Implicit Neural Image Stitching
With Enhanced and Blended Feature Reconstruction

Minsu Kim\textsuperscript{1}, Jaewon Lee\textsuperscript{1}, Byeonghun Lee\textsuperscript{2}, Sunghoon Im\textsuperscript{1}, and Kyong Hwan Jin\textsuperscript{2}\textsuperscript{*}
\textsuperscript{1}DGIST, Republic of Korea  \textsuperscript{2}Korea University, Republic of Korea
\textsuperscript{1}\{axin.kim, ljw3136, sunghoonim\}@dgist.ac.kr  \textsuperscript{2}\{byeonghun.lee, kyong.jin\}@korea.ac.kr

Abstract

Existing frameworks for image stitching often provide visually reasonable stitchings. However, they suffer from blurry artifacts and disparities in illumination, depth level, etc. Although the recent learning-based stitchings relax such disparities, the required methods impose sacrifice of image qualities failing to capture high-frequency details for stitched images. To address the problem, we propose a novel approach, implicit Neural Image Stitching (NIS) that extends arbitrary-scale super-resolution. Our method estimates Fourier coefficients of images for quality-enhancing warps. Then, the suggested model blends color mismatches and misalignment in the latent space and decodes the features into RGB values of stitched images. Our experiments show that our approach achieves improvement in resolving the low-definition imaging of the previous deep image stitching with favorable accelerated image-enhancing methods. Our source code is available at https://github.com/minshu-kim/NIS.

1. Introduction

Image stitching aims to generate a wider field-of-view panorama from multiple scenes with arbitrary views. They provide rich visual information for various fields that require panoramic images including autonomous driving, virtual reality, and medical imaging.

Depending on the existence of a fixed grid transformation, image stitching is categorized as view-fixed [17,22,44] or view-free methods [10,21,26,33,48]. Previous view-free stitchings [10,21,26,32–35,48] align multiple views without the priors of inter-relationship between given scenes. In contrast, view-fixed approaches [17,22,44] use pre-defined grid transformations to stitch different views. Among them, a trainable stitching method [33] with a neural network shows good qualities at capturing color mismatches and blending misalignment in latent space. Because the method is fast and automatic, it demonstrates its practicality for real-time applications like virtual reality (VR) [2,15], autonomous driving [17,44]. However, the existing methods struggle with low image qualities caused by large deformations and a lack of constraints for reasonable trade-offs between image quality and blending.

The recent successful demonstrations of arbitrary-scale super-resolution (SR) with implicit neural representation (INR) [5,20] shed light on restoring such damaged low-definition images. Because the warp of an image using a grid is equivalent to the arbitrary-scale up-and-down sampling, the warped images can be enhanced by an extension of the arbitrary-scale SR. Following the idea, we propose a novel approach, implicit neural image stitching (NIS), which enables enhanced image stitching.
With the assumption that aligning transformations are given, NIS predicts warped Fourier features with high-frequency details. Then, a CNN-based blender combines two aligned features into one latent variable in order to alleviate color mismatch and misalignment. Afterward, a decoder provides a representation of a stitched image over the extracted features. Our experiments show that the suggested INR restores local textures while keeping the estimation of a blended feature as the previous method [33].

In summary, our main contributions are as follows:

• We propose an implicit neural representation for image stitching that restores high-frequency details.
• We extend the concept of arbitrary-scale super-resolution into image stitching.
• Our model simplifies the image stitching pipeline performing various tasks into inference including warping paired images, blending misalignment and parallax errors, and relaxing blurred effects.

2. Related Work

Homography Estimation The feature-based approaches estimate homography using the direct linear transformation (DLT) [11] given feature correspondences [3, 27, 39]. Although those approaches can infer reasonable aligning transformations, they often fail to compute homography in challenging conditions where a pair of images is captured in different environments, such as day-to-night scenes and scenes with dynamic objects. To overcome the limitations, various deep homography estimators with superior feature extractors were suggested [4, 7, 8, 18, 31, 35, 50, 52, 53]. The first supervised and unsupervised deep homography estimators are proposed by Detone et al. and Nguyen et al. [7, 31], respectively. The proposed methods commonly estimate 4 corner displacements between two images. Then, using DLT algorithm and predicted displacements, they compute an aligning homography. Inspired by these works, various advanced methods are suggested, including variations of VGG-style networks [8, 18, 53], the moving content-aware model [50], extracting one-channel of Lucas-Kanade feature map [52], and iterative architectures for inferencing a homography [4, 35].

Implicit Neural Representation Implicit neural representation [12] approximates continuous signals such as 2D images and 3D shapes. Thanks to the property of INR, various tasks have been proposed including arbitrary image super-resolution (SR) [5, 20], SR for image warping [19, 40, 47], view synthesis [28], etc. Among them, local INR [5, 19, 20] uses both feature maps from a CNN encoder and relative coordinates showing the robustness in generalization to out-of-scale datasets. Lee et al. [19] proposed an INR architecture, Local Texture Estimator for Warping (LTEW). Compared to the previous work [40], it shows the robustness of the generalization performances by demonstrating the model under unseen grid transformations, including out-of-scale homographies, and Equirectangular projection (ERP). Motivated by their successful application of arbitrary-scale SR for image warping, we propose an implicit neural function for image stitching.

Image Blending Blending techniques combine the overlapping regions of semantically aligned images as naturally as possible. There is a number of blending methods, including gradient-domain smoothing of color (a.k.a. Poisson blending) [1, 9, 36, 42], alpha blending [37], multi-band blending [49], and deep blending [45, 51]. Among the approaches, the introduction of deep blending techniques paved the way for the previous learning-based image stitching [33].

Image Stitching There are two branches under image stitching, view-fixed [17, 22, 41, 44], and view-free tasks [10, 14, 21, 26, 32, 33, 35, 48]. View-fixed scheme stitches images with given fixed views which are free from estimating an aligning transformation. View-fixed image stitching is often applied for specific tasks like autonomous driving [17, 44] and surveillance videos [22] that use cameras with fixed locations. In contrast, view-free image stitching is applied to images with arbitrary views. It estimates geometric relations under dynamically distributed views. Then,
it aligns the inputs and merges them. Gao et al. [10] proposed a model that estimates a dual homography to align two globally dominant frames in given images. Lin et al. [26] proposed a model that estimates a dual homography to align two globally dominant frames in given images. Zaragoza et al. [48] suggested moving DLT to infer as-projective-as-possible (APAP) spatially weighted homography. Li et al. [21] computed thin plate spline (TPS) to finely optimize per-pixel deformations. Liao et al. [23] proposed single-perspective warps (SPW) and emphasizes image alignment using dual-feature (point + line) for structure-preserving image stitching. Jia et al. [13] suggested leveraging line-point-consistence (LPC) which preserves the geometric structures of given scenes. Recently, Nie et al. [33] proposed an unsupervised deep image stitching method (UDIS) which blends two warped images in latent space and decodes them to a stitched image. The approach shows favorable results in correcting illumination differences and relaxing parallax and misalignment. Inspired by such achievements, we propose a neural architecture for enhanced image stitching based on arbitrary-scale SR.

3. Problem Formulation

In our formulations, we define \( x \) and \( y \in \mathbb{R}^2 \) as an input frame and a warped frame coordinate, respectively. \( A[x] \) denotes the nearest neighbor interpolation for a signal \( A \) using a pixel coordinate. Given a reference \( (I_r) \) and a target image \( (I_t) \) where \( I_r \in \mathbb{R}^{h \times w} \), we formulate an implicit neural function as

\[
N_\theta : (y_r, y_t, I_r[x_r], I_t[x_t]) \mapsto (R, G, B). \tag{1}
\]

The coordinates are obtained by transformation estimators [4, 21] and NIS leverages them with input images. For detailed modularization of NIS, we decompose \( N_\theta \) as

\[
N_\theta = G_\theta \circ B_\eta \circ g, \tag{2}
\]

where \( G_\theta, B_\eta, \) and \( g \) are an implicit neural representation, a blender, and a displacement-dependent learnable warp, respectively. Our modular decomposition enables us to keep the property of the implicit neural representation that provides continuous RGB values of an implicitly stitched feature. Because the blending and reconstruction of high-frequency details are conflicting tasks, we suggest thorough strategies in Sec. 4.

3.1. Homography Estimation

As described in Sec. 5.2, we use a deep homography estimator IHN [4] to train NIS. Because IHN recursively updates displacement vectors in the pre-fixed number of iterations, we design an additional formulation for it. Specifically, we formulate an unsupervised training for IHN [4] as follows:

\[
\hat{D} = \arg \min_D \sum_{k=1}^{K} \alpha^{K-k} \cdot \left\| I_r - W(I_t; H_k) \right\|_1, \tag{3}
\]

where \( H_k = f_t(D_k, c) \), \( D = \sum_k D_k \), \( f_t \) denotes the Direct Linear Transform (DLT) [11]. \( W(A; B) \) means warping \( A \) using a homography \( B \). \( K \) is the total number of iterations in IHN [4]. \( c \) is a set of 4 corner coordinates of a target image. \( D_k \) and \( H_k \) are a k-th estimated displacement vector and a homography computed by Direct Linear Transform (DLT) [11]. \( \alpha \) is the weight of the objective function set as 0.85 in our implementation.

3.2. Implicit Neural Image Stitching

Estimation of Detail-aware Feature We suggest Neural Warping (NW, \( g \)) for the estimation of warped features with enhanced textures. To this end, NW uses a vector containing direction and distance from the nearest referenceable coordinate. Specifically, we use a relative coordinate \( (c_m \in \mathbb{R}^2) \) as a prior and a CNN-based filter that contains a learnable
After NIS prepares a pair of phase and frequency estimators, respectively. Inspired by previous work on super-resolution [5, 19, 20, 40], we design supervised learning using data synthesis methods [32]. We minimize L1 loss between ground truth and the estimated stitched images as

$$\mathcal{L}_1(\Theta) = \arg\min_{\Theta} \sum_{y_u} \| C[y_u] - \hat{C}[y_u] \|_1. \quad (9)$$

Learning Enhanced Details Inspired by previous work on super-resolution [5, 19, 20, 40], we design supervised learning using data synthesis methods [32]. We minimize L1 loss between ground truth and the estimated stitched images as

$$\mathcal{L}_1(\Theta) = \arg\min_{\Theta} \sum_{y_u} \| C[y_u] - \hat{C}[y_u] \|_1. \quad (9)$$

Learning Blended Features The concatenation of two warped features causes inevitable differences between the distribution of overlapped and non-overlapped regions. The blending layer ($B_{\eta}$) has to be trained to combine the two latent spaces into a single space correcting color mismatches and hiding parallax errors. To this end, the blender is constrained with a photometric seam loss [33] as follows:

$$\mathcal{L}_{seam}(\Theta) = \arg\min_{\Theta} \sum_{n} \| \hat{S}_n - S_n \|_1, \quad (10)$$

where $\hat{S}_n = M_n \odot \hat{C}$,

$$S_n = M_n \odot N_{\Theta}(y_{rt}, y_t, I_n, \bar{0}), \quad n \in \{r,t\},$$

$M_{\cdot}()$ is a mask of seam regions between the target and reference. $\bar{0} \in \mathbb{R}^{h \times w}$ is a blank image with the same size as an input image $I_n$ domain. Overall procedures are provided in Fig. 4.
### 4.2. Training Details

**Enhancement** We apply an on-the-fly data generation that warps a mini-batch with the same homography. It provides synthesized paired images and ground truth stitched images per a mini-batch. To generate synthetic datasets, we follow the whole pipeline of [32]. We set the ratio of 4 corner offsets ($\Delta x, \Delta y$) as random values within 25% of cropped image resolutions $h_c, w_c$. For training, we randomly sample M spatial query points from valid regions and minimize the $L_1$ loss between RGB prediction and ground truth as in [19]. This augmentation strategy is as follows:

$$ (\Delta x, \Delta y) = (a_x \cdot w_c, a_y \cdot h_c) \in \mathbb{R}^2, \quad (11) $$

where $a_l \sim \mathcal{U}(0, 0.25)$, $x, y \in l$.

**Blending** In this stage, NIS learns to correct color mismatches and misalignments. Because there is no given transformation prior in the real images, we use a homography estimator IHN [4], which was trained with the suggested unsupervised training methods [33]. Our network is trained only with the queries extracted from the seam regions of two images. We freeze NIS except for the decoding INR and fine-tune it as a representation provider for blended stitched images. Because no ground truth or reference exists for predicted seam region RGBs, we generate warped target and reference images to use them as reference images. By forwarding an image (target or reference image) and a blank image into a frozen, scratch-trained NIS, we get the warped images and use them as references for predicted samples. The elements of a blank image are set to 0.

### 5. Experiment

#### 5.1. Dataset

We use two datasets: MS-COCO [25] and UDIS-D [33]. The first stage uses synthetic MS-COCO which is free from parallax errors. In the second stage, we use UDIS-D that includes various degrees of parallax errors.

#### 5.2. Implementation Details

**Estimation of Alignment** We employ a deep homography estimator IHN [4] and robust ELA [21] for estimation of transformation to align images. We train a 2-scale IHN on UDIS-D and then use the estimated transformation to train NIS in our experiments. To check the performance of the trained model for unseen elastic warps, we use robust ELA to obtain aligning grids.

**Neural Image Stitching** The blender ($B_p$) and the encoder ($E_p$) of the neural warping use EDSR [24] without upsampling module. The decoder ($G_p$) is a 4-layer MLP with ReLUs, whose hidden dimension is 256. The amplitude, frequency, and phase estimators in neural warping are implemented with a single convolution layer without activation function. The amplitude and frequency estimators use a 256-channel $3 \times 3$ convolution layer and the phase estimation layer uses a 128-channel $1 \times 1$ convolution layer.

#### 5.3. Evaluation

**Quantitative result** In Tab. 1a, we evaluate stitched image qualities on the synthetic MS-COCO dataset under synthe-
sized ground-truth alignment. We find that the image stitching of UDIS negatively affects image quality as reported in Tab. 1a. Although the method employs bilinear warps, the UNet-based architectures cause such harmed imaging with jagging and blurred artifacts. However, the table validates that NIS successfully resolves the problem with 2.44, and 3.91 mPSNR gains compared to the bicubic warp, and bilinear warp, respectively. Tab. 1b shows the summarized performances of image stitching quality. Since there is no ground truth for UDIS-D real dataset, we report NIQE [30], PIQE [43], and BRISQUE [29]. We denote the fine-tuned NIS with (F). As shown in the table, the fine-tuned NIS that is capable of correcting color mismatches and parallax relaxing shows superior performances compared to the model without the training of feature blending. In addition, the comparison between UDIS and ‘LPC + Graph Cut’ indicates that the learnable image stitching contributes the better image quality. Furthermore, the performance gains from a comparison of ‘IHN+NIS’ to UDIS implies that our method for image stitching achieves significant image en-
Qualitative result Qualitative comparisons on UDIS-D real dataset are shown in Fig. 5. We compare NIS with the feature-based [48], [21] approaches, and a learning-based approach [33]. APAP and robust ELA show parallax-tolerant image warping while UDIS and NIS blend parallax errors correcting the illumination difference. The qualitative comparisons of image qualities between UDIS and ours are provided in Fig. 6. The comparison supports the performance gains in Tab. 1a indicating that our method captures high-frequency details while keeping the image blending capability.

Inference under Elastic warp Our model is trained under rigid transformations (or homography). To check if this training configuration can limit model performances for elastically transformed grids, we explore an experiment as in Fig. 7. As shown in the figure, we demonstrate an image aligned by a Thin-plate Spline grid and stitched by NIS. We use robust ELA [21] to estimate the elastic grid and generate the stitched image using NIS. As shown in the figure, our implicit neural representation for image stitching recovers high-frequency details on the totally unseen aligning grids.

Cost Effectiveness In Tab. 2, we compare the specifications of stitching methods for 3 different resolutions. To evaluate the model, we forward a common pair of images with 100 iterations, repeatedly. Then, we report max memory consumption (GB) and the average computation time (ms). As shown in the table, while NIS is cost-efficient for 192\(^2\) and 784\(^2\) sizes but, it shows weakness on high-resolution images 1536\(^2\) compared to the UDIS.

5.4. Ablation Study

Fourier Features To explore the contribution of Fourier features for image stitching on synthetic MS-COCO, we revisit the models with 3 configurations: 1) removal of amplitude, 2) frequency, and 3) phase estimator, respectively. We train all the cases with the first stage to clarify model performances on image enhancements. Note that the experiment using NIS without blending training shows different scale mPSNR as the model is fine-tuned on the other dataset, UDIS-D. In Tab. 3, ‘w/o Amp.’ removes the amplitude estimator by setting all the amplitudes of frequencies as 1. ‘w/o Freq.’ estimates 128 frequencies whose size corresponds to half of the frequency estimator in NIS. The ‘w/o Phase’ removes the phase estimator. As shown in Tab. 3, the ‘w/o Amp.’ model shows damaged Fourier features with significant performance degradation. The ‘w/o Freq.’ model shows the reduced number of samples from the Fourier distribution. The result provided in the table emphasizes the importance of frequency priors. The ‘w/o Phase’ model keeps the number of Fourier samples but shows a worse Fourier distribution. The observation has a negative effect as shown in the table.

Learning Strategy We investigate the contributions of our training methods in Fig. 8. The first row contains learning procedures for enhanced feature reconstruction. The second row provides the observations during the fine-tuning stage. As in the figure, our learning methods are helpful for capturing high-frequency details and correcting color mismatches and misalignment. Despite the estimation of blended signals that may cause a worse image quality, we notice that our learning strategy for blended feature reconstruction provides visually pleasing stitched images.

6. Discussion

Efficiency of Neural Warping LTEW’s pipeline for local INR prevents over/undershoot and blocky artifacts for more than \(\times 8\) upscale factors. In contrast, our pipeline leverages displacement vector (\(c_m\)) as a prior for the representation of
stitched images with accelerated computations. In Tab. 4, we compare Neural Warping (NW) with LTEW, which is the state-of-the-art method for high-definition warp to verify our method. Following the recent research of arbitrary-scale super-resolution for Equirectangular projection (ERP SR), we employ the same evaluation configurations as previous works [6, 47]. The measurement of computation time is conducted under a fixed resolution to focus on complexity checking. After we train Neural Warping with $G_θ$ and LTE on ODI-SR $×4$ scale dataset, we test the models on ODI-SR test dataset [6] and SUN 360 [46] dataset. As shown in Tab. 4, we see that our method shows favorable performances on both computational complexity and super-resolution.

**NIS with Seam cutting** In Fig. 9, we demonstrate a comparison of multi-view stitching to explore NIS with seam cutting. We prepare seam masks using OpenCV ‘DpSeamFinder’. For image reconstruction with NIS, we determine a reference image $I_r$ and two target images $I_t1, I_t2$. After that, using seam masks, we obtain two latent variables $C'_1, C'_2$ from $z_r, z_{t1}$ and $z_r, z_{t2}$, respectively. Then NIS estimates the stitched image by blending $C'_1, C'_2$ and decoding the output. As shown in Fig. 9, The enhancement of NIS demonstrates the potential to be used with seam cutting for advanced blending.

### 7. Conclusion

We proposed NIS, a novel end-to-end implicit neural reconstructor for image stitching. Our model predicted the dominant frequencies in warped domains from a pair of images to represent high-frequency details in a stitched image. Furthermore, our method successfully blended color mismatches and misalignments relaxing parallax errors. Our framework shows significant gains in synthetically stitched datasets over traditional methods for stitching, including bilinear and bicubic warps. Qualitative results on real images support that we successfully achieved high-frequency details for view-free image stitching compared to existing image stitching methods. On the other hand, we fail to simplify the training pipeline for neural image stitching. To design full end-to-end deep image stitching from image alignment to reconstruction, additional strategies for training would be required.

**Acknowledgment** This work was partly supported by the grants of the grant of the DGIST R&D program of the Ministry of Science and ICT of KOREA (22-KUJoint-02), Smart HealthCare Program(www.kipot.or.kr) funded by the Korean National Police Agency(KNPA) (No. 230222M01), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).
References


[48] 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. 1, 2, 3, 5, 6, 7


