# Image Stitching in Adverse Condition: A Bidirectional-Consistency Learning Framework and Benchmark

**Zengxi Zhang**
The University of Tokyo
cyouzoukyuu@gmail.com

**Junchen Ge**
Tsinghua University
junchen54ge@gmail.com

**Zhiying Jiang**
Dalian Martime University
zyjiang0630@gmail.com

**Miao Zhang**
Dalian University of Technology
miaozhang@dlut.edu.cn

**Jinyuan Liu***
Dalian University of Technology
atlantis918@hotmail.com

## Abstract

Deep learning-based image stitching methods have achieved promising performance on conventional stitching datasets. However, real-world scenarios may introduce challenges such as complex weather conditions, illumination variations, and dynamic scene motion, which severely degrade image quality and lead to significant misalignment in stitching results. To solve this problem, we propose an adverse condition-tolerant image stitching network, dubbed ACDIS. We first introduce a bidirectional consistency learning framework, which ensures reliable alignment through an iterative optimization paradigm that integrates differentiable image restoration and Gaussian-distribute encoded homography estimation. Subsequently, we incorporate motion constraints into the seamless composition network to produce robust stitching results without interference from moving scenes. We further propose the first adverse scene image stitching dataset, which covers diverse parallax and scenes under low-light, haze, and underwater environments. Extensive experiments show that the proposed method can generate visually pleasing stitched images under adverse conditions, outperforming state-of-the-art methods. Code and benchmark are available at https://github.com/ZengxiZhang/ACDIS.

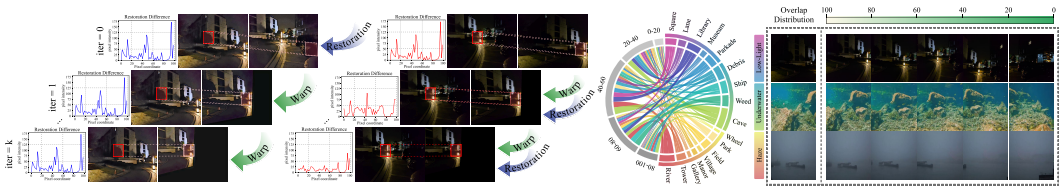
## 1 Introduction

Image stitching aims to construct a wide field-of-view (FoV) scene from multiple images captured in different viewpoints. It is widely used in various applications such as autonomous driving [1], map construction [2] and virtual reality [3]. Although it has achieved rapid development in recent years, aligning images over large disparity views, visually degraded environments and dynamic scenes still remains challenging.

Early stitching methods [4, 5, 6] predominantly relied on SIFT-based feature detection [7] to estimate parametric warping models. However, they fall short in low-texture regions or scenes with repetitive geometric patterns. In recent years, learning-based methods [8, 9, 10] replace handcrafted feature

*Corresponding Author.

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




(a) Conventional alignment strategy (b) Our alignment strategy (c) The proposed Adverse Scene Image Stitching Dataset

Figure 1: (a) and (b) compare the alignment process of the traditional approach and our proposed strategy in challenging environments. The progression from top to bottom illustrates the refinement of alignment from coarse to fine, where the quadrilateral boxes denote regions corresponding to the same scene. The line charts on the left depict the current intensity difference of the restoration effect within the respective boxes. (c) presents the baseline and scene distribution of the proposed dataset.

extraction with deep semantic representations, enabling more robust alignment. Nevertheless, they still struggle to achieve satisfactory results in adverse environments (e.g., low-light, haze, underwater, and dynamic scenes). These conditions corrupt the image pairs through noise, color casts, or scene variation, leading to information distortion, which in turn affects feature matching and reconstruction.

To address this issue, some methods [11, 12, 13] employ image stitching by introducing environment-insensitive multi-modal data to alleviate the impact of degradation factors. However, these cannot be applied in general scenarios due to their strong data dependency. Furthermore, few methods mitigate image degradation in adverse environments without introducing additional information. On the other hand, existing stitching datasets [14, 9] are mainly collected under ideal natural light conditions and lack guiding reference, making it difficult to evaluate stitching tasks in degraded scenes.

To overcome the above limitations, in this paper, we propose a robust deep image stitching network for adverse conditions (ACDIS), which consists of two stages: In the first stage, we introduce a bidirectional-consistency learning framework to achieve robust image alignment under harsh environment. Specifically, considering that conventional homography estimation methods with deterministic displacement outputs often produce unreliable predictions in textureless or blurry regions, we begin by proposing a recursive parameterized homography estimation module that models displacements in a Gaussian-distributed transformation space. By jointly predicting the mean and variance, the model explicitly encodes uncertainty and mitigates overconfident regression. A Jensen–Shannon divergence–based optimization further refines the displacement distribution, yielding stable and reliable homography estimation.

Traditional perception methods in adverse conditions typically treat visual enhancement [15, 16, 17, 18] as a preprocessing step, as illustrated in Fig. 1 (a). However, as depicted in the line chart, this straightforward concatenation strategy may introduce irreducible restoration discrepancies, ultimately disrupting feature matching. To address this, as depicted in Fig. 1 (b), we embed a differentiable restoration module within the iterative homography estimation pipeline, enabling bidirectional optimization of both restoration and alignment. As shown in the corresponding line chart, this mechanism progressively harmonizes the scene structure, enhancing restoration consistency in iterative deformation processes, thereby achieving more reliable alignment.

In the second stage, we propose a motion-tolerant seamless composition network, which introduces motion loss to reduce the impact of moving object displacement caused by inconsistent image capturing times while generating clear stitched images. For evaluating the performance of stitching methods in harsh environments, we additionally propose the first adverse scene image stitching dataset (ASIS), which covers three harsh environments: low-light, haze, and underwater. ASIS includes 17 scenes with a total of 2,250 pairs of images, which are derived from manual capturing and network collection. All the image pairs are far from the plane structure and under a wide parallax range. In addition, we pre-align all image pairs to obtain the reference homography for a more comprehensive evaluation. In summary, our contributions are as follows:

- A recursive parameterized homography estimation module is proposed that iteratively encodes displacement within a Gaussian-distributed transformation space, progressively refining a reliable homography transformation.

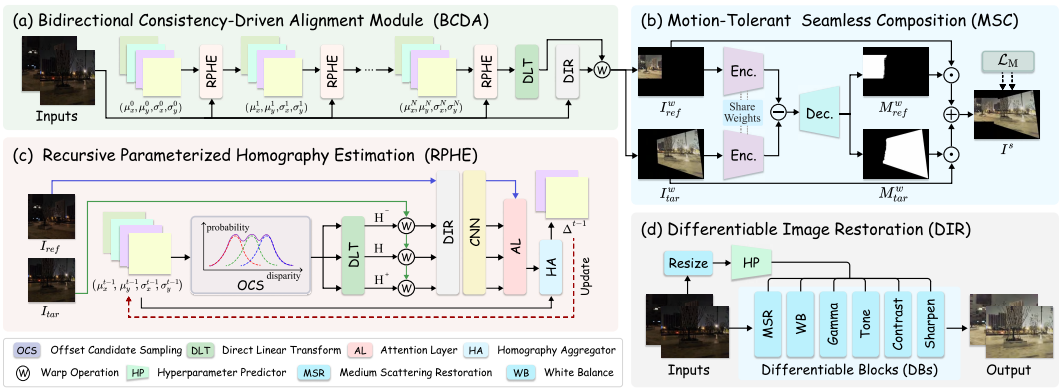



Figure 2: Workflow of the proposed Method.

- We propose a bidirectional-consistency learning framework, which achieves adverse condition-tolerant homography estimation by embedding lightweight differentiable restoration blocks into iterative alignment process.
- A motion-tolerant seamless composition network is introduced that generates visually pleasing stitching images while avoiding the interference of moving objects on temporally inconsistent image pairs.
- We release the first real-world adverse environment image stitching dataset, which contains 2,250 degraded image pairs with homography reference to evaluate the effectiveness of stitching methods in adverse conditions.

## 2 Related Work

### 2.1 Image Stitching Methods

Early traditional stitching works [4, 5, 6, 14, 19, 20, 21, 22] constructed global adaptive warping by using SIFT [7] to extract key points. Brown *et al.* [4] used multi-image matching technology in image stitching tasks and combined multi-band mixing to generate panoramas. Zaragoza *et al.* [14] fine-tuned the warping projection by moving direct linear transformations to reduce artifacts during warping, thereby reducing the reliance on ghost concealment algorithms. Peng *et al.* [23] introduced edge detection and sampled edges to construct triangles representing geometric structures. Then, a similarity transformation is performed by combining the geometric structure preserving energy term.

In recent years, some works have applied deep learning networks [8, 9, 11, 24, 12, 10] to the stitching task. Nie *et al.* [9] proposed an unsupervised image stitching method and introduced seam loss during the reconstruction process to eliminate the pixel-level misalignment. Then, they [10] employed a seam-driven mask as an alternative, effectively preserving structural coherence across the stitched regions. Although these methods have proven effective, in practice, immeasurable environmental factors often affect the captured images, thus greatly affecting the stitching effect.

### 2.2 Image Stitching Datasets

In recent years, many datasets [5, 21, 14, 19, 25] have been proposed to more comprehensively evaluate the effect of image stitching. Nie *et al.* [8] synthesized a stitching dataset with ground-truth by cropping and warping existing image datasets. Subsequently, considering the lack of disparity in synthetic datasets, they [9] additionally provided a real-world unsupervised dataset UDIS-D. However, as the source data are captured instead of synthesized, the ground truth of corresponding FoV scenes cannot be obtained, limiting quantitative evaluation. Kweon *et al.* [26] then proposed a stitching dataset with pixel-level warping and ground-truth stitching results, which simulated the real-world stitching by collected data from 3D virtual scenes. However, except for a small number of low-light images provided by UDIS-D, there is currently no stitching dataset specifically used to evaluate the robustness of the image stitching network in harsh environments.

## 3 The Proposed Method

Our method consists of two stages: Bidirectional Consistency-Driven Alignment Module (BCDA) and Motion-Tolerant Seamless Composition (MSC), as shown in Fig. 2. $\{I_{ref}, I_{tar}\}$ denotes images captured in adverse environments, where $ref$ and $tar$ represent the reference and target viewpoint.

BCDA comprises Recursive Parameterized Homography Estimation Module (RPHE) and Differentiable Image Restoration (DIR). RPHE alleviates the challenge of spatially limited dynamic cost volume with large baseline image pairs by sampling offset candidate values by mean and variance in 2D directions. DIR is designed to mitigate the interference of degradation factors on feature extraction. It applies CNN-encoded hyperparameters on differentiable restoration blocks to achieve lightweight and robust image restoration. Rather than applying the restoration module as a preprocessing step for alignment, we establish a Bidirectional-Consistency Learning Framework (BCLF), where DIR is embedded within the coarse-to-fine deformation estimation process, forming a progressive bidirectional optimization pipeline. Ultimately, BCDA achieves region-asymptotically consistent image pair restoration and robust image alignment. After obtaining $\{I^w_{ref}, I^w_{tar}\}$ by warping restored images, we input them into MSC to generate motion-tolerant wide field-of-view scene $I^s$. Next, we provide detailed information about each module.

### 3.1 Recursive Parameterized Homography Estimation

Some traditional stitching methods [9, 27, 28, 29] apply dynamic cost volume $\mathcal{C}(x) = c^r_{x,y}$ with search redius $r$ to match difference between $(x, y)$ of $I_{ref}$ and $(x \pm r, y \pm r)$ of $I_{tar}$ for reducing the memory consumption. However, spatial limited cost volume is not conducive to convergence in large baseline scenarios. Furthermore, these methods based on deterministic displacement outputs tend to produce unreliable predictions in textureless or blurry regions, and perform poorly in image stitching tasks under adverse conditions such as heavy fog or low light.

To address this, modeling displacement space $d$ as a Gaussian distribution provides a more robust alternative. The proposed model predicts not only the mean (displacement) but also the variance (as an uncertainty estimate [30]), allowing it to express high uncertainty in ambiguous or ill-posed regions and avoid overconfident regression to potentially incorrect values. Specifically, the cost volume will sample candidate offsets by mean $\mu$ and standard deviation $\sigma$ from each offset:

$$\mathcal{C}(i, \mu_i, \sigma_i) = c_i^{d(\mu_i,\sigma_i)},\ d(\mu_i, \sigma_i) \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), \tag{1}$$

where $\sim$ represents the sampling from Gaussian distribution $\mathcal{N}$. $i \in \{x, y\}$, $x$ and $y$ denotes the horizontal and vertical displacements.

As shown in Fig. 2, we first sampling candidate offsets $\mu_i$, $\mu_i + \sigma_i$ and $\mu_i - \sigma_i$ from the initialized offsets in each iteration. Then, we transform them all into the homography $H$, $H^+$ and $H^-$ via Direct Linear Transformation (DLT) for warping the original images respectively. Subsequently, we input the warped images into the DIR (3.2) and pyramid CNN to extract features resilient to adverse environments. Through Attention Layer (AL) [31] and Homography Aggregation Module (HA) [31], we finally obtain the residual $\mu$ for updating. The detailed structure of CNN, AL and HA is illustrated in supplementary materials.

In order to effectively learn the parameters $\theta = \{\mu, \sigma\}$ of the Gaussian distribution, we use JS divergence [32] based optimization $\mathcal{J}$, which forces $\theta$ to gradually approach the Gaussian distribution during the optimization process:

$$\mathcal{J}_i = \frac{1}{2}\left(F\left(\mathcal{N}_{\mathrm{GT}} \| \mathcal{N}_i\right) + F\left(\mathcal{N}_i \| \mathcal{N}_{\mathrm{GT}}\right)\right), \tag{2}$$

where $\mathcal{N}_i$ is the short form of $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. $F(\cdot \| \cdot)$ represents KL divergence [33]. $\mathcal{N}_{\mathrm{GT}} = \mathcal{N}\left(\mu_{\mathrm{GT}}, \sigma^2_{\mathrm{GT}}\right)$ denotes predefined Gaussian distribution, where $\mu_{\mathrm{GT}}$ represents the ground truth offset. We can then update $\theta$ via feedforward gradient optimization [34, 35] in each iteration $t$:

$$\begin{aligned}
\sigma_i^t &= \sigma_i^{t-1} - \frac{1}{2}\left(\frac{(\sigma_i^{t-1})^2 - \sigma^2_{\mathrm{GT}} - (\mu_{\mathrm{GT}} - \mu_i^{t-1})^2}{(\sigma_i^{t-1})^3} - \frac{1}{\sigma_i^{t-1}} + \frac{\sigma_i^{t-1}}{\sigma^2_{\mathrm{GT}}}\right), \\
\mu_i^t &= \mu_i^{t-1} - \left(-\frac{\mu_{\mathrm{GT}} - \mu_i^{t-1}}{2}\left(\frac{1}{(\sigma_i^{t-1})^2} + \frac{1}{\sigma^2_{\mathrm{GT}}}\right)\right).
\end{aligned} \tag{3}$$

Considering that $\mu_{\mathrm{GT}}$ is not available during inference, we replaced the $\mu_{\mathrm{GT}}$ by approximate the optimizing step $\Delta^{t-1} = \mu_{\mathrm{GT}} - \mu_i^{t-1}$, which is estimated in each iteration. By updating $\theta$ in $N$ iterations, we can finally estimate a robust homography that is suitable for large baseline scenes in harsh conditions. L1 loss is used during the training stage with the ground truth of the offset in $N$ iterations, which can be described as:

$$L_1 = \sum_{t=1}^{N} \lambda_1^{(N-t)} \left| \mu^t - \mu_{\mathrm{GT}} \right|. \tag{4}$$

### 3.2 Bidirectional-Consistency Learning Framework

When conducting computer vision tasks in adverse conditions, previous works [36, 37, 38] usually directly employ restoration networks as the preprocessing step for the current task. However, when faced with stitching tasks, restoration networks encounter two main challenges: (1) Data-driven networks often overlook the physical imaging priors of images under adverse scenes, resulting in visually insensitive artifacts that implicitly disrupt the geometry structure of the images. (2) Restoration networks may exhibit varying degrees of effectiveness in restoring images from different perspectives. Both of these challenges can affect the feature matching between image pairs, thereby influencing warping estimation.

To address the first challenge, we propose a Differentiable Image Restoration Module (DIR), which includes a Hyperparameter Predictor (HP) and Differentiable Restoration Blocks (DBs). First, HP learns the global information of the resized input image to obtain hyperparameters. Then, we used the obtained hyperparameters as weights in DBs to achieve adaptive image restoration.

DBs consists of Medium Scattering Restoration (MSR), White Balance (WB), Gamma, Hue, Contrast and Sharpen. MSR estimates the atmospheric light $A$ and the transmission map $t(x)$ through the atmospheric light imaging model [39] to restore the absorption and scattering of light in the gas or liquid medium, which is expressed as $I(x) = J(x)t(x) + A(1 - t(x))$ , where $I$ and $J$ denotes the degraded image and its clear counterpart. $A$ can be obtained by calculating the first 1000 brightest pixels in the dark channel and averaging the pixels at the corresponding positions of the original image. $t(x)$ can be obtained according to the Beer-Lambert law [40, 41], which is described as:

$$t(x, \omega) = 1 - \omega \min_{C} \left( \min_{y \in \Omega(x)} \frac{I^C(y)}{A^C} \right), \tag{5}$$

where $\omega$ is an hyperparameter that controls the degree of restoration, which is obtained by HP.

WB and Gamma are pixel-level restoration blocks. Among them, $\{\omega_r, \omega_g, \omega_b\}$ are used as hyperparameters of HP for mapping of three channel pixels. $G$ is the hyperparameter of Gamma for the weight of the power function. Tone restoration can be expressed as a channel-independent monotonic piecewise linear function [42]. The point $(k/M,\, T_k/T_M)$ on the tone curve is represented by M hyperparameters $\{t_0, ..., t_{M-1}\}$, where $T_k = \sum_{i=0}^{k-1} t_i$. The mapping process can be expressed as $I_o = \frac{1}{T_M} \sum_{j=0}^{M-1} \mathrm{clip}\left(L \cdot I_i - j, 0, 1\right) t_k$, where $I_o$ and $I_i$ represent the output and the input image. Contrast restoration determines the linear difference between the $I_o$ and $I_i$ through the hyperparameter $\alpha$, which is expressed as $I_o = \alpha \cdot \mathrm{En}\left(I_i\right) + (1 - \alpha) \cdot I_i$, where En is the mapping process [42]. The sharpening [43] process is described as $F(x, \lambda) = I(x) + \lambda(I(x) - G(I(x)))$, where $G$ represents Gaussian filter. The hyperparameter $\lambda$ is a positive scaling factor to control the prominence of details.

In summary, different from traditional data-driven networks, DIR leverages physical priors to adaptively restore images in various adverse environments while minimizing interference, ensuring more accurate feature matching. We use L1 loss $\mathcal{L}_1$ and SSIM loss $\mathcal{L}_{\mathrm{s}}$ with the ground truth image $I_{\mathrm{GT}}$ in the training process, which can be described as:

$$\mathcal{L}_{\text{res}} = \lambda_2 \mathcal{L}_1(I_o, I_{\mathrm{GT}}) + \lambda_3 \mathcal{L}_{\mathrm{s}}(I_o, I_{\mathrm{GT}}). \tag{6}$$

Although artifacts can be reduced with physical prior, consistent restoration of images from different perspective still remains challenging. Therefore, we further build a Bidirectional-Consistency Learning Framework. Specifically, we use DIR to restore the warped images at each iteration, rather than just once as a preprocessing means only before homography estimation process, achieving in equivariant restoration. This framework can be expressed as:

$$E_{tar}^n = \Phi\left(\mathcal{W}\left(I_{tar}; \hat{\mathbf{H}}^n\right)\right), \hat{\mathbf{H}}^{n+1} = \Psi\left(\mathbf{I}_{ref}; E_{tar}^n\right), \tag{7}$$

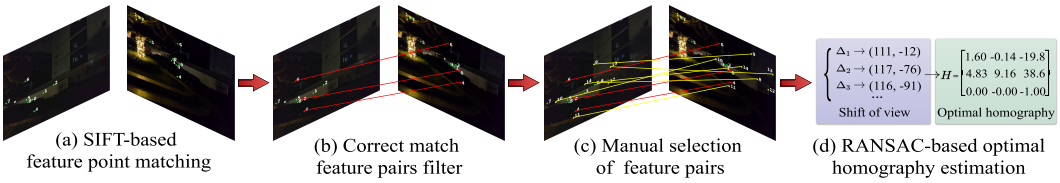



Figure 4: Reference generation process of the newly proposed ASIS dataset.

where $\Phi$ and $\Psi$ denote the restoration and the homography estimation and module. $E^n_{tar}$ and $\mathbf{H}^n$ represent the enhanced warping images and current homography in the $n$-th iteration. Since the computation cost of DIR is only 0.07 GFlops at each restoration, it will not significantly increase the time consumption in embedding it in the iteration. As the iteration proceeds, $I^w_{tar}$ becomes more aligned with $I_{ref}$, enabling DIR to increasingly achieve consistent restoration of overlapping regions, which further improves homography estimation.

### 3.3 Motion-Tolerant Seamless Composition

Inspired by UDIS2 [10], this stage generates a soft mask with floating point numbers and synthesizes a seamless stitched image $I^s$ by $I^s = M^w_{ref} \times I^w_{ref} + M^w_{tar} \times I^w_{tar}$. U-Net [44] is applied as the network structure for generating composition masks $\{M^w_{ref}, M^w_{tar}\}$. Since the image pairs are usually captured at different times, the presence of moving objects can affect the visual appearance of the stitched image, especially in the seam region. As shown in (a) of Fig. 3, taking a human subject as an example, when capturing images from two viewpoints, the man in the scene may move during the interval, causing him to appear in different positions across the two viewpoints. In order to avoid unsmooth stitching result with ghost caused by above challenges, we further introduce a motion loss $\mathcal{L}_{\mathrm{M}}$ in the training process of MSC to improve the robustness to moving objects, which can be described as follows:

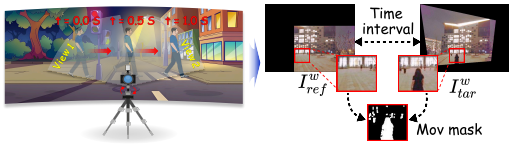



Figure 3: The generation process of the motion mask during image capturing.

$$\mathcal{L}_{\mathrm{M}} = \omega_1 \left\| M^m I^s - M^m I^w_{ref} \right\|_1 + \omega_2 \left\| M^m I^s - M^m I^w_{tar} \right\|_1, \tag{8}$$

where $\{\omega_1, \omega_2\} = \mathcal{S}(\mathcal{D}(M^s_{ref} \cdot I^s), \mathcal{D}(M^s_{tar} \cdot I^s))$, which controls the composition weight of MSC to the region of moving objects between $I^w_{tar}$ and $I^w_{ref}$. $M^s$ denotes the seam mask, which is introduced in supplementary materials. $\mathcal{S}$ denotes the Softmax operation. $\mathcal{D}$ represents the 2D adjacent difference, which can be illustrated as follows:

$$\mathcal{D}(I) = \sum_{i,j} |I(i, j+1) - I(i, j)| + \sum_{i,j} |I(i+1, j) - I(i, j)|, \tag{9}$$

where $i$ and $j$ represent the location index in x and y axis. $M^m$ denotes the motion mask of the warping images, which can be described as follows:

$$M^m = \mathcal{M}\left(\mathcal{B}\left(|I^w_{ref} - I^w_{tar}|, \phi\right), \kappa\right), \tag{10}$$

where $\mathcal{B}$ denotes the binarization operation with the threshold $\phi$. $\mathcal{M}$ represents the morphology filter. $\kappa$ is the kernel of $\mathcal{M}$ with the size of 3. Therefore, under the constraints of $\mathcal{L}_{\mathrm{M}}$, MSC can adaptively ignore the interference of moving objects appearing in overlapping areas on seamless composition, thereby generating visually pleasing stitched images. We additionally apply boundary loss [10] $\mathcal{L}_{\mathrm{B}}$ and smoothness loss [10] $\mathcal{L}_{\mathrm{S}}$ to make the stitched images smooth and clear. The final composition loss is expressed as follows:

$$\mathcal{L}_{\mathrm{com}} = \lambda_4 \mathcal{L}_{\mathrm{B}} + \lambda_5 \mathcal{L}_{\mathrm{S}} + \lambda_6 \mathcal{L}_{\mathrm{M}}. \tag{11}$$

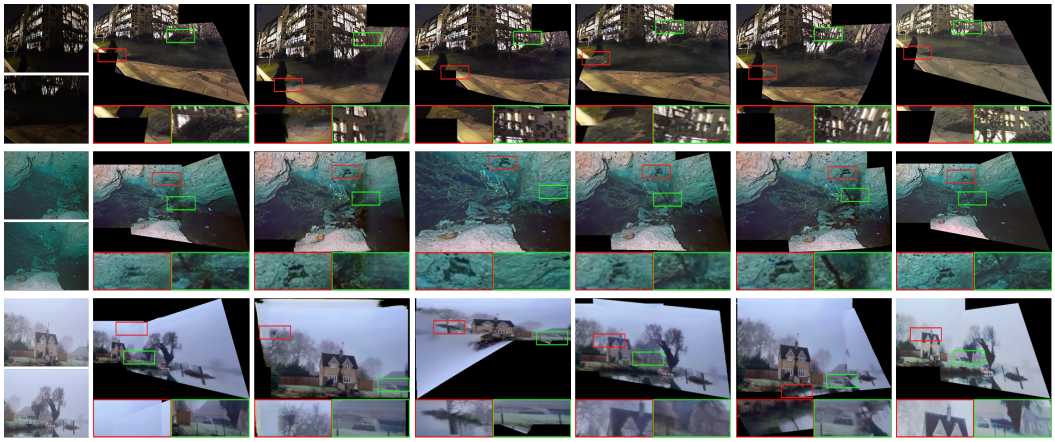



Figure 5: Visualization results on the ASIS Dataset.

## 4 The Proposed Dataset

In order to comprehensively evaluate image stitching tasks in various harsh environments, we released an Adverse Scene Image Stitching Dataset (ASIS) that integrates in low light, haze and underwater environments. A visual representation of the dataset is shown in (c) of Fig. 1. ASIS contains 2,250 pairs of images, including 750 images each of low-light, underwater, and haze environments, covering 17 scenes such as caves, wrecks and fields. It is worth mentioning that some of the sources of these data come from the Internet and some come from independent photography. The captured images are far from planar structures, which ensures the disparity diversity of the proposed ASIS dataset. More detailed information of ASIS is demonstrated in supplementary materials.

To verify the warping performance of different methods, we provided the homography reference for ASIS. Inspired by [45], the generation process of the reference is illustrated in Fig. 4. We first match image pairs using SIFT [46], however, due to image degradation caused by adverse conditions, it usually suffers from low number and accuracy of matched feature points. Therefore, we manually filter correct matching pairs to prevent the generation of outliers. Moreover, as shown in Fig. 4 (b), where the number of correctly matched pairs is insufficient to generate a reliable homography. Thus, we further select new matching pairs to achieve a more robust homography generation. After obtaining matching pairs, we apply Random sample consensus (RANSAC) [47] to generate the optimal homography and perform preliminary distortion to verify the performance of the reference.

## 5 Experiments

### 5.1 Implement Details

The training process is divided into three steps. First, we pre-train the DIR module with a learning rate of 1e-4 and an epoch of 200. LSRW [48] and UIEBD [49] are used as training datasets for low-light and underwater image enhancement tasks respectively. We also synthesize the training dataset for single image dehazing from VOC [50] according to the atmospheric scattering model [51]. Subsequently, we trained the BCDA module with DIR with an epoch of 160. We then use the above datasets to synthesize the homography training datasets through the synthesis method of VFIS [8]. During the training process, the $\mu$ and $\sigma$ are initialized to 0 and 32 respectively. M, N, $\sigma_{\mathrm{GT}}$, $\phi$ are set to 8, 6, 2 and 20. Batch size and learning rate are set to 16 and 5e-5. Finally, we trained the MSC with a learning rate of 1e-4 and an epoch of 100. $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ ,$\lambda_5$ and $\lambda_6$ are set to 2, 1, 100, 100, 1 and 1. All the experiments are conducted on PyTorch with NVIDIA RTX 4090 GPU.

Table 1: Quantitative comparison for image stitching under adverse environment. ↑ indicates that higher values correspond to superior outcomes. The top-performing and second-best results are highlighted in red and blue.

| **Low-light Environment** | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Low-light Image | | | | | HBLALS | | | | | NeRCo | | | | | Ours |
| | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | |
| PSNR↑ | 25.792 | 23.002 | 24.465 | 24.910 | 24.742 | 27.461 | 22.048 | 25.346 | 25.867 | 25.890 | 28.244 | 22.244 | 25.717 | 26.340 | 26.053 | 30.506 |
| SSIM↑ | 0.876 | 0.869 | 0.879 | 0.914 | 0.905 | 0.897 | 0.835 | 0.887 | 0.919 | 0.914 | 0.906 | 0.863 | 0.883 | 0.935 | 0.927 | 0.939 |
| SIQE↑ | 10.352 | 12.329 | 10.861 | 11.372 | 11.493 | 8.786 | 11.897 | 11.062 | 10.197 | 10.437 | 8.507 | 12.578 | 10.837 | 11.089 | 9.688 | 12.854 |
| NIQE↓ | 4.402 | 3.671 | 4.402 | 4.261 | 4.679 | 4.500 | 3.312 | 3.655 | 3.596 | 4.234 | 4.168 | 3.263 | 4.350 | 4.178 | 5.103 | 3.186 |
| **Underwater Environment** | | | | | | | | | | | | | | | | |
| Method | Underwater Image | | | | | HBLALS | | | | | WaterFlow | | | | | Ours |
| | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | |
| PSNR ↑ | 24.631 | 23.740 | 23.054 | 25.105 | 25.354 | 25.192 | 22.399 | 23.299 | 24.851 | 24.897 | 24.251 | 21.276 | 22.458 | 23.706 | 24.776 | 28.607 |
| SSIM↑ | 0.856 | 0.838 | 0.826 | 0.876 | 0.889 | 0.863 | 0.796 | 0.823 | 0.863 | 0.868 | 0.862 | 0.787 | 0.826 | 0.806 | 0.861 | 0.918 |
| SIQE ↑ | 5.657 | 5.668 | 4.012 | 5.646 | 6.115 | 5.456 | 6.089 | 5.313 | 5.901 | 6.152 | 4.934 | 5.998 | 5.919 | 5.086 | 6.057 | 6.444 |
| NIQE ↓ | 4.210 | 3.998 | 4.510 | 4.941 | 6.444 | 3.918 | 3.531 | 4.000 | 4.496 | 5.849 | 4.108 | 3.715 | 4.203 | 4.586 | 6.067 | 4.202 |
| **Haze Environment** | | | | | | | | | | | | | | | | |
| Method | Haze Image | | | | | HBLALS | | | | | WGWS | | | | | Ours |
| | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | APAP | VFIS | LPC | UDIS | UDIS2 | |
| PSNR↑ | 25.065 | 22.037 | 23.657 | 23.532 | 24.690 | 27.615 | 23.969 | 26.135 | 26.009 | 27.065 | 27.530 | 24.038 | 26.380 | 25.836 | 27.245 | 31.136 |
| SSIM ↑ | 0.911 | 0.866 | 0.900 | 0.909 | 0.928 | 0.941 | 0.869 | 0.928 | 0.921 | 0.939 | 0.936 | 0.878 | 0.929 | 0.922 | 0.941 | 0.963 |
| SIQE↑ | 7.785 | 6.817 | 6.251 | 6.970 | 7.637 | 6.885 | 7.524 | 6.798 | 7.009 | 7.120 | 7.908 | 8.000 | 7.361 | 7.470 | 7.510 | 8.057 |
| NIQE ↓ | 5.632 | 5.297 | 5.681 | 5.629 | 6.224 | 5.255 | 4.972 | 5.311 | 5.434 | 5.990 | 5.687 | 5.160 | 5.747 | 5.547 | 6.190 | 4.970 |

### 5.2 Comparison with Existing Methods

This method mainly studies the stitching algorithm in low-light, haze and underwater environments. For a fair comparison, we first use a unified restoration framework HBLALS [52] and environment-specific restoration methods NeRCo [53], WaterFlow [54] and WGWS [55] respectively to obtain images closer to ideal conditions. We then applied the these clear images to representative stitching algorithms, including traditional APAP[14], LPC [56], and deep learning-based VFIS [8], UDIS [9], and UDIS2 [10]. It is worth mentioning that we also applied the stitching algorithm to the original images without restoration to verify the effectiveness of the restoration method for image stitching.

The first row in Fig. 5 shows the stitching comparison of images restored using HBLALS in low-light scenes. Given the resolution differences across stitching methods, we resize them just for a consistent and aesthetic display. APAP fails to achieve a smooth transition, resulting in obvious seams. VFIS, UDIS and UDIS2 are severely misaligned in the green zoom-in region. LPC breaks the consistency of the road structure in the red zoom-in region. The second row shows the stitching results in the underwater environment, where severe color distortions and similar rock textures impact the stitching performance. LPC causes the loss of distinctive details from the reference image. VFIS and UDIS2 disrupt the reef structures in the overlapping region, resulting in visually unappealing seams. In UDIS, the information within the red box exhibits noticeable blurring, while APAP introduces prominent artifacts in the red box and visible seams in the green box. The third row shows the visualization in the haze environment. Most methods suffer from significant misalignment during the stitching process, resulting in a loss of scene information. Only APAP correctly estimates the image distortion, however introducing significant visual interference at the seam region.

Tab. 1 shows the quantitative results of all methods. For a comprehensive comparison of the proposed methods, we verify the alignment effects by calculating PSNR and SSIM on the overlapping regions. Larger values represent better warping effects. Subsequently, the stitching effect is evaluated through the stitched image quality evaluator (SIQE) [57]. We additionally quoted the Natural Image Quality Evaluator (NIQE)[58] to evaluate the quality of the reconstructed image. A higher SIQE index indicates better stitching effect, and a lower NIQE index indicates better image quality.

It is evident that the proposed method outperforms all others in PSNR and SSIM, validating its effectiveness in warping estimation. Additionally, the SIQE of the proposed method also exceeds other methods in all adverse environments, confirming that our method can greatly reduce the impact of image degradation during stitching. Our NIQE surpasses all methods in low-light and

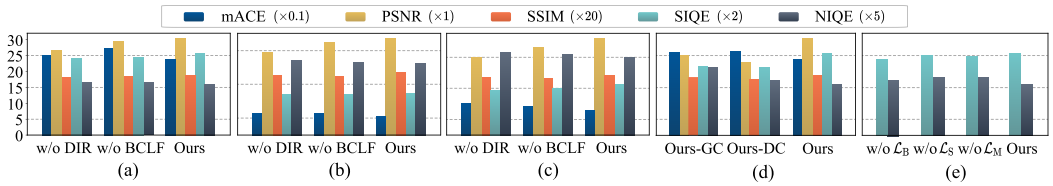



Figure 7: Evaluation on ablation experiment.

haze environments but does not exhibit advantages in underwater environment. This is because underwater scene images are mostly captured in close range with large disparities between image pairs. Therefore, correct warping leads to excessive invalid black regions in the generated images, affecting the evaluation of image quality metrics. Considering all metrics, our method achieves significant advantages in warping and stitching performance under adverse conditions.

## 5.3 Ablation Study

### 5.3.1 Study on Bidirectional-Consistency Learning Framework:

We conducted ablation studies across different environments to verify the effectiveness of the learning framework. Specifically, we first abandoned DIR (w/o DIR) to verify the effect of image restoration on the proposed method. Subsequently, instead of employing BCLF, we only concatenate DIR with homography estimation module (w/o BCLF) to verify the effect of our learning framework. We additionally utilized the mean Average Corner Error (mACE) [59] to validate the performance of the homography estimation. A smaller mACE indicates a more accurate estimation. Fig. 6 illustrates the performance in three adverse environments. It is worth noting that fine details in degradation scenes are difficult to discern by the human eye. Therefore, we additionally restored the results of the unenhanced version in qualitative comparison to facilitate readers in comparing the stitching effects.

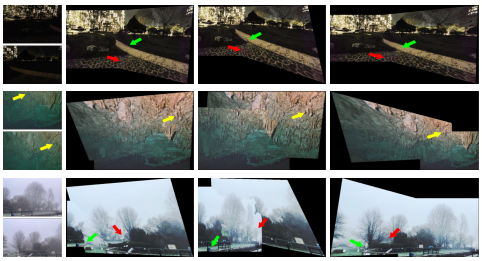



Figure 6: Visualization results for the study on DIR and BCLF.

In the low-light example, w/o DIR produces artifacts on the floor tiles pointed by the red arrows, destroying the original texture. w/o BCLF suffers from a clear misalignment on the path pointed by the green arrow. In the underwater environment, facing a reef group with a similar structure, neither w/o DIR nor w/o BCLF considers of the correspondence between the input pairs, resulting in a misalignment of the stitching results. In the haze environment, both w/o DIR and w/o BCLF misestimate the deformation, ultimately causing the results to lose a large amount of scene information. (a)-(c) of Fig. 7 depict the quantitative results of the scenarios mentioned above. In summary, the proposed enhancement method demonstrates significant effectiveness under adverse environments.

### 5.3.2 Study on Recursive Parameterized Homography Estimation:

We further conducted ablation studies on RPHE to validate its performance in homography estimation. Specifically, we adopted traditional global correlation (Ours-GC) [9] and dynamic correlation (Ours-DC) [10] as a replacement for the parameterized cost volume. Fig. 8 presents the visual results in low-light environment. The yellow rectangle indicates the overlap ratio of image pairs. The image pairs in the first example face the challenge of a low overlap ratio. Compared with the ours stitching result, both results of Ours-GC and Ours-GC fails on matching the limited scene struc-

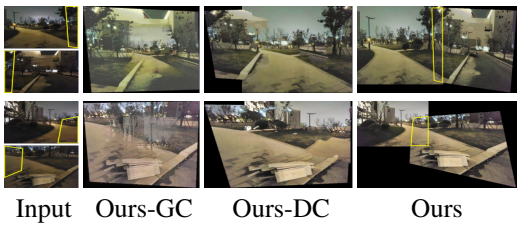



Figure 8: Visualization results for the ablation study on the RPHE.

ture. In the second example, the overlapping region mainly consists of ground areas with fewer fine details, posing significant challenges for stitching. Ours-GC suffers from a significant loss of scene information, while noticeable misalignment occurs in Ours-DC. Only the proposed method achieves accurate stitching results in limited effective areas. Quantitative results of this ablation study are shown in Fig. 7 (d), which also validate the advantages of the proposed homography estimation strategy in all metrics.

#### 5.3.3 Study on Motion-Tolerant Seamless Composition:

We ablate $\mathcal{L}_{\mathrm{B}}$, $\mathcal{L}_{\mathrm{S}}$, and $\mathcal{L}_{\mathrm{M}}$ respectively to verify the effect of losses on the image composition stage. The visualization results are shown in Fig. 9. In this example, due to the time difference in image capturing, the character in the red frame has moved from the back of the car to the front of the car. The reconstructed scenes of w/o $\mathcal{L}_{\mathrm{B}}$, w/o $\mathcal{L}_{\mathrm{S}}$ and w/o $\mathcal{L}_{\mathrm{M}}$ all show the same person at different times in the stitched images. At the same time, both w/o $\mathcal{L}_{\mathrm{S}}$ and w/o $\mathcal{L}_{\mathrm{M}}$ suffer from serious information loss in the human leg. Furthermore, as shown in the green frame, obvious seam differences appears in the results of w/o $\mathcal{L}_{\mathrm{B}}$ and w/o $\mathcal{L}_{\mathrm{S}}$. Only the proposed method ensures the temporal uniformity while smoothing the seam difference. (e) of Fig. 7 shows the qualitative metrics of the reconstruction module. Since all experiments are performed under the same image warping, we only use image quality metrics for evaluation. Both qualitative and quantitative experiments demonstrate the effectiveness of the proposed constraints for reconstructing the network.

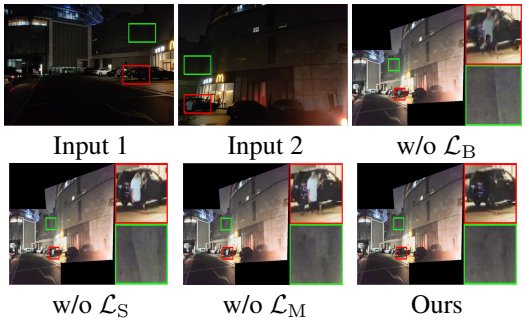



Figure 9: Visualization results for the ablation study on the MSC.

## 6 Limitations

The proposed motion constraint only improves stitching performance for small-range movements in dynamic scenes. For instance, when a pedestrian moves within the overlapping region of two images, our method effectively suppresses visual ghosting. However, if the movement extends beyond the overlap, the network cannot recognize it as the same object, leading to duplicate appearances in the final stitched image. This limitation arises because the motion loss focuses solely on movements within the overlapping area and fails to handle large displacements. Addressing fast, large-scale motions while ensuring non-redundant targets in the output remains challenging and may require integrating explicit object detection with a generative model to reconstruct occluded backgrounds around duplicated targets.

## 7 Conclusion

We propose an adverse condition-tolerant image stitching network. It features a bidirectional consistency learning framework, which iteratively optimizes differentiable restoration and Gaussian-encoded homography estimation for reliable alignment. Additionally, motion constraint is proposed in the seamless composition stage, suppressing artifacts from moving objects. We also construct the first adverse scene image stitching dataset, covering diverse low-light, haze, and underwater scenarios. Extensive experiments on the proposed dataset demonstrate the stitching performance of our method under adverse conditions.

## Acknowledgments

This work is partially supported by the China Postdoctoral Science Foundation (No. 2023M730741)); in part by the National Natural Science Foundation of China (Nos. 62302078, 62372080); and in part by the Fundamental Research Funds for the Central Universities (No. 3132025276).

## References


[1] Miao Liao, Feixiang Lu, Dingfu Zhou, Sibo Zhang, Wei Li, editor="Vedaldi Andrea Yang, Ruigang", Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Dvi: Depth guided video inpainting for autonomous driving. In *Computer Vision – ECCV 2020*, pages 1–17, 2020.

[2] Hongbo Jiang, Wenping Liu, Guoyin Jiang, Yufu Jia, Xingjun Liu, Zhicheng Lui, Xiaofei Liao, Jing Xing, and Daibo Liu. Fly-navi: A novel indoor navigation system with on-the-fly map generation. *IEEE Transactions on Mobile Computing*, 20(9):2820–2834, 2020.

[3] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020.

[4] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74:59–73, 2007.

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

[6] Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. Smoothly varying affine stitching. In *CVPR 2011*, pages 345–352. IEEE, 2011.

[7] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[8] 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.

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

[10] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7399–7408, 2023.

[11] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3783–3791, 2022.

[12] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multi-spectral image stitching via spatial graph reasoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 472–480, 2023.

[13] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multispectral image stitching via global-aware quadrature pyramid regression. *IEEE Transactions on Image Processing*, 2024.

[14] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2339–2346, 2013.

[15] Zengxi Zhang, Zhiying Jiang, Long Ma, Jinyuan Liu, Xin Fan, and Risheng Liu. Hupe: Heuristic underwater perceptual enhancement with semantic collaborative learning. *International Journal of Computer Vision*, pages 1–19, 2025.

[16] 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 Journal of Automatica Sinica*, 2024.

[17] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025.

[18] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025.

[19] Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1163, 2015.

[20] Yu-Sheng Chen and Yung-Yu Chuang. Natural image stitching with the global similarity prior. In *European conference on computer vision*, pages 186–201. Springer, 2016.

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

[22] Jing Li, Baosong Deng, Rongfu Tang, Zhengming Wang, and Ye Yan. Local-adaptive image alignment based on triangular facet approximation. *IEEE Transactions on Image Processing*, 29:2356–2369, 2019.

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

[24] Erman Nghonda Tchinda, Maximillian Kealoha Panoff, Danielle Tchuinkou Kwadjo, and Christophe Bobda. Semi-supervised image stitching from unstructured camera arrays. *Sensors*, 23(23):9481, 2023.

[25] Kaimo Lin, Nianjuan Jiang, Loong-Fah Cheong, Minh Do, editor="Leibe Bastian Lu, Jiangbo", Jiri Matas, Nicu Sebe, and Max Welling. Seagull: Seam-guided local alignment for parallax-tolerant image stitching. In *Computer Vision – ECCV 2016*, pages 370–385, 2016.

[26] Hyeokjun Kweon, Hyeonseong Kim, Yoonsu Kang, Youngho Yoon, Wooseong Jeong, and Kuk-Jin Yoon. Pixel-wise warping for deep image stitching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1196–1204, 2023.

[27] Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao. Learning edge-preserved image stitching from multi-scale deep homography. *Neurocomputing*, 491:533–543, 2022.

[28] Zhiying Jiang, Zengxi Zhang, and Jinyuan Liu. Harmonized domain enabled alternate search for infrared and visible image alignment. *IEEE Transactions on Image Processing*, 2025.

[29] Zeru Shi, Zengxi Zhang, Kemeng Cui, Ruizhe An, Jinyuan Liu, and Zhiying Jiang. Sefenet: Robust deep homography estimation via semantic-driven feature enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[30] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17235–17244, 2023.

[31] Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9833–9842, 2023.

[32] ML Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[33] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[35] Jiaxi Zeng, Chengtang Yao, Lidong Yu, Yuwei Wu, and Yunde Jia. Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18347–18357, 2023.

[36] Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. Exploring image enhancement for salient object detection in low light images. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(1s):1–19, 2021.

[37] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1792–1800, 2022.

[38] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023.

[39] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE transactions on image processing*, 21(4):1756–1769, 2011.

[40] Pierre Bouguer. *Essai d'optique sur la gradation de la lumière*. Claude Jombert, 1729.

[41] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[42] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018.

[43] A. Polesel, G. Ramponi, and V.J. Mathews. Image enhancement via adaptive unsharp masking. *IEEE Transactions on Image Processing*, 9(3):505–510, 2000.

[44] Olaf Ronneberger, Philipp Fischer, editor="Navab Nassir Brox, Thomas", Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.

[45] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 653–669. Springer, 2020.

[46] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

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

[48] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023.

[49] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2019.

[50] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[51] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *International journal of computer vision*, 48:233–254, 2002.

[52] Guijing Zhu, Long Ma, Xin Fan, and Risheng Liu. Hierarchical bilevel learning with architecture and loss search for hadamard-based image restoration. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1757–1764, 7 2022.

[53] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12918–12927, October 2023.

[54] Zengxi Zhang, Zhiying Jiang, Jinyuan Liu, Xin Fan, and Risheng Liu. Waterflow: Heuristic normalizing flow for underwater image enhancement and beyond. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7314–7323, 2023.

[55] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[56] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2021.

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

[58] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[59] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1879–1888, June 2022.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the direction and specific contributions of the proposed work.

   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: The proposed method does not yet take into account all harsh environments, which has been explained in the manuscript.

   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: [Yes]

   Justification: Each experiment provided in the paper is evaluated under a fair experimental setting. At the same time, all theorems and formulas have strict numbering and logical order.

   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: The detailed network structure of the paper has been fully described in the main text and supplementary materials, and the experimental details including the collection sources of the provided datasets are also marked.

   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?

   Answer: [No]

   Justification: After the paper is accepted, we will release the code, model parameters and dataset of the proposed method.

   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: All training parameters, including hyperparameters, training frameworks, and iterator choices are described in the experimental details.

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: [No]

Justification: We provide detailed experimental data, but do not record detailed error bars.

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 provide clear instructions for the experimental setup used for training and testing.

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: This research complies with the ethical standards specified by NeurIPS

   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: Our work takes into account some blank areas in the current field, so it will have a certain positive impact on the overall development of the industry, which is explained in the article.

    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.
    - 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: There is no risk of the paper being misused.

    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 all content are appropriately acknowledged and the licenses and terms of use are clearly mentioned and strictly adhered to.

    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: [Yes]

    Justification: We provide detailed information of the proposed dataset, and the dataset will be made publicly available after the paper is accepted.

    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: Our research does not include any human-related experiments.

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: Our research does not include any human-related experiments.

    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.

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: The core method of this study does not involve LLM related technologies

    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.

# A Technical Appendices and Supplementary Material

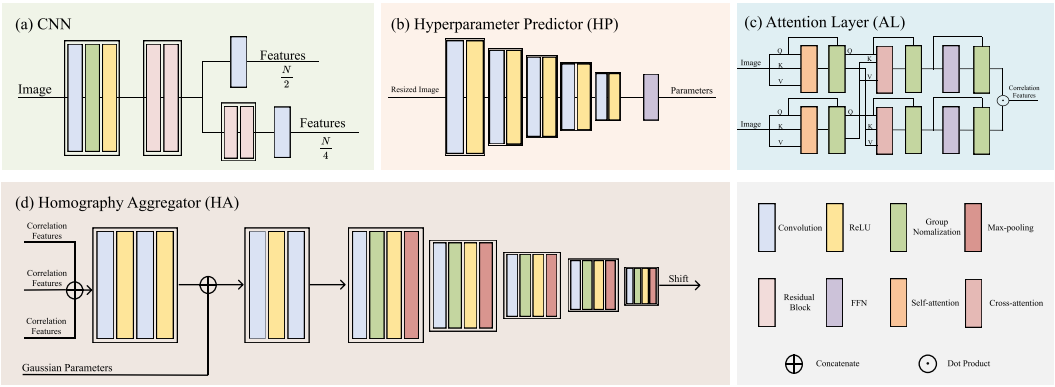



Figure 10: Detailed structure of CNN (a), Hyperparameter Predictor (b), Attention Layer (c) and Homography Aggregator (HA).

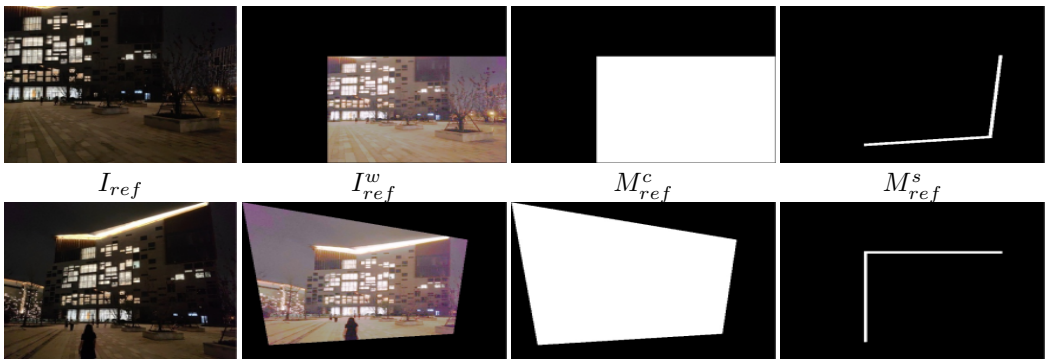



Figure 11: Visualization of the seam mask.

### A.1 Detailed Network Architecture

Fig. 10 demonstrates the detailed network structure of the proposed Pyramid-CNN (a), Hyperparameter Predictor (b), Attention Layer (c) and Homography Aggregator (d).

### A.2 Details in MSC

The generation process of the introduced seam mask $\{M^s_{ref}, M^s_{tar}\}$ can be formulated as follows:

$$\begin{aligned} \nabla M^c_{ref} &= \left|M^c_{ref,i,j} - M^c_{ref,i-1,j}\right| + \left|M^c_{ref,i,j} - M^c_{ref,i,j-1}\right|, \\ \nabla M^c_{tar} &= \left|M^c_{tar,i,j} - M^c_{tar,i-1,j}\right| + \left|M^c_{tar,i,j} - M^c_{tar,i,j-1}\right|, \end{aligned} \tag{12}$$

$$\begin{aligned} M^s_{ref} &= \mathcal{C}(\mathcal{E}(\mathcal{E}(\mathcal{E}(\mathcal{E}(\nabla M^c_{tar}))))) \odot M^c_{ref}, \\ M^s_{tar} &= \mathcal{C}(\mathcal{E}(\mathcal{E}(\mathcal{E}(\mathcal{E}(\nabla M^c_{ref}))))) \odot M^c_{tar}, \end{aligned} \tag{13}$$

where $i, j$ are index of the Cartesian coordinates. $\{M^c_{ref}, M^c_{tar}\}$ are the content mask, which replace the original pixel of image into all-in-one matrix. $\mathcal{E}$ is the convolution layer with 3×3 SOBEL filters. $\mathcal{C}$ clip image to 0-1. The visualization of the $\{M^c_{ref}, M^c_{tar}, M^s_{ref}, M^s_{tar}\}$ is also shown in Fig. 11.

### A.3 More Demonstration on ASIS Dataset

We additionally provide examples of the dataset in different scenarios under adverse environments, which is demonsterated in Fig. 12. It is worth mentioning that the low-light scene image pairs are

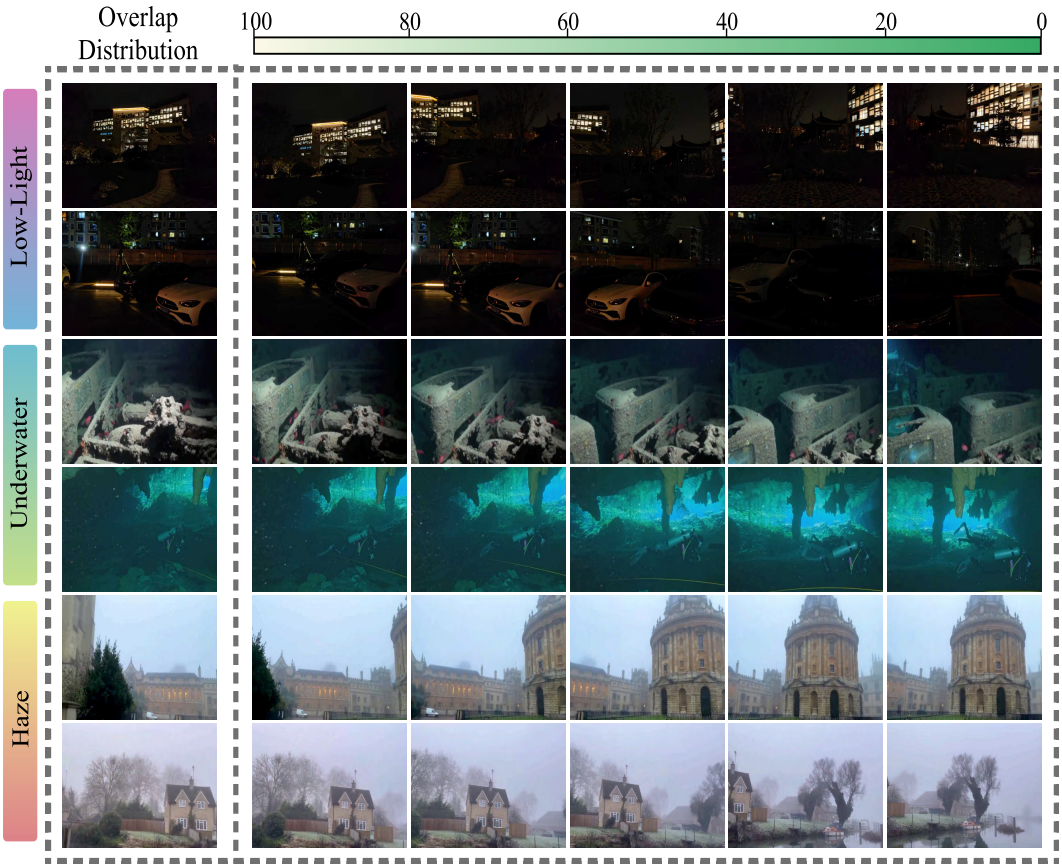



Figure 12: Examples of image pairs under three adverse environment in order of baseline from small to large.

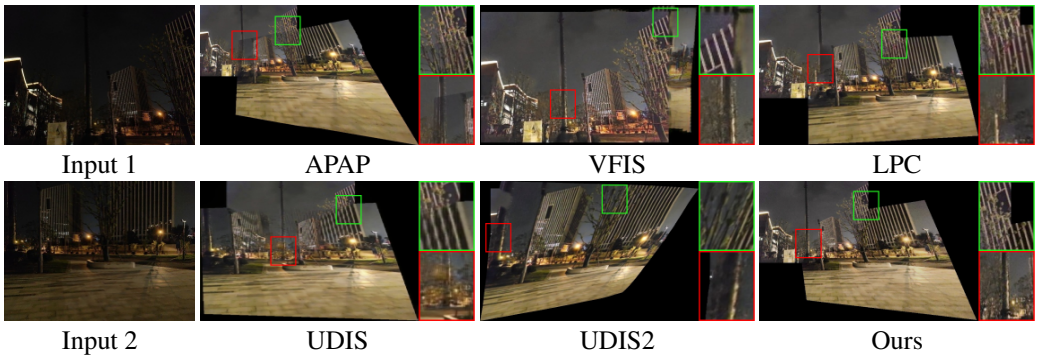



Figure 13: Visualization results on the low-light environment.

collected manually, and the image pairs of the underwater and haze environment are collected from the Internet [2]

[2] https://www.youtube.com/watch?v=93bWdgI69To
https://www.youtube.com/watch?v=dBMrRJWrFEU
https://www.youtube.com/watch?v=G5Mr3NuHjaI
https://www.youtube.com/watch?v=D1KsEOUqCEU
https://www.youtube.com/watch?v=yCoNJHqWYnU&list=RDyCoNJHqWYnU&start_radio=1&t=73s

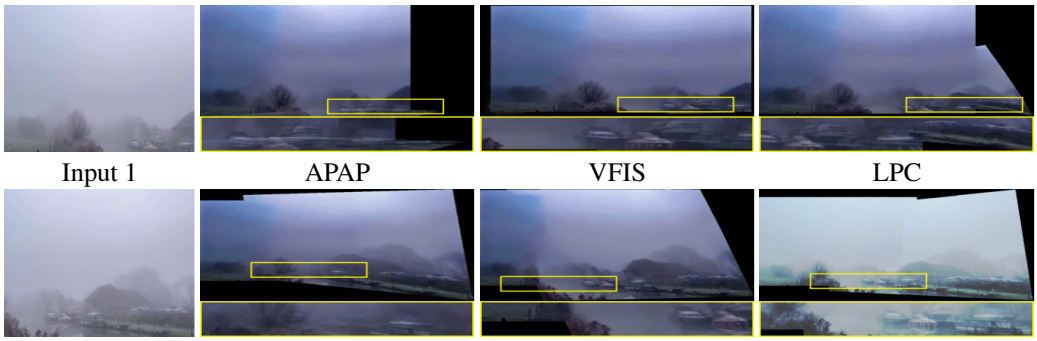


Figure 14: Visualization results on the haze environment.

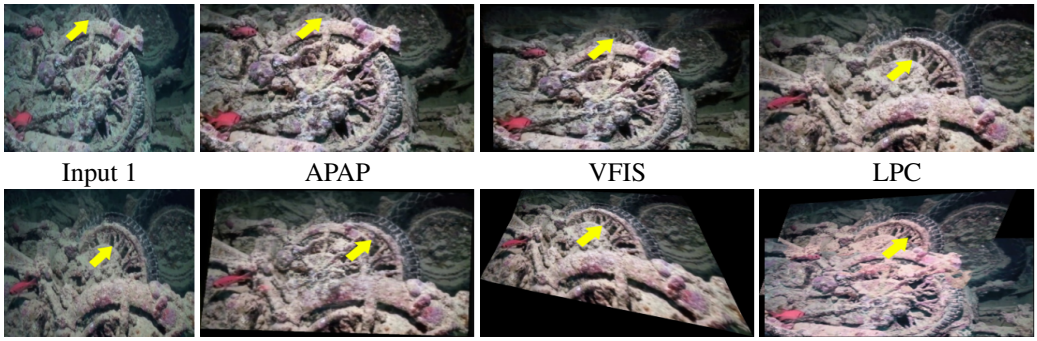


Figure 15: Visualization results on the underwater environment.

### A.4 More Visual Comparisons

We additionally provide examples for evaluating the performance in low-light,haze and underwater environment. We use the unified restortation framework HBALALS [52] combined with APAP [14], VFIS [8], LPC [56], UDIS [9] and UDIS2 [10] to compare with our method. The qualitative results are shown in Fig. 13, Fig. 14 and Fig. 15.