistency priors to align targets across varying depth ranges. Secondly, a soft-seam region diffusion strategy is introduced to accurately identify transition regions, enabling natural and smooth fusion while mitigating ghosting and misalignment issues. Additionally, the reparameterization strategy for shift regression enhances the adaptability and reduces computational overhead. Extensive experiments validate the effectiveness of the proposed method. Although our method can improve multi-view alignment and fusion performance by leveraging depth consistency guidance, the presence of dynamic elements in the scene poses challenges for obtaining accurate depth information. In the future, we will focus on robust stitching under dynamic conditions to further enhance alignment robustness in such scenarios.

7 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62302078, in part by the Fundamental Research Funds for the Central Universities under Grant 3132025276, and in part by China Postdoctoral Science Foundation under Grant 2023M730741.

References

[1] Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *IEEE TIM*, 71:1–34, 2022.

[2] Zeru Shi, Zengxi Zhang, Kemeng Cui, Ruizhe An, Jinyuan Liu, and Zhiying Jiang. Sefenet: Robust deep homography estimation via semantic-driven feature enhancement. *IEEE TCSVT*, 2025.

[3] Zhiying Jiang, Zhuoxiao Li, Shuzhou Yang, Xin Fan, and Risheng Liu. Target oriented perceptual adversarial fusion network for underwater image enhancement. *IEEE TCSVT*, 32(10):6584–6598, 2022.

[4] Zengxi Zhang, Zhiying Jiang, Long Ma, Jinyuan Liu, Xin Fan, and Risheng Liu. Hupe: Heuristic underwater perceptual enhancement with semantic collaborative learning. *IJCV*, 133:3259–3277, 2025.

[5] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *IJCV*, 132(5):1748–1775, 2024.

[6] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE TPAMI*, 2024.

[7] Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE TVCG*, 29(4):1992–2004, 2021.

[8] Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA JAS*, 2024.

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[10] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pages 2564–2571. Ieee, 2011.

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

[12] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *CVPR*, pages 3254–3261, 2014.

[13] Weiqing Yan, Guanghui Yue, Jindong Xu, Yanwei Yu, Kai Wang, Chang Tang, and Xiangrong Tong. Shape-optimizing mesh warping method for stereoscopic panorama stitching. *Information Sciences*, 511:58–73, 2020.

[14] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multi-spectral image stitching via spatial graph reasoning. In *ACMMM*, pages 472–480, 2023.

[15] Nan Li, Tianli Liao, and Chao Wang. Perception-based seam cutting for image stitching. *Signal, Image and Video Processing*, 12:967–974, 2018.

[16] Junhong Gao, Yu Li, Tat-Jun Chin, and Michael S Brown. Seam-driven image stitching. In *Eurographics*, pages 45–48. Girona, 2013.

[17] Lang Nie, Chunyu Lin, Kang Liao, Meiqin Liu, and Yao Zhao. A view-free image stitching network based on global homography. *Journal of Visual Communication and Image Representation*, 73:102950, 2020.

[18] Canwei Shen, Xiangyang Ji, and Changlong Miao. Real-time image stitching with convolutional neural networks. In *RCAR*, pages 192–197. IEEE, 2019.

 https://doi.org/10.52202/085713-4339

[19] Wei-Sheng Lai, Orazio Gallo, Jinwei Gu, Deqing Sun, Ming-Hsuan Yang, and Jan Kautz. Video stitching for linear camera arrays. *arXiv preprint arXiv:1907.13622*, 2019.

[20] Hyeokjun Kweon, Hyeonseong Kim, Yoonsu Kang, Youngho Yoon, Wooseong Jeong, and Kuk-Jin Yoon. Pixel-wise deep image stitching. *arXiv preprint arXiv:2112.06171*, 2021.

[21] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching. In *ICCV*, pages 7399–7408, 2023.

[22] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *IJCV*, 74:59–73, 2007.

[23] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. Parallax-tolerant image stitching based on robust elastic warping. *IEEE TMM*, 20(7):1672–1687, 2017.

[24] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *CVPR*, pages 49–56. IEEE, 2011.

[25] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *CVPR*, pages 2339–2346, 2013.

[26] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *CVPR*, pages 3262–3269, 2014.

[27] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM TOG*, 22(3):277–286, 2003.

[28] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, pages 224–236, 2018.

[29] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021.

[30] Zhiying Jiang, Zengxi Zhang, and Jinyuan Liu. Harmonized domain enabled alternate search for infrared and visible image alignment. *IEEE TIP*, 2025.

[31] Dae-Young Song, Geonsoo Lee, HeeKyung Lee, Gi-Mun Um, and Donghyeon Cho. Weakly-supervised stitching network for real-world panoramic image generation. In *ECCV*, pages 54–71. Springer, 2022.

[32] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE TIP*, 30:6184–6197, 2021.

[33] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In *NeurIPS*, volume 18, 2005.

[34] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE TPAMI*, 36(11):2144–2158, 2014.

[35] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, pages 1119–1127, 2015.

[36] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE TPAMI*, 38(10):2024–2039, 2015.

[37] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[38] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248. IEEE, 2016.

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024.

[40] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multispectral image stitching via global-aware quadrature pyramid regression. *IEEE TIP*, 2024.

[41] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[43] Zhiying Jiang, Risheng Liu, Shuzhou Yang, Zengxi Zhang, and Xin Fan. Drnet: Learning a dynamic recursion network for chaotic rain streak removal. *Pattern Recognition*, 158:111004, 2025.

[44] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, June 2021.

[45] Tao Huang, Shan You, Bohan Zhang, Yuxuan Du, Fei Wang, Chen Qian, and Chang Xu. Dyrep: Bootstrapping training with dynamic re-parameterization. In *CVPR*, pages 588–597, June 2022.

[46] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pages 7464–7475, June 2023.

[47] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

[48] Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, and Longin Jan Latecki. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *CVPR*, pages 12186–12195, 2021.

[49] Tianli Liao and Nan Li. Single-perspective warps in natural image stitching. *IEEE TIP*, 29:724–735, 2019.

[50] Zhiying Jiang, Xingyuan Li, Jinyuan Liu, Xin Fan, and Risheng Liu. Towards robust image stitching: An adaptive resistance learning against compatible attacks. In *AAAI*, volume 38, pages 2589–2597, 2024.

[51] Ziqi Xie, Weidong Zhao, Jian Zhao, and Ning Jia. Reconstructing the image stitching pipeline: Integrating fusion and rectangling into a unified inpainting model. In *NeurIPS*, volume 37, pages 123812–123845, 2024.

[52] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE TIP*, 15(11):3440–3451, 2006.

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[54] Pavan Chennagiri Madhusudana and Rajiv Soundararajan. Subjective and objective quality assessment of stitched images for virtual reality. *IEEE TIP*, 28(11):5620–5635, 2019.

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

 https://doi.org/10.52202/085713-4339

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: We have claimed the contribution and scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: In the conclusion section, we discuss the limitations of the proposed method and potential directions for future improvements.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: We provide a detailed description of the implementation details required to reproduce this algorithm within the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/085713-4339

Answer: [Yes] .

Justification: We are willing to open access our source data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: We provide a detailed description of the training and testing procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: We have provided an explanation and clarification of the accuracy of the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

   Justification: We have provided a detailed description of the computational resources required for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: We conform with the NeurIPS Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: We have discussed and analyzed the positive impact of the proposed method on subsequent visual applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/085713-4339

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper does not involve data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] .

    Justification: The creators or original owners of the assets used in the paper, including code, data, and models, are properly credited. The licenses and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/085713-4339

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: This paper did not employ LLMs in our methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.