Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction

Minsu Kim[1], Jaewon Lee[1], Byeonghun Lee[2], Sunghoon Im[1], and Kyong Hwan Jin[2*]
[1]DGIST, Republic of Korea [2]Korea University, Republic of Korea
[1]{axin.kim, ljw3136, sunghoonim}@dgist.ac.kr [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.

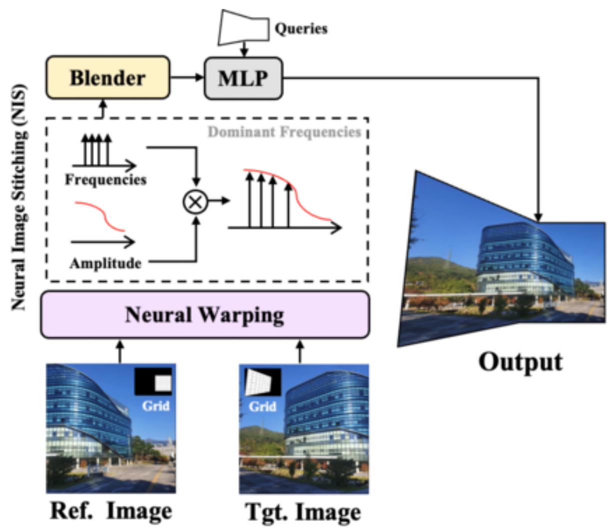


Figure 1. **Overview of our proposed image stitching.** Our model, NIS, estimates stitched images given a pair of target & reference images and transformed grids. Neural warping extracts high-frequency-aware 2D features and aligns them according to the given grids. Blender takes the warped features and merges them into a feature map. A decoder (MLP) predicts an RGB signal at a coordinate (x, y) of stitched image domain from a blended feature.

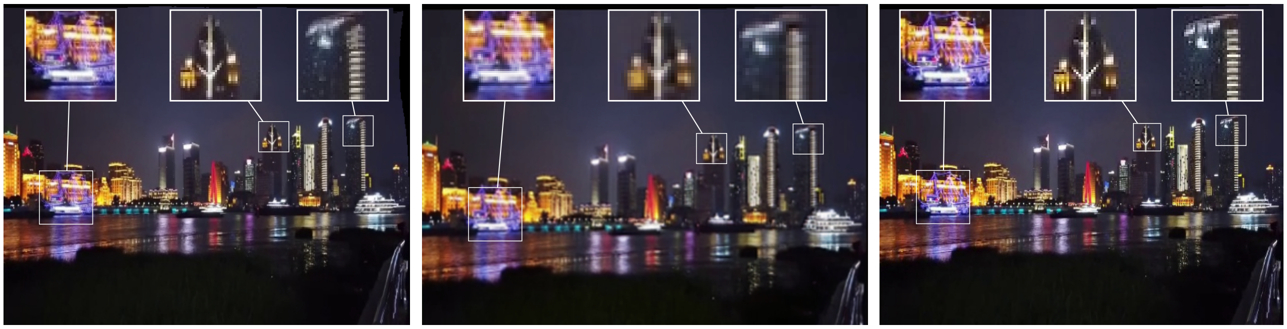
1. Introduction

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.

*Corresponding author.



APAP [48] UDIS [33] *Ours*

Figure 2. **Visual Demonstration**. two-view image stitching with single perspective Homography.

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 combines 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,

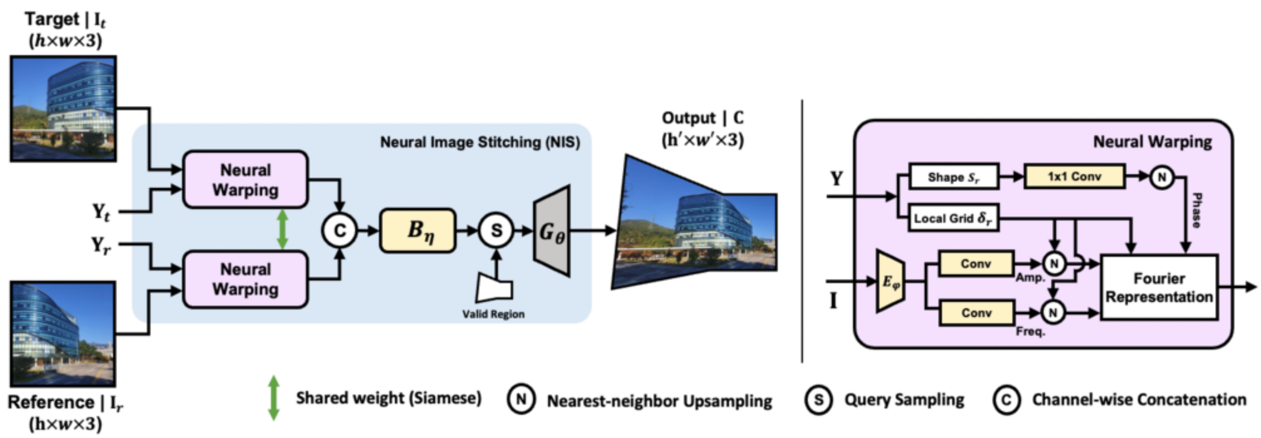


Figure 3. **Overall Flowchart of our Neural Image Stitching framework (NIS).** Our NIS consists of a neural warping, a blender ($\boldsymbol{B}_\eta$), and a decoder ($\boldsymbol{G}_\theta$). We employ a homography estimator, IHN [4], for image alignment.

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 method of estimating smoothly varying affine fields. Zaragoza *et al.* [48] suggested moving DLT to infer as-projective-as-possible (APAP) spatially weighted homography. Li *et al.* [21] computed *thin plate splinte* (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 $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^2$ as an input frame and a warped frame coordinate, respectively. $\mathbf{A}[\mathbf{x}]$ denotes the nearest neighbor interpolation for a signal $\mathbf{A}$ using a pixel coordinate. Given a reference ($\mathbf{I}_r$) and a target image ($\mathbf{I}_t$) where $\mathbf{I}_. \in \mathbb{R}^{h\times w}$, we formulate an implicit neural function as

$$\mathcal{N}_\Theta : (\mathbf{y}_r, \mathbf{y}_t, \mathbf{I}_r[\mathbf{x}_r], \mathbf{I}_t[\mathbf{x}_t]) \mapsto (R, G, B). \tag{1}$$

The coordinates are obtained by transformation estimators [4, 21] and NIS leverages them with input images. For detailed modularization of NIS, we decompose $\mathcal{N}_\Theta$ as

$$\mathcal{N}_\Theta = \mathbf{G}_\theta \circ \mathbf{B}_\eta \circ \boldsymbol{g}, \tag{2}$$

where $\mathbf{G}_\theta, \mathbf{B}_\eta$, and $\boldsymbol{g}$ 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:

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}} \sum_{k=1}^{K} \alpha^{K-k} \cdot \|\mathbf{I}_r - W(\mathbf{I}_t; \mathbf{H}_k)\|_1, \tag{3}$$
$$\text{where} \quad \mathbf{H}_k = f_t(\mathbf{D}_k, \mathbf{c}), \;\; \mathbf{D} = \sum_k \mathbf{D}_k,$$

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]. $\mathbf{c}$ is a set of 4 corner coordinates of a target image. $\mathbf{D}_k$ and $\mathbf{H}_k$ are a k-th estimated displacement vector and a homography computed by Direct Linear Transform (DLT) [11]. α 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, $\boldsymbol{g}$) 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 ($\mathbf{c}_m \in \mathbb{R}^2$) as a prior and a CNN-based filter that contains a learnable

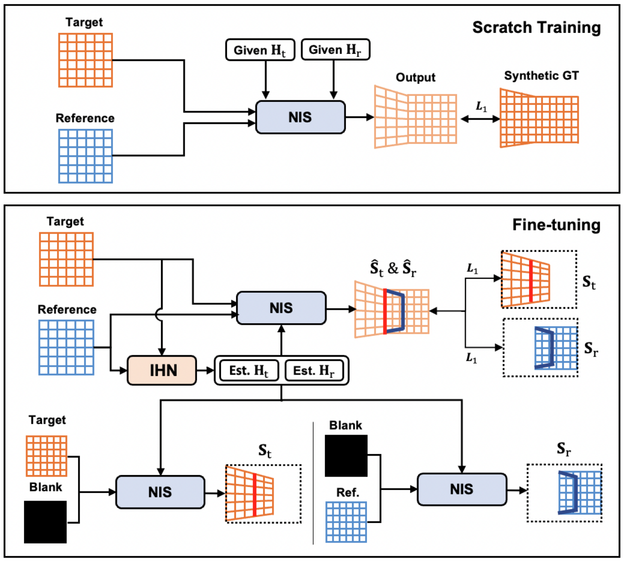


Figure 4. **Training Strategy.** In scratch training, we synthesize the aligning transformations and a stitched GT. In fine-tuning, we use IHN to estimate a homography $\mathbf{H}_t$ and train NIS in an unsupervised manner. Because there are no reference seam regions for $\hat{\mathbf{S}}_t$ and $\hat{\mathbf{S}}_r$, we estimate them with scratch-trained NIS.

warping module restoring local textures in a warped feature $\mathbf{z} \in \mathbb{R}^{h' \times w' \times 4C}$ as

$$\mathbf{z}[\mathbf{y}] = \boldsymbol{g}(\mathbf{E}_\varphi(\mathbf{I}^{IN})[\mathbf{x}], \mathbf{c}_m), \quad \text{where} \;\; \mathbf{c}_m = \mathbf{y} - \mathbf{y}', \;\; \mathbf{y}' = \mathbf{X}[\mathbf{y}], \tag{4}$$

$\mathbf{E}_\varphi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^C$ is a CNN encoder [24]. $\mathbf{X} := \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^2\}$ is a grid of an input image $\mathbf{I}^{IN} \in \mathbb{R}^{h \times w}$ frame.
Fourier Coefficients for Stitching INR Inspired by the successful demonstration of LTE [20] for resolving spectral bias [38], NW leverages Fourier coefficients given a feature $\mathbf{z}' = \mathbf{E}_\varphi(\mathbf{I}) \in \mathbb{R}^{h \times w \times C}$ and a cell $\mathbf{s} \in \mathbb{R}^{h' \times w' \times 10}$ that contains numerical Jacobian and Hessian tensors [19] to predict the density distribution of a warped grid (or a cell decoding) as

$$\mathbf{A} = \boldsymbol{g}_a(\mathbf{z}'), \;\; \mathbf{F} = \boldsymbol{g}_f(\mathbf{z}'), \;\; \mathbf{P} = \boldsymbol{g}_p(\mathbf{s}; \mathbf{y}), \tag{5}$$

where $\boldsymbol{g}_a(\cdot)$ and $\boldsymbol{g}_f(\cdot) : \mathbb{R}^C \mapsto \mathbb{R}^{4C}$ are an amplitude and a frequency estimator, respectively. $\boldsymbol{g}_p(\cdot) : \mathbb{R}^{10} \mapsto \mathbb{R}^{2C}$ is a phase estimator. After the estimation of the coefficients, a detail-aware feature ($\mathbf{z}$) that represents a high-dimensional stitched image is associated as

$$\mathbf{z}[\mathbf{y}] = \mathbf{A}[\mathbf{y}] \cdot (\cos\; \mathcal{F} + \sin\; \mathcal{F}), \quad \text{where} \;\; \mathcal{F} = \pi(\langle \mathbf{F}[\mathbf{y}], \mathbf{c}_m \rangle + \mathbf{P}[\mathbf{y}]). \tag{6}$$

Implicit Neural Representation After NIS prepares a pair of features of a target ($\mathbf{z}_t$) and a reference ($\mathbf{z}_r$) images, it stitches them in a merged signal $\mathbf{C}' \in \mathbb{R}^{h' \times w' \in C}$. To this end, we design a blender $\mathbf{B}_\eta(\cdot, \cdot) : (\mathbb{R}^{4C}, \mathbb{R}^{4C}) \mapsto \mathbb{R}^C$ that learns a high-dimensional representation of a seamlessly stitched image as

$$\mathbf{C}'(\mathbf{y}) = \mathbf{B}_\eta(\mathbf{z}_t, \mathbf{z}_r). \tag{7}$$

Then a decoding INR $\mathbf{G}_\theta(\cdot) : \mathbb{R}^C \mapsto \mathbb{R}^3$ takes a latent vector and provides an RGB of a stitched image ($\widehat{\mathbf{C}}$) as

$$\widehat{\mathbf{C}}[\mathbf{y}_u] = \boldsymbol{G}_{\boldsymbol{\theta}}(\mathbf{C}', \mathbf{y}_u), \tag{8}$$

where $\mathbf{y}_u \in \mathbf{Y}_u \subset \mathbf{U}$ denotes a valid region coordinate, $\mathbf{U} := [0, h) \times [0, w)$ is a uniform grid.

4. Method

NIS consists of a neural warping ($\boldsymbol{g}$), blender ($\mathbf{B}_\eta$), and a decoding INR ($\mathbf{G}_\theta$). The model is an end-to-end learnable model and infers a stitched image, automatically. This section provides detailed methods for training.

4.1. Training Strategy

We suggest 2 stage training to make the model strictly focus on each task (i.e., enhancing and blending). NIS learns high-frequency details in the first stage. After then, the model learns blended features in the second stage.
Learning Enhanced Details Inspired by previous arbitrary-scale super-resolution [5, 19, 20, 40], we design supervised learning using data synthesize methods [32]. We minimize L1 loss between ground truth and the estimated stitched images as

$$\mathcal{L}_1(\mathbf{\Theta}) = \arg\min_{\mathbf{\Theta}} \sum_{\mathbf{y}_u} \|\mathbf{C}[\mathbf{y}_u] - \widehat{\mathbf{C}}[\mathbf{y}_u]\|_1. \tag{9}$$

Learning Blended Features The concatenation of two warped features causes inevitable differences between the distribution of overlapped and non-overlapped regions. The blending layer ($\mathbf{B}_\eta$) 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:

$$\mathcal{L}_{seam}(\mathbf{\Theta}) = \arg\min_{\mathbf{\Theta}} \sum_n \|\hat{\mathbf{S}}_n - \mathbf{S}_n\|_1, \tag{10}$$
$$\text{where} \;\; \hat{\mathbf{S}}_n = \mathbf{M}_n \odot \widehat{\mathbf{C}},$$
$$\mathbf{S}_n = \mathbf{M}_n \odot \mathcal{N}_\Theta(\mathbf{y}_r, \mathbf{y}_t, \mathbf{I}_n, \vec{\mathbf{0}}), \;\; n \in \{r, t\},$$

$\mathbf{M}_{(\cdot)}$ is a mask of seam regions between the target and reference. $\vec{\mathbf{0}} \in \mathbb{R}^{h \times w}$ is a blank image with the same size as an input image $\mathbf{I}_n$ domain. Overall procedures are provided in Fig. 4.

Method	mPSNR (↑)	mSSIM (↑)	#Params.
Bilinear	34.78	0.96	-
Bicubic	36.25	0.97	-
UDIS [33]	33.45	0.97	8.0M
NIS (*ours*)	38.69	0.98	3.2M

(a) Evaluation on Synthetic Images.

Method	NIQE (↓)	PIQE (↓)	BRISQUE (↓)
APAP [48]	3.30	46.95	34.72
Robust ELA [21]	3.59	53.67	37.78
SPW [23]	3.39	51.68	36.84
LPC [13]	3.37	50.81	37.15
LPC + Graph Cut	3.50	50.63	37.14
UDIS [33]	3.43	50.01	36.71
IHN [4]+NIS	3.28	46.21	33.17
IHN [4]+NIS (F)	3.15	43.05	31.14

(b) Evaluation on Real Images.

Table 1. **Quantitative Comparison** on Stitching performance. Red and Blue colors indicate the best and the second-best performances, respectively.

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 (Δ_x, Δ_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:

$$(\Delta_x, \Delta_y) = (a_x \cdot w_c,\ a_y \cdot h_c) \in \mathbb{R}^2, \quad \text{where } a_l \sim \mathcal{U}(0, 0.25) \ \ x, y \in l. \tag{11}$$

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.

Method	Size		Mem.	Time	Resource
	Input	Stitched	(GB)	(ms)	
LPC [13]	128^2	192^2	-	109	AMD Ryzen
	512^2	784^2	-	4,883	4800H
	1024^2	1536^2	-	28,446	(CPU)
UDIS† [33]	128^2	192^2	9.1	74	NVIDIA
	512^2	784^2	9.3	128	RTX 3090
	1024^2	1536^2	9.7	253	(GPU)
NIS	128^2	192^2	0.5	58	NVIDIA
	512^2	784^2	5.4	384	RTX 3090
	1024^2	1536^2	17.3	1,944	(GPU)

Table 2. **Specifications of Stitching Methods.** † denotes a third-party implementation.

4.3. Training Configurations

In the first stage of training, we train NIS with a batch size of 20 and 250,000 iterations. We choose the crop size $h_c \times w_c$ as 48×48. In the second stage, we fine-tune the model with a single batch size. We set the number of iterations as 60,000 and 300,000, respectively. For the second stage, the resolution of input images is downsampled to 128×128 by bilinear downsampling. We use Adam optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rates are initialized as 1×10^{-4} for the first and second stages. The learning rates are exponentially decayed by a factor of 0.98 for every epoch.

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.

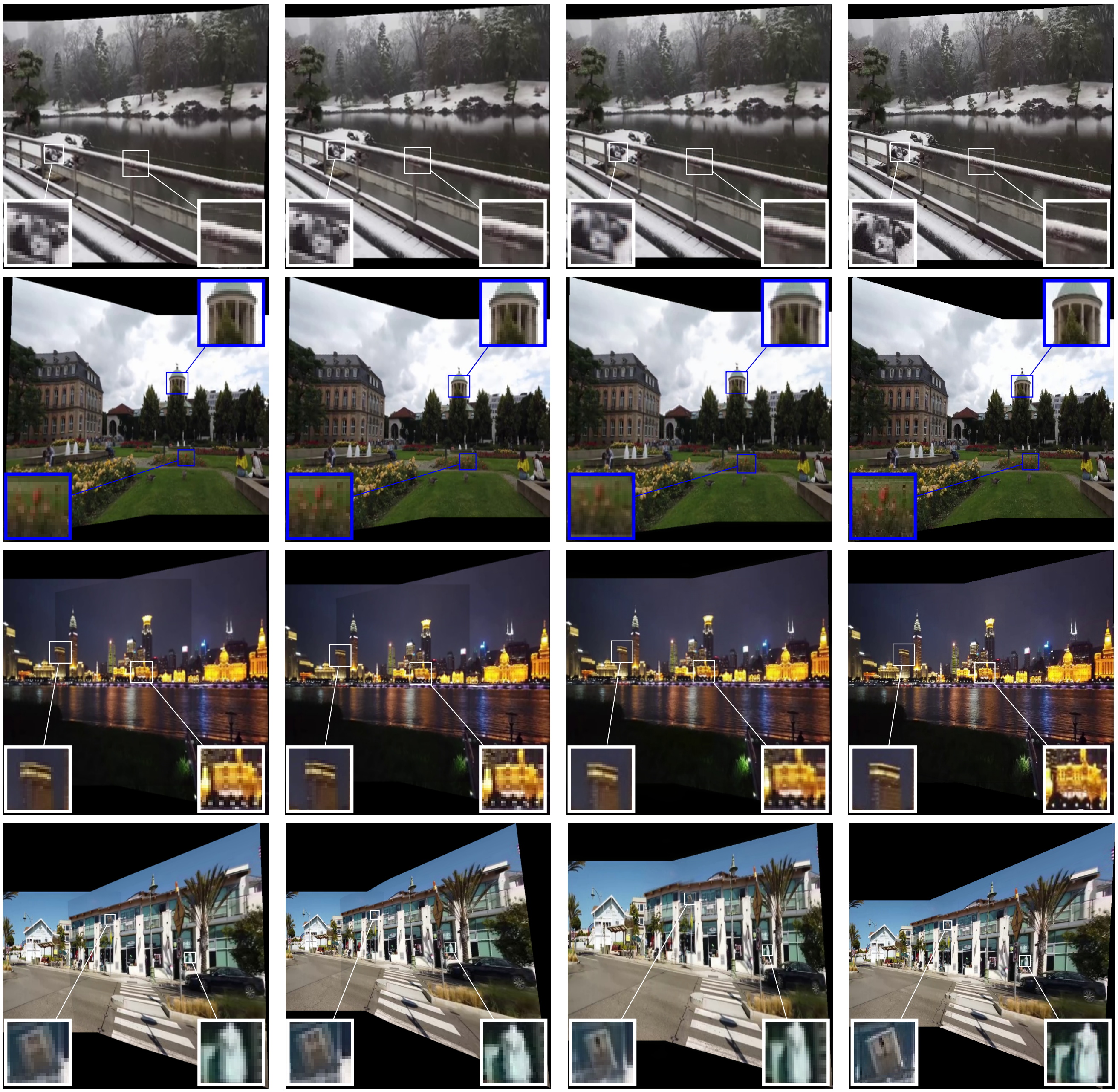
5.2. Implementation Details

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 $(\boldsymbol{B_\eta})$ and the encoder $(\boldsymbol{E_\varphi})$ of the neural warping use EDSR [24] without upscaling module. The decoder $(\boldsymbol{G_\theta})$ 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 3×3 convolution layer and the phase estimation layer uses a 128-channel 1×1 convolution layer.

5.3. Evaluation

Quantitative result In Tab. 1a, we evaluate stitched image qualities on the synthetic MS-COCO dataset under synthe-



Robust ELA [21] APAP [48] UDIS [33] NIS (*ours*)

Figure 5. Qualitative comparisons to other **image stitching approaches** on the UDIS-D real dataset [33].

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-

Figure 6. **Qualitative comparison.** On the diagonal, the left side is the stitched image of UDIS [33] and the right side is our image.

hancement.

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.

Figure 7. **Stitching with pre-computed Elastic Transformation grid.** Our pre-trained model is successfully applied on an elastically warped grid (unseen grid). We use robust ELA [21] to estimate the elastic grids.

Method	mPSNR (↑)	mSSIM (↑)
w/o Amp.	20.33	0.767
w/o Freq.	43.23	0.990
w/o Phase	44.07	0.990
NIS (*ours*)	**44.16**	**0.991**

Table 3. **Quantitative Ablation Study of NIS.**

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 ($\mathbf{c}_m$) as a prior for the representation of

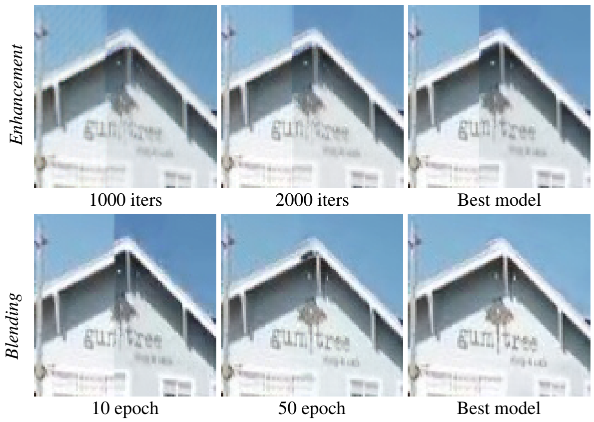


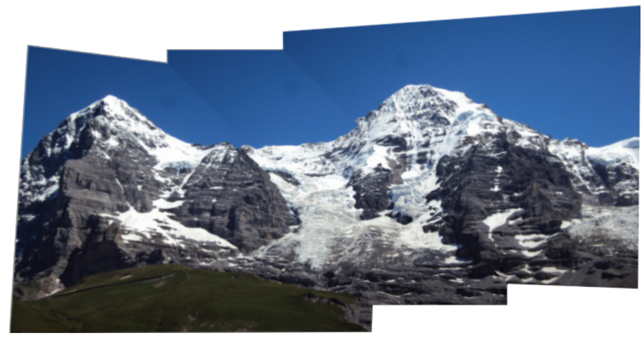
Figure 8. **Ablation study of our training method.**

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 $\mathbf{G}_\theta$ and LTE on ODI-SR $\times 4$ 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 $\mathbf{I}_r$ and two target images $\mathbf{I}_{t_1}, \mathbf{I}_{t_2}$. After that, using seam masks, we obtain two latent variables $\mathbf{C}'_1$ and $\mathbf{C}'_2$ from $\mathbf{z}_r$, $\mathbf{z}_{t_1}$ and $\mathbf{z}_r$, $\mathbf{z}_{t_2}$, respectively. Then NIS estimates the stitched image by blending $\mathbf{C}'_1, \mathbf{C}'_2$ 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



[Bilinear + Seam cutting] NIQE/PIQE/BRISQUE: 3.17/32.30/26.68

[NIS + Seam cutting] NIQE/PIQE/BRISQUE: 2.90/27.55/21.98

Figure 9. **Comparison on the stitching of Multiple images.**

Dataset	Method	WS-PSNR	
		$\times 8$	$\times 16$
ODI-SR	LTEW	25.53	23.91
	Neural Warping	**25.56**	**23.93**
SUN 360	LTEW	25.60	23.59
	Neural Warping	**25.62**	**23.60**

(a) Performance on ERP SR.

Size	Method	Memory (GB)	Time (ms)
256×256	LTEW	0.36	20.50
	Neural Warping	0.24 ($\downarrow$ **33.3%**)	5.37 ($\downarrow$ **73.8%**)

(b) Cost-efficiency.

Table 4. **Comparison of Neural Warping to LTEW.**

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

[1] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive digital photomontage. In *ACM SIGGRAPH 2004 Papers*, pages 294–302, 2004. 2

[2] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. Jump: virtual reality video. *ACM Transactions on Graphics (TOG)*, 35(6):1–13, 2016. 1

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[4] 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*, pages 1879–1888, 2022. 2, 3, 5

[5] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1, 2, 4

[6] Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9189–9198, 2021. 8

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 2

[8] Farzan Erlik Nowruzi, Robert Laganiere, and Nathalie Japkowicz. Homography estimation from image pairs with hierarchical convolutional networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 913–920, 2017. 2

[9] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 249–256, 2002. 2

[10] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *CVPR 2011*, pages 49–56. IEEE, 2011. 1, 2, 3

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3

[12] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 2

[13] 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. 3, 5

[14] 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. 2

[15] Hak Gu Kim, Heoun-Taek Lim, and Yong Man Ro. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):917–928, 2019. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Wei-Sheng Lai, Orazio Gallo, Jinwei Gu, Deqing Sun, Ming-Hsuan Yang, and Jan Kautz. Video stitching for linear camera arrays. *arXiv preprint arXiv:1907.13622*, 2019. 1, 2

[18] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7652–7661, 2020. 2

[19] Jaewon Lee, Kwang Pyo Choi, and Kyong Hwan Jin. Learning local implicit fourier representation for image warping. In *ECCV*, 2022. 2, 4, 5

[20] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1929–1938, 2022. 1, 2, 4

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

[22] Jia Li, Yifan Zhao, Weihua Ye, Kaiwen Yu, and Shiming Ge. Attentive deep stitching and quality assessment for 360° omnidirectional images. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):209–221, 2019. 1, 2

[23] Tianli Liao and Nan Li. Single-perspective warps in natural image stitching. *IEEE transactions on image processing*, 29:724–735, 2019. 3, 5

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 4, 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[26] 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. 1, 2, 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 6

[30] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[31] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. 2

[32] 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. 1, 2, 4, 5

[33] 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. 1, 2, 3, 4, 5, 6, 7

[34] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Learning thin-plate spline motion and seamless composition for parallax-tolerant unsupervised deep image stitching. *arXiv preprint arXiv:2302.08207*, 2023. 1

[35] Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao. Learning edge-preserved image stitching from large-baseline deep homography. *arXiv preprint arXiv:2012.06194*, 2020. 1, 2

[36] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318, 2003. 2

[37] Thomas Porter and Tom Duff. Compositing digital images. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 253–259, 1984. 2

[38] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 4

[39] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[40] Sanghyun Son and Kyoung Mu Lee. SRWarp: Generalized image super-resolution under arbitrary transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2021. 2, 4

[41] Dae-Young Song, Geonsoo Lee, HeeKyung Lee, Gi-Mun Um, and Donghyeon Cho. Weakly-supervised stitching network for real-world panoramic image generation. In *European Conference on Computer Vision*, pages 54–71. Springer, 2022. 2

[42] Matthew Uyttendaele, Ashley Eden, and Richard Skeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 2

[43] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015. 6

[44] Lang Wang, Wen Yu, and Bao Li. Multi-scenes image stitching based on autonomous driving. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 694–698. IEEE, 2020. 1, 2

[45] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019. 2

[46] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. 8

[47] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. SphereSR: 360deg image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5677–5686, 2022. 2, 8

[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

[49] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3269, 2014. 2

[50] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *European Conference on Computer Vision*, pages 653–669. Springer, 2020. 2

[51] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 231–240, 2020. 2

[52] Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep Lucas-Kanade homography for multimodal image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15950–15959, 2021. 2

[53] Qiang Zhou and Xin Li. STN-homography: Direct estimation of homography parameters for image pairs. *Applied Sciences*, 9(23):5187, 2019. 2