Pixel-Wise Warping for Deep Image Stitching

Hyeokjun Kweon*, Hyeonseong Kim*, Yoonsu Kang*, Youngho Yoon*, Wooseong Jeong and Kuk-Jin Yoon

Korean Advanced Institute of Science and Technology
{0327june, brian617, gzgzys9887, dudgh1732, stk14570, kjyoon}@kaist.ac.kr

Abstract

Existing image stitching approaches based on global or local homography estimation are not free from the parallax problem and suffer from undesired artifacts. In this paper, instead of relying on the homography-based warp, we propose a novel deep image stitching framework exploiting the pixel-wise warp field to handle the large-parallax problem. The proposed deep image stitching framework consists of a Pixel-wise Warping Module (PWM) and a Stitched Image Generating Module (SIGMo). For PWM, we obtain pixel-wise warp in a similar manner as estimating an optical flow (OF). In the stitching scenario, the input images usually include non-overlap (NOV) regions of which warp cannot be directly estimated, unlike the overlap (OV) regions. To help the PWM predict a reasonable warp on the NOV region, we impose two geometrical constraints: an epipolar loss and a line-preservation loss. With the obtained warp field, we relocate the pixels of the target image using forward warping. Finally, the SIGMo is trained by the proposed multi-branch training framework to generate a stitched image from a reference image and a warped target image. For training and evaluating the proposed framework, we build and publish a novel dataset including image pairs with corresponding pixel-wise ground truth warp and stitched result images. We show that the results of the proposed framework are qualitatively and quantitatively superior to those of the conventional methods.

Introduction

Image stitching is a classic computer vision task widely used in diverse applications such as robot navigation or 360° image acquisition. Image stitching aligns multiple images taken from different viewpoints into an image from a specific viewpoint with a wider field of view. A general pipeline of image stitching methods is as follows: 1) obtaining transformation between images based on the correspondences, 2) warping image with the transformation, and 3) blending the warped images while reducing unpleasant artifacts.

As far as we know, most image stitching studies formulate the transformation between images as a homography (or affine) matrix. However, the approach relies on the planar scene assumption, which cannot be established when the camera and the target scene are close and/or scenes include abrupt depth changes. In these "large-parallax" cases, the ghosting effect degrades the quality of the stitched results. Several studies (Gao, Kim, and Brown 2011; Zaragoza et al. 2013; Zheng et al. 2019) reduced dependency on the planar scene assumption, by optimizing the homography-based transformation on each subregion (e.g. grid or superpixel). However, there is still a trade-off between accuracy and convergence of warping functions when using these approaches. If we divide the image into smaller subregions to make them nearly planar, then the correspondences on such subregions would be scarce and therefore difficult to optimize the regional warp on these regions. On the other hand, if we use larger subregions, it would provide sufficient correspondences but the planar assumption could be weakened.

To resolve this issue, we propose a deep image stitching framework, formulating the transformation between images as a pixel-wise warp field. In our framework, the 2D warp vector of each pixel is directly estimated, instead of optimizing the transformation function shared for each subregion. Since our approach defines neither subregions nor regionally shared function, it is free from the aforementioned

Figure 1: Qualitative comparison on large parallax images. Reference image (left top), target image (right top), stitched image of APAP (Zaragoza et al. 2013) (left bottom), and stitched image of our method (right bottom).

*These authors contributed equally.

trade-off and therefore can handle large-parallax scenes.

To train and evaluate our framework, we build and publish a novel **Pixel-wise Deep Image Stitching (PDIS) dataset**. Note that our dataset is the first large-scale dataset including GT pixel-wise warps, overlap region masks, and GT stitched images, in addition to the input images. We believe that our dataset could largely contribute to the community.

In our framework, estimating the pixel-wise warp field is performed by a **Pixel-wise Warping Module (PWM)**. PWM extracts visual features from the images and matches them, similar to estimating stereo depth or optical flow. The 2D warp vectors can be directly estimated on the overlapped (OV) regions of the input images. However, in the stitching scenario, the input images usually include non-overlapped (NOV) regions, of which warp cannot be directly obtained due to the lack of correspondences. Predicting NOV warp is indeed a severely ill-defined problem since there can be multiple possible scene structures. Nevertheless, NOV region should follow perspective priors such as epipolar geometry or line preservation. We impose these geometric constraints on PWM to make it reasonably predict the NOV warp, while preventing over-fitting on the training dataset. With the obtained pixel-wise warp, PWM relocates the pixels of the target images onto the reference image plane.

To blend the warped target image and the reference image, we devise a **Stitched Image Generating Module (SIGMo)**. Although our dataset provides the GT stitched result corresponding to the GT warp, making the SIGMo reconstruct the GT result with the estimated warp (which can be far different from the GT warp) leads to blurred results, due to the ill-posed nature of the NOV region. To address this, we devise a multi-branch learning strategy. For the Ground-truth Warp (GW) branch, the input target image is warped using the GT warp, and the direct reconstruction loss is applied. In the Prediction Warp (PW) branch, we feed the target image warped with the predicted warp. Here, we apply input-reconstruction loss and adversarial loss to handle undesirable artifacts such as blurriness or seam. Finally, in the Domain Adaptation (DA) branch, we use real images as input to help the model adapt to the real scene distribution.

Our approach obtains quantitatively and qualitatively superior results compared to the existing image stitching methods on both the proposed dataset and the real images. As shown in Fig. 1, we observed that the proposed method effectively performs image stitching, for the scenes containing large parallax or non-planar objects where the existing approaches usually have failed.

In summary, our contributions are as follows:

- We develop a novel deep image stitching framework that estimates pixel-wise warp and blends the warped target image with the reference image, which can address the large-parallax problem in the image stitching task.
- We propose and publish a large-scale PDIS dataset for the training and evaluation of the image stitching task.
- Our method obtains quantitatively and qualitatively superior stitching results on both the proposed dataset and real images compared to the existing methods, especially for the scenes including large-parallax.

Related Works

Warp Estimation

To relieve the parallax problem, existing image stitching approaches have proposed to divide an input image into subregions such as grid (Zaragoza et al. 2013), triangular projective-consistent planes (Zheng et al. 2019), and superpixels (Lee and Sim 2020). Then, a transformation matrix for each cell is optimized while assuming that the subregions are planar. However, using a regional warp for each subregion induces a trade-off between the accuracy and convergence of the optimization for warp. Plenty of studies have proposed to refine the homography-based warp with various methods: spatially-varying affine transformation (Lin et al. 2011), meshed image plane (Lee and Sim 2020), and line-point constraint (Jia et al. 2021). These works, however, cannot fully resolve the drawback of using homography-based warp since the refined warp is a locally adjusted version of the initial warp. Several prior works have suggested using optical flow for refining the artifacts caused by homography-based methods. Video stitching methods (Krishna et al. 2021; Peleg et al. 2000) use optical flow between the frames for stitching, but they mainly targets the overlapped regions. Panorama stitching methods (Li et al. 2017; Meng and Liu 2020) first stitch the images based on feature matching or known pose, and then refine the misalignments in the overlap region using optical flow. However, the methods less consider about the NOV regions due to the nature of panorama synthesis. Recently, deep-learning-based stitching methods are actively researched. However, the works are based on deep homography estimation (Hoang et al. 2020; Shi et al. 2020; Nie et al. 2020a,b, 2021), still sharing the limitation of homography-based methods. Also, CNN-based generation approaches (Li et al. 2019) are not view-free. Compared to them, this paper proposes a first pixel-wise deep image stitching framework that directly estimates the pixel-wise warp from the alignment stage, while handling the NOV warp with geometrical constraints.

Image Blending

Blending is crucial to create natural stitching results without artifacts. Existing methods exploit seam cost functions (Levin et al. 2004), seam-cutting (Zhang and Liu 2014), or iterative enhancement (Lin et al. 2016). (Nie et al. 2020a) was the first to propose a fully deep framework for image stitching, especially targeting view-free stitching. Attempts such as an edge-preserved deformation branch (Nie et al. 2020b), content loss (Nie et al. 2020a) and seam loss (Nie et al. 2021) have also been made. We also devise SIGMo to target the artifacts in the blending stage. We especially consider the holes, inevitably caused by forward warping in our framework. Since the holes cannot be distinguished from the dummy region, image inpainting techniques (Chaohao Xie 2019; Guilin Liu 2018, 2020) cannot fully address them without a manual mask explicitly indicating the area to be filled. Instead, for SIGMo, we propose a multi-branch learning scheme that can effectively remove seams and holes while obtaining clear stitching results.

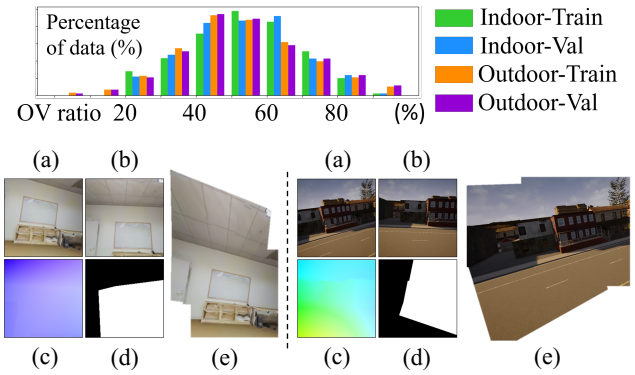


Figure 2: Overlap ratio distribution (top) and sample data of the proposed PDIS dataset (bottom). (a) Reference image, (b) Target image, (c) GT warp from the target image to the reference image, (d) mask for overlap region on the target image, (e) stitched result. Left scene and right scene are from S2D3D and CARLA simulator, respectively.

Proposed Dataset

In introducing deep learning to the field of image stitching, the largest obstacle is the absence of a dataset including ground truth (GT). Although there exist several datasets with pairs of images, none of them provide GT that can be used for training or evaluation of the image stitching method. Therefore, we build and propose PDIS dataset, a novel large-scale dataset for deep image stitching.

We utilize two distinct 3D virtual environments: the S2D3D dataset (Armeni et al. 2016) for indoor scenes and the CARLA simulator (Dosovitskiy et al. 2017) for outdoor scenes. We render image pairs from multiple scenes with virtual cameras at two different views: the reference image and the target image. For each camera, we provide both intrinsic and extrinsic parameters. Here, to mimic the situations where image stitching is generally applied, we enforce a pair of rendered images to have overlapping regions. With the projective camera matrices of the virtual cameras and z-buffer (i.e. depth map) obtained during the rendering process, the warp field of the target image (to the reference image) is obtained with respect to the reference image. Using the GT warp field, our dataset also provides a binary mask indicating which pixels of the target image are overlapped with the reference image. In addition, the proposed dataset also includes GT stitching results constructed from the reference image and the warped target image (according to the obtained GT warp). Fig. 2 shows statistics on the OV ratio, as well as some sample data of the proposed dataset. More sample data and a detailed explanation regarding the proposed dataset can be found in *Supplementary Material*.

Proposed Method

The proposed framework consists of two modules: Pixel-wise Warping Module (PWM) and Stitched Image Generating Module (SIGMo), as shown in Fig. 3. Since our goal is to stitch the target image into the reference image plane in a pixel-wise manner, we first reposition each pixel in the target image to obtain the warped target image using PWM. Then, SIGMo blends the warped target image with the reference image to generate a resulting stitched image. In this section, we will introduce each stage in detail.

Pixel-wise Warping Module (PWM)

Instead of using a homography-based regional warp, we formulate the transformation between images as a pixel-wise warp field. Similar to estimating stereo depth or optical flow (OF), PWM extracts visual features from the input images and matches them. Our pixel-wise warp estimation and OF estimation have a similar formulation: estimating pixel-wise 2D offset vector from one image to the other image. Therefore, we borrow the network architectures from recent OF estimation research (Teed and Deng 2020) as our backbone.

However, this does not mean that the proposed pixel-wise warp estimation for image stitching is identical to the OF estimation. In usual stitching scenarios, the portion of the OV region over the whole image is much smaller than that of the usual settings for OF estimation (e.g. frame-based datasets). Unlike estimating the warp on the OV region, predicting the pixel-wise warp field of the NOV region is an ill-defined problem due to the absence of correspondences.

At first, as a baseline, we directly train the PWM to predict the GT warp field (W_{gt}) from the input image pairs (I^R and I^T), using our dataset. To consider the ill-posed nature of predicting NOV warp, we regularize the loss for the pixels on the NOV region, instead of applying an identical loss function for all pixels on the image. With the GT masks indicating the OV and NOV regions ($M_{gt,ov}$ and $M_{gt,nov}$, respectively), we define the **warp estimation loss** as follows:

$$\begin{aligned}\mathcal{L}_{warp} = \quad & \| M_{gt,ov} \odot W_{pr} - M_{gt,ov} \odot W_{gt} \|_1 \\ & + \beta \| M_{gt,nov} \odot W_{pr} - M_{gt,nov} \odot W_{gt} \|_1,\end{aligned} \tag{1}$$

where W_{pr} is the warp estimated by PWM and β is the weighting factor. Surprisingly, with this loss, we observe that the PWM can predict the reasonable warp field to some degree, even in the NOV region. It implies that providing GT warp as supervision for the NOV region can work as rough guidance. We think that the network could learn how to predict the NOV warp based on the contexts of the OV region, since the existing OF models generally have a wide receptive field to cover both OV and NOV regions.

However, excessively forcing the model to reduce the loss between the predicted warp and GT warp on the NOV region lead to overfitting on the trained dataset, similar to that of the monocular depth estimation task. Although the result shows some feasibility, the predicted NOV warp is not sufficient to achieve plausible image stitching results. To help the PWM predict proper NOV warp in the perspective of image stitching, we explore additional priors that can be imposed on NOV region. In this paper, we focus on the fact that the NOV regions of the input images should follow the geometrical rules induced by perspective projection, even though they do not include correspondences with each other. Under the perspective prior, we devise two geometrical constraints: an epipolar loss and a line-preservation loss.

The **epipolar loss** is based on an epipolar geometry, a

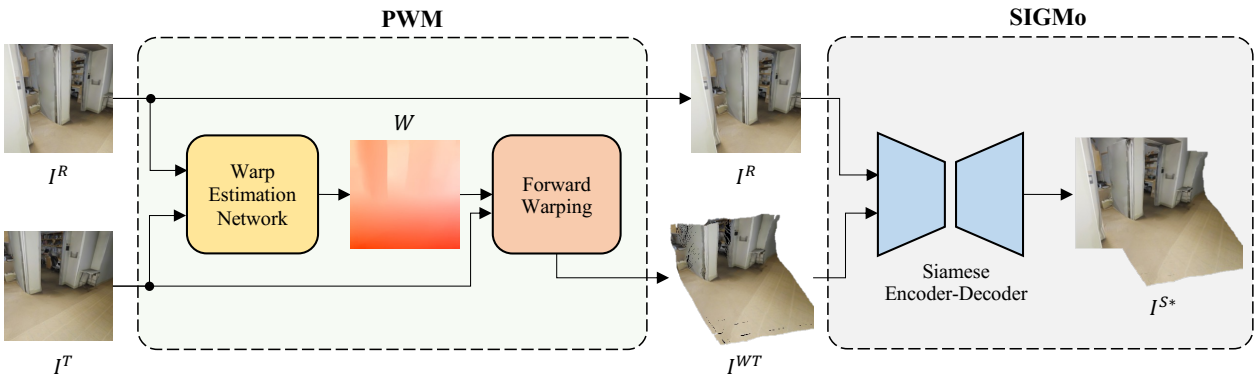


Figure 3: Overall framework for pixel-wise deep image stitching. Given I^R and I^T, PWM relocates the pixels in I^T to the I^R domain using the estimated warp field W in order to obtain I^{WT}. Then, I^R and I^{WT} are padded or cropped to match the predefined size of the stitched image. Finally, I^R and I^{WT} are fed into the SIGMo to obtain a stitched image I^S.

pose-dependent relationship between the pixels of multi-view images. In specific, a pixel of the target image should be relocated to the epipolar line corresponding to that pixel. Although it is difficult to explicitly use the epipolar geometry at the time of inference (since the relative pose is unknown and it is inaccurate to infer), we can use it as a geometry-aware guideline for training PWM. We formulate the epipolar loss as a form of Sampson distance error to constrain the warped pixel p_i^W to be on the epipolar line of p_i, the pixel before being warped. Using the GT fundamental matrix F computed from the projective matrices of the virtual cameras, we define the epipolar loss $\mathcal{L}_{epi}$ as follows:

$$\mathcal{L}_{epi} = \sum_{i=1}^{N} \frac{(p_i^W \cdot Fp_i)^2}{(Fp_i)_1^2 + (Fp_i)_2^2 + (F^T p_i^W)_1^2 + (F^T p_i^W)_2^2}, \tag{2}$$

where N is the number of pixels in the target image, and $(Fp_i)_k$ represents k-th entry of the vector Fp_i.

In addition, we impose the **line-preservation loss** on PWM. The pixels located on a line should also form a line after warping, regardless of whether the pixels are in the OV region or NOV region. Also, from the perspective of the image stitching, preserving lines of target images during warping is crucial for the perceptual quality of stitched results. To realize the line-preservation loss, we first detect line segments $L = \{l_1, \ldots, l_n\}$ from the target image using ULSD (Li et al. 2020). For each detected line l_i, we build a set of pixels $P_i = \{p_{i1}, \ldots, p_{im_i}\}$ that are located on the line. In ideal, on the warped target image, the set of the warped pixels $P_i^W = \{p_{i1}^W, \ldots, p_{im_i}^W\}$ also should form a line. To formulate this constraint as a back-propagable loss function, we fit a line l_i^W for each warped set of pixels P_i^W. The distance between the line l_i^W and the pixel p_{ij}^W is computed as

$$d(l_i^W, p_{ij}^W) = \frac{|a_i x_{ij} + b_i y_{ij} + c_i|}{\sqrt{a_i^2 + b_i^2}} \tag{3}$$

where $l_i^W : a_i x + b_i y + c_i = 0$ and $p_{ij}^W = (x_{ij}, y_{ij})$. Finally, we define the line-preservation loss $\mathcal{L}_{line}$ as follows:

$$\mathcal{L}_{line} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} d(l_i^W, p_{ij}^W)}{\sum_{i=1}^{n} m_i} \tag{4}$$

In summary, the total loss function for training the PWM is defined as follows:

$$\mathcal{L}_{PWM} = \lambda_{warp}\mathcal{L}_{warp} + \lambda_{epi}\mathcal{L}_{epi} + \lambda_{line}\mathcal{L}_{line}. \tag{5}$$

With the predicted pixel-wise warp field, we obtain a warped target image by relocating each pixel in the target image. Here, we use softmax splatting (Niklaus and Liu 2020), which is a well-known forward warping method in video frame interpolation. We denote the warping process as a function $\mathfrak{W}$, which gets the target image I^T and warp W and returns the warped target image I^{WT} as follows:

$$I^{WT} = \mathfrak{W}(I^T, W). \tag{6}$$

Stitched Image Generating Module (SIGMo)

We propose the SIGMo (denoted as $\mathfrak{F}$) to generate a stitching result I^{S*} by blending the reference image I^R and the warped target image I^{WT} as follows:

$$I^{S*} = \mathfrak{F}(I^R, I^{WT}). \tag{7}$$

Aforementioned in the dataset section, the proposed dataset provides stitching results I_{gt}^S for each given pair of the reference image and the target image. I_{gt}^S is constructed from the target image warped by the GT warp field W_{gt}. However, we cannot assure that the warp estimated by PWM (W_{pr}) is identical to the GT warp. Therefore, enforcing SIGMo to follow the GT stitching results could weaken its blending capability. Indeed, when we directly trained the SIGMo using I_{GT}^S as the reconstruction target, we observed that SIGMo generates severely blurred results.

To address this, we propose a novel multi-branch learning framework for SIGMo, as depicted in Fig 4. In the **GT Warp (GW) branch**, we obtain the GT warped target image $I_{gt}^{WT} = \mathfrak{W}(I^T, W_{gt})$ with the W_{gt} provided by our dataset. Then, we feed I^R and I_{gt}^{WT} to the SIGMo. Since the I_{gt}^{WT}

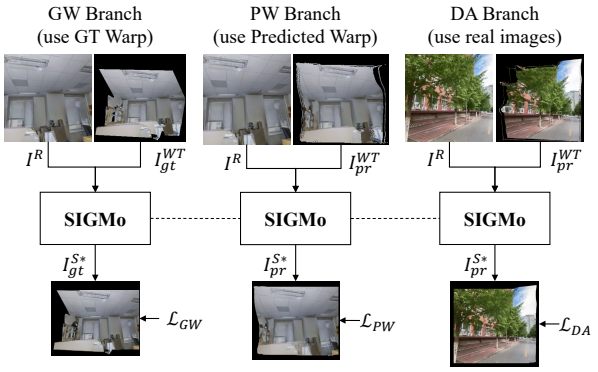


Figure 4: Illustration of a multi-branch training framework for the SIGMo. It shows the GT Warp (GW), Prediction Warp (PW), and Domain Adaptation (DA) branches. The dashed lines represent weight sharing between the SIGMo.

perfectly fits with the GT stitched image I^S_{gt}, we can directly train the SIGMo to reconstruct the I^S_{gt} from the reference image I^R and I^{WT}_{gt}. We apply the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) and the L1 loss as follows:

$$\begin{aligned}\mathcal{L}_{GW} = \quad & \lambda_{GW,per} \parallel V(I^{S*}_{gt}) - V(I^S_{gt}) \parallel_1 \\ & + \lambda_{GW,rec} \parallel I^{S*}_{gt} - I^S_{gt} \parallel_1,\end{aligned} \tag{8}$$

where $I^{S*}_{gt} = \mathfrak{F}(I^R, I^{WT}_{gt})$ is the output of SIGMo when using I^{WT}_{gt}, and V is the perceptual embedding function. Through the GW branch, we can help the network implicitly handle the occlusion and hole caused by forward warping.

On the other hand, in a **Predicted Warp (PW) branch**, we obtain $I^{WT}_{pr} = \mathfrak{W}(I^T, W_{pr})$, where W_{pr} is the warp predicted by PWM. Instead of using I^S_{gt} as the reconstruction target, we make SIGMo preserve the pixels of inputs (I^R and I^{WT}_{pr}) while blending them. To address the hole of I^{WT}_{pr}, we obtain binary occupancy masks M^R and M^{WT} for the reference image and the warped target image, respectively. Then, according to the masks, we impose a reconstruction loss only to the pixels existing in I^R and I^{WT}, as follows:

$$\begin{aligned}\mathcal{L}_{PW,rec} = \quad & \parallel M^R \odot I^{S*}_{pr} - M^R \odot I^R \parallel_1 \\ & + \parallel M^{WT} \odot I^{S*}_{pr} - M^{WT} \odot I^{WT}_{pr} \parallel_1,\end{aligned} \tag{9}$$

where $I^{S*}_{pr} = \mathfrak{F}(I^R, I^{WT}_{pr})$ is the result of SIGMo.

Furthermore, to handle the other undesired artifacts such as seams or misalignments, we impose an adversarial loss for the output of SIGMo. We employ an unconditional patch discriminator (Isola et al. 2017), denoted as D, where I^{gt}_S is used as a real sample. The adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_D \quad &= \mathbb{E}_{I^S_{gt}}[(D(I^S_{gt}) - 1)^2] + \mathbb{E}_{I^{S*}_{pr}}[(D(I^{S*}_{pr}))^2] \\ \mathcal{L}_{PW,adv} \quad &= \mathbb{E}_{I^{S*}_{pr}}[(D(I^{S*}_{pr}) - 1)^2],\end{aligned} \tag{10}$$

The total loss function for the PW branch is defined as $\mathcal{L}_{PW} = \lambda_{pw,rec}\mathcal{L}_{PW,rec} + \lambda_{pw,adv}\mathcal{L}_{PW,adv}$.

Although the proposed GW and PW branches enable the SIGMo to generate plausible results on our dataset, the generalization issue for in-the-wild images still remains a crucial problem. To address this, we additionally devise a **Domain Adaptation (DA) branch**, which helps SIGMo adapt to the distribution of the real images. Since we have neither GT warp nor GT stitched results for the real images, we only impose input-reconstruction loss for $\mathcal{L}_{DA}$, similar to Eq.9.

In summary, with the aforementioned three branches (GW, PW, and DA), the total loss function for training the SIGMo is defined as follows:

$$\mathcal{L}_{SIGMo} = \mathcal{L}_{GW} + \mathcal{L}_{PW} + \mathcal{L}_{DA}. \tag{11}$$

Experiments

We train mainly using our PDIS dataset, and additionally use UDIS dataset (Nie et al. 2021) for real images. In PWM, we borrow the architecture of (Teed and Deng 2020), which has shown promising results in OF estimation. We first train PWM alone, and then train the SIGMo with fixed PWM. We use AdamW optimizer (Loshchilov and Hutter 2017) (β_1 = 0.5, β_2 = 0.999, and lr=1e-4), with batch size of 8. For perceptual loss (Eq. 8), we use layer relu2_2 from the VGG16 (Simonyan and Zisserman 2014) pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012).

Ablations

Geometrical Losses for PWM We experimentally verify the effect of losses for PWM by ablating the proposed epipolar loss and line-preservation loss, while using the warp estimation loss as a baseline. For evaluation, we employ three metrics: L1 loss between the predicted warp and the GT warp, Average Epipolar Distance (AED), and Line Preservation loss (LP). We evaluate each setting with the metrics on the whole, OV, and NOV regions of the target image. In PDIS dataset, AED was measured using the GT fundamental matrix. However, since the GT fundamental matrix is not accessible for the UDIS dataset, we used the fundamental matrix calculated from the estimated OF on the OV region. Detailed process for acquiring the reliable fundamental matrix from the OF can be found in *Supplementary Material*. As shown in Table 1, using the epipolar loss reduces AED, as well as the Warp L1 loss, in both OV and NOV regions. Furthermore, with additional line-preservation loss, PWM achieves even better results for all metrics and all regions. In particular, the improvement in the NOV region is significantly higher than in the OV region. With the results, we confirm that the proposed epipolar and line-preservation losses not only help PWM predict better warp in terms of L1 loss, but also effectively provide geometrical priors, as we intended. Also, in Fig 5, we provide qualitative comparison between the stitching results obtained by ablating the epipolar and line-preservation losses. Both qualitative and quantitative results strongly support that the proposed losses enable PWM to provide a more accurate and geometrically plausible warp, which is our key contribution to achieve pixel-wise warp estimation for image stitching.

Losses			PDIS dataset									UDIS dataset					
			Warp L1			AED			LP			AED			LP		
WP	Epi	Line	OV	NOV	Total	OV	NOV	Total	OV	NOV	Total	OV	NOV	Total	OV	NOV	Total
✓			11.1	19.7	13.8	6.51	11.4	8.86	1.43	1.62	1.48	2.98	6.44	4.36	1.34	1.63	1.42
✓	✓		10.1	19.0	12.9	6.04	10.28	8.08	1.49	1.57	1.51	2.64	5.38	3.69	1.32	1.54	1.38
✓	✓	✓	**8.76**	**16.9**	**11.4**	**4.60**	**6.65**	**5.53**	**1.26**	**1.29**	**1.27**	**2.30**	**4.86**	**3.18**	**1.10**	**1.25**	**1.14**

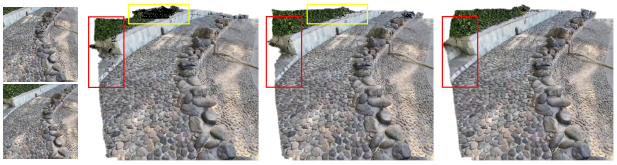
Table 1: Quantitative results of the ablation study regarding the loss functions ($\mathcal{L}_{warp}$, $\mathcal{L}_{epi}$, and $\mathcal{L}_{line}$) for AED and LP denote Average Epipolar Distance and Line-preservation loss, respectively. Bold numbers represent the best results.

Branch			PDIS dataset (Whole region)				UDIS dataset (Reference region)			
Pred	GT	DA	LPIPS(↓)	PSNR(↑)	SSIM(↑)	BRISQUE(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	BRISQUE(↓)
✓			0.1746	25.22	0.7493	48.40	0.0419	33.83	0.9714	50.87
✓	✓		0.1220	25.44	0.8810	47.66	0.0419	32.05	0.9623	48.77
✓	✓	✓	**0.1009**	**26.31**	**0.9148**	**43.80**	**0.0270**	**35.53**	**0.9804**	**44.53**

Table 2: Quantitative results of the ablation study regarding the multi-branch learning framework for SIGMo.

Method	PDIS dataset			UDIS dataset
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)
UDIS	31.4430	0.9399	0.0341	21.1715
Ours	**34.8244**	**0.9746**	**0.0066**	**23.3109**

Table 3: Quantitative comparison of the proposed method with UDIS (Nie et al. 2021) on PDIS and UDIS datasets.



(a) (b) (c) (d)

Figure 5: (a) Input images, (b) Result with warp estimation loss, (c) Result with additional epipolar loss, (d) Result with additional epipolar loss and line-preservation loss.

Multi-branch Learning We conduct an ablation experiment to observe the effect of each branch of the proposed multi-branch learning framework. As shown in Table 2, we set the PW branch as the baseline and compared the performance by adding the GW branch and the DA branch. Please refer to the *Supplementary material* for implementation details for each model used in the ablation experiments. LPIPS (Zhang et al. 2018), PSNR, and SSIM (Wang et al. 2004) are used as metrics. For PDIS dataset, the metrics are measured between the GT stitched image and the stitched image generated using the GT warp. However, since GT stitched images do not exist in the UDIS dataset, we evaluate the image quality of the reference image region only. In addition, we use BRISQUE (Mittal, Moorthy, and Bovik 2012) metric to evaluate no-reference image quality. Since GT images are not required to measure the BRISQUE score, we directly evaluate the entire stitched images in both datasets. When the GW branch is added to the baseline, LPIPS, PSNR, and SSIM metrics show that the performance

(a) (b) (c) (d)

Figure 6: (a) Input images, (b) Result with PW, (c) Result with PW+GW (d) Result with PW+GW+DA.

is improved for PDIS images. In Fig.6, it can be observed that adding the GW branch helps SIGMo fill the holes created by forward warping through direct GT reconstruction loss as we intended. However, since the L1 reconstruction loss is used for the GW branch, the image is slightly blurred as shown in Fig. 6 (a). Finally, the model with both GW and DA branches achieves the best for all metrics on both datasets. Comparing Fig. 6 (c) with Fig. 6 (a), it can be seen that the overall clarity of the real image is improved. Also, the blurriness caused by the GW branch becomes clearer. These are because the model adapts to the distribution of the real scene through the DA branch during training. In summary, with all branches of the proposed blender framework, we can obtain clear stitching results with minimal artifacts.

Comparisons with Existing Methods

To clarify the superiority of the proposed image stitching method based on pixel-wise warp estimation, we quantitatively compare our method with the UDIS (Nie et al. 2021), which is based on the deep homography estimation. We conduct the experiment on the validation set of PDIS dataset (synthetic) and the UDIS dataset (Nie et al. 2021) (real). For the PDIS dataset, we evaluate the PSNR, SSIM, and LPIPS between the OV region of the stitched result and the GT stitched image. The OV mask and the GT stitched image are provided by our dataset. For the UDIS dataset, we only use the PSNR (which can be evaluated at pixel-level) between the reference image and the warped target image, since there

Figure 7: Qualitative comparisons of the stitched images of various methods. From left to right: (a) input images (reference ↑, target ↓), (b) APAP (Zaragoza et al. 2013), (c) AANAP (Lin et al. 2015), (d) UDIS (Nie et al. 2021), (e) the proposed method. More comparative experiments can be found in the *Supplementary material*.

Figure 8: Case of small OV region. Left (blue) is the result of UDIS (Nie et al. 2021) where the right (red) is ours.

is no GT mask for the OV region. As shown in Table 3, the proposed method outperforms the homography-based UDIS method both on the PDIS and UDIS datasets. Although our framework is trained on the proposed synthetic dataset, the superior performance of the proposed method on the real images (UDIS dataset) supports the generalizability of the proposed framework. Also, it is worth noting that the proposed method's inference time (0.13s) is significantly shorter than the other methods (APAP: 4.2s and UDIS: 0.24s). Fig. 7 shows the qualitative comparison between our method and the conventional methods (Zaragoza et al. 2013; Lin et al. 2015; Nie et al. 2021) on the PDIS and UDIS dataset. As intended, the proposed method robustly stitches the input images without misalignments and severe distortion, while the other methods often fail to converge or produce misalignments. It supports that our pixel-wise deep image stitching method generates substantial stitching results. We have also tested our method in more challenging cases: stitching of images having a small OV region (Fig. 8).

Conclusions

Most existing studies in image stitching have exploited the homography-based warp, which causes parallax problems. In this paper, to address the large parallax problem, we propose to estimate the pixel-wise warp instead of the homography-based warp. We devise a novel pixel-wise deep image stitching framework composed of a Pixel-wise Warping Module (PWM) and a Stitched Image Generating Module (SIGMo). PWM estimates the 2D warp field from a target image to the reference image and warps the target image with the predicted warp. To handle the ill-posed nature of predicting warp on the non-overlapped region, we train PWM with the proposed epipolar loss and line-preservation loss. Then, SIGMo generates plausible stitching results by blending the reference image and the warped target image, with the help of the proposed multi-branch training strategy. To train and evaluate the proposed framework, we also build a large-scale synthetic dataset including GT warp and GT stitched images, which can serve as a benchmark for the field of image stitching. As a result, with the proposed framework and the dataset, we obtain impressive stitching results. Our method can handle challenging scenes with large parallax as we intended. It supports the superiority of the proposed pixel-wise warp estimation approach compared to the homography-based approaches.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636).

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Chaohao Xie, C. L. M.-M. C.-W. Z. X. L. S. W. E. D., Shaohui Liu. 2019. Image inpainting with learnable bidirectional attention maps. *IEEE/CVF International Conference on Computer Vision*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Gao, J.; Kim, S. J.; and Brown, M. S. 2011. Constructing image panoramas using dual-homography warping. In *CVPR 2011*, 49–56. IEEE.

Guilin Liu, K. J. S. T.-C. W. A. T. B. C., Fitsum A. Reda. 2018. Image inpainting for irregular holes using partial convolutions. *European conference on computer vision*.

Guilin Liu, K. J. S. T.-C. W. A. T. B. C., Fitsum A. Reda. 2020. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *European conference on computer vision*.

Hoang, V.-D.; Tran, D.-P.; Nhu, N. G.; Pham, V.-H.; et al. 2020. Deep feature extraction for panoramic image stitching. In *Asian Conference on Intelligent Information and Database Systems*, 141–151. Springer.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jia, Q.; Li, Z.; Fan, X.; Zhao, H.; Teng, S.; Ye, X.; and Latecki, L. J. 2021. 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*, 12186–12195.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Krishna, V.; Joshi, A.; Vrabac, D.; Bulterys, P.; Yang, E.; Fernandez-Pol, S.; Ng, A. Y.; and Rajpurkar, P. 2021. GloFlow: Whole Slide Image Stitching from Video Using Optical Flow and Global Image Alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 519–528. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lee, K.-Y.; and Sim, J.-Y. 2020. Warping residual based image stitching for large parallax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8198–8206.

Levin, A.; Zomet, A.; Peleg, S.; and Weiss, Y. 2004. Seamless image stitching in the gradient domain. In *European conference on computer vision*, 377–389. Springer.

Li, H.; Yu, H.; Yang, W.; Yu, L.; and Scherer, S. 2020. ULSD: Unified Line Segment Detection across Pinhole, Fisheye, and Spherical Cameras. arXiv:2011.03174.

Li, J.; Zhao, Y.; Ye, W.; Yu, K.; and Ge, S. 2019. Attentive Deep Stitching and Quality Assessment for 360º Omnidirectional Images. *IEEE Journal of Selected Topics in Signal Processing*, 14(1): 209–221.

Li, L.; Yao, J.; Xie, R.; Xia, M.; and Zhang, W. 2017. A unified framework for street-view panorama stitching. *Sensors*, 17(1): 1.

Lin, C.-C.; Pankanti, S. U.; Natesan Ramamurthy, K.; and Aravkin, A. Y. 2015. Adaptive as-natural-as-possible image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1155–1163.

Lin, K.; Jiang, N.; Cheong, L.-F.; Do, M.; and Lu, J. 2016. Seagull: Seam-guided local alignment for parallax-tolerant image stitching. In *European conference on computer vision*, 370–385. Springer.

Lin, W.-Y.; Liu, S.; Matsushita, Y.; Ng, T.-T.; and Cheong, L.-F. 2011. Smoothly varying affine stitching. In *CVPR 2011*, 345–352. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Meng, M.; and Liu, S. 2020. High-quality Panorama Stitching based on Asymmetric Bidirectional Optical Flow. In *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, 118–122. IEEE.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. In *IEEE Transactions on image processing*, volume 21, 4695–4708. IEEE.

Nie, L.; Lin, C.; Liao, K.; Liu, M.; and Zhao, Y. 2020a. A view-free image stitching network based on global homography. *Journal of Visual Communication and Image Representation*, 73: 102950.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2021. Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images. *IEEE Transactions on Image Processing*, 30: 6184–6197.

Nie, L.; Lin, C.; Liao, K.; and Zhao, Y. 2020b. Learning edge-preserved image stitching from large-baseline deep homography. *arXiv preprint arXiv:2012.06194*.

Niklaus, S.; and Liu, F. 2020. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5437–5446.

Peleg, S.; Rousso, B.; Rav-Acha, A.; and Zomet, A. 2000. Mosaicing on adaptive manifolds. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10): 1144–1154.

Shi, Z.; Li, H.; Cao, Q.; Ren, H.; and Fan, B. 2020. An image mosaic method based on convolutional neural network semantic features extraction. *Journal of Signal Processing Systems*, 92(4): 435–444.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Zaragoza, J.; Chin, T.-J.; Brown, M. S.; and Suter, D. 2013. As-projective-as-possible image stitching with moving DLT. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2339–2346.

Zhang, F.; and Liu, F. 2014. Parallax-tolerant image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3262–3269.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, J.; Wang, Y.; Wang, H.; Li, B.; and Hu, H.-M. 2019. A novel projective-consistent plane based image stitching method. *IEEE Transactions on Multimedia*, 21(10): 2561–2575.