arXiv:2410.11506v3 [cs.CV] 5 Aug 2025

Spatio-Temporal Distortion Aware Omnidirectional Video Super-Resolution

Hongyu An[1], Xinfeng Zhang[1*], Shijie Zhao[2], Li Zhang[2], Ruiqin Xiong[3]

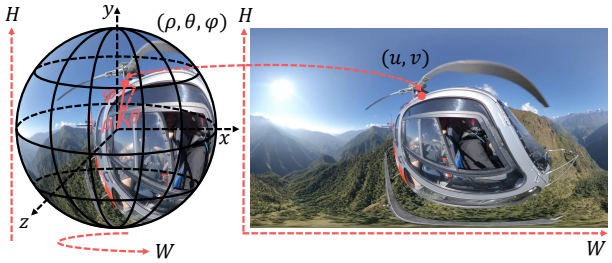
[1]University of Chinese Academy of Sciences, [1]Bytedance Inc, [1]Peking University
anhongyu22@mails.ucas.ac.cn, xfzhang@ucas.ac.cn,
{zhaoshijie.0526, lizhang.idm}@bytedance.com, rqxiong@pku.edu.cn

Abstract

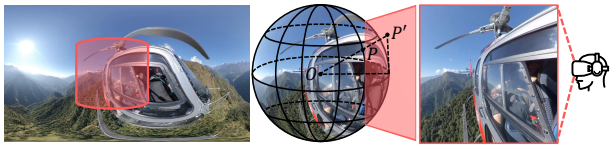
Omnidirectional videos (ODVs) provide an immersive visual experience by capturing the 360° scene. With the rapid advancements in virtual/augmented reality, metaverse, and generative artificial intelligence, the demand for high-quality ODVs is surging. However, ODVs often suffer from low resolution due to their wide field of view and limitations in capturing devices and transmission bandwidth. Although video super-resolution (SR) is a capable video quality enhancement technique, the performance ceiling and practical generalization of existing methods are limited when applied to ODVs due to their unique attributes. To alleviate spatial projection distortions and temporal flickering of ODVs, we propose a Spatio-Temporal Distortion Aware Network (STDAN) with joint spatio-temporal alignment and reconstruction. Specifically, we incorporate a spatio-temporal continuous alignment (STCA) to mitigate discrete geometric artifacts in parallel with temporal alignment. Subsequently, we introduce an interlaced multi-frame reconstruction (IMFR) to enhance temporal consistency. Furthermore, we employ latitude-saliency adaptive (LSA) weights to focus on regions with higher texture complexity and human-watching interest. By exploring a spatio-temporal jointly framework and real-world viewing strategies, STDAN effectively reinforces spatio-temporal coherence on a novel *ODV-SR* dataset and ensures affordable computational costs. Extensive experimental results demonstrate that STDAN outperforms state-of-the-art methods in improving visual fidelity and dynamic smoothness of ODVs.

Introduction

Omnidirectional videos (ODVs), also known as 360° or panoramic videos, cover a full 360° field of view by stitching multiple images onto a spherical surface. This expansive vision enables engaging immersive interaction, making ODVs valuable in various fields, like entertainment, creativity, advertising, intelligent driving, and video conferencing. To achieve realistic visuals, ODVs ideally require high resolutions such as 4K, 8K, or higher (Elbamby et al. 2018). Nevertheless, the trade-off between display quality and transmission costs is challenging. Video super-resolution (VSR) offers a solution by reconstructing high-resolution (HR) frames from low-resolution (LR) ones. Compared with general videos, ODVs necessitate an additional projection step from sphere to plane for subsequent processing, which causes uneven stretching and broken boundaries. These issues render traditional VSR models ineffective for ODVs.



(a) Equirectangular Projection (ERP).



(b) ERP viewpoint conversion.

Figure 1: Geometric transformations: (a) ERP. (b) A practically observed ERP viewpoint. To align with viewing preferences, attractive viewpoints should be emphasized.

Practical ODV projection methods from 3D to 2D include Equirectangular Projection (ERP), Cubemap Projection (CMP), Icosahedral Projection (ISP), and Equi-Angular Cubemap Projection (EAC), *etc.* Among these, ERP is the most widely used due to its low computation complexity and extensive adaptability. As illustrated in Fig. 1(a)[1], spherical coordinates are defined as (ρ, θ, ϕ), where $\theta \in (0, 2\pi)$ and $\phi \in (0, \pi)$ represent longitude and latitude, respectively. The planar horizontal and vertical coordinates are defined as (u, v). However, ERP notably introduces latitude-related distortions, particularly severe in polar areas. Simultaneously, ERP creates discontinuities at rectangular edges, disrupting the continuity of adjacent pixels on the sphere.

Deep learning-based VSR models have replaced traditional interpolation methods, achieving remarkable success. Unfortunately, these techniques fall short when applied to ODVs with spatio-temporal distortions. As for omnidirectional image (ODI) super-resolution (SR), existing methods focused on reducing ERP distortions through latitude correlation (Ozcinar et al. 2019; Deng et al. 2021; An and

*Corresponding author.

[1]Detailed derivative processes can be found in Appendix 1.

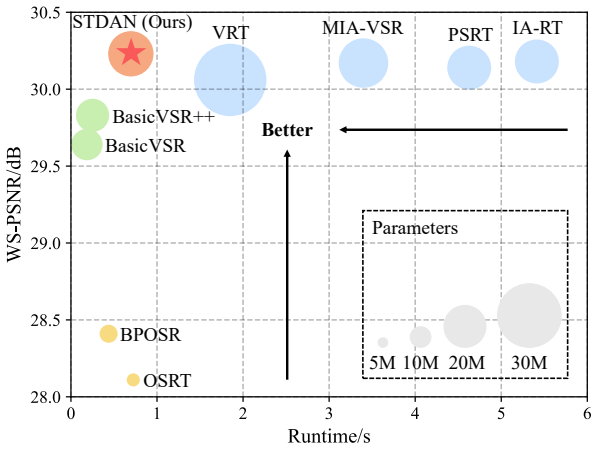


Figure 2: WS-PSNR(dB) and Runtime(s) comparison. Our STDAN outperforms other methods and strikes a balance between performance and computational efficiency.

Zhang 2023). OSRT (Yu et al. 2023) designed feature-level offset and alignment. FATO (An et al. 2024) explored frequency distribution. Research on ODV-SR remains limited, with current works (Liu et al. 2024; Baniya et al. 2023; Li and Liu 2024) concentrating on latitude-related distortions without considering spatial discontinuity and temporal inconsistency. Moreover, the aforementioned models overlook the human visual pattern. As demonstrated in Fig. 1(b), the highlighted region stands for the real viewpoint of interest.

This paper proposes a Spatio-Temporal Distortion Aware Network (STDAN) to address the unique challenges of ODV-SR. While general VSR methods exploit inter-frame similarity through temporal alignment, extending beyond spatial feature extraction and up-sampling reconstruction, they often fail to handle the complex spatio-temporal degradations in ODVs. STDAN optimizes single-dimensional spatial or temporal modules and integrates joint spatio-temporal alignment and reconstruction. Specifically, it leverages spatial alignment with omni-positional encoding priors to modulate projection stretching and discontinuous boundaries, and introduces an interlaced multi-frame reconstruction module to maintain temporal stability. Moreover, latitude-saliency adaptive weights are computed within the loss function to concentrate on high-heat regions. We also present a new ODV-SR dataset with diverse scenarios. Both objective metrics and subjective evaluations demonstrate the superior performance of STDAN over existing methods. The main contributions can be summarized as follows:

1) We propose STDAN to fully exploit the spatio-temporal properties of ODVs. The spatio-temporal continuous alignment (STCA) incorporates positional information to mitigate projection distortions effectively.
2) An interlaced multi-frame reconstruction (IMFR) is designed based on a spatio-temporal sequence rescheduling mechanism, guaranteeing temporal consistency.
3) A latitude-saliency adaptive (LSA) loss is adopted to enhance regions with high texture complexity and watching interest, thereby enhancing perceptual quality.
4) We develop the first ODV-SR network tailored for practical applications, emphasizing computational efficiency and focusing on actual viewpoints, as shown in Fig. 2. Additionally, a variety of ODV-SR dataset is introduced to improve the generalization capability of STDAN.

Related Work

Video Super-Resolution

Traditional VSR networks are broadly divided into two categories: sliding-window and recurrent frameworks. The former (Kappeler et al. 2016; Caballero et al. 2017; Wang et al. 2019; Tian et al. 2020) leverages the temporal correlation between a reference frame and its neighboring frames within a short temporal window. The latter (Sajjadi, Vemulapalli, and Brown 2018; Haris, Shakhnarovich, and Ukita 2019; Chan et al. 2021, 2022) employs a recurrent mechanism to capture long-term relationships across frames. Benefiting from the long-term modeling capability, Transformer (Vaswani et al. 2017) has been integrated into VSR (Liang et al. 2022; Shi et al. 2022; Zhou et al. 2024; Xu et al. 2024; Kim et al. 2025). However, the complex attention calculations in Transformer-based VSR approaches pose significant challenges for practical deployment, particularly when it comes to enhancing ODVs to at least 2K resolution.

Omnidirectional Image Super-Resolution

Intuitively, Ozcinar *et al.* (Ozcinar, Rana, and Smolic 2019) designed a latitude-related loss to alleviate ERP distortions. LAU-Net (Deng et al. 2021) designed a framework with multiple latitude levels. SphereSR (Yoon et al. 2022) proposed an icosahedron-based feature extraction module. OSRT (Yu et al. 2023) presented a distortion-aware Transformer oriented to real-world fisheye down-sampling. FATO (An et al. 2024) reduced ODI distortions by analyzing the frequency domain distribution. BPOSR (Wang et al. 2024) applied a bi-projection network to facilitate the fusion between different projections. FAOR (Shen et al. 2025) adapted the implicit image function from the planar domain to the ERP domain by integrating spherical geometric priors.

Omnidirectional Video Super-Resolution

Despite the impressive progress in VSR techniques, ODV-SR has not garnered the attention it deserves. SMFN (Liu et al. 2024) introduced a dual single-frame and multi-frame network. A weighted loss function is also provided to make the network pay more attention to equatorial regions. S3PO (Baniya et al. 2023) adopted a spherical distortion feature extractor in an attention-recurrent framework, unbound from conventional VSR alignment. VertexShuffle (Li and Liu 2024) utilized a focused icosahedral mesh to represent local sphere regions and further constructed matrices to rotate spherical content over the entire sphere. Although referenced ODI/ODV-SR models modulated latitude-related ERP distortions effectively, the spatial discontinuity inherent in projection and temporal inconsistency remain underexplored. Besides, the tendency of human visual attention in ODVs is hardly involved. In this paper, we propose an efficient spatio-temporal aggregation framework with watching inclination to enhance the immersive experience.

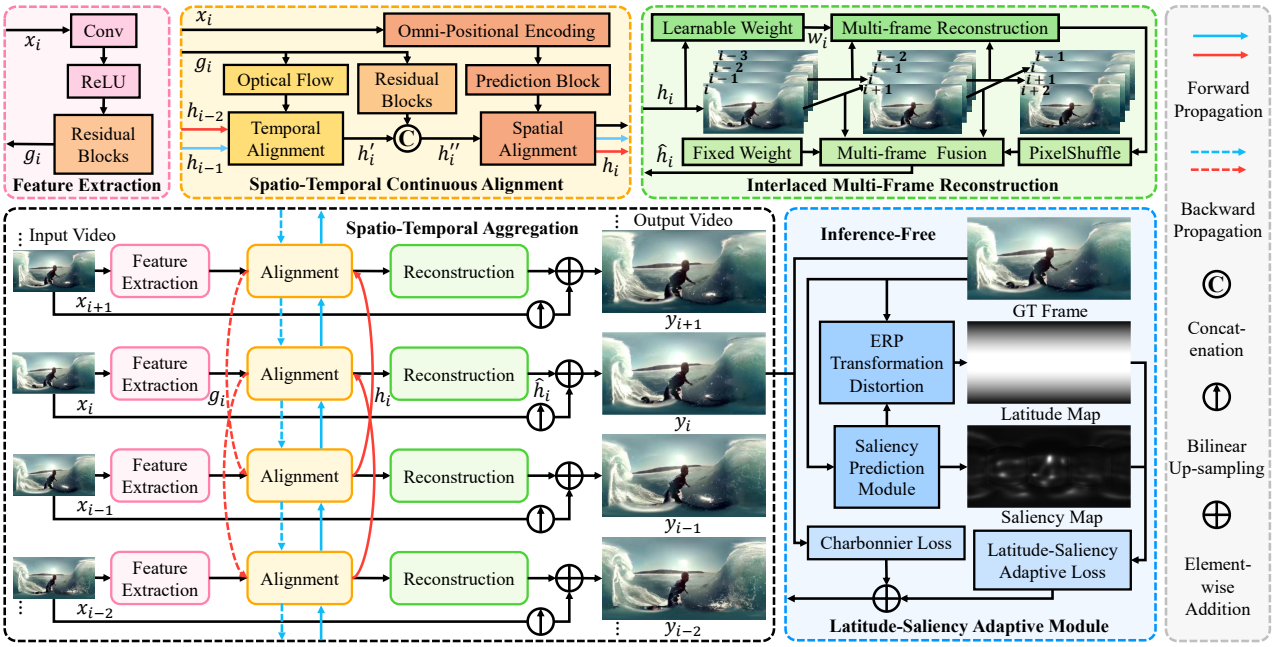


Figure 3: Overview of the proposed Spatio-Temporal Distortion Aware Network (STDAN).

Methodology

Overview

The overall framework of STDAN is depicted in Fig. 3. It is built upon the bi-directional second-order grid propagation framework (Chan et al. 2022) and comprises: spatio-temporal continuous alignment (STCA), interlaced multi-frame reconstruction (IMFR), and a latitude-saliency adaptive (LSA) module. Given a LR video sequence $\{x_1, ..., x_i, ..., x_n\}$, we first extract the feature g_i from each frame x_i. This feature g_i is then fed into STCA, where it undergoes spatio-temporal alignment to produce an enhanced feature h_i. The process is formulated as follows:

$$h_i = \text{SA}(\text{C}(\text{TA}(h_{i-2}, h_{i-1}, g_i), \text{Res}(g_i)), \quad (1)$$

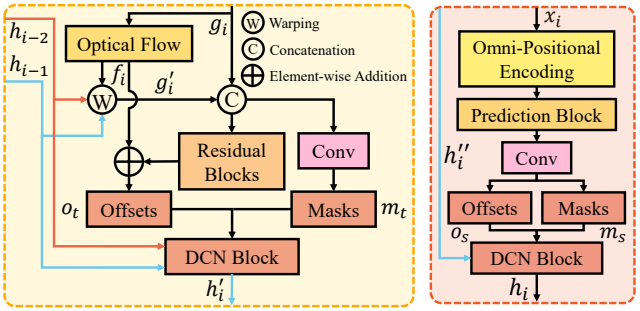
where TA(·) and SA(·) refer to temporal and spatial alignments, C(·) stands for concatenation along channel dimension, and Res(·) denotes residual blocks. Subsequently, the enhanced feature h_i is processed by the IMFR module for interlaced spatio-temporal reconstruction, as described by:

$$\widehat{h}_i = \text{MFF}(\text{PixelShuffle}(\text{MFR}(h_i, w_i)), w_i^{'}), \quad (2)$$

where MFR(·) and MFF(·) indicate the multi-frame reconstruction and fusion processes. The parameter w_i is a learnable weight that guides interaction across frames. After MFR, we employ PixelShuffle (Shi et al. 2016) to up-sample reconstructed features and further smooth outputs using a fixed-weight $w_i^{'}$. The final HR frame y_i is generated via a global residual connection. Notably, we estimate the LSA loss to emphasize practically appealing regions.

Spatio-Temporal Continuous Alignment (STCA)

The ERP inherently introduces distortions and discontinuities, which pose challenges for ODV-SR. To address these issues, we propose STCA that integrates omni-positional encoding (OPE) to achieve joint spatio-temporal alignment.

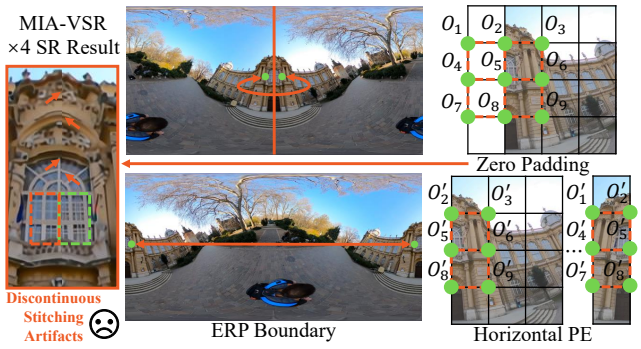


(a) Temporal Alignment (b) Spatial Alignment

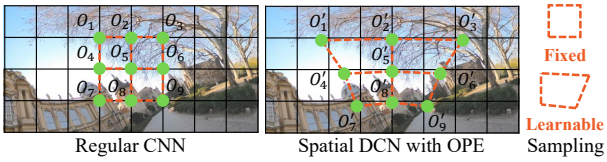
Figure 4: DCN-based temporal correlation alignment (a) and spatial distortion alignment (b). Optical flows and omni-positional encoded offsets supply spatio-temporal priors.

Dimension Expansion Alignment DCN-based (Dai et al. 2017) temporal alignment has been widely adopted to compensate for motions. Following BasicVSR++ (Chan et al. 2022), we apply SpyNet (Ranjan and Black 2017) to compute the optical flow f_i as temporal guidance. Based on this optical flow, we calculate the offset o_t and mask m_t of the warped feature $g_i^{'}$, which serve as the initialization for temporal DCN_T. Fig. 4 (a) inspects the forward alignment, with the backward progress following a similar procedure. To tackle geometric distortions in ODVs, we extend the alignment range. Unlike previous feature-level warping (Yu et al. 2023), we design a spatio-temporal joint framework with an ODV-oriented OPE to achieve multi-dimensional continuous modulation. As shown in Fig. 4 (b), we incorporate positional information and estimate distortions through a prediction block, which consists of a 1×1 convolution with two hidden layers. The proposed STCA is defined as follows:

$$o_t = \text{Res}(\text{C}(g_i, g_i^{'})) + f_i, m_t = \text{Conv}(\text{C}(g_i, g_i^{'})), \quad (3)$$



(a) ERP boundary and horizontal PE.



(b) Sampling location of spatial DCN.

Figure 5: Spatial alignment with OPE. Extra positional information enables continuous sampling.

$$h_i^{'} = \text{DCN}_\text{T}(\text{C}(h_{i-1}, h_{i-2}), o_t, m_t), h_i^{''} = \text{C}(h_i^{'}, \text{Res}(g_i)), \quad (4)$$

$$o_s = \text{Conv}(\text{P}(\text{OPE}(x_i))), m_s = \text{Conv}(\text{P}(\text{OPE}(x_i))), \quad (5)$$

$$h_i = \text{DCN}_\text{S}(h_i^{''}, o_s, m_s), \quad (6)$$

where OPE$(\cdot)$ means omni-positional encoding, and P$(\cdot)$ represents prediction block. The prediction block allows positional priors to adapt dynamically to the content distribution, thereby enhancing the flexibility of spatial DCN.

Omni-Positional Encoding In contrast to the complicated cylinder-style convolution proposed for ODI inpainting (Liao et al. 2023), we develop a tight spatial alignment with OPE to exploit geometric priors. Specifically, we capture ERP positional cues to predict offsets for a learnable spatial DCN, achieving precise distortion modeling. For ODVs, ERP introduces quantifiable latitude-dependent distortions (Appendix 1), hence, we adopt cosine-based vertical PE at each latitude. Additionally, the ERP left-right boundary separates adjacent pixels and introduces stitching artifacts. As depicted in Fig. 5 (a), boundary windows exhibit severe mismatches, degrading visual quality. Since polar regions are free of discontinuities, we introduce a horizontal PE that mimics the unrolling of a cylinder:

$$[sin(1/10000^{2k/d}u), cos(1/10000^{2k/d}u), \dots], \quad (7)$$

where d indicates one-half of the encoding dimensions. This handcrafted encoding explicitly facilitates interaction across boundaries. Finally, we concatenate vertical and horizontal PE to obtain the complete OPE. After OPE-guided prediction and learnable spatial DCN, STCA adaptively reallocates sampling positions to modulate ERP distortions and restore spatial coherence, as illustrated in Fig. 5 (b).

Interlaced Multi-Frame Reconstruction (IMFR)

The memory constraint of HR ODVs limits the sequence length that can be processed. It is intuitive to divide the entire sequence into short subsequences and enhance them sequentially. However, truncated frames risk flickering. To improve visual consistency, we explore the relationship across restricted frames in subsequences and propose an interlaced multi-frame reconstruction (IMFR). In brief, IMFR reconstructs three temporally interlaced sequences and gets target frames through sliding interaction among corresponding frames in three sequences.

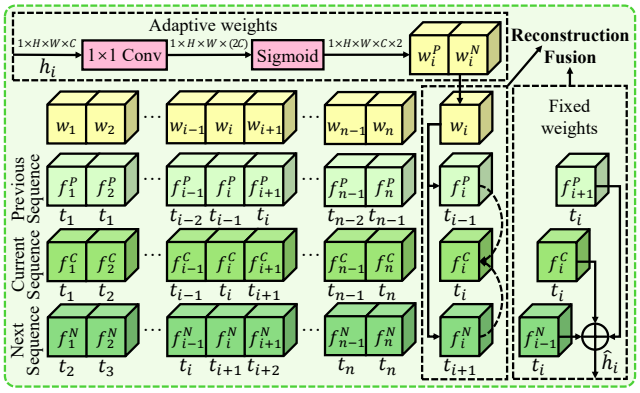


Figure 6: Interlaced multi-frame reconstruction. The target frame f_i^C is reconstructed with weighted f_i^P and f_i^N and fused with f_{i+1}^P and f_{i-1}^N to ensure temporal consistency.

To leverage temporal correlation among limited frames, we triple the channel of intermediate features and split them along the temporal axis into three subsequences: previous f^P, current f^C, and next f^N. Each index i thus corresponds to prior frame, target frame, and subsequent frame. To unify sequence lengths, we pad f^P and f^N by replicating their first and last frames. As demonstrated in Fig. 6, the target frame f_i^C is jointly reconstructed from its neighbors f_i^P and f_i^N. The aligned feature h_i is fed into a 1×1 convolution to predict adaptive weights w_i^P and w_i^N that modulate the contributions of f_i^P and f_i^N. A Sigmoid followed by a scaling factor of 0.5 constrains the range of weights.

Thereafter, we up-sample three sequences and fuse the target frame f_i^C with its temporal counterparts f_{i+1}^P and f_{i-1}^N, enforcing a weighted smoothing across the three subsequences. Formally, the fused frame is computed via:

$$f_i = \text{Norm}(\alpha_1 f_{i+1}^P + f_i^C + \beta_1 f_{i-1}^N), \quad (8)$$

where Norm$(\cdot)$ indexes normalization, and weights α_1 and β_1 are preset to 0.01. The IMFR propagates temporal context into every frame and smooths it, yielding a temporally coherent video sequence.

Latitude-Saliency Adaptive Module (LSAM)

Considering practical ODV watching habits, we introduce a latitude-saliency adaptive module (LSAM) to counteract latitude distortions while respecting viewing preferences. LSAM offline estimates latitude projection and saliency attention maps to assign weights in the loss function. Instead of fusing them as priors in the network, LSAM incurs zero additional cost at inference, making it deployment-friendly.

Dataset	*ODV-SR* Dataset				
Method	PSNR (dB)↑ / SSIM↑	WS-PSNR (dB)↑ / WS-SSIM↑	E^*_{warp}↓	VMAF↑	Top-5 PSNR (dB)↑ / SSIM↑
Bicubic	28.01 / 0.8218	27.20 / 0.7933	8.25	44.85	26.37 / 0.7562
OSRT	27.98 / 0.8419	28.11 / 0.8314	9.84	70.00	28.01 / 0.8204
BPOSR	27.58 / 0.8024	28.41 / 0.8294	9.65	71.58	28.27 / 0.8148
EDVR	30.40 / 0.8765	29.61 / 0.8569	11.69	81.56	29.87 / 0.8580
BasicVSR	30.53 / 0.8785	29.64 / 0.8591	8.42	82.54	29.76 / 0.8581
BasicVSR++	30.72 / 0.8807	29.83 / 0.8621	8.44	83.39	30.10 / 0.8640
VRT	30.92 / 0.8832	30.06 / 0.8649	8.74	83.88	30.44 / 0.8665
PSRT	31.01 / 0.8834	30.14 / 0.8653	8.68	83.77	31.53 / 0.8732
MIA-VSR	31.09 / 0.8843	30.17 / 0.8660	8.71	83.57	31.48 / 0.8729
IA-RT	31.11 / 0.8845	30.18 / 0.8659	8.69	83.75	31.47 / 0.8734
STDAN(Ours)	**31.13** / **0.8876**	**30.23** / **0.8698**	**7.60**	**85.13**	**30.61** / **0.8736**
Dataset	360VDS Dataset		MiG Dataset		
Method	PSNR(dB)↑ / SSIM↑	WS-PSNR(dB)↑ / WS-SSIM↑	PSNR(dB)↑ / SSIM ↑		WS-PSNR(dB)↑ / WS-SSIM↑
Bicubic	25.11 / 0.7518	24.37 / 0.7164	28.97 / 0.8215		28.78 / 0.7985
OSRT	25.35 / 0.7926	25.00 / 0.7676	30.63 / 0.8618		30.13 / 0.8472
EDVR	26.79 / 0.8252	25.99 / 0.7994	30.30 / 0.8534		29.93 / 0.8340
SFMN	-	-	30.56 / 0.8505		30.13 / 0.8381
BasicVSR	27.10 / 0.8313	26.25 / 0.8057	30.71 / 0.8588		30.10 / 0.8397
BasicVSR++	27.21 / 0.8358	26.36 / 0.8110	30.71 / 0.8578		30.23 / 0.8387
S3PO	27.24 / 0.8225	26.31 / 0.8026	31.16 / 0.8565		30.42 / 0.8453
VRT	27.23 / 0.8359	26.43 / 0.8124	30.85 / 0.8636		30.45 / 0.8461
PSRT	27.30 / 0.8383	26.48 / 0.8148	30.97 / 0.8625		30.42 / 0.8443
MIA-VSR	27.28 / 0.8380	26.48 / 0.8146	31.03 / 0.8619		30.38 / 0.8433
IA-RT	27.34 / 0.8397	**26.53** / 0.8164	**31.24** / 0.8655		**30.50** / 0.8472
STDAN(Ours)	**27.38** / **0.8408**	26.51 / **0.8165**	31.21 / **0.8669**		**30.50** / **0.8488**

Table 1: Quantitative ×4 ODV-SR comparison on *ODV-SR*, 360VDS, and MiG Panorama Video datasets. **Bold** and underlined values indicate the best and second-best results.

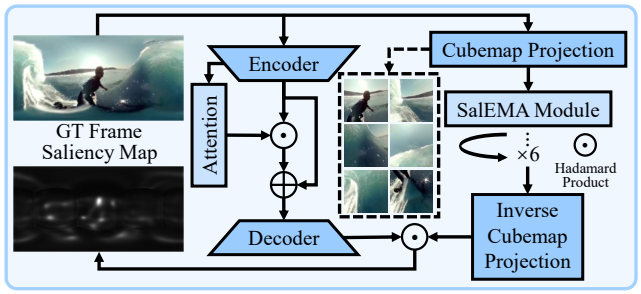


Figure 7: Saliency prediction module. The attention branch captures global information and the projection branch exploits local viewpoint details dynamically.

Latitude-Related Projection Map As shown in Fig. 1, ERP pixels are non-uniform. We quantify this distortion by the stretching ratio (STR) (Sun, Lu, and Yu 2017), which measures the local area variation after projection. From Appendix 1, the ERP STR is derived as:

$$\mathrm{STR}(u, v) = \cos(\phi) = \cos(v), \tag{9}$$

Based on Eqs. 9, we measure ERP distortions through latitude-related pixel weights ranging from 0 to 1:

$$W_{lat}(u, v) = \cos((v - (H/2) + 0.5)\pi/H). \tag{10}$$

Saliency-Oriented Attention Map While ODVs span 360°, human views fixate on limited regions. We therefore enhance high-frequency observed viewpoints to improve perception. Built upon SalEMA (Linardos et al. 2019; Dahou et al. 2021), we utilize a two-branch network (Fig. 7) to capture global context and regress viewpoint likelihood for predicting ODV saliency. Afterward, we multiply parallel results to get the saliency-oriented weight $W_{sal}(u, v)$.



Figure 8: Examples of the proposed *ODV-SR* dataset, including dynamic and static scenes.

Latitude-Saliency Adaptive Loss Function As shown in Fig. 3, LSAM is pre-computed to yield pixel-level weights. The composite loss steers the network toward attractive areas. Loss L_{lat} and L_{sal} are formulated as follows:

$$L_{lat/Sal} = \mathbb{E}(W_{lat/Sal} \odot |I_{HR} - I_{SR}|), \tag{11}$$

where I_{HR} and I_{SR} denote ground-truth HR and reconstructed SR patches. $\mathbb{E}(\cdot)$ is matrix averaging. The latitude-saliency adaptive loss function L is defined as:

$$L = \sqrt{\|I_{HR} - I_{SR}\|^2 + \varepsilon^2} + \alpha_2 L_{lat} + \beta_2 L_{sal}, \tag{12}$$

where the first item represents Charbonnier loss (Lai et al. 2017) and ε is assigned as 10^{-3}.

Experiments

Experimental Details

Prior ODV datasets have few testing sequences (4, MiG (Liu et al. 2024)) or limited resolution (360 × 480, 360VDS(Baniya et al. 2023)), which are unconvincing. Although NTIRE (Cao et al. 2023) released a high-quality ODV360 dataset, its test set remains undisclosed. To facilitate training and testing, we gather a new ODV-SR

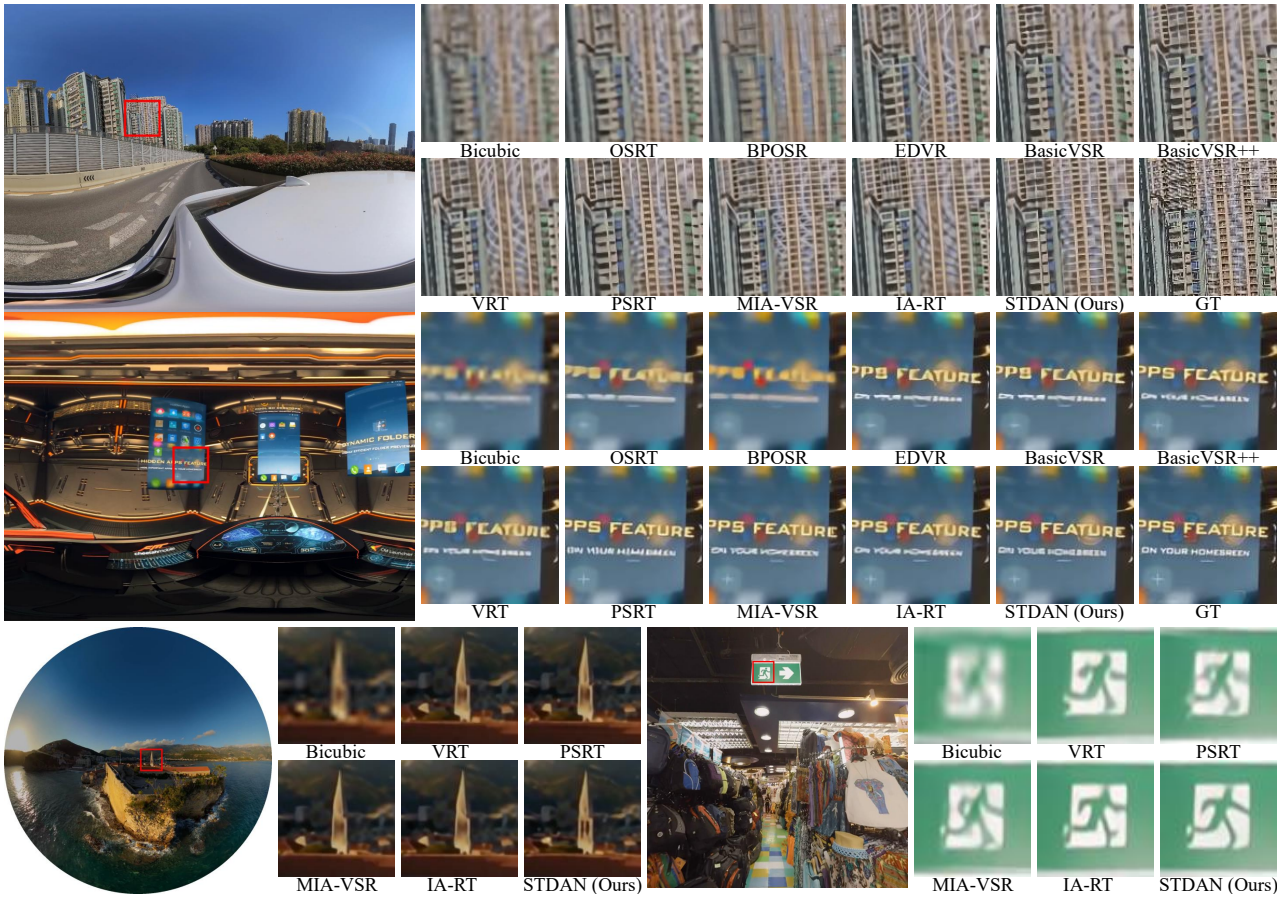


Figure 9: Qualitative $\times 4$ ODV-SR comparison on ERP, fisheye, and perspective *ODV-SR* dataset.

dataset sourced and reorganized from public websites and the ODV360 dataset. The proposed *ODV-SR* dataset comprises 270 train, 20 val, and 25 test clips at 1080 $\times$ 2160 resolution. As illustrated in Fig. 8, the collection spans diverse scenes: landscapes, concerts, narration, sports, driving, movies, and virtual games, animation ODVs.

During training, HR frames are cropped to 256 $\times$ 256 patches and augmented with rotation and flip. Each clip contains 20 frames with a batch size of 2. STCA embeds 7 residual blocks. We train STDAN with Adam ($lr = 10^{-4}$) and CosineRestart for 300K iterations. The SpyNet is jointly fine-tuned at $\times 0.25$ the STDAN learning rate. Training is conducted on 8 A800 GPUs in PyTorch.

Comparison Results

Interpolation method Bicubic, ODI-SR methods OSRT and BPOSR (Yu et al. 2023; Wang et al. 2024), CNN-based VSR methods EDVR, BasicVSR, and BasicVSR++ (Wang et al. 2019; Chan et al. 2021, 2022), Transformer-based VSR method, VRT, PSRT, IA-RT, and MIA-VSR, (Liang et al. 2022; Shi et al. 2022; Xu et al. 2024; Zhou et al. 2024), ODV-SR methods SMFN and S3PO (Liu et al. 2024; Baniya et al. 2023), and our STDAN are comprehensively evaluated to compare ODV-SR performance. PSNR, SSIM, and ERP-specific WS-PSNR (Sun, Lu, and Yu 2017), WS-SSIM (Zhou et al. 2018) are calculated on the luma (Y) channel. We fairly retrain reproducible video models on the ODV360 dataset, otherwise, results are migrated from the 360VDS dataset and MiG dataset as reported in papers.

Computational Efficiency Comparison Computational efficiency on the *ODV-SR* dataset is summarized in Fig. 2. Relative to Transformer-based VSR approaches, VRT, PSRT, MIA-VSR, and IA-RT, STDAN achieves a faster running speed while preserving the best SR quality, striking an effective complexity-performance trade-off. Detailed parameters, FLOPs, and runtime are reported in Appendix 2.1.

Quantitative Comparison As displayed in Tab. 1, our STDAN achieves the best or second-best quantitative metrics on all benchmarks, demonstrating its superior capacity. Notably, the HR-grade *ODV-SR* dataset offers results that are closer to real-world conditions. While Transformer-based MIA-VSR and IA-RT yield competitive performance, they incur markedly higher computational overhead.

Qualitative Comparison Fig. 9 presents visual results from representative VSR methods. Upon inspection of zoomed-in regions, only STDAN restores grids and stripes of buildings clearly. In the second scene, the small white letters "YOUR" become most legible in the STDAN reconstruction. Beyond the ERP format, we further visualize results on fisheye and perspective ODVs. On such formats, STDAN continues to rebuild visually more pleasing tex-

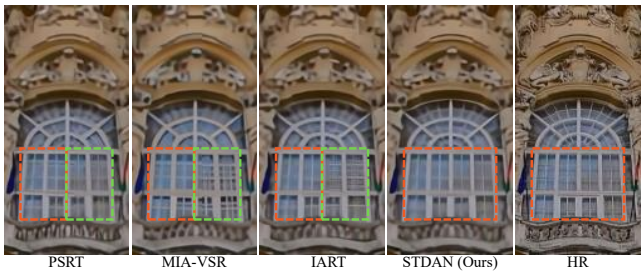


Figure 10: Qualitative $\times 4$ ODV-SR comparison on ERP boundaries in *ODV-SR* dataset.

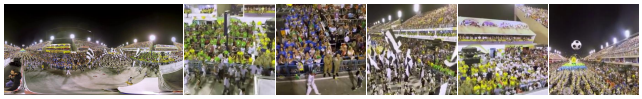


Figure 11: Top-5 viewpoints of an ERP ODV. The observed attractive viewpoints are projected from ERP to perspective.

tures. These observations corroborate that STDAN reconstructs finer structural details and surpasses current methods. More comparison results can be found in Appendix 2.2.

Spatio-Temporal Consistency Comparison As discussed in Sec. STCA, we introduce spatio-temporal joint alignment to modulate discontinuities flexibly. First, we evaluate spatial consistency across ERP boundaries. According to Fig. 5 (a) and 10, we stich the left and right margins into a unified perspective view. Within the highlighted rectangular regions, it is clearly evident that STDAN maintains identical window structures on both sides of the seam, whereas competing approaches reconstruct dissimilar neighboring windows with obvious visual discontinuities.

The warping-error metric E^*_{warp} (Lai et al. 2018) quantifies temporal consistency. As illustrated in Tab. 1, STDAN attains the lowest E^*_{warp}, evidencing its generation of smooth and flicker-free ODVs. Moreover, VMAF (Blog 2016) is a widely adopted video quality assessment that jointly measures spatial fidelity and temporal motion coherence. We exploit it to examine the spatio-temporal performance of ODV-SR. From Tab. 1, STDAN receives the highest VMAF score. To intuitively indicate the temporal stability improvement of STDAN, we include temporal profiles of different algorithms in Appendix 2.3. We also conduct a user study in Appendix 2.4.

Viewpoint Comparison When watching 360° ODVs, viewers naturally focus on appealing local regions. To assess practical perceptual quality, we sample the Top-5 high-frequency observed viewpoints with a SURF-based detector (Xu, Zhou, and Chen 2020) and calculate their PSNR and SSIM. An illustrative example is provided in Fig. 11. By achieving 0.14 dB PSNR gains over IA-RT in Tab. 1, STDAN performs better on salient viewpoints where it matters most to human observers.

Ablation Study

In this section, we perform ablation studies to validate each proposed module, comparing the $\times 4$ SR baseline against variants that incrementally include individual components.

	Models with Proposed Modules	*ODV-SR* Dataset WS-PSNR / WS-SSIM
0	Baseline	29.83 / 0.8621
1	+Latitude Distortion Map	29.93 / 0.8643
2	+Saliency Attention Map	29.91 / 0.8641
3	+Spatio Distortion Alignment	29.85 / 0.8635
4	+Cylinder Position Encoding	29.88 / 0.8639
5	+Interlaced Multi-frame Reconstruction	30.06 / 0.8665
6	+Interlaced Multi-frame Fusion	30.02 / 0.8667
7	+Data Augmentation	30.04 / 0.8653

Table 2: $\times 4$ ODV-SR ablation study on different modules.

Fusion and Weight Parameters		*ODV-SR Dataset* WS-PSNR(dB) / WS-SSIM	
$\alpha_1, \beta_1 = 0$		29.83 / 0.8621	
$\alpha_1, \beta_1 = 0.001$		29.98 / 0.8657	
$\alpha_1, \beta_1 = 0.01$		**30.02** / 0.8667	
$\alpha_1, \beta_1 = 0.1$		29.69 / **0.8669**	
$\alpha_2 = 0$	$\beta_2 = 0$	29.83 / 0.8621	29.83 / 0.8621
$\alpha_2 = 0.01$	$\beta_2 = 0.01$	29.90 / **0.8644**	29.86 / 0.8637
$\alpha_2 = 0.1$	$\beta_2 = 0.1$	**29.93** / 0.8643	**29.91** / **0.8641**
$\alpha_2 = 1$	$\beta_2 = 1$	29.81 / 0.8610	29.83 / 0.8615

Table 3: $\times 4$ ODV-SR comparison of different fusion and weight parameters.

Effects of Proposed Modules

Tab. 2 summarizes the ablation results. Model 0 is the baseline and Models 1-7 add one module at a time. Specifically, Models 1 and 2 with the refined loss bring about 0.1 dB WS-PSNR gains. Model 3 exceeds the baseline by 0.13 dB, and its optimized variant Model 4 gains another 0.03 dB. Models 5 and 6 achieve about 0.2 dB improvements with iteration in three temporal sequences. A similar conclusion can be observed in WS-SSIM. These results indicate the effectiveness of the proposed modules, which can enhance ODV-SR performance. Besides, Model 7 boots WS-PSNR by 0.21 dB, confirming the value of the proposed *ODV-SR* dataset. Moreover, we visualize $\times 4$ SR results of Models 1-7 in Appendix 2.5 to provide an intuitive comparison.

Effects of Different Fusion and Weight Parameters

As discussed in IMFR and LSAM Section, we preset fusion parameters $\alpha_1, \beta_1 = 0.01$ and balance latitude and saliency weights with $\alpha_2, \beta_2 = 0.1$. As shown in Tab. 3, these default values are from a grid search and yield the best performance.

Conclusion

In this paper, we proposed a Spatio-Temporal Distortion Aware Network (STDAN) for omnidirectional video super-resolution (ODV-SR) in practical scenarios. First, we designed a spatio-temporal continuous alignment module with omni-positional coding to offset spatio-temporal distortions and discontinuities. Next, we presented an interlaced multi-frame reconstruction module to enhance the temporal consistency of restored ODVs. Finally, we introduced a latitude-saliency adaptive module to weigh regions with visually sensitive texture and human-watching interest. We further collected a diverse *ODV-SR* dataset and developed comprehensive experiments on it. Extensive experimental results on different datasets demonstrated that our STDAN achieves the best ODV-SR performance and is application-friendly with faster speed and real-world viewpoint enhancement.

References

An, H.; Zhang, X.; Zhao, S.; and Zhang, L. 2024. FATO: Frequency Attention Transformer for Omnidirectional Image Super-Resolution. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, 1–7.

Baniya, A. A.; Lee, T.-K.; Eklund, P. W.; and Aryal, S. 2023. Omnidirectional Video Super-Resolution using Deep Learning. *IEEE Transactions on Multimedia*.

Blog, N. 2016. Toward a practical perceptual video quality metric.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4778–4787.

Cao, M.; Mou, C.; Yu, F.; Wang, X.; Zheng, Y.; Zhang, J.; Dong, C.; Li, G.; Shan, Y.; Timofte, R.; et al. 2023. NTIRE 2023 Challenge on 360deg Omnidirectional Image and Video Super-Resolution: Datasets, Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1731–1745.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4947–4956.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5972–5981.

Dahou, Y.; Tliba, M.; McGuinness, K.; and O'Connor, N. 2021. ATSal: An Attention Based Architecture for Saliency Prediction in 360° Videos. In *International Conference on Pattern Recognition*, 305–320. Springer.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Deng, X.; Wang, H.; Xu, M.; Guo, Y.; Song, Y.; and Yang, L. 2021. LAU-Net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9189–9198.

Elbamby, M. S.; Perfecto, C.; Bennis, M.; and Doppler, K. 2018. Toward low-latency and ultra-reliable virtual reality. *IEEE Network*, 32(2): 78–84.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3897–3906.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016. Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging*, 2(2): 109–122.

Kim, E.; Kim, H.; Jin, K. H.; and Yoo, J. 2025. BF-STVSR: B-Splines and Fourier—Best Friends for High Fidelity Spatial-Temporal Video Super-Resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28009–28018.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 624–632.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 170–185.

Li, N.; and Liu, Y. 2024. VertexShuffle-Based Spherical Super-Resolution for 360-Degree Videos. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2022. VRT: A video restoration transformer. *arXiv preprint arXiv:2201.12288*.

Liao, K.; Xu, X.; Lin, C.; Ren, W.; Wei, Y.; and Zhao, Y. 2023. Cylin-Painting: Seamless 360 panoramic image outpainting and beyond. *IEEE Transactions on Image Processing*, 33: 382–394.

Linardos, P.; Mohedano, E.; Nieto, J. J.; O'Connor, N. E.; Giró-i Nieto, X.; and McGuinness, K. 2019. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*.

Liu, H.; Ma, W.; Ruan, Z.; Fang, C.; Shang, F.; Liu, Y.; Wang, L.; Wang, C.; and Jiang, D. 2024. A single frame and multi-frame joint network for 360-degree panorama video super-resolution. *Engineering Applications of Artificial Intelligence*, 134: 108601.

Ozcinar, C.; Rana, A.; and Smolic, A. 2019. Super-resolution of omnidirectional images using adversarial learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. IEEE.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.

Sajjadi, M. S.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6626–6634.

Shen, X.; Wang, Y.; Zheng, S.; Xiao, K.; Yang, W.; and Wang, X. 2025. Fast Omni-Directional Image Super-Resolution: Adapting the Implicit Image Function with Pixel and Semantic-Wise Spherical Geometric Priors. *arXiv preprint arXiv:2502.05902*.

Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; and Dong, C. 2022. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 35: 36081–36093.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time

single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Sun, Y.; Lu, A.; and Yu, L. 2017. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9): 1408–1412.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. TDAN: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3360–3369.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Cui, Y.; Li, Y.; Ren, W.; and Cao, X. 2024. Omnidirectional Image Super-resolution via Bi-projection Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5454–5462.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 1954–1963.

Xu, J.; Zhou, W.; and Chen, Z. 2020. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1724–1737.

Xu, K.; Yu, Z.; Wang, X.; Mi, M. B.; and Yao, A. 2024. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2546–2555.

Yoon, Y.; Chung, I.; Wang, L.; and Yoon, K.-J. 2022. SphereSR: 360° image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5677–5686.

Yu, F.; Wang, X.; Cao, M.; Li, G.; Shan, Y.; and Dong, C. 2023. OSRT: Omnidirectional image super-resolution with distortion-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13283–13292.

Zhou, X.; Zhang, L.; Zhao, X.; Wang, K.; Li, L.; and Gu, S. 2024. Video super-resolution transformer with masked inter&intra-frame attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25399–25408.

Zhou, Y.; Yu, M.; Ma, H.; Shao, H.; and Jiang, G. 2018. Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, 54–57. IEEE.

Spatio-Temporal Distortion Aware Omnidirectional Video Super-Resolution Technical Appendix

Hongyu An[1], Xinfeng Zhang[1*], Shijie Zhao[2], Li Zhang[2], Ruiqin Xiong[3]

[1]University of Chinese Academy of Sciences, [1]Bytedance Inc, [1]Peking University
anhongyu22@mails.ucas.ac.cn, xfzhang@ucas.ac.cn,
{zhaoshijie.0526, lizhang.idm}@bytedance.com, rqxiong@pku.edu.cn

Due to the lack of space in the main paper, we provide more details of the proposed Spatio-Temporal Distortion Aware Network (STDAN) in the supplementary material. In Section 1, we present the Equirectangular Projection (ERP) transformation relationships from 3D sphere to 2D plane and the formula derivation of the projection distortion quantization. More experimental results and visual comparisons can be found in Section 2.

1. Equirectangular Projection

In this section, we follow the coordinate representation in the main paper and provide a rigorous derivation of the ERP and its associated distortion analysis. More precisely, spherical coordinates are defined as (ρ, θ, ϕ), where $\theta \in (0, 2\pi)$ and $\phi \in (0, \pi)$ denote longitude and latitude, respectively. The corresponding planar horizontal and vertical coordinates are parameterized as (u, v).

1.1 Geometric Transformation The coordinate transformation between 3D sphere and 2D plane is defined as:

$$\begin{cases} \rho = \sqrt{x^2 + y^2 + z^2}, \\ \theta = \arctan(y/x), \\ \phi = \arcsin(z/\rho). \end{cases} \quad \begin{cases} x = \rho\cos(\phi)\cos(\theta), \\ y = \rho\cos(\phi)\sin(\theta), \\ z = \rho\sin(\phi). \end{cases} \tag{1}$$

$$\begin{cases} \theta = (u/W - 0.5)2\pi, \\ \phi = (0.5 - v/H)\pi. \end{cases} \quad \begin{cases} u = (\theta/2\pi + 0.5)W, \\ v = (0.5 - \phi/\pi)H. \end{cases} \tag{2}$$

where (*x*, *y*, *z*) are world coordinates responding to spherical coordinates (ρ, θ, ϕ), *H* and *W* indicate the height and width (in pixels) of the ERP omnidirectional image (ODI). Fig. 1 provides an illustrative example of the forward ERP projection and its exact inverse back-projection between the spherical and planar domains.

1.2 Projection Distortion Inspired by (Sun, Lu, and Yu 2017), we quantify the projection distortion via the local stretching ratio (STR) induced by the mapping from the spherical surface to the ERP plane. As shown in Fig. 1, assuming a spherical element $d\theta d\phi$ whose area is $\delta S(\theta, \phi)$. After ERP, the planar area element $dudv$ with a center point (u, v) is defined as $\delta P(u, v)$. The differential relation between element $d\theta d\phi$ and $\delta P(u, v)$ is captured by the Jacobian determinant $dudv = J(\theta, \phi)d\theta d\phi$ as follows:

*Corresponding author.

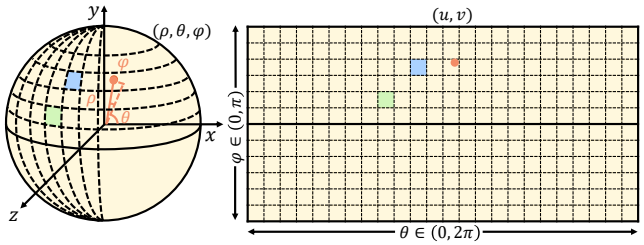


Figure 1: Geometric explanation of transforming between the ideal spherical surface and ERP plane.

$$J(\theta, \phi) = \frac{\partial(u, v)}{\partial(\theta, \phi)} = \begin{vmatrix} \frac{\partial(u)}{\partial(\theta)} & \frac{\partial(v)}{\partial(\phi)} \\ \frac{\partial(u)}{\partial(\theta)} & \frac{\partial(v)}{\partial(\phi)} \end{vmatrix}. \tag{3}$$

Furthermore, the area $\delta P(u, v)$ and $\delta S(\theta, \phi)$ are equal to $|dudv|$ and $d\theta d\phi\cos(\phi)$, respectively. Therefor, the area stretching ratio $\text{STR}(u, v)$ can be derived as:

$$\begin{aligned} \text{STR}(u, v) &= \frac{\delta S(\theta, \phi)}{\delta P(u, v)} = \frac{|d\theta d\phi|\cos(\phi)}{|dudv|} = \frac{\cos(\phi)}{\left|\frac{\partial(u,v)}{\partial(\theta,\phi)}\right|} \\ &= \frac{\cos(\phi)}{|J(\theta, \phi)|}. \end{aligned} \tag{4}$$

From Eq. 2, we can conclude that $|J(\theta, \phi)| = 1$, thus obtaining $\text{STR}(u, v) = \cos(\phi)$ for ERP. As a result, ERP distortions are fully characterized by latitudes. As latitude increases, STR decreases, and distortion becomes more severe. Therefore, we formulate STR as latitude-related pixel weights ranging from 0 to 1:

$$W_{lat}(u, v) = \cos((v - (H/2) + 0.5)\pi/H). \tag{5}$$

2. Experimental Results

As mentioned in the main paper, more comparisons of computational efficiency, visualization, spatio-temporal consistency, user study, and ablation study are provided in this section. To provide a thorough demonstration of the effectiveness of STDAN on the real-world ODV-SR task, we examine various aspects in practical application scenarios.

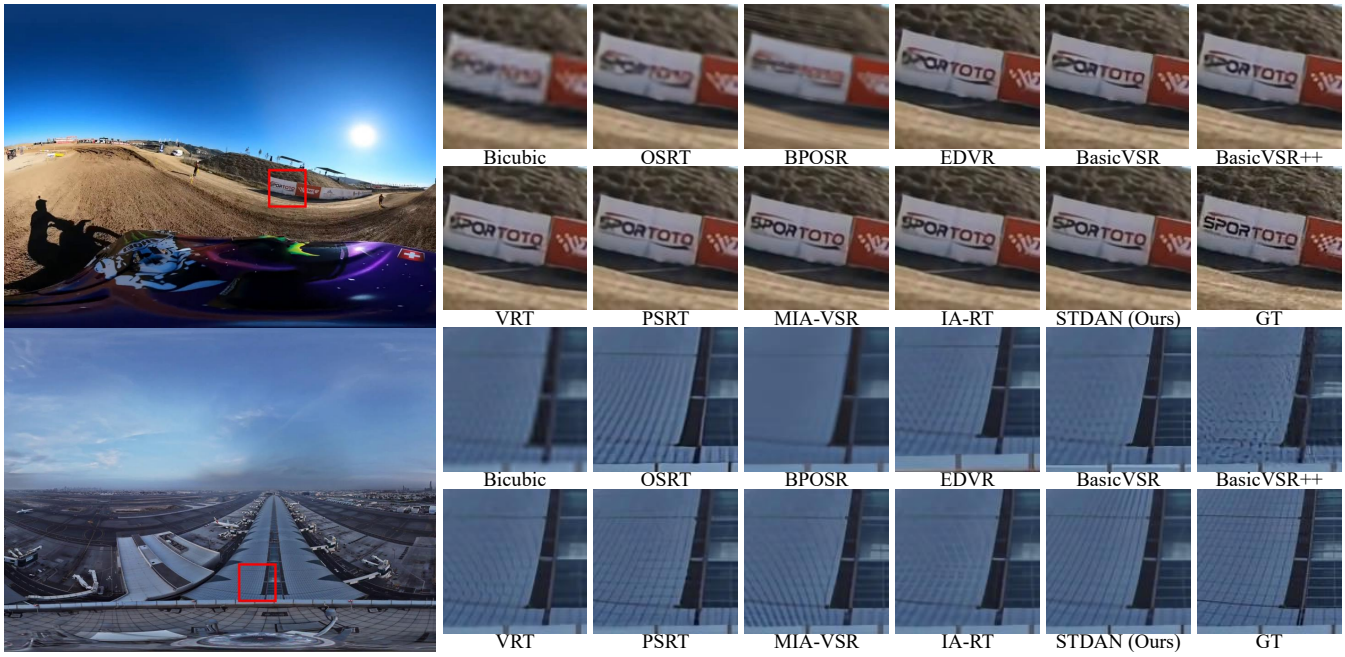


Figure 2: Qualitative $\times 4$ ODV-SR comparison on ERP *ODV-SR* dataset.

Method	OSRT	BPOSR	BasicVSR	BasicVSR++	VRT	PSRT	MIA-VSR	IA-RT	STDAN(Ours)
Parameters(M)	1.09	2.07	6.29	7.03	33.44	12.31	15.64	12.36	13.12
FLOPs(T)	0.1597	0.3507	0.8968	0.9619	3.2097	4.8029	2.8384	4.9163	1.8621
Runtime(s)	0.7243	0.4378	0.1859	0.2512	1.8507	4.6278	3.3991	5.4140	0.6960
WS-PSNR(dB)	28.11	28.41	29.64	29.83	30.06	30.14	30.17	30.18	30.23

Table 1: Computational efficiency comparison of $\times 4$ ODV-SR on *ODV-SR* dataset.

2.1 Computational Efficiency Comparison The computational efficiency is evaluated and shown in Tab. 1. As the number of FLOPs(T) and Runtime(s) is related to the resolution of input video sequences, we therefore report the average per-frame results on the proposed *ODV-SR* dataset with the practical high-resolution (HR) need of ODVs. We also adopt this approach to facilitate fair comparison between single-frame ODI-SR methods (OSRT and BPOSR) and multi-frame models (BasicVSR, BasicVSR++, VRT, PSRT, MIA-VSR, and IA-RT). From Tab. 1, STDAN achieves a faster running speed and fewer FLOPs compared to SOTA Transformer-based VSR methods VRT, PSRT, MIA-VSR, and IA-RT, primarily due to the lightweight network architecture design. Meanwhile, STDAN performs best in the term of WS-PSNR by incorporating ODV-specified spatio-temporal modulation modules. Notably, compared to IA-RT, our STDAN has a similar model size but operates much faster with better reconstruction performance.

2.2 Qualitative Comparison We visualize more comparison results of different VSR methods in Fig. 2. In the first case, only STDAN restores the letters "S" without wrong branches. As for curved architectural texture structures, STDAN reconstructs continuous and consistent curves and matches the texture distribution in the GT image. Although PSRT generates a grid-like structure successfully, the lines within it are twisted and disconnected. More intuitively, STDAN clearly rebuilds the exact count of glass panes while SOTA MIA-VSR and IA-RT introduce superfluous white window dividers.

2.3 Spatio-Temporal Consistency Comparison As described in the main paper, we visualize temporal profiles of different algorithms in Fig. 3. The temporal profiles are obtained by stitching horizontal rows at the same location in consistent frames. It is observed that multi-frame based methods significantly outperform single-frame based approaches (Bicubic, BPOSR, and OSRT). Moreover, the reconstruction results of STDAN avoid over-smoothing and are closer to the ground truth.

2.4 User Study To intuitively compare the spatio-temporal quality of ODVs reconstructed by different methods, we perform a user study to select the better one from the results of STDAN and other approaches. In detail, we invite 15 participants to compare STDAN with BasicVSR++, VRT, PSRT, MIA-VSR, and IA-RT in pairs, respectively. As depicted in Fig. 5, volunteers prefer the performance of our STDAN.

2.5 Ablation Study As mentioned in the main paper, extensive ablation studies are conducted to demonstrate the ef-

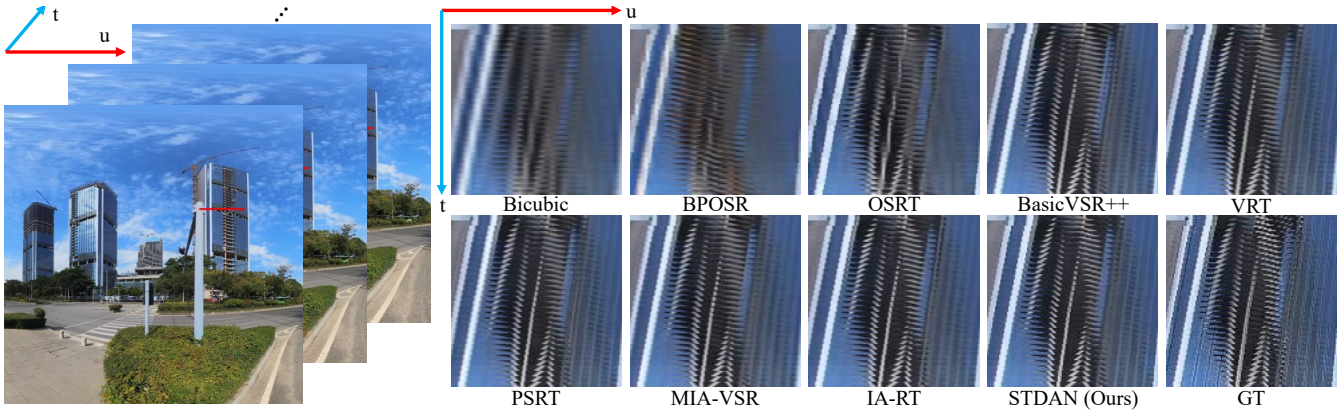


Figure 3: Temporal profile comparison of $\times 4$ ODV-SR on ERP *ODV-SR* dataset.

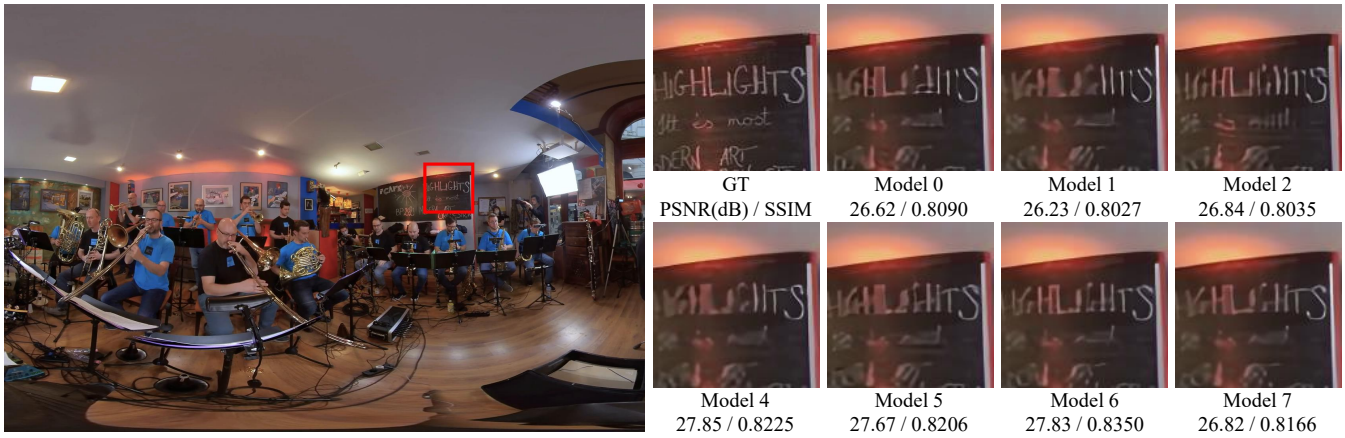


Figure 4: Subjective ablation study of $\times 4$ ODV-SR with different modules on ODV-SR dataset. The PSNR and SSIM of the local regions are provided for a more intuitive comparison.

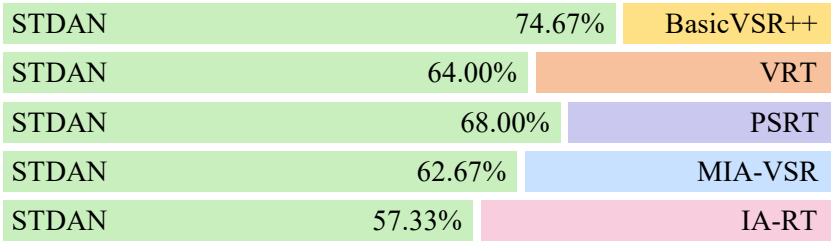


Figure 5: User study results of $\times 4$ ODV-SR on *ODV-SR* dataset. The values represent the percentage of the participants preferring the results of STDAN compared with other methods.

fectiveness of the introduced spatio-temporal modules. Furthermore, we visualize $\times 4$ ODV-SR results of models with proposed modules. As presented in Fig. 4, the improved models reconstruct sharper frames, especially Models 4, 5, 6, and 7, which successfully recover the letters "GHTS" that are blurred in the baseline Model 0. Compared with the baseline network, incorporating the proposed modules can produce more accurate texture details with a pleasing visual feeling. In terms of patch PSNR and SSIM, Model 1 degrades since it inclines to emphasize equator areas. Other models all make progress, and Model 6 has an obviously better SSIM, which improves the visual quality.

References

Sun, Y.; Lu, A.; and Yu, L. 2017. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9): 1408–1412.