Geometry Grounding: Elevating Blind Distortion Correction with 3D Structural Priors

Xiang Li[1], Weimin Shi[1,2], Qichuan Geng[3] *, and Zhong Zhou[1,2]*

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] Zhongguancun Laboratory, Beijing, China
{lx_7342, shiwm, zz}@buaa.edu.cn
[3] The Information Engineering College, Capital Normal University, Beijing, China
gengqichuan1989@cnu.edu.cn

Abstract. Geometric distortion breaks the projective mapping and geometric relations in images, which degrades the performance and reliability of computer vision tasks. Although deep network-based methods for single-image distortion correction have achieved remarkable progress, most existing approaches optimize only for 2D pixel-level alignment, such as optical flow losses or image reconstruction errors. Without explicit 3D geometric constraints, such models may produce visually plausible rectifications yet fail to enforce 3D-consistent structure. To address this limitation, this paper proposes a 3D geometry guided framework for single image distortion correction, named Geometry-Grounded Distortion Correction (G2DC). G2DC introduces a pretrained 3D foundation model to extract multi-dimensional geometric representations as supervision, and incorporates projective geometry priors into feature learning. Specifically, G2DC adds a geometric constraint scheme and jointly optimizes a structural alignment loss and a physical alignment loss. The former minimizes feature mismatch to ensure that the corrected output is structurally close to the undistorted scene. The latter enforces strict 3D constraints using the predicted point map, depth topology, and camera parameter regularization, so the model preserves 3D structure consistency and plausibility while restoring 2D appearance. Experiments on several widely used distortion benchmark datasets show that G2DC outperforms state-of-the-art methods, especially on metrics that are sensitive to 3D geometric consistency. Project webpage: https://gitee.com/VR_NAVE/g2dc.git.

Keywords: Distortion Correction · 3D Vision · Geometric Grounding

1 Introduction

Geometric distortion represents a persistent and fundamental challenge in digital imaging. It primarily arises from the physical mapping deviations of non-linear optical systems, such as radial and tangential distortions, and the diversification

* Corresponding authors.

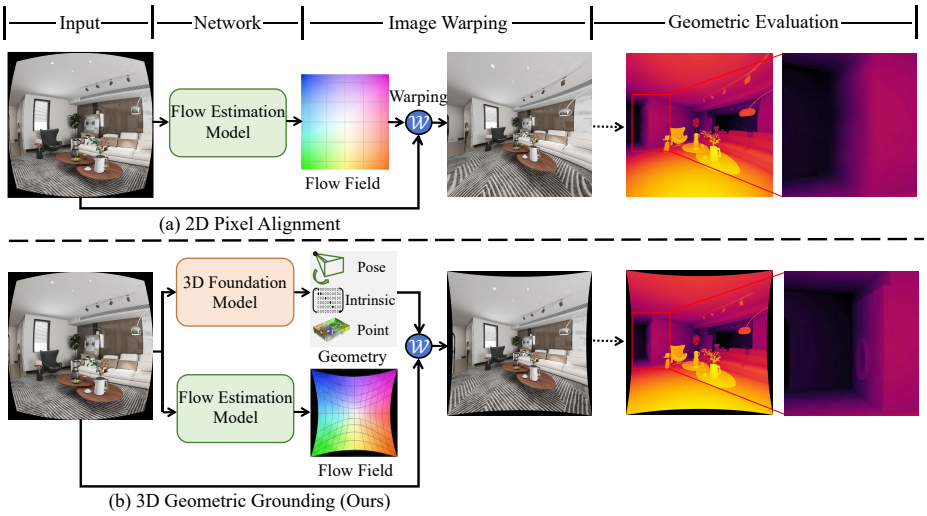

Fig. 1: Comparison between conventional 2D rectification and the proposed Geometry-Grounded Distortion Correction (G2DC) framework. (*a*) 2D Pixel Alignment: Existing methods regress a displacement flow ϕ to warp the distorted input I_d into the rectified output I_r via a warping operation $\mathcal{W}$. However, relying solely on surface-level 2D appearance mapping frequently leads to severe 3D structural collapse, as evidenced by the non-physical stretching and fractured topology in the predicted depth map $\mathbf{D}$ (*highlighted in the red box*). (*b*) 3D Geometric Grounding (Ours): The G2DC framework introduces a frozen 3D foundation model to enforce explicit geometric constraints during training. By transitioning to deep structural consistency modeling, the proposed approach preserves essential 3D properties, resulting in physically valid rectifications with logical depth transitions.

of imaging perspectives, including perspective and shear distortions. These distortions violate the geometric relationships of ideal imaging models and induce the non-linear warping of object shapes, thereby compromising the performance and reliability of fundamental computer vision tasks. For instance, distorted images interfere with the regression of bounding boxes in object detection [11, 20], lead to the misalignment of pixel classification in semantic segmentation [3, 6], and undermine the geometric constraints essential for 3D pose estimation [2, 21].

Since the digital imaging process adheres to the rigorous laws of projective geometry, early methods for distortion correction typically relied on camera calibration to inversely solve for distortion parameters [14, 24], or utilized structural cues extracted from images, such as lines, curves, and vanishing points, for automated rectification [7, 32]. Benefiting from mature photogrammetric models, these methods generally achieve high precision under controlled conditions. However, the dependence on preset calibration patterns and explicit geometric cues limits the applicability and robustness of these methods in scenarios where only a single image is provided without reliable structural constraints.

In recent years, researchers have increasingly employed deep neural networks trained via supervised learning to address the problem of single-image rectification. These methods leverage the capability of deep neural networks to mine rich geometric and semantic cues from large-scale data. Existing approaches fall into two primary categories: regression-based methods and generation-based methods [29, 30]. The former predict distortion parameters and parse them into pixel-level displacement fields, whereas the latter directly synthesize rectified images through end-to-end architectures.

Despite these advancements, the prohibitive cost of acquiring precise geometric ground truth in real-world scenes makes the introduction of explicit geometric supervision difficult during training. Consequently, the optimization objectives of existing methods typically focus on two-dimensional pixel alignment, such as L_1 and L_2 flow losses or image reconstruction errors. The resulting rectifications reflect pixel rearrangements at the appearance level. While these results appear natural in texture, they fail to recover the underlying 3D scene geometry. To verify this issue, we integrate a DPT prediction head into a state-of-the-art rectification model to map intermediate features to corresponding depth maps. As illustrated in Fig. 1, features learned solely from raw data and two-dimensional reconstruction targets struggle to generate meaningful geometric representations, often leading to severe structural collapse and fractured topology. This observation indicates that rectification models trained without explicit 3D supervision remain deficient in fundamental geometric understanding.

To address this problem, we propose the Geometry-Grounded Distortion Correction (G2DC) framework for single-image rectification, which introduces projective geometry priors into the optimization process. Specifically, G2DC incorporates geometric representations from a pre-trained 3D-aware foundation model, such as VGGT, as auxiliary supervision signals, injecting projective geometry priors into feature learning and guiding the network from surface-level pixel alignment to structural consistency modeling. The framework jointly optimizes two complementary objectives: the Structural Alignment and the Physical Alignment. The former enforces structural faithfulness by minimizing feature discrepancies between the rectified image and the ground-truth undistorted image in a high-dimensional latent space, while the latter imposes explicit physical constraints using predicted point maps, depth topology, and camera parameters. Unlike prior work that relies on hand-crafted geometric primitives (e.g., plumb-line, arc, and line cues) and often fails in texture-less scenes, G2DC introduces constraints derived from dense 3D model priors without requiring primitive detection, thereby enforcing all three constraints. These constraints jointly regularize rectification under projection relationships and local structural continuity, encouraging the predicted displacement field to maintain 3D geometric consistency while restoring 2D appearance.

The primary contributions are summarized as follows. First, we propose the G2DC framework, a novel approach that integrates explicit 3D geometric guidance to effectively overcome the inherent 2D-centric limitations of previous methods. Second, we design a comprehensive geometric alignment supervision scheme

consisting of two synergistic components: the Structural Alignment Loss and the Physical Alignment Loss. This scheme innovatively utilizes the multi-dimensional outputs of the VGGT model to impose explicit physical and topological constraints. Finally, extensive experimental results on various distortion benchmark datasets demonstrate that the proposed framework significantly outperforms existing state-of-the-art methods, particularly achieving superior performance on evaluation metrics sensitive to high-precision 3D geometric consistency.

2 Related Work

The field of distortion rectification has undergone a systematic transition from traditional geometry-driven techniques to advanced architectures based on deep learning. This section reviews the evolution of traditional solutions and discusses contemporary approaches based on representation learning, emphasizing the shift from explicit parameter estimation to dense mapping, as well as the emerging necessity for 3D structural grounding.

2.1 Traditional Distortion Correction

Early research on the rectification of lens distortion is primarily grounded in established geometric primitives and optical models. Conventional techniques [14, 23,24,33] estimate lens parameters by capturing multiple views of planar calibration targets to obtain closed-form solutions. While these methods provide high precision, they require specialized calibration equipment and assume a static environment that is pre-calibrated. To mitigate the need for external targets, self-calibration strategies [8, 13] have been introduced. However, these approaches often necessitate multiple frames or specific camera motions, which limits the efficacy of these methods for the correction of single images.

To achieve automated rectification, researchers have explored the utilization of geometric invariants. Bukhari and Jawahar [5] utilized a single-parameter lens model [10] to extract circular arcs for the estimation of distortion. Similarly, Aleman-Flores et al. [1] leveraged the plumb-line constraint, which assumes that straight lines in 3D space must project as straight lines in an undistorted image. Despite the theoretical elegance of these automated methods, they rely heavily on the presence of salient artificial features, such as clear lines or arcs. Consequently, the performance of these methods often degrades in natural or texture-less scenes where such structural cues are unavailable.

2.2 Learning-Based Distortion Correction

The development of deep learning has transformed the field by enabling robust representation learning without requiring manual feature detection. Rong et al. [22] initiated this shift by formulating the estimation of distortion parameters as a classification task based on convolutional neural networks. Subsequent studies, such as DeepCalib [4] and FisheyeRecNet [31], extended this approach

by predicting focal lengths or incorporating high-level semantic information. To reconcile pixel-level accuracy with geometric consistency, Xue et al. [27] introduced line-aware constraints, while Li et al. [15] proposed a unified framework for diverse geometric distortions.

Recent methodologies have shifted toward model-free rectification and iterative refinement. Liao et al. [17] introduced distortion distribution maps for flexible rectification, and the DR-GAN model [18] utilized generative adversarial networks to reduce visual artifacts. Progressive frameworks, such as PCN [30], and multi-level curriculum strategies, such as MLC [16], have further stabilized the rectification process. Current state-of-the-art approaches integrate advanced priors; for instance, SimFIR [9] employs self-supervised representation learning, RDTR [26] combines geometric priors with backward warping, and DDA [29] adopts a dual diffusion framework.

As indicated by the analysis of the DPT prediction head, these methods primarily optimize for two-dimensional pixel alignment and lack explicit grounding in the 3D structural consistency of the scene. This limitation often results in rectifications that, while visually plausible in terms of texture, fail to preserve the underlying 3D geometric consistency. The proposed G2DC framework addresses this gap by incorporating the Structural Alignment Loss and the Physical Alignment Loss to enforce higher-dimensional geometric constraints.

3 Preliminaries

This section establishes the mathematical formulation for geometric distortion and introduces the 3D foundation model that provides authoritative structural guidance for the proposed rectification framework.

3.1 Geometric Distortion Model

Lens distortion, particularly in wide-angle and fisheye lenses, originates from the non-linear mapping between 3D scene points and the 2D image plane. While the ideal pinhole model adheres to the relationship $d = f \tan\theta$, where d denotes the distance from the principal point and θ represents the incidence angle, real-world optical systems often deliberately introduce non-linearity to achieve a larger field of view.

To model this phenomenon without the computational complexity of precise angle calculations, we adopt the polynomial model [33], which is widely utilized in blind distortion correction due to its numerical stability. The relationship between a distorted coordinate $\mathbf{p}_d = (x', y')^\top$ and the corresponding rectified coordinate $\mathbf{p}_r = (x, y)^\top$ is defined as follows:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \left(1 + \sum_{n=1}^{N} \lambda_n (r')^{2n}\right) \begin{bmatrix} x' \\ y' \end{bmatrix} \tag{1}$$

In this formulation, λ_n represents the distortion coefficients that determine the degree of deformation, and r' is the distortion radius calculated as the Euclidean distance from (x', y') to the distortion center (x_c, y_c). In the G2DC framework, the objective of the rectification network is to predict a dense displacement field ϕ such that for every pixel in the distorted image I_d, its undistorted position can be identified. The rectified image I_r is subsequently obtained via a differentiable warping operation defined as $I_r = \mathcal{W}(I_d, \phi)$.

3.2 Visual Geometry Grounded Transformer

The Visual Geometry Grounded Transformer (VGGT) [25] serves as the geometric cornerstone of the proposed approach. Unlike traditional pipelines for structure-from-motion that require iterative optimization, VGGT is a feed-forward foundation model capable of directly regressing comprehensive 3D attributes from one or more 2D views.

The architecture of VGGT consists of a Transformer-based backbone utilizing an alternating attention mechanism. This mechanism interleaves frame-wise self-attention to capture intra-frame spatial structures with global self-attention to model contextual geometry. Given a rectified image I_r as input, the model generates a unified latent representation, which is then processed by specialized prediction heads to output three primary components. First, it produces a point map ($\mathbf{X}$), which is a dense 3D coordinate map representing the spatial position of each pixel in camera coordinates. Second, it generates a depth map ($\mathbf{D}$) representing the canonical z-buffer depth, signifying the distance of scene elements from the optical center. Finally, it estimates camera parameters ($\mathbf{K}$), including intrinsic parameters such as focal length and the principal point. In the G2DC framework, both the high-dimensional latent features from the Transformer backbone and the explicit 3D outputs $(\mathbf{X}, \mathbf{D}, \mathbf{K})$ are utilized as robust regulators to ensure the geometric integrity of the rectification process.

4 Methodology

This section describes the detailed architecture of Geometry-Grounded Distortion Correction (G2DC). In contrast to conventional methods that treat rectification as a standard two-dimensional image-to-image translation task, the proposed framework internalizes the underlying 3D structural priors of the physical world. This internalization is achieved by leveraging a frozen geometric foundation model as an authoritative regulator to guide the learning process.

4.1 Framework Overview

The objective of G2DC is to learn a non-linear mapping that transforms a distorted image $I_d \in \mathbb{R}^{H \times W \times 3}$ into a rectified perspective image I_r. As illustrated in Fig. 2, the pipeline consists of two primary components: a trainable rectification network and a frozen geometric regulator based on the Visual Geometry Grounded Transformer (VGGT).

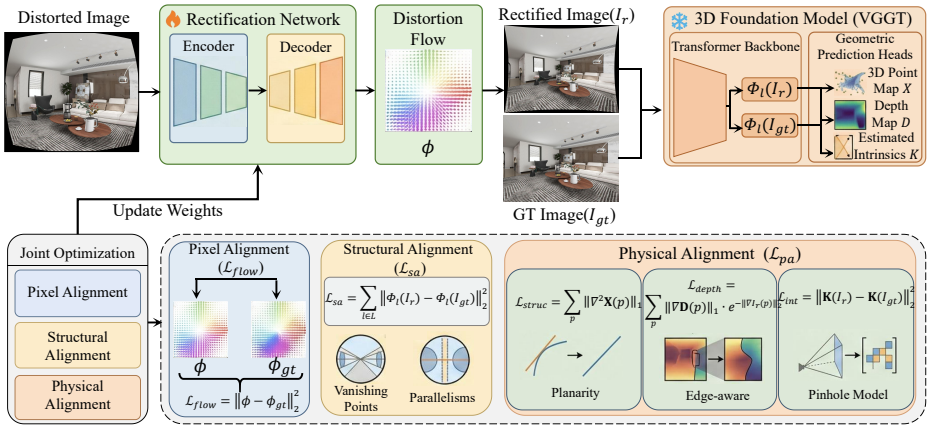

Fig. 2: Overview of the proposed Geometry-Grounded Distortion Correction (G2DC) framework. The trainable Rectification Network regresses a dense distortion flow ϕ to warp the input image. To transcend simple 2D pixel alignment ($\mathcal{L}_{\text{flow}}$), both the rectified image and ground truth are fed into a frozen 3D foundation model (VGGT). We jointly optimize a Structural Alignment Loss ($\mathcal{L}_{\text{sa}}$) to enforce feature consistency within a high-dimensional geometric latent space, and a Physical Alignment Loss ($\mathcal{L}_{\text{pa}}$) leveraging VGGT's specialized prediction heads to impose explicit 3D constraints: collinearity preservation ($\mathcal{L}_{\text{struc}}$), depth topology consistency ($\mathcal{L}_{\text{depth}}$), and intrinsic parameter regularization ($\mathcal{L}_{\text{int}}$).

The rectification network utilizes an encoder-decoder architecture as the backbone. The encoder extracts multi-scale semantic and geometric cues from the distorted input I_d, while the decoder regresses a dense displacement field, or flow field, denoted as $\phi \in \mathbb{R}^{H \times W \times 2}$. To avoid the blurry artifacts often associated with direct pixel prediction, the network predicts the pixel-wise mapping $\phi(p) = p_d - p$, where p represents the coordinate in the rectified plane. The final rectified image is generated via a differentiable bilinear warping operation:

$$I_r(p) = I_d(p + \phi(p)) \tag{2}$$

To bridge the gap between two-dimensional pixel alignment and 3D physical reality, the rectified output I_r and the corresponding ground-truth I_{gt} are processed by the frozen VGGT, which provides a geometric audit of the rectification quality. The total training objective is formulated as a multi-task optimization problem:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_1 \mathcal{L}_{\text{sa}} + \lambda_2 \mathcal{L}_{\text{pa}}, \tag{3}$$

where $\mathcal{L}_{\text{flow}} = \|\phi - \phi_{gt}\|_2^2$ ensures accurate displacement prediction. The terms $\mathcal{L}_{\text{sa}}$ and $\mathcal{L}_{\text{pa}}$ represent the Structural Alignment and the Physical Alignment, respectively. To balance the $\sim$10$\times$ raw-scale gap between $\mathcal{L}_{sa}$ and $\mathcal{L}_{pa}$, we simply set λ_1 and λ_2 to 1.0 and 0.1, without bells and whistles.

4.2 Structural Alignment

While low-level flow losses ensure local pixel proximity, they frequently fail to preserve the global structural coherence necessary for high-quality rectification. We introduce the Structural Alignment Loss ($\mathcal{L}_{\text{sa}}$) to enforce consistency within the high-dimensional latent manifold of the foundation model.

Let $\Phi_l(\cdot)$ denote the latent feature representations extracted from the l-th layer of the frozen Transformer backbone of VGGT. Since deeper encoder features encode higher-level geometric structures while shallow features mainly capture local textures, we use the last VGGT encoder layer for supervision. The structural alignment loss is defined as:

$$\mathcal{L}_{\text{sa}} = \|\Phi_{\text{last}}(I_r) - \Phi_{\text{last}}(I_{gt})\|_2^2. \tag{4}$$

The rationale for utilizing features from VGGT, rather than standard architectures such as VGG or ResNet, lies in the geometric task-specificity of the model. Conventional perceptual losses focus primarily on texture and style; in contrast, VGGT is pre-trained on 3D primitives, including point tracking and spatial occupancy. Consequently, $\mathcal{L}_{\text{sa}}$ compels the rectification network to internalize the geometric essence of the scene, ensuring that rectified structures, such as vanishing points and parallelisms, remain consistent with a perspective view.

4.3 Physical Alignment

The cornerstone of the proposed approach is the explicit grounding of the rectification process in 3D space through the Physical Alignment Loss ($\mathcal{L}_{\text{pa}}$). By utilizing the specialized prediction heads of VGGT, we impose three types of physical constraints, which are jointly formulated as $\mathcal{L}_{\text{pa}} = \mathcal{L}_{\text{struc}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{int}}$.

Structural integrity ($\mathcal{L}_{\text{struc}}$) is maintained by addressing the preservation of collinearity, a fundamental property of perspective projection where 3D straight lines must project as 2D straight lines. This property is typically violated in distorted images. We utilize the 3D point map $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ predicted by VGGT to penalize local bending of the scene geometry via a curvature-based constraint:

$$\mathcal{L}_{\text{struc}} = \sum_p \|\nabla^2 \mathbf{X}(p)\|_1 \tag{5}$$

where ∇^2 denotes the discrete Laplacian operator. By minimizing the second-order spatial gradients of the 3D coordinates, the model is encouraged to recover the planarity and collinearity of scene surfaces.

Depth topology consistency ($\mathcal{L}_{\text{depth}}$) is enforced to ensure the rectified image possesses a valid 3D topology. Geometric distortion often introduces non-physical stretching and shearing. We employ an edge-aware smoothness loss on the predicted depth map $\mathbf{D} \in \mathbb{R}^{H\times W\times 1}$ to ensure that depth transitions are logical and aligned with image boundaries:

$$\mathcal{L}_{\text{depth}} = \sum_p \|\nabla \mathbf{D}(p)\|_1 \cdot e^{-\|\nabla I_r(p)\|_2} \tag{6}$$

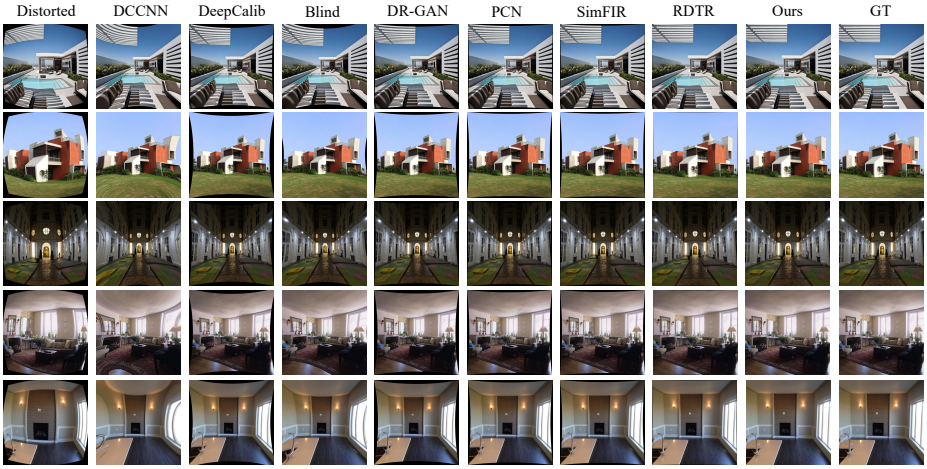

Fig. 3: Qualitative comparisons of image rectification. Compared with existing methods that still exhibit residual curvature along structural edges, our G2DC effectively restores straight lines and improves geometric consistency.

This loss prevents artifacts where the underlying spatial continuity is fractured despite pixel alignment. Finally, intrinsic parameter regularization ($\mathcal{L}_{\text{int}}$) ensures the output I_r adheres to the pinhole model. This constraint is defined as $\mathcal{L}_{\text{int}} = ||\mathbf{K}_{I_r} - \mathbf{K}_{gt}||_2^2$, where $\mathbf{K} = [f_x, f_y, c_x, c_y]$ represents the estimated focal lengths and principal point. This regularization ties the dense displacement field to a globally consistent projection model, ensuring the rectified imagery is geometrically valid for downstream 3D vision tasks.

To reduce noisy VGGT supervision, we adopt warm-up and loss truncation. Warm-up first stabilizes the rectifier with flow supervision, while truncation prevents high-loss samples from dominating the update. This limits the influence of unreliable VGGT signals during optimization.

5 Experiments

This section evaluates the effectiveness of the proposed Geometry-Grounded Distortion Correction (G2DC) framework. The experimental setup is detailed first, followed by comprehensive quantitative and qualitative comparisons with existing state-of-the-art methods across standard and perceptual metrics. Subsequently, the computational efficiency of the framework is analyzed. Finally, ablation studies are conducted to validate the contribution of each geometric constraint within the proposed optimization scheme.

Table 1: Quantitative comparison with state-of-the-art methods using standard metrics on the test set.

Comparison		Metrics					
Methods	Type	PSNR ↑	SSIM ↑	MS-SSIM ↑	FID ↓	LPIPS-Alex ↓	LPIPS-Vgg ↓
DCCNN [22]	Regression	15.12	0.48	0.38	191.5	0.292	0.344
DeepCalib [4]	Regression	20.84	0.78	0.80	72.3	0.141	0.198
Blind [15]	Generation	17.52	0.65	0.73	148.6	0.315	0.328
DR-GAN [18]	Generation	23.15	0.82	0.87	62.1	0.128	0.182
PCN [30]	Generation	24.56	0.85	0.89	57.5	0.112	0.158
SimFIR [9]	Generation	22.51	0.86	0.88	25.8	0.088	0.124
RDTR [26]	Generation	27.92	0.93	0.94	21.4	0.075	0.116
G2DC	**Generation**	**28.68**	**0.95**	**0.96**	**19.8**	**0.066**	**0.105**

5.1 Implementation Setup

The evaluation of the proposed framework follows the established protocols from prior studies [22, 26, 30]. The distorted image datasets are derived from two primary sources: the Wireframe dataset [12] and the Places365 dataset [35].

Specifically, 5,000 images are selected for training and 462 images for testing from the Wireframe dataset. Following the experimental configuration of previous studies [22], additional images are randomly sampled from the Places365 dataset to enlarge scene diversity. In total, the constructed dataset contains 55,000 training images and 6,000 testing images.

To synthesize distorted samples, we adopt both the division and polynomial distortion models, using the first four parameters of the polynomial model together with the division model to generate diverse distortion profiles. This mixed distortion strategy, combined with training on Places365 and Wireframe datasets to cover both natural scenes and structural cues, provides broader coverage of real-world lens distortions and improves the robustness of the learned rectification model.

In the context of image distortion correction, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) serve as the primary evaluation metrics. PSNR measures the pixel-level differences between the corrected image and the distortion-free reference, whereas SSIM quantifies the similarity of structural information. The Multi-Scale Structural Similarity Index Measure (MS-SSIM) is also adopted to evaluate structural fidelity across multiple image scales.

Although rectification is primarily a geometric correction task, perceptual metrics are also informative for evaluating the visual realism of the restored images. Therefore, the Fréchet Inception Distance (FID) and the Learned Perceptual Image Patch Similarity (LPIPS), computed using both AlexNet and VGG backbones, are additionally reported to assess distribution-level similarity and perceptual consistency between the rectified results and ground-truth images.

Table 2: 3D evaluation results with different supervision.

Baseline	Supervisor	PSNR↑	SSIM↑	Depth-MAE↓	Str. Score↑
Blind [15]	-	17.52	0.65	0.420	0.1268
Ours	DINOv2 [19]	24.61	0.84	0.318	0.2416
Ours	VGGT [25]	**28.68**	**0.95**	**0.098**	**0.4321**

5.2 Comparison with State-of-the-Art Methods

To demonstrate the superiority of the proposed approach, the G2DC framework is compared with several representative image rectification methods. Based on their optimization paradigms, the evaluated baselines are categorized into two groups. The first group consists of regression-based approaches, including DC-CNN [22] and DeepCalib [4]. The second group includes generation-based architectures that directly synthesize dense correction fields or rectified images, including Blind [15], DR-GAN [18], PCN [30], SimFIR [9], and RDTR [26].

The quantitative results on standard metrics are summarized in Table 1. The proposed G2DC framework consistently achieves the highest performance across most evaluation metrics. Specifically, G2DC achieves a PSNR of 28.68 and an SSIM of 0.95, outperforming the strong RDTR baseline. In addition, G2DC obtains the lowest perceptual distance scores, achieving an FID of 19.8 and LPIPS values of 0.066 and 0.105 using AlexNet and VGG features, respectively. These results show that incorporating explicit geometric priors helps the network better preserve structural consistency while maintaining high perceptual quality.

Unlike existing methods that mainly focus on two-dimensional pixel alignment, the proposed framework introduces explicit geometric supervision within a higher-dimensional latent manifold. By leveraging the frozen VGGT model to provide geometric cues such as structural alignment and depth consistency, the network learns to produce rectified images that are both visually coherent and geometrically plausible. For 3D evaluation, we adopt two metrics: Depth-MAE [28] and Str. Score [34]. We further compare VGGT supervision with DINOv2 under the same setting. The results in Table 2 show that our method benefits from the powerful geometric priors brought by pretrained 3D foundation models and achieves significant improvements over the baseline in 3D evaluation metrics.

5.3 Ablation Study

Computational Efficiency. The computational efficiency of the proposed framework is evaluated in terms of parameter count, floating-point operations (FLOPs), and inference speed (FPS). The VGGT foundation model is used during training to provide geometric supervision, and is completely removed during inference. All FPS measurements are conducted using batch size 1 on an NVIDIA RTX 2080Ti GPU to ensure fair comparison. As shown in Table 3, the deployed rectification network remains lightweight while maintaining competitive computational complexity. The proposed G2DC contains 15.72M parameters and requires 18.45G

Table 3: Computational efficiency analysis during inference.

Method	Params (M) ↓	FLOPs (G) ↓	FPS ↑
DCCNN [22]	12.16	**4.95**	3.88
DeepCalib [4]	29.77	8.34	3.41
Blind [15]	31.78	42.46	7.23
DR-GAN [18]	27.46	37.82	5.96
PCN [30]	35.64	12.30	9.17
SimFIR [9]	19.12	24.03	7.95
RDTR [26]	15.94	20.35	8.76
G2DC	**15.72**	18.45	**10.42**

Table 4: Progressive ablation study on the core alignment objectives. The synergistic combination of structural ($\mathcal{L}_{\text{sa}}$) and physical ($\mathcal{L}_{\text{pa}}$) alignment losses yields the best overall performance, demonstrating the necessity of each 3D constraint.

		Physical Alignment ($\mathcal{L}_{\text{pa}}$)						
$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_{\text{sa}}$	$\mathcal{L}_{\text{struc}}$	$\mathcal{L}_{\text{depth}}$	$\mathcal{L}_{\text{int}}$	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
✓					17.65	0.66	145.2	0.312
✓	✓				24.12	0.84	58.4	0.145
✓	✓	✓			28.25	0.94	34.2	0.092
✓	✓	✓	✓		**28.95**	**0.96**	25.8	0.078
✓	✓	✓	✓	✓	28.68	0.95	**19.8**	**0.066**

FLOPs, while achieving an inference speed of 10.42 FPS. Among generative methods, RDTR is the most lightweight and fastest baseline, while ours is faster and smaller.

Core Components. We conduct a comprehensive ablation study to validate the core components of the proposed G2DC framework, including the Structural Alignment Loss ($\mathcal{L}$sa) and the sub-components of the Physical Alignment Loss ($\mathcal{L}$pa). As shown in Table 4, the baseline trained with only the pixel-level flow loss ($\mathcal{L}$flow) yields suboptimal performance (e.g., PSNR of 17.65), as it mainly focuses on local pixel alignment while ignoring global geometric structure. Adding $\mathcal{L}$sa significantly improves performance, increasing PSNR to 24.12 and reducing FID to 58.4. Building on this, we progressively introduce the three constraints of $\mathcal{L}$pa. The structural integrity constraint ($\mathcal{L}$struc) improves geometric consistency (PSNR 28.25). The depth topology constraint ($\mathcal{L}$depth) further enhances pixel-level alignment (PSNR 28.95, SSIM 0.96) by enforcing depth-aware transitions. Finally, the intrinsic parameter regularization ($\mathcal{L}$int) anchors the deformation field to a physically valid projection model. Although it slightly affects pixel-wise metrics (PSNR 28.68), it significantly improves perceptual and structural quality (FID 19.8, LPIPS 0.066). Overall, the results demonstrate that each component contributes to progressively stronger geometric consistency and

Table 5: Performance gain in PSNR by incorporating the proposed loss into different baseline methods.

Baseline	Original PSNR	+ Proposed Loss	Gain
DeepCalib	20.84	25.50	+4.66
PCN	24.56	28.40	+3.84
RDTR	27.92	29.95	+2.03
G2DC	–	28.68	–

that their synergistic integration is crucial for achieving physically faithful and visually robust 3D-aware rectification.

Loss Generalization. To evaluate the generalizability of the proposed loss, we incorporate it into multiple baseline methods, as shown in Table 5. Our method achieves consistent improvements across different baselines, demonstrating its effectiveness and generalizability.

6 Conclusion

In this paper, we presented Geometry-Grounded Distortion Correction (G2DC), a novel framework that overcomes the inherent 2D-centric limitations of existing methods for single-image rectification. By internalizing explicit 3D projective priors through a frozen geometric foundation model, the proposed approach transitions from superficial pixel alignment to deep structural consistency modeling. Specifically, the joint optimization of the Structural Alignment Loss and the Physical Alignment Loss ensures that the rectified imagery maintains strict geometric consistency, including the preservation of collinearity and valid depth topology. Extensive evaluations demonstrate that the proposed framework establishes new state-of-the-art performance, particularly in metrics that demand high-precision 3D geometric consistency.

7 Limitations

We acknowledge that real-world fisheye images and extreme out-of-distribution cases remain challenging, as the main limitation lies in synthetic-distortion training, which still leaves a gap to real lenses. Future work will introduce real fisheye data and integrate 3D models for end-to-end training and optimization.

Acknowledgements

This paper is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM123).

References

1. Alemán-Flores, M., Alvarez, L., Gomez, L., Santana-Cedrés, D.: Automatic lens distortion correction using one-parameter division models. Image Processing On Line **4**, 327–343 (2014)
2. Aso, K., Hwang, D.H., Koike, H.: Portable 3d human pose estimation for human-human interaction using a chest-mounted fisheye camera. In: Proceedings of the Augmented Humans International Conference. pp. 116–120 (2021)
3. Blott, G., Takami, M., Heipke, C.: Semantic segmentation of fisheye images. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
4. Bogdan, O., Eckstein, V., Rameau, F., Bazin, J.C.: Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In: Proceedings of the ACM SIGGRAPH European Conference on Visual Media Production. pp. 1–10 (2018)
5. Bukhari, F., Dailey, M.N.: Automatic radial distortion estimation from a single image. Journal of Mathematical Imaging and Vision **45**, 31–45 (2013)
6. Deng, L., Yang, M., Qian, Y., Wang, C., Wang, B.: Cnn based semantic segmentation for urban traffic scenes using fisheye camera. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). pp. 231–236. IEEE (2017)
7. Devernay, F., Faugeras, O.: Straight lines have to be straight. Machine Vision and Applications **13**, 14–24 (2001)
8. Espuny, F., Burgos Gil, J.I.: Generic self-calibration of central cameras from two rotational flows. International Journal of Computer Vision **91**, 131–145 (2011)
9. Feng, H., Wang, W., Deng, J., Zhou, W., Li, L., Li, H.: Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12418–12427 (2023)
10. Fitzgibbon, A.W.: Simultaneous linear estimation of multiple view geometry and lens distortion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. I–I. IEEE (2001)
11. Goodarzi, P., Stellmacher, M., Paetzold, M., Hussein, A., Matthes, E.: Optimization of a cnn-based object detector for fisheye cameras. In: Proceedings of the IEEE International Conference on Vehicular Electronics and Safety (ICVES). pp. 1–7. IEEE (2019)
12. Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y.: Learning to parse wireframes in images of man-made environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 626–635 (2018)
13. Kang, S.B.: Catadioptric self-calibration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 201–207. IEEE (2000)
14. Kannala, J., Brandt, S.S.: 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 (2006)
15. Li, X., Zhang, B., Sander, P.V., Liao, J.: Blind geometric distortion correction on images through deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4855–4864 (2019)
16. Liao, K., Lin, C., Liao, L., Zhao, Y., Lin, W.: Multi-level curriculum for training a distortion-aware barrel distortion rectification model. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4389–4398 (2021)
17. Liao, K., Lin, C., Zhao, Y., Xu, M.: Model-free distortion rectification framework bridged by distortion distribution map. IEEE Transactions on Image Processing **29**, 3707–3718 (2020)
18. Liao, K., Lin, C.s., Zhao, Y., Gabbouj, M.: Dr-gan: Automatic radial distortion rectification using conditional gan in real-time. IEEE Transactions on Circuits and Systems for Video Technology **30**(3), 725–733 (2019)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Plaut, E., Ben Yaacov, E., El Shlomo, B.: 3d object detection from a single fisheye image without a single fisheye training image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3659–3667 (2021)
21. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016)
22. Rong, J., Huang, S., Shang, Z., Ying, X.: Radial lens distortion correction using convolutional neural networks trained with synthesized images. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 35–49. Springer (2017)
23. Toepfer, C., Ehlgen, T.: A unifying omnidirectional camera model and its applications. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1–5. IEEE (2007)
24. Tsai, R.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. IEEE Journal on Robotics and Automation **3**(4), 323–344 (1987)
25. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5306 (2025)
26. Wang, W., Feng, H., Zhou, W., Liao, Z., Li, H.: Model-aware pre-training for radial distortion rectification. IEEE Transactions on Image Processing **32**, 5764–5778 (2023)
27. Xue, Z., Xue, N., Xia, G.S., Shen, W.: Learning to calibrate straight lines for fisheye image rectification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1643–1651 (2019)
28. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
29. Yang, S., Lin, C., Liao, K., Zhao, Y.: Innovating real fisheye image correction with dual diffusion architecture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12699–12708 (2023)
30. Yang, S.w., Lin, C., Liao, K., Zhang, C., Zhao, Y.: Progressively complementary network for fisheye image rectification using appearance flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6348–6357 (2021)
31. Yin, X., Wang, X., Yu, J.s., Zhang, M., Fua, P., Tao, D.: Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 469–484 (2018)
32. Zhang, M., Yao, J., Xia, M.r., Li, K., Zhang, Y., Liu, Y.: Line-based multi-label energy optimization for fisheye image rectification and calibration. In: Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4137–4145 (2015)
33. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1330–1334 (2002)
34. Zhao, J., Wei, S., Chang, Y., Ruan, T., Zhao, Y.: Model-free rectification via cascaded distortion model and enhanced backward flow network. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 6291–6302 (2024)
35. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2017)