Rectification Reimagined: A Unified Mamba Model for Image Correction and Rectangling with Prompts

Linwei Qiu[1,2,3,4], **Gongzhe Li**[5], **Xiaozhe Zhang**[1,2,3], **Qilin Sun**[5], **Fengying Xie**[1,2,3*]

[1]Tianmushan Laboratory, Beihang University, Hangzhou, China
[2]State Key Laboratory of High-Efficiency Reusable Aerospace Transportation Technology, Beijing, China
[3]School of Astronautics, Beihang University, Beijing, China
[4]School of Aerospace Engineering, Huazhong University of Science and Technology, Wuhan, China
[5]School of Data Science, The Chinese University of Hong Kong, Shenzhen
{qiulinwei, xiaozhe_zhang, xfy_73}@buaa.edu.cn, gongzheli1@link.cuhk.edu.cn, sunqilin@cuhk.edu.cn

Abstract

Image correction and rectangling are valuable tasks in practical photography systems such as smartphones. Recent remarkable advancements in deep learning have undeniably brought about substantial performance improvements in these fields. Nevertheless, existing methods mainly rely on task-specific architectures. This significantly restricts their generalization ability and effective application across a wide range of different tasks. In this paper, we introduce the **Unified Rectification Framework (UniRect)**, a comprehensive approach that addresses these practical tasks from a consistent distortion rectification perspective. Our approach incorporates various task-specific inverse problems into a general distortion model by simulating different types of lenses. To handle diverse distortions, UniRect adopts one task-agnostic rectification framework with a dual-component structure: a Deformation Module, which utilizes a novel Residual Progressive Thin-Plate Spline (RP-TPS) model to address complex geometric deformations, and a subsequent Restoration Module, which employs Residual Mamba Blocks (RMBs) to counteract the degradation caused by the deformation process and enhance the fidelity of the output image. Moreover, a Sparse Mixture-of-Experts (SMoEs) structure is designed to circumvent heavy task competition in multi-task learning due to varying distortions. Extensive experiments demonstrate that our models have achieved state-of-the-art performance compared with other up-to-date methods.

Code — https://github.com/yyywxk/UniRect

Introduction

With the rapid development of photography system, there has been a surging interest in low-level tasks, which play pivotal roles in enhancing image quality. Image correction usually aims at estimating and correcting distortions (Li et al. 2019; Zhao et al. 2018), and has a very wide range of applications (Zhou, Cao, and Sun 2018; Zhuang et al. 2019; Xue et al. 2019; Zhan and Lu 2019). Image rectangling is a special type of post-processing task (He, Chang, and Sun 2013b) to acquire user-friendly rectangular images for photos with irregular boundaries while maintaining high content fidelity. Previous studies have addressed these problems by devising diverse specific architectures to learn different types of processing. For instance, Tan et al. (2021) proposed a cascaded network and progressively correct portrait distortion. Nie et al. (2022) optimized a mesh model for stitched image rectangling. Nearly all of these methods concentrate on a single task and depend on task-specific networks. Besides, we note that these practical tasks may be required within one smartphone device and associated with specific cameras or lenses, as depicted in Fig. 1a. The former indicates that identifying a unified task-agnostic model could be highly beneficial for edge devices with limited computation capabilities and memory resources. The latter suggests that potential distortion relationships may exist in other tasks beyond the portrait and rotation correction tasks when considering all tasks as one general rectification task.

This work considers these problems from **a consistent rectification perspective**. We rethink the inverse processes of these problems as the distortion and demonstrate that they can be encompassed within a general distortion model in Fig. 1b. Here, rectified wide-angle image rectangling and stitched image rectangling can be viewed as a unified rectification task addressing wide-angle image rectified distortion and stitched distortion respectively. These two types of distortion lead to irregular boundaries, which may impair image quality and vision understanding. Subsequently, we propose a unified rectification framework (UniRect) to estimate such unified distortion and accomplish image correction and rectangling missions (including four specific tasks in Fig. 1a and obtaining four-by-one in Fig. 1c.2) by rectifying the input image. It consists of two main components: a deformation module based on residual progressive thin-plate spline (RP-TPS) model, which approximates the distortion by progressively predicting the locations of control points; a restoration module based on Residual Mamba Blocks (RMBs) to counteract the degradation caused by previous deformation process. Moreover, to address the heavy task competition in multi-task learning, we design a Sparse Mixture-of-Experts (SMoEs) strategy and perform four-in-one in Fig. 1c.3. This approach uses a gating network to assign weights to ex-

*Corresponding Author

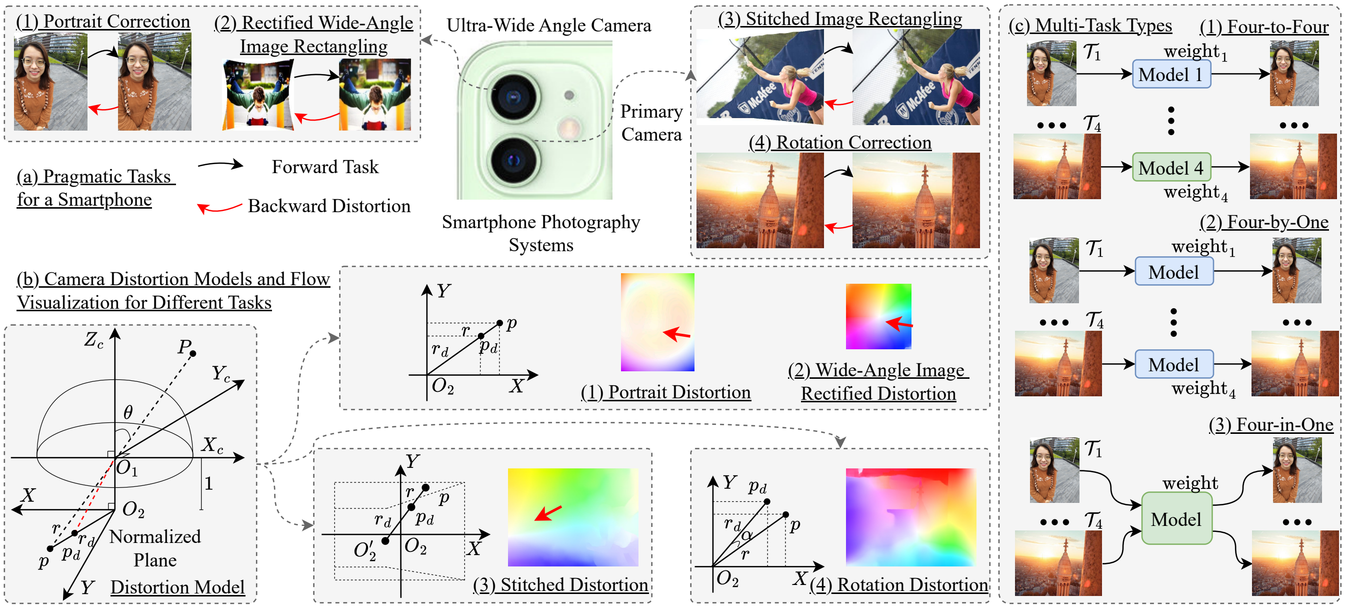

Figure 1: (a) Pragmatic tasks for a smartphone. We consider four detailed tasks for image correction and rectangling, which are closely related to two types of common cameras on a mainstream mobile phone. Portrait correction ($\mathcal{T}_1$) and rectified wide-angle image rectangling ($\mathcal{T}_2$) often need to process pictures taken with a wide-angle lens. Stitched image rectangling ($\mathcal{T}_3$) and rotation correction ($\mathcal{T}_4$) are two practical tasks for daily life. (b) Camera distortion models and flow visualization for different tasks. On a normalized plane, the image of the point P is p_d whereas it would be p without distortion following a pinhole camera model (Kannala and Brandt 2006). The optical flows of backward distortions are generated through the RAFT (Teed and Deng 2020), which is the one of the most powerful tools for optical flow estimation. (c) Multi-task types. (1) four-to-four. Previous studies often use four disparate models to accomplish four tasks separately. (2) four-by-one. four tasks are achieved by one model structure but sharing different even counteractive network weights. (3) four-in-one. four tasks are employed in one model which can handle diverse tasks at the same time. four-by-one and four-in-one are all task-agnostic models. We obtain the **four-by-one** and **four-in-one** in this paper.

pert networks, thereby alleviating performance degradation. Our models have attained state-of-the-art performance levels comparable to those of current methods across four task scenarios.

In summary, our main contributions are as follows:

- We take a novel view of image correction and rectangling from a consistent rectification perspective and establish a general distortion model mathematically to describe their backward distortion processes.
- We propose a Unified Rectification Framework (UniRect) with prompts to effectively address four typical tasks (four-by-one) and design a Sparse Mixture-of-Experts (SMoEs) structure for multi-task learning (four-in-one).
- We conduct extensive experiments on various datasets to validate the superior of our models compared to up-to-date methods across different tasks.

Related Work

Image Correction. Portrait distortion is a common issue encountered in smartphone photography. Shih, Lai, and Liang (2019) proposed an algorithm based on mesh optimization to restore the wrapped content. Tan et al. (2021) introduced the first end-to-end method based on deep learning and provided a supervised dataset and Zhu et al. (2022) proposed a semi-supervised method to further improve results. Rotation distortion is caused by arbitrary rotation angle. Direct rotation requires angle-prior and also destroys the rectangular boundaries. Some previous methods preserve image structure (edge or line) to ensure the naturalness of image content (He, Chang, and Sun 2013a; Von Gioi et al. 2008). Nie et al. (2023) proposed a mesh transformation network to correct content and boundaries. However, all these networks need task-specific designs.

Image Rectangling. To obtain rectangular images after stitching, simple inner rectangle clipping leads to the loss of pixel semantics, while certain image inpainting or completion methods (Suvorov et al. 2022; Teterwak et al. 2019; Liao et al. 2021) pose the risk of introducing uncertain information. He, Chang, and Sun (2013b) first introduced a content-aware algorithm that utilizes an energy function to optimize local mesh deformation. Recent DNN-based methods (Nie et al. 2022; Zhou et al. 2024; Qiu et al. 2024) estimated the wrapping meshes or motion fields to handle non-rectangular borders. Liao et al. (2023) constructed a win-win representation for rectified wide-angle image rectangling. In this work, we regard these tasks as a unified rectification task.

Multi-Task Learning. Multi-task learning methods can be roughly divided into task balancing and multi-task architecture. Tasking balancing (Kendall, Gal, and Cipolla 2018; Guo et al. 2018) methods focus on weights on task and gradient conflicts by re-weighting the loss or manipulating the gradient. This type of method can successfully address multi-task learning for similar problems (Liao et al. 2025), such as simultaneous image denoising and deblurring (Zamir et al. 2021; Potlapalli et al. 2024). However, it tends to underperform when there is a significant disparity between different tasks. Besides, some other approaches (Bragman et al. 2019; Ruder et al. 2019; Gao et al. 2019) propose the advanced architecture for parameter sharing. Gao et al. (2019) employed different decoders for different tasks, while Ruder et al. (2019) proposed a gating mechanism as soft parameters sharing between different task networks. We try to mitigate task competition and achieve the four-in-one in this paper.

Methodology

General Distortion Model

We initiate the construction of distortion models for these tasks by leveraging optical flow. The inverse problems $\mathcal{T}^{-1}$ of these tasks $\mathcal{T}$ are backward distortion procedures, which are depicted in Fig. 1b. $X_cO_1Y_c$ is a camera coordinate system and XO_2Y is the scaled coordinate system. For the image point $p = [x, y]^T$ and its distortion point $p_d = [x_d, y_d]^T$, r and r_d are distances between them and the principal point O_2 respectively. We categorize the backward distortion into four types, based on tasks and different flow visualization maps.

1) For portrait distortion in $\mathcal{T}_1^{-1}$, it has been fully studied by previous researchers and the Kanala-Brandt model (Kannala and Brandt 2006) is widely used to approximate the distortion of a wide-angle lens: $r_d = \sum_{j=1}^{N} k_j\theta^{2j-1}, N = 1, 2, 3, \cdots$, where θ denotes the angle between the incoming ray and the optical axis and k_j is the distortion parameter. **2)** For wide-angle image rectified distortion in $\mathcal{T}_2^{-1}$, we observe that its flow appears a radial distribution as indicated by the red arrow in Fig. 1b.2 and a traditional Brown-Conrady model (Weng et al. 1992) in introduced here: $r_d = \sum_{j=1}^{N} k_j r^{2j-1}, N = 1, 2, 3, \cdots$. **3)** The stitched distortion in $\mathcal{T}_3^{-1}$ is formed by supposing that the rectangular image is transmuted into a image with irregular boundaries through an amorphous and vibrant lens. We discover that the flow in Fig. 1b.3 presents an approximate radial distribution around a particular point O_2', which is affected by irregular borders. Thus, we biuld a variant of Brown-Conrady model to describe this phenomenon:

$$\left[\begin{array}{c} x_d \\ y_d \end{array}\right] = \frac{r_d}{r}\left[\begin{array}{c} x \\ y \end{array}\right] + T_0 = \frac{1}{r}\left(\sum_{j=1}^{N} k_j r^{2j-1}\right)\left[\begin{array}{c} x \\ y \end{array}\right] + T_0, \tag{1}$$

where $T_0 = [x_0, y_0]^T$ penalizes the decentered influence of O_2'. **4)** The rotation distortion in $\mathcal{T}_4^{-1}$ need to consider the rotation angle α of the coordinates:

$$\left[\begin{array}{c} x_d \\ y_d \end{array}\right] = \frac{r_d}{r}\mathcal{R}_\alpha\left[\begin{array}{c} x \\ y \end{array}\right] = \frac{1}{r}\left(\sum_{i=1}^{N} k_j r^{2j-1}\right)\mathcal{R}_\alpha\left[\begin{array}{c} x \\ y \end{array}\right], \tag{2}$$

where $\mathcal{R}_\alpha$ is the rotation matrix.

Considering all of the aforementioned, the final general distortion model is shown below

$$\left[\begin{array}{c} x_d \\ y_d \end{array}\right] = \frac{1}{r}\sum_{j=1}^{N}\left(k_j\theta^{2j-1} + k_j' r^{2j-1}\right)\mathcal{R}_\alpha\left[\begin{array}{c} x \\ y \end{array}\right] + T_0, \tag{3}$$

where k_j, k_j' are distortion parameters. From the unified distortion model in Eq. (3), we discover that the backward distortion of some tasks has a specific coupling relationship.

Framework of UniRect

By describing inverse problems from a unified distortion perspective, we can view all these tasks as a unified distortion rectification task. Our UniRect is depicted in Fig. 2. Here, the Deformation Module (DM) is designed to estimate task-specific distortions and carry out rectification. Concurrently, the Restoration Module (RM) is employed to compensate for potential degradation resulting from estimation errors and interpolation within the DM.

Residual Progressive Thin-Plate Spline Model. As shown in Eq. (3), this unified distortion process is inconstant and usually non-linear, which cannot be approximated by simple transformations. Therefore, we utilize Thin-Plate Spline (TPS) model to simulate aforementioned distortion. Give a set of basic control points $\mathbf{P} = \{p_1, \cdots, p_{N_c}\}$ on a rectified image and their corresponding control points $\mathbf{P}' = \{p_1', \cdots, p_{N_c}'\}$ on an image with distortion in Fig. 2f, the minimum approximation of TPS is formulated (Bookstein 1989):

$$\min \sum_{j=1}^{N_c} \left\| \Phi\left(p_j\right) - {p'}_j \right\| + \lambda \iint_{\mathbb{R}^2} \left(\left(\frac{\partial \Phi}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 \Phi}{\partial x \partial y}\right)^2 + \left(\frac{\partial \Phi}{\partial y^2}\right)^2\right)^2 dxdy, \tag{4}$$

where Φ denotes the transformation and λ is a weight to balance two terms. The only closed-form solution (Kent and Mardia 1994) of Eq. (4) can be derived as

$$\Phi\left(p\right) = A_\Phi \left[\begin{array}{cc} p & 1 \end{array}\right]^T + \sum_{j=1}^{N_c} w_j U\left(\|p - p_j\|\right), \tag{5}$$

where p is any point on the rectified image and U is the radial basis function. $A_\Phi \in \mathbb{R}^{2\times 3}, w_i \in \mathbb{R}^{2\times 1}$ are parameters which can be calculated by control point pairs. Therefore, if the basic control points are set to be evenly distributed throughout the rectified image, we can compute them from the distortion image by Eq. (5) so long as the locations of corresponding control points are given. This process and related matrices are fixed when the image resolution and N_c are preset.

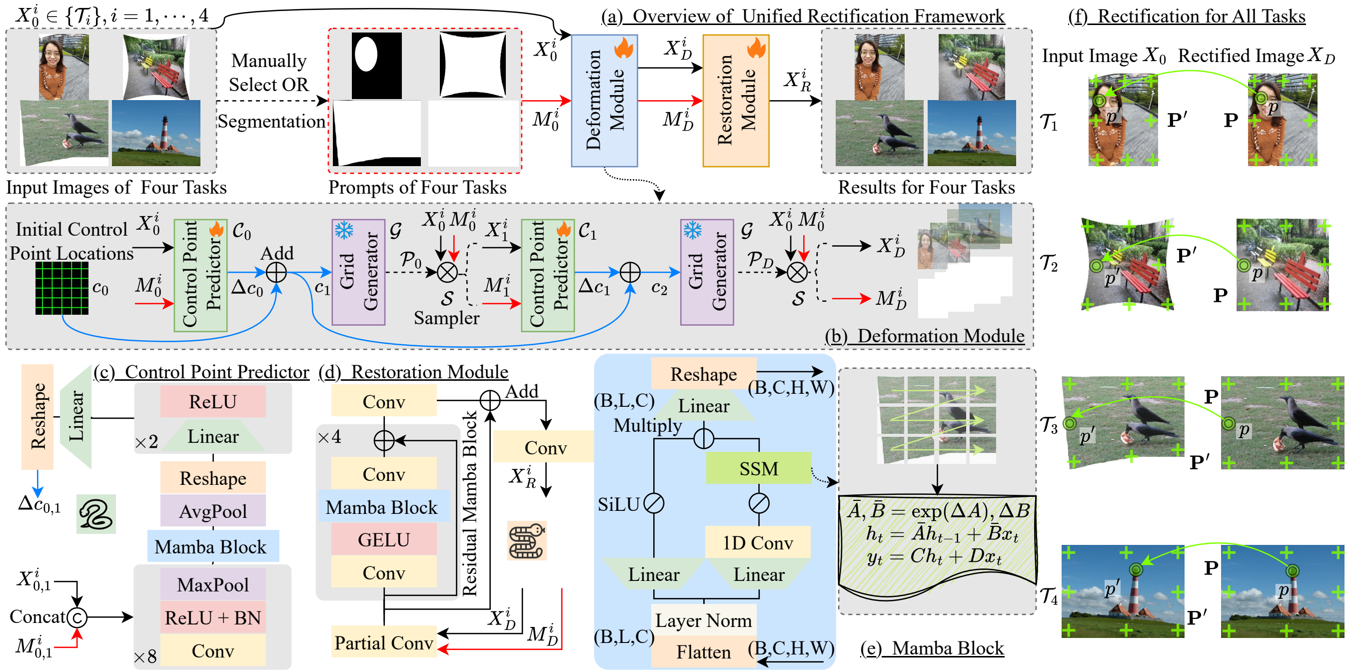

Figure 2: (a) Framework of our Unified Rectification. It mainly consists of a (b) deformation module (DM) and a (d) restoration module (RM), which are trained simultaneously. An image X_0^i from the image set of task $\mathcal{T}_i$ and a corresponding visual prompt M_0^i indicating which task to perform are set as input for UniRect, which yields the final result X_R^i. In deformation module (ignoring our residual progressive setting for simplicity), the (c) control point predictor $\mathcal{C}$ can predict the locations of a set of control points, *i.e.* c , with which the grid generator $\mathcal{G}$ produces a sampling grid $\mathcal{P}$ by Eq. (5). The sampler $\mathcal{S}$ then samples from X_0^i and M_0^i with the restriction of $\mathcal{P}$, resulting in the rectified image X_D^i and its new prompt M_D^i. For some tasks without boundary changes like $\mathcal{T}_1$ and $\mathcal{T}_4$, M_D^i will be changed into the all-one matrix. (e) Mamba block is applied in the control point predictor to obtain geometric information like borders by its scan characteristic (Gu et al. 2021; Gu and Dao 2023). RM is composed of residual mamba blocks (RMBs), which effectively captures global connections for restoration. (f) Rectification for all tasks. Our model treats all tasks as rectification tasks and subsequently incorporates them into a unified framework.

However, we discover that simply applying the TPS model does not perform well on some tasks. By introducing the initial control point locations c_0, we propose the residual progressive TPS (RP-TPS) model to progressively enhance the approximation and avoid intermediate interpolation errors. The flow of our deformation module based on RP-TPS is exhibited in Fig. 2b. Our initial control point locations c_0 are equal to our pre-defined basic control points. Two control point predictors $\mathcal{C}_0, \mathcal{C}_1$ with same structures are designed to predict the deviations of control points based on their inputs. In summary, the overall procedures through RP-TPS are formulated by

$$c_1 = c_0 + \mathcal{C}_0\left(X_0^i, M_0^i\right), X_1^i = \mathcal{S}\left[\mathcal{G}\left(c_1\right); X_0^i\right], \tag{6}$$

$$c_2 = c_1 + \mathcal{C}_1\left(X_1^i, M_1^i\right), X_D^i = \mathcal{S}\left[\mathcal{G}\left(c_2\right); X_0^i\right], \tag{7}$$

where $\mathcal{G}$ is a grid generator which produces a grid $\mathcal{P}$ on the coordinate space of X_0^i with distortions by iterating all points in the rectified image. $\mathcal{S}$ is the differentiable bilinear sampler. Sampling processes in Eq. (6) and Eq. (7) are only from the foremost inputs X_0^i and visual prompts M_0^i, which effectively alleviates the interpolation degradation of the intermediate process.

Prompt Designs. We elaborate different visual prompts for different tasks, which is also a reflection of Eq. (3) and helps the network focus on task-specific distortion. Specifically, for $\mathcal{T}_1$, M_0^1 is the face mask since the face distortion is import for a portrait photo in this task. M_0^2 and M_0^3 indicate the irregular boundaries, thereby enhancing the perception of these areas, and they also contribute to the loss function (Eq. (8)). For rotation correction, the prompt M_0^4 is a white image as the model should pay attention to the whole content of the input image to perceive the subtle rotation distortion.

Network Structures. The structure of Control Point Predictor $\mathcal{C}_{0,1}$ is shown in Fig. 2c. The image $X_{0,1}^i$ and its prompt $M_{0,1}^i$ are concatenated and fed into $\mathcal{C}_{0,1}$ to generate the deviation locations $\Delta c_{0,1} \in \mathbb{R}^{B \times N_c \times 2}$ of control points P' in $X_{0,1}^i$. Inspired by recent burgeoning development of Mamba (Gu et al. 2021; Gu and Dao 2023), a mamba block is introduced in the latent space to scan the geometric information and handle insidious long-range dependencies of distortions. The structure of Restoration Module is shown in Fig. 2d, which is mainly composed of four Residual Mamba Blocks (RMBs). In practice, each RMB has 32 channels in convolution layers and a same mamba block is in the core of RMB

Figure 3: Qualitative comparison for our UniRect on four tasks. Zoom in for best view.

to capture non-local connections effectively like many vision tasks (Liu et al. 2024; Zhu et al. 2024; Guo et al. 2025). X_D^i and M_D^i first pass through a partial convolution layer (Liu et al. 2018) as there may be some irregular borders in X_D^i for boundary-changing task like $\mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_3$ due to the inaccurate estimation of the positions of the control points in the previous module. For $\mathcal{T}_1, \mathcal{T}_4$, the partial convolution is equivalent to a standard convolution layer since the M_D^i is set to all-one matrix.

Loss Functions

For deformation module, we propose a boundary loss $\mathcal{L}_b$ to facilitate the outermost points $P'+$ of the control points P' approaching the boundary. Given a prompt M indicating boundaries (for $\mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_3$), the $\mathcal{L}_b$ can be calculated by $\mathcal{L}_b = \sum_{p' \in P'+} LS\left(p'\right) / |P' +|$, where $LS(\cdot)$ is defined as

$$LS(p') = \begin{cases} +\inf\limits_{q \in \partial M} \|p' - q\|_2, p' \in M_{in} \\ 0, p' \in \partial M \\ +2 \inf\limits_{q \in \partial M} \|p' - q\|_2, p' \in M_{out}, \end{cases} \tag{8}$$

where ∂M is the zero level set. M_{in} and M_{out} represent the inside area and outside area for M. We also use the appearance loss $\mathcal{L}_a$ (*i.e.* L1 loss), line and shape penalty $\mathcal{L}_p$ (Nie et al. 2022; Liao et al. 2023) on the mesh formed by P' to prevent excessive rectification, and the gradient loss $\mathcal{L}_g$ to align local textures. Finally, the all loss $\mathcal{L}_{DM}$ can be expressed by

$$\mathcal{L}_{DM} = \sum_{j=0}^{1} \gamma^j (\mathcal{L}_a^j + \alpha_1 \mathcal{L}_b^j + \alpha_2 \mathcal{L}_p^j + \alpha_3 \mathcal{L}_g^j), \tag{9}$$

where $j = 0$ for X_1^i and $j = 1$ for X_D^i. $\gamma, \alpha_1, \alpha_2, \alpha_3$ are loss weights. Note that we only apply $\mathcal{L}_a$ for $\mathcal{T}_1, \mathcal{T}_4$ since these two tasks have no border changes during rectification. For restoration module, we only use appearance loss and perceptual loss like some super-resolution works (Wang et al. 2018; Johnson, Alahi, and Fei-Fei 2016).

Sparse Mixture-of-Experts

After obtaining four-by-one, we can accommodate four-in-one model. Sharing a same network is meaningful since it would be more convenient to further design model compression and inference acceleration algorithms for deployment in devices with limited computing resources. However, we still find it difficult to jointly train our UniRect on four task datasets since there are heavy **task competitions** and degradations observed in Fig. 4a-d and Tab. 2. This could be attributed to the conflicts in the optimization direction resulting from the complex distortions associated with different tasks. Therefore, we design a Spare Mixture-of-Experts (SMoEs) structure to aggregate our UniRect and circumvent task competitions:

$$SMoEs(X_0^i, M_0^i) = \sum_{j=1}^{5} G(X_0^i)_j E_j(X_0^i, M_0^i), \quad (10)$$

where $G(\cdot)$ is a gating network and E_j is the j-th expert network (*i.e.* our UniRect). This gating network can apportion the specific input task among E_j with a 5-dimension vector weights. Specifically, to reduce computation, G with network weights W_G has a additional top-k operator, enforcing the invalid value to be zero (Shazeer et al. 2017), *i.e.*

$$G(X_0^i) = \text{SoftMax}\left(\text{Top k}\left(X_0^i \cdot W_G, k\right)\right), \\ \text{Topk}(x,k)_j = \begin{cases} x_j, \text{if } x_j \text{ in Topk elements} \\ -\infty, \text{otherwise.} \end{cases} \quad (11)$$

We can train the gating network along with the rest of the model. Gradient flow can also back-propagate through the gating network it inputs. In our experiments, we choose $k = 1$, which can achieve a better balance between prediction and computation.

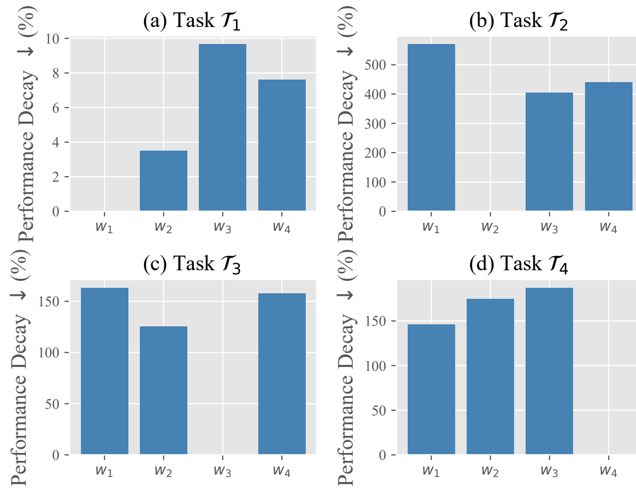

Figure 4: Cross-Task Degradation. Models trained on four tasks (different weights w_i) are evaluated for one task $\mathcal{T}_j$. For $\mathcal{T}_1$, we use LineACC to quantify the performance, and FID is for remaining tasks.

Experiments

Experimental Settings

Datasets. We conduct experiments on public representative benchmarks including portrait correction dataset (Tan et al. 2021), rectified wide-angle image rectangling dataset (Liao et al. 2023), stitched image rectangling dataset (Nie et al. 2022), and rotation correction dataset (Nie et al. 2023).

Implementation Details. Our networks are based on the PyTorch framework with four NVIDIA Tesla V100 GPUs. For single-task learning, it takes one to three days to train our model on each dataset with a batch size of 4 and 200 epochs. For multi-task learning, it takes about 7 days to train on four datasets jointly with a bacth size of 10 and 200 epochs. We adopt the poly policy of the learning rate, the factor of which is 0.96, and the initial learning rate is $1e-5$ for rotation correction and $1e-4$ for the rest tasks. Adam (Kingma and Ba 2014) optimizer is selected and the weight decay is $1e-5$. For RP-TPS, the number of control points is set to 12×10. For Eq. (9), we set $\gamma, \alpha_1, \alpha_2, \alpha_3$ as $0.9, 1e-2, 1.0, 1e-2$ respectively. For SMoEs, we utilize a simple ResNet18 (He et al. 2016) as the backbone of G. For computation complexity, the amounts of parameters (Params), floating point operations (FLOPs), and frames per second (FPS) are 357.9M, 62.98G, and 35.8 respectively. More analysis for computations can be seen in the supplementary files.

Comparison with State-of-the-Art Methods

We adopt the evaluation setting from the previous studies, utilizing the PSNR, SSIM (Wang et al. 2004), FID (Heusel et al. 2017) and LPIPS (Zhang et al. 2018) to assess these methods for $\mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4$. The first two are objective metrics while the following are image perceptual metrics. For portrait correction, Line Straightness Metric (LineACC) and Shape Congruence Metric (ShapeACC) are suggested in (Tan et al. 2021) to measure the correction results of salient lines and faces. Besides, many studies (Zhang et al. 2023; Wu et al. 2024; Xie et al. 2024) have proposed exploiting MLLMs to assess the quality of unlabeled images, which has led to remarkable progress in the field. Owing to the limitations [1] of these two evaluation metrics, we further utilize Multimodal Large Language Models (MLLMs)-based metrics, namely LineACC-LLM and ShapeACC-LLM. By capitalizing on the potent capabilities of MLLMs, our objective is to carry out an assessment of image quality in terms of line straightness and shape congruence. Further details about these metrics can be found in the supplementary files.

Quantitative results are presented in Tab. 1. Our proposed UniRect is compatible for four different types of tasks in a unified network structure and achieves promising results compared with methods for a single task. Specifically, we have made significant progress on rectified wide-angle image stitching, stitched image stitching, and rotation correction. For portrait correction, we also acquire satisfactory results. Moreover, our UniRect offers a novel potential technological path to solve the relevant low-level problems from the distortion rectification angle. Qualitative results are shown in Fig. 3 for four tasks. More results and specific analysis can be seen in the supplementary files.

Multi-Task Learning Strategy

Task Competitions. We investigate task competitions and showcase the efficacy of our proposed SMoEs strategy when applied to unified networks, *i.e.* our UniRect, and the exper-

[1] In the test set of this dataset, only the coordinates of the corresponding points in the reference and input distorted images are provided for metric calculation. Nevertheless, in our approach, because of the characteristics of the TPS model (irreversibility), it is arduous to acquire the precise coordinates of the key points on the distorted image after rectification, as depicted in Fig. 2f. Hence, we resort to an estimation method, resulting in a substantial reduction in the metrics of us.

Tasks	Methods	Line-ACC ↑	Shape-ACC ↑	LineACC-LLM ↑	ShapeACC-LLM ↑
Portrait correction $\mathcal{T}_1$	Shih, Lai, and Liang (2019)	66.143	97.253	-	-
	Tan et al. (2021)	66.784	97.490	-	-
	Zhu et al. (2022)	**66.825**	**97.491**	7.082	8.082
	UniRect (Ours)	66.523	97.454	**7.168**	**8.152**
Tasks	**Methods**	**PSNR ↑**	**SSIM ↑**	**FID ↓**	**LPIPS ↓**
Rectified wide-angle image rectangling $\mathcal{T}_2$	ROP (Liao et al. 2021)	13.90	0.3516	-	-
	RecRecNet (Liao et al. 2023)	18.68	0.5450	**19.01**	**0.1136**
	UniRect (Ours)	**19.90**	**0.5721**	27.02	0.1245
Stitched image rectangling $\mathcal{T}_3$	He, Chang, and Sun (2013b)	14.70	0.3775	38.19	0.2846
	Nie et al. (2022)	21.28	0.7141	21.77	0.1557
	MOWA (Liao et al. 2025)	20.42	0.6307	-	-
	UniRect (Ours)	**25.10**	**0.7526**	**19.59**	**0.1120**
Rotation correction $\mathcal{T}_4$	He, Chang, and Sun (2013a)	21.69	0.6460	8.51	0.2120
	DRC (Nie et al. 2023)	21.02	0.6280	7.12	0.2050
	CoupledTPS (Nie et al. 2024)	22.29	0.6790	7.90	0.1970
	UniRect (Ours)	**23.16**	**0.7179**	**6.55**	**0.0873**

Table 1: The quantitative results of the proposed UniRect and other solutions of different tasks on five datasets. In the test dataset of $\mathcal{T}_1$, no image ground-truth is provided. Metrics such as PSNR, SSIM, FID, and LPIPS are not applicable to this task. Due to these limitations, LLM-based metrics are also introduced. The best performance is in **bold**.

Strategy	$\mathcal{T}_1$ (ShapeACC↑)	$\mathcal{T}_2$ (PSNR↑)	$\mathcal{T}_3$ (PSNR↑)	$\mathcal{T}_4$ (PSNR↑)	Parms
ML	97.223	13.07	15.74	21.73	357.9M
SL (1-4-3-2)	97.401	15.23	16.94	21.51	357.9M
SL (2-3-4-1)	97.375	15.72	16.59	20.73	357.9M
SL (3-2-4-1)	**97.409**	17.54	23.55	15.10	357.9M
SL (4-1-3-2)	97.385	15.00	16.81	22.93	357.9M
UniRect (SMoEs)	97.390	**19.90**	**25.07**	**23.16**	369.6M

Table 2: Studies on learning strategy. For sequential learning, the number means the sequence in which training tasks are added. We have trained model using a longer 250 epochs for mixed learning and all sequential learning strategies. 'ML' denotes mixed learning and 'SL' represents sequential learning. The number after the 'SL' is task order.

iments are presented in Fig. 4a-d and Tab. 2. As can be observed from the experiments, learning multiple tasks leads to a degradation in the performance of any individual task. In other words, directly using a model trained on other tasks can result in significant performance degradation, as different tasks often involve distinct types of distortions. We analyze this phenomenon in the supplementary files from the perspectives of distortion and data distribution.

ML vs. SL vs. SMoEs. Mixed learning (ML) is a simple approach to multi-task learning, which involves training a model using all task datasets simultaneously. Still, we find learning multiple tasks does not contribute to an overall improvement in performance. Another strategy is sequential learning (SL), aiming at providing good starting points for subsequent tasks. However, the first task in sequential learning always becomes a predominant task. Thus, we propose employing SMoEs to adaptively switch different tasks, which can mitigate performance degradation when handling multi-task learning. Our SMoEs are nearly close to the single task learning for an effective gating network. More details and analysis can be seen in the supplementary files.

Strategy	$\mathcal{T}_1$ (↑)	$\mathcal{T}_2$ (↑)	$\mathcal{T}_3$ (↑)	$\mathcal{T}_4$ (↑)
w/o prompts	96.324	19.84	23.82	23.15
w prompts	**97.454**	**19.90**	**25.10**	**23.16**

Table 3: Results of the model with and without prompts. $\mathcal{T}_1$ uses ShapeACC, while the remaining tasks uses PSNR.

Task	RP-TPS		Metrics			
	w/o	w	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
$\mathcal{T}_5$	✓		21.56	0.6502	7.21	0.0972
		✓	**23.16**	**0.7179**	**6.55**	**0.0873**

Table 4: The ablation study on the influence of RP-TPS.

Ablation Studies

We evaluate the components of UniRect on diverse tasks. More ablation experiments are in the supplementary files.

Prompts Designs. According to our experiments in Tab. 3, without prompt, the results of task $\mathcal{T}_1$ and task $\mathcal{T}_3$ were significantly reduced. They have a slight impact on task $\mathcal{T}_2$ and no impact on task $\mathcal{T}_4$. Therefore, prompts have a significant effect on the network performance. **Why not Task IDs?** The

Task	Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓	Param ↓
$\mathcal{T}_4$	CNN	24.12	0.7123	21.03	0.1250	508.9M
	Transformer	24.23	0.7226	21.27	0.1320	1.168G
	Mamba	**25.10**	**0.7526**	**19.59**	**0.1120**	**357.9M**
$\mathcal{T}_5$	CNN	22.31	0.6796	7.20	0.0972	508.9M
	Transformer	**23.23**	0.6884	8.59	0.1546	1.168G
	Mamba	23.16	**0.7179**	**6.55**	**0.0873**	**357.9M**

Table 5: The ablation study on the architectures of pure convolutions (CNN), Transformer, and Mamba.

Task	Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
$\mathcal{T}_2$	w/o RM	**19.99**	**0.5791**	28.69	0.1296
	w RM	19.90	0.5721	**12.51**	**0.1245**
$\mathcal{T}_4$	w/o RM	22.84	0.6901	7.88	0.1560
	w RM	**23.16**	**0.7179**	**6.55**	**0.0873**

Table 6: Study on the role of our restoration module (RM).

original datasets for $\mathcal{T}_1$ and $\mathcal{T}_3$ already have masks: one indicating the face location and the other indicating the effective area. Thus, prompts are designed for other tasks to align input channels. Visual prompt is of paramount significance for our network as it takes part in the calculation of the loss function and can also be utilized to determine and even control the rectification position. However, task-ID is hard to indicate the location and enhance the performance.

RS-TPS. We tested the validity of the RP-TPS model as seen in Tab. 4. With the help of it, distortion can be estimated more accurately, leading to a better performance. We also discover that recursively repeating $\mathcal{G}$ and $\mathcal{C}$ resulted in no obvious improvement and introduce more parameters due to the matrix inversion operator in the recursive manner.

Comparisons with CNN/Transformer. Mamba has shown impressive performance on many vision tasks due to its long-range modeling capability and efficient computing complexity. Thus, we also try to explore its ability of modeling geometric information in this paper. The ablation results are shown in Tab. 5. Compared with CNN and Transformer, Mamba is a strong choice for balancing both accuracy and computational efficiency.

Restoration Module. The contribution of our restoration module differs in different tasks (Tab. 6). For $\mathcal{T}_4$, our restoration module can greatly improve the final image quality and compensate for the sampling degradation in the deformation module. However, for $\mathcal{T}_2$, our deformation module dominates the improvement. Actually, this module is extremely important, since the restoration module cannot learn accurate mappings if the former module gives wrong and chaotic pixel locations.

More Applications

We found that the introduction of prompts enables the implementation of some interesting applications. Fig. 5 represents a scenario in which two forms of distortion are concurrently present within a single image. By providing different prompts for the same input under different tasks, our model

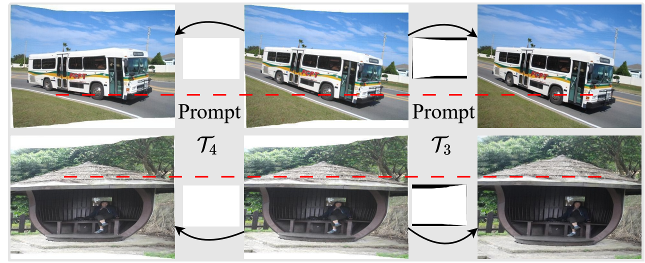

Figure 5: Results of the same input under $\mathcal{T}_3$ and $\mathcal{T}_4$ prompts.

can accomplish the corresponding tasks and achieve controllable rectification of images (designated tasks and areas). Furthermore, the cross-domain results in real-world scenes can be found in the supplementary files. This adequately demonstrates the generalization capability of our model.

Conclusion

In this paper, we study four tasks from a novel consistent distortion perspective. We build a general distortion model to unify different photography distortion problem, and propose a unified rectification framework based on RP-TPS to handle these tasks in the same structure designs. Moreover, we introduce sparse MoEs to address task competitions of multi-task learning. It is also intriguing to consider another complex scenario in which a model can apply multiple rectifications to the same sample simultaneously. Since all existing datasets contain only one type of distortion, conducting an experimental verification at this stage is quite challenging. We look forward to exploring this highly interesting issue in our subsequent work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62475006 and 62125102, in part by the National Key Research and Development Program of China under Grant 2022ZD0160401.

References

Bookstein, F. 1989. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6): 567–585.

Bragman, F. J.; Tanno, R.; Ourselin, S.; Alexander, D. C.; and Cardoso, J. 2019. Stochastic filter groups for multi-task cnns: Learning specialist and generalist convolution kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1385–1394.

Gao, Y.; Ma, J.; Zhao, M.; Liu, W.; and Yuille, A. L. 2019. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3205–3214.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2025. Mambair: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, 222–241. Springer.

Guo, M.; Haque, A.; Huang, D.-A.; Yeung, S.; and Fei-Fei, L. 2018. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision*, 270–287.

He, K.; Chang, H.; and Sun, J. 2013a. Content-aware rotation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 553–560.

He, K.; Chang, H.; and Sun, J. 2013b. Rectangling panoramic images via warping. *ACM Transactions on Graphics (TOG)*, 32(4): 1–10.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, 694–711. Springer.

Kannala, J.; and Brandt, S. S. 2006. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8): 1335–1340.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7482–7491.

Kent, J.; and Mardia, K. 1994. The link between kriging and thin-plate splines. *Probability, Statistics and Optimization*, 326–339.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, X.; Zhang, B.; Sander, P. V.; and Liao, J. 2019. Blind geometric distortion correction on images through deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4855–4864.

Liao, K.; Lin, C.; Wei, Y.; Li, F.; Yang, S.; and Zhao, Y. 2021. Towards complete scene and regular shape for distortion rectification by curve-aware extrapolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14569–14578.

Liao, K.; Nie, L.; Lin, C.; Zheng, Z.; and Zhao, Y. 2023. RecRecNet: Rectangling rectified wide-angle images by thin-plate spline model and DoF-based curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10800–10809.

Liao, K.; Yue, Z.; Wu, Z.; and Loy, C. C. 2025. MOWA: Multiple-in-One Image Warping Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Proceedings of the European Conference on Computer Vision*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2022. Deep rectangling for image stitching: a learning baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5740–5748.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2023. Deep rotation correction without angle prior. *IEEE Transactions on Image Processing*.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2024. Semi-Supervised Coupled Thin-Plate Spline Model for Rotation Correction and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13.

Potlapalli, V.; Zamir, S. W.; Khan, S. H.; and Shahbaz Khan, F. 2024. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36.

Qiu, L.; Xie, F.; Liu, C.; Wang, K.; Song, X.; and Shi, Z. 2024. Remote Sensing Image Rectangling with Iterative Warping Kernel Self-correction Transformer. *IEEE Transactions on Geoscience and Remote Sensing*.

Ruder, S.; Bingel, J.; Augenstein, I.; and Søgaard, A. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 4822–4829.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shih, Y.; Lai, W.-S.; and Liang, C.-K. 2019. Distortion-free wide-angle portraits on camera phones. *ACM Transactions on Graphics (TOG)*, 38(4): 1–12.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Tan, J.; Zhao, S.; Xiong, P.; Liu, J.; Fan, H.; and Liu, S. 2021. Practical wide-angle portraits correction with deep structured models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3498–3506.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, 402–419. Springer.

Teterwak, P.; Sarna, A.; Krishnan, D.; Maschinot, A.; Belanger, D.; Liu, C.; and Freeman, W. T. 2019. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10521–10530.

Von Gioi, R. G.; Jakubowicz, J.; Morel, J.-M.; and Randall, G. 2008. LSD: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4): 722–732.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Loy, C. C.; Qiao, Y.; and Tang, X. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision Workshops*.

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: 600–612.

Weng, J.; Cohen, P.; Herniou, M.; et al. 1992. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on pattern analysis and machine intelligence*, 14(10): 965–980.

Wu, T.; Ma, K.; Liang, J.; Yang, Y.; and Zhang, L. 2024. A comprehensive study of multimodal large language models for image quality assessment. In *Proceedings of the European Conference on Computer Vision*, 143–160. Springer.

Xie, Z.; Lai, X.; Zhao, W.; Liu, X.; and Hou, W. 2024. Modification Takes Courage: Seamless Image Stitching via Reference-Driven Inpainting. *arXiv preprint arXiv:2411.10309*.

Xue, Z.; Xue, N.; Xia, G.-S.; and Shen, W. 2019. Learning to calibrate straight lines for fisheye image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1643–1651.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14821–14831.

Zhan, F.; and Lu, S. 2019. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2059–2068.

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/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, X.; Lu, Y.; Wang, W.; Yan, A.; Yan, J.; Qin, L.; Wang, H.; Yan, X.; Wang, W. Y.; and Petzold, L. R. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Zhao, Q.; Zhu, C.; Dai, F.; Ma, Y.; Jin, G.; and Zhang, Y. 2018. Distortion-aware CNNs for Spherical Images. In *IJCAI*, 1198–1204.

Zhou, E.; Cao, Z.; and Sun, J. 2018. Gridface: Face rectification via learning local homography transformations. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Zhou, T.; Li, H.; Wang, Z.; Luo, A.; Zhang, C.-L.; Li, J.; Zeng, B.; and Liu, S. 2024. RecDiffusion: Rectangling for Image Stitching with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2692–2701.

Zhu, F.; Zhao, S.; Wang, P.; Wang, H.; Yan, H.; and Liu, S. 2022. Semi-supervised wide-angle portraits correction by multi-scale transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19689–19698.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zhuang, B.; Tran, Q.-H.; Ji, P.; Cheong, L.-F.; and Chandraker, M. 2019. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4551–4560.