SpatioTemporal Difference Network for Video Depth Super-Resolution

Zhengxue Wang[1], Yuan Wu[1], Xiang Li[2], Zhiqiang Yan[3*], Jian Yang[1*]

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Nankai University
[3]National University of Singapore
{zxwang, wuyuan, csjyang}@njust.edu.cn, xiang.li.implus@nankai.edu.cn, yanzq@nus.edu.sg

Abstract

Depth super-resolution has achieved impressive performance, and the incorporation of multi-frame information further enhances reconstruction quality. Nevertheless, statistical analyses reveal that video depth super-resolution remains affected by pronounced long-tailed distributions, with the long-tailed effects primarily manifesting in spatial non-smooth regions and temporal variation zones. To address these challenges, we propose a novel SpatioTemporal Difference Network (STDNet) comprising two core branches: a spatial difference branch and a temporal difference branch. In the spatial difference branch, we introduce a spatial difference mechanism to mitigate the long-tailed issues in spatial non-smooth regions. This mechanism dynamically aligns RGB features with learned spatial difference representations, enabling intra-frame RGB-D aggregation for depth calibration. In the temporal difference branch, we further design a temporal difference strategy that preferentially propagates temporal variation information from adjacent RGB and depth frames to the current depth frame, leveraging temporal difference representations to achieve precise motion compensation in temporal long-tailed areas. Extensive experimental results across multiple datasets demonstrate the effectiveness of our STDNet, outperforming existing approaches.

Code — https://github.com/yanzq95/STDNet

Introduction

Depth data constitutes a fundamental component in various fields, including 3D reconstruction (Im et al. 2018; Lian et al. 2025; Yan et al. 2025c), virtual reality (Yan et al. 2022b; Lian et al. 2023; Zhou et al. 2023; Yan et al. 2024), and augmented reality (Song et al. 2020; Yan et al. 2025a; Yin et al. 2023). Recently, numerous depth super-resolution (DSR) methods (Guo et al. 2018; Wang et al. 2023a,b) have been proposed to reconstruct high-resolution (HR) depth from low-resolution (LR) inputs, achieving remarkable performance. Furthermore, (Sun et al. 2023) introduce a video depth super-resolution (VDSR) framework that effectively aggregates multi-frame RGB-D features, demonstrating substantial improvements over single-frame approaches.

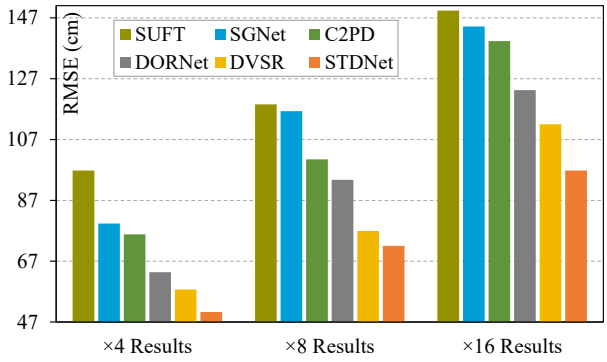
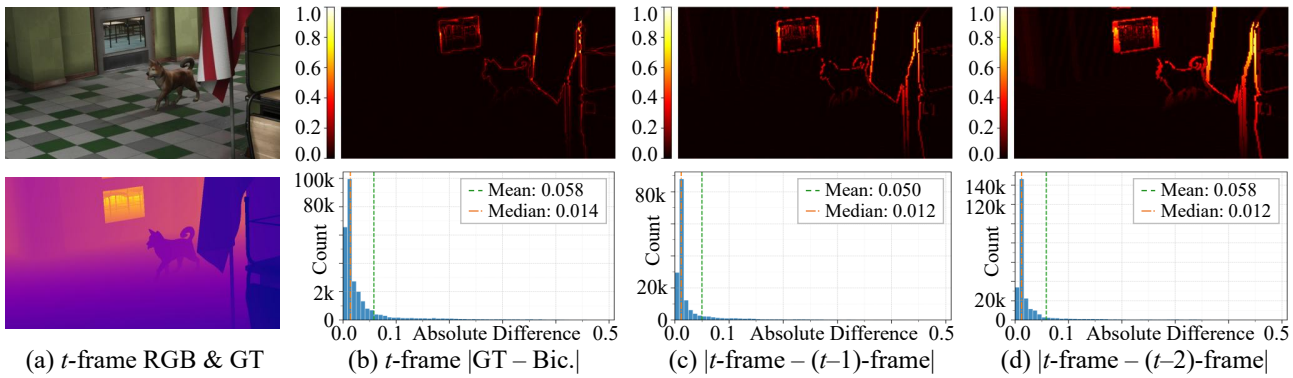


Figure 1: Quantitative comparisons between our STDNet and previous state-of-the-art methods on TarTanAir dataset.

However, as shown in Fig. 2, VDSR manifests long-tailed distributions across both spatial and temporal dimensions. Specifically, Fig. 2(b) quantifies the spatial difference between the ground truth (GT) depth and the upsampled LR depth. Statistical results indicate that spatial non-smooth regions exhibit distinct long-tailed issues, accounting for only a small fraction of the overall depth data. These regions pose substantially greater reconstruction challenges than the dominant smooth areas. Furthermore, Figs. 2(c) and (d) respectively present the difference results between consecutive and cross depth frames, demonstrating that temporal variation zones (e.g., dynamic objects, edge contours, and occlusion areas) in the depth videos are primarily concentrated in the long-tailed portion of the distributions.

Building upon the above statistical analysis, we propose a spatiotemporal difference network (STDNet) that focuses on handling long-tailed distributions in VDSR. The STDNet mainly consists of two dedicated branches: spatial difference branch and temporal difference branch. To effectively mitigate the spatial long-tailed distribution issues, our spatial difference branch implements a spatial difference mechanism. This mechanism uses learned spatial difference representations to precisely align intra-frame RGB features with non-smooth depth regions. These aligned RGB features are selectively aggregated to enhance the depth prediction. In the temporal difference branch, we prioritize regions with significant temporal variations, which typically exhibit long-

*Corresponding authors



(a) t-frame RGB & GT (b) t-frame |GT – Bic.| (c) |t-frame – (t–1)-frame| (d) |t-frame – (t–2)-frame|

Figure 2: Visualization of (a) RGB and GT depth at frame t, (b) absolute difference representations (top) and corresponding histogram distribution (bottom) between GT depth and bicubic-upsampled LR depth (Bic.). (c) shows the error analysis between consecutive frames (t and $t-1$ bicubic-upsampled LR depth), while (d) presents cross frame results between frames t and $t-2$.

tailed characteristics. To this end, we first estimate the temporal difference representations between consecutive frames and cross frames in the depth videos. Then, we develop a temporal difference strategy that propagates information from multiple adjacent RGB and depth frames to the current depth frame. In this strategy, spatial and temporal difference representations are jointly employed to facilitate the alignment of multi-frame and multi-modal RGB-D data, enhancing temporal consistency in the predicted depth videos. Furthermore, we introduce a difference regularization comprising both spatial and temporal difference terms to optimize the learning of spatiotemporal difference representations.

Owing to these innovative designs, our method successfully restores accurate HR depth videos. As shown in Fig. 1, STDNet outperforms five state-of-the-art approaches by 32.6% ($\times$4), 28.8% ($\times$8), and 27.6% ($\times$16) in average.

In summary, our contributions are as follows:

- Based on statistical analysis, we introduce a novel VDSR perspective that exploits the spatiotemporal long-tailed characteristics to enhance depth videos.
- We propose a novel framework termed STDNet, which comprises dual spatiotemporal difference branches. The spatial difference branch focuses on mitigating the long-tailed effects in spatial non-smooth regions, while the temporal difference branch prioritizes multi-frame RGB-D aggregation in temporal variation areas.
- Extensive experiments demonstrate that our STDNet effectively recovers high-quality depth videos, achieving state-of-the-art performance.

Related Work

Depth Super-Resolution

Recently, single-frame DSR methods (Ye et al. 2020; De Lutio et al. 2022; Chen et al. 2024) have made remarkable progress. Existing approaches can be broadly categorized into filtering-based methods (Metzger, Daudt, and Schindler 2023; Zhong et al. 2023; Wang et al. 2024), multi-modal fusion-based methods (Zhong et al. 2021; Wang et al. 2022; Zhao et al. 2023), multi-task collaborative methods (Sun et al. 2021; Yan et al. 2022a), and structure-oriented methods (Yan et al. 2025b; Bi et al. 2025a; Zheng, Han, and Shen 2025; Bi et al. 2025b). For example, (Kim, Ponce, and Ham 2021) integrate deformable networks with joint image filtering to adaptively transfer RGB information to depth features. To effectively aggregate RGB-D, (Zhao et al. 2022) introduce a discrete cosine network to disentangle the shared and private information present in RGB and depth features. Additionally, (Sun et al. 2021) and (Tang et al. 2021) employ an auxiliary depth estimation network to effectively fuse RGB and depth features. More recently, several methods have focused on reconstructing high-frequency information. For instance, (Yuan et al. 2023) develop a recurrent structure attention to perform frequency decomposition and edge restoration. Unlike these single-frame approaches, our method focuses more on addressing the spatiotemporal long-tailed distribution issues inherent in VDSR, enabling robust reconstruction of temporally consistent HR depth videos.

Video RGB Super-Resolution

Single-modal video RGB super-resolution (VSR) aims to restore HR RGB videos from the corresponding LR inputs. Existing methods (Isobe et al. 2022; Gao et al. 2022; Shi et al. 2022) can be categorized into sliding window-based and recurrent-based approaches. Specifically, sliding window-based methods (Wang et al. 2019; Li et al. 2020; Cao et al. 2021) employ a temporal sliding window to align neighboring frames with the current frame within the window, achieving impressive performance. However, such methods are typically limited by their fixed window size and can only integrate information from a restricted number of nearby frames. To overcome this constraint, recent work has introduced bidirectional recurrent architectures (Chan et al. 2021; Hu et al. 2025; Zhou et al. 2024) that can aggregate information across the entire video sequence for feature enhancement. In contrast to VSR, VDSR presents unique challenges in establishing multi-frame and multi-modal correspondences between RGB and depth videos. These fundamental discrepancy hinder the effectiveness of existing advanced VSR methods in reconstructing depth videos.

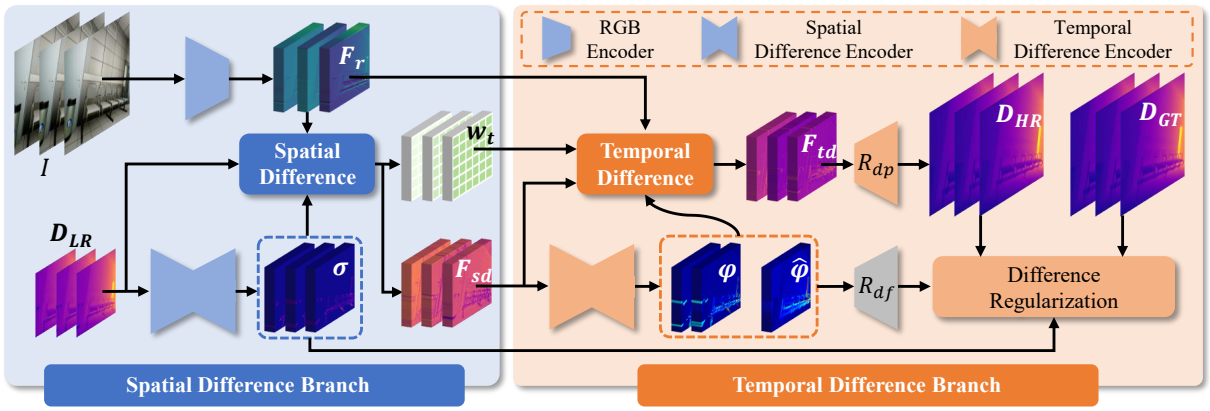


Figure 3: Overview of STDNet. Given $\boldsymbol{D}_{LR}$, we first predict its spatial difference representation $\boldsymbol{\sigma}$. Then, $\boldsymbol{D}_{LR}$, $\boldsymbol{I}$, and $\boldsymbol{\sigma}$ are jointly fed into the spatial difference to enhance non-smooth regions, producing $\boldsymbol{F}_{sd}$. Next, we estimate the temporal difference representations for consecutive frames and cross frames, generating $\boldsymbol{\varphi}$ and $\hat{\boldsymbol{\varphi}}$. These difference representations are used to propagate adjacent RGB and depth frames to the current depth frame, generating HR depth video $\boldsymbol{D}_{HR}$. Finally, a degradation regularization takes $\boldsymbol{D}_{HR}$, $\boldsymbol{D}_{GT}$, $\boldsymbol{\sigma}$, $\boldsymbol{\varphi}$, and $\hat{\boldsymbol{\varphi}}$ as inputs to optimize the learning of spatiotemporal difference representations.

Video Depth Enhancement

Although single-frame depth enhancement methods have demonstrated success in recovering high-quality depth from degraded inputs, they often exhibit suboptimal performance and temporal inconsistency when applied to depth video sequences. To address these limitations, recent advances in video depth enhancement have investigated temporal fusion strategies that integrate information from both current and adjacent frames. For example, (Sun et al. 2023) propose the first dToF-based VDSR method, which effectively harnesses rich information from multi-frame videos sequences to mitigate spatial ambiguity in LR depth videos. In addition, (Dong et al. 2024) develop a multi-frame depth denoising network that models the intra-scene geometric correlations and inter-frame noise distribution correlations, effectively suppressing multi-path interference and shot noise. More recently, (Zhu et al. 2025) introduce a video depth completion framework that integrates multi-frame features through an adaptive frequency selection fusion module. Different from them, we focus on leveraging the long-tailed distribution characteristics of depth videos to enhance both non-smooth regions and temporal variation zones.

Method

Overall Architecture

As illustrated in Fig. 3, our STDNet mainly comprises two branches: a spatial difference branch and a temporal difference branch. In the spatial difference branch, we first predict spatial difference representation $\boldsymbol{\sigma} \in \mathbb{R}^{T\times h\times w\times c}$ from the LR depth video $\boldsymbol{D}_{LR} \in \mathbb{R}^{T\times h\times w\times 1}$. c, h, w, and T are the channels, height, width, and the number of frames respectively. Then, $\boldsymbol{D}_{LR}$, $\boldsymbol{\sigma}$, and RGB video $\boldsymbol{I} \in \mathbb{R}^{T\times sh\times sw\times 3}$ are fed into the spatial difference, which adaptively performs intra-frame RGB-D aggregation to enhance spatial non-smooth of LR depth video, producing enhanced depth feature $\boldsymbol{F}_{sd}$ and spatial difference weights $\boldsymbol{w}_t$. s is upsampling factor. In the temporal difference branch, we utilize $\boldsymbol{F}_{sd}$ to estimate temporal difference representations for both consecutive and cross frames, yielding $\boldsymbol{\varphi} \in \mathbb{R}^{(T-1)\times h\times w\times c}$ and $\hat{\boldsymbol{\varphi}} \in \mathbb{R}^{(T-2)\times h\times w\times c}$, respectively. Subsequently, the temporal difference takes $\boldsymbol{F}_{sd}$, $\boldsymbol{w}_t$, $\boldsymbol{\varphi}$, $\hat{\boldsymbol{\varphi}}$, and RGB feature $\boldsymbol{F}_r$ as inputs to selectively transform adjacent RGB-D frames into current depth frame, optimizing the temporal variation zones in the depth video. Finally, the reconstructed HR depth video $\boldsymbol{D}_{HR} \in \mathbb{R}^{T\times sh\times sw\times 1}$, GT depth video $\boldsymbol{D}_{GT} \in \mathbb{R}^{T\times sh\times sw\times 1}$, and spatiotemporal difference representations ($\boldsymbol{\sigma}$, $\boldsymbol{\varphi}$, and $\hat{\boldsymbol{\varphi}}$) are input into the difference regularization to facilitate difference learning.

Spatial Difference Branch

As shown in the blue part of Fig. 3, our spatial difference branch is designed to accurately enhance spatial non-smooth regions. Specifically, we first encode LR depth video $\boldsymbol{D}_{LR}$ to depth feature $\boldsymbol{F}_d$, and then predict its spatial difference representation $\boldsymbol{\sigma}$:

$$\boldsymbol{\sigma} = |\boldsymbol{F}_d - f_{bu}(f_{bd}(\boldsymbol{F}_d))|, \tag{1}$$

where f_{bd} and f_{bu} are bilinear downsampling and upsampling operations ($\times 2$), respectively. Subsequently, the RGB feature $\boldsymbol{F}_r$, depth feature $\boldsymbol{F}_d$, and $\boldsymbol{\sigma}$ are fed into the proposed spatial difference for intra-frame RGB-D aggregation.

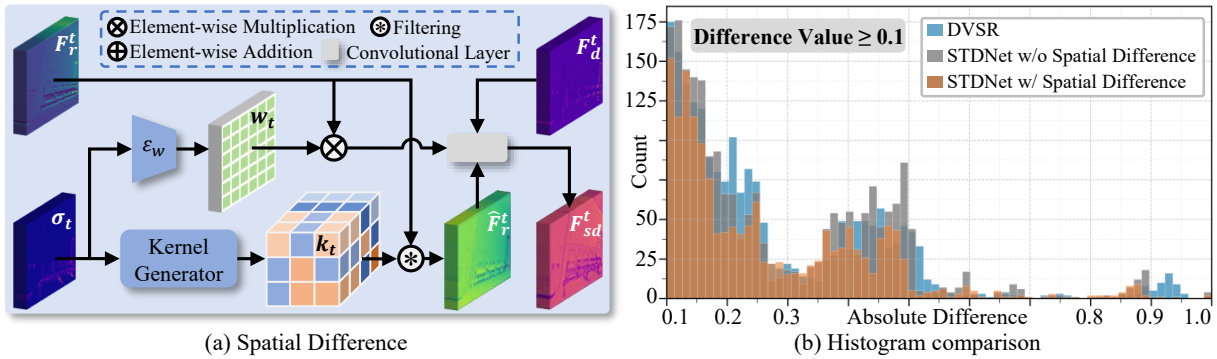
Spatial Difference. Our spatial difference is depicted in the Fig. 4(a). Given t-th frame difference representation $\boldsymbol{\sigma}_t$, we first generate filtering kernel $\boldsymbol{k}_t$ through a kernel generator $\mathcal{G}$, which will be applied to filter the RGB features, ensuring their alignment with spatial non-smooth regions:

$$\boldsymbol{k}_t = \mathcal{G}(\boldsymbol{\sigma}_t), \tag{2}$$

where generator $\mathcal{G}$ is composed of multiple convolutional layers and activation function layers.

Then, we employ encoder $\mathcal{E}_w$ transforms the $\boldsymbol{\sigma}_t$ into adaptive weight $\boldsymbol{w}_t$:

$$\boldsymbol{w}_t = \mathcal{E}_w(\boldsymbol{\sigma}_t), \tag{3}$$



(a) Spatial Difference (b) Histogram comparison

Figure 4: Details of (a) spatial difference, and (b) histogram comparison between our STDNet and DVSR (Sun et al. 2023).

where $\mathcal{E}_w$ consists of a 3×3 convolutional layer, maximum function, mean function, and sigmoid function.

Finally, the predicted filtering kernel $\boldsymbol{k}_t$ and weight $\boldsymbol{w}_t$ are leveraged to selectively propagate the aligned t-th frame RGB features to the depth feature $\boldsymbol{F}_d^t$, yielding the calibrated t-th frame depth feature $\boldsymbol{F}_{sd}^t$:

$$\boldsymbol{F}_{sd}^t = f_c(\boldsymbol{F}_d^t, \boldsymbol{w}_t \otimes \boldsymbol{F}_r^t, \widehat{\boldsymbol{F}}_r^t), \tag{4}$$

where $\widehat{\boldsymbol{F}}_r^t = \mathcal{F}(\boldsymbol{F}_r^t, \boldsymbol{k}_t)$. f_c is a convolutional layer, while $\otimes$ and $\mathcal{F}$ are the element-wise multiplication and filtering operation, respectively. Fig. 4(b) compares the histogram distributions in the long-tailed regions (difference value ≥ 0.1). Compared to state-of-the-art DVSR, our spatial difference effectively mitigates spatial long-tailed issues, thereby enhancing non-smooth regions in LR depth videos.

Temporal Difference Branch

Statistical results reveal that depth videos exhibit significant long-tailed distributions along the temporal dimension. Unlike existing multi-frame fusion methods, our temporal difference strategy prioritizes motion compensation in temporal variation regions. As illustrated in the orange part of Fig. 3, we first employ a temporal difference encoder to predict two difference representations: $\boldsymbol{\varphi}$ for consecutive frames and $\widehat{\boldsymbol{\varphi}}$ for cross frames:

$$\begin{aligned} \boldsymbol{\varphi}_t &= |\boldsymbol{F}_{sd}^t - \boldsymbol{F}_{sd}^{t+1}|, \forall t \in \{1, 2, \cdots, T-1\}, \\ \widehat{\boldsymbol{\varphi}}_t &= |\boldsymbol{F}_{sd}^t - \boldsymbol{F}_{sd}^{t+2}|, \forall t \in \{1, 2, \cdots, T-2\}. \end{aligned} \tag{5}$$

Given $\boldsymbol{F}_r$, $\boldsymbol{w}_t$, $\boldsymbol{F}_{sd}$, $\boldsymbol{\varphi}$ and $\widehat{\boldsymbol{\varphi}}$, we then follow a common bidirectional iterative scheme (Chan et al. 2022; Sun et al. 2023) to execute the proposed temporal difference strategy, fully aggregating information from adjacent RGB and depth frames. Please see our appendix for more iterative details.

Next, all forward and backward output features are aggregated to generate the temporally enhanced depth feature sequence $\boldsymbol{F}_{td}$. Finally, the HR depth videos are reconstructed through the depth reconstruction module $\mathcal{R}_{dp}$, which consists of convolutional layers and pixel shuffle layers:

$$\boldsymbol{D}_{HR} = \mathcal{R}_{dp}(\boldsymbol{F}_{td}). \tag{6}$$

Temporal Difference. Our temporal difference is delineated in Fig. 5(a), consists of adjacent frame fusion and cross frame fusion. Taking forward propagation as an example, we use RGB and depth features from adjacent frames ($t-1$ and $t-2$) to enhance temporal variation regions in the t-th frame depth, guided by temporal difference representations $\boldsymbol{\varphi}_{t-1}$ and $\widehat{\boldsymbol{\varphi}}_{t-2}$. Conversely, the backward propagation employs subsequent RGB-D frames ($t+1$ and $t+2$) for complementary refinement of t-th frame depth.

In the adjacent frame fusion stage, we utilize a temporal difference encoder $\mathcal{E}_\varphi$ (composed of convolutional layers and sigmoid function) to project $\boldsymbol{\varphi}_{t-1}$ into offset δ_{t-1} and modulation scalar m_{t-1}. Deformable convolution $\mathcal{D}$ (Zhu et al. 2019) is then introduced to dynamically sample temporal variation information matched with the temporal difference representations. Additionally, adaptive weight $\boldsymbol{w}_t$ derived from the spatial difference branch is employed to mitigate cross-modal discrepancies between adjacent RGB frames and the current depth frames. The enhanced depth feature $\boldsymbol{F}_f^{t-1,t}$ for frame t can be expressed as:

$$\boldsymbol{F}_f^{t-1,t} = f_c(\boldsymbol{F}_f^t, \boldsymbol{F}_{f,dc}^{t-1}, \boldsymbol{w}_t \otimes \boldsymbol{F}_{r,dc}^{t-1}), \tag{7}$$

where intermediate feature $\boldsymbol{F}_{f,dc}^{t-1} = \mathcal{D}(\boldsymbol{F}_f^{t-1}, \delta_{t-1}, m_{t-1})$, $\boldsymbol{F}_{r,dc}^{t-1} = \mathcal{D}(\boldsymbol{F}_r^{t-1}, \delta_{t-1}, m_{t-1})$. $\boldsymbol{F}_f^t$ and $\boldsymbol{F}_f^{t-1}$ are the depth features of frame t and frame $t-1$ during forward iteration.

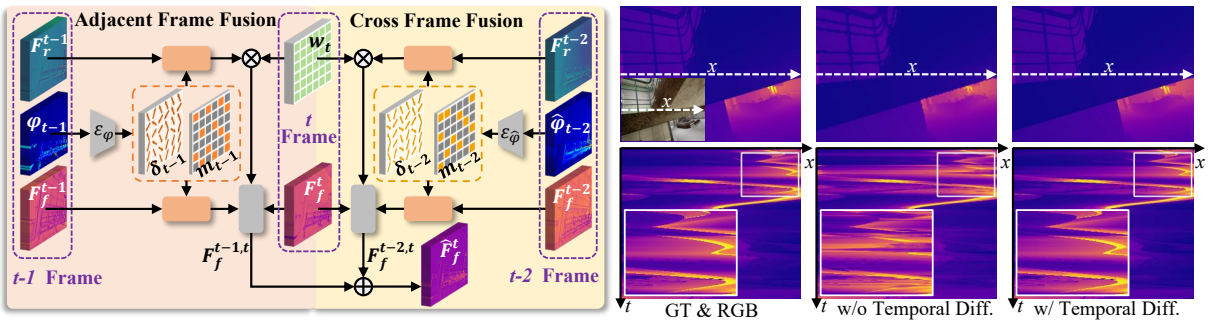
The cross frame fusion stage follows an analogous procedure to produce feature $\boldsymbol{F}_f^{t-2,t}$. Finally, the temporally-refined feature $\widehat{\boldsymbol{F}}_f^t$ for frame t are generated through the integration of adjacent frame fusion and cross frame fusion:

$$\widehat{\boldsymbol{F}}_f^t = \boldsymbol{F}_f^{t-1,t} + \boldsymbol{F}_f^{t-2,t}. \tag{8}$$

Fig. 5(b) visualizes central slices (white dashed line) across all frames of the predicted depth videos, demonstrating that our temporal difference strategy contributes to more stable temporal predictions, particularly in regions with temporal variations that exhibit long-tailed distributions. These results confirm that STDNet effectively enhances the temporal consistency of depth videos.

Loss Function

Given the GT depth video $\boldsymbol{D}_{GT}$ and predicted HR depth video $\boldsymbol{D}_{HR}$, we follow prior work (Sun et al. 2023) by applying Charbonnier regularization (Charbonnier et al. 1994)



(a) Temporal Difference (b) Comparison of temporal consistency

Figure 5: Details of (a) temporal difference, and (b) temporal consistency visualization for x–t slices (along dashed line). Diff.: Difference. Orange rectangular boxes are the deformable convolutional layers (Zhu et al. 2019).

Methods	×4			×8			×16			Venue
	RMSE↓	MAE↓	TEPE↓	RMSE↓	MAE↓	TEPE↓	RMSE↓	MAE↓	TEPE↓	
DJFR	75.56	10.59	10.19	105.45	18.43	14.15	141.14	31.22	20.27	PAMI 2019
CUNet	89.38	14.11	11.64	122.56	22.82	15.93	155.00	38.56	21.30	PAMI 2020
DKN	82.69	11.73	10.83	110.10	18.78	14.49	153.56	33.21	21.93	IJCV 2021
FDKN	79.39	11.14	10.66	109.10	18.48	14.79	147.61	29.31	19.77	IJCV 2021
FDSR	80.18	13.34	11.79	104.77	19.12	14.52	132.52	29.09	19.28	CVPR 2021
SUFT	96.80	15.87	11.93	118.57	22.09	15.45	149.40	34.72	19.46	MM 2022
SGNet	79.40	11.36	9.40	116.33	23.15	12.83	144.17	34.34	20.14	AAAI 2024
C2PD	75.83	13.12	10.36	100.48	18.70	13.03	139.36	40.86	19.32	AAAI 2025
DORNet	63.38	8.60	8.00	93.75	13.96	11.90	123.24	23.59	16.40	CVPR 2025
DVSR	57.72	4.40	5.33	76.96	7.74	8.04	112.04	14.39	11.06	CVPR 2023
STDNet	**50.28**	**3.73**	**4.58**	**72.03**	**6.75**	**6.54**	**96.80**	**12.01**	**8.90**	-

Table 1: Quantitative comparisons with existing state-of-the-art methods on the TarTanAir dataset.

as reconstruction loss $\mathcal{L}_{rec}$ to constrain our STDNet:

$$\mathcal{L}_{rec} = \sum_{q\in\mathbb{Q}} \sqrt{(\boldsymbol{D}^q_{GT} - \boldsymbol{D}^q_{HR})^2 + \epsilon}, \tag{9}$$

where $\mathbb{Q}$ is the set of valid pixels of $\boldsymbol{D}_{GT}$. $\epsilon = 1 \times 10^{-12}$.

Degradation Regularization. To optimize the learned spatiotemporal difference representations, we introduce difference regularization $\mathcal{L}_{diff}$ comprising two terms: spatial difference term $\mathcal{L}_{sd}$ and temporal difference term $\mathcal{L}_{td}$:

$$\mathcal{L}_{diff} = \alpha_1 \mathcal{L}_{sd} + \alpha_2 \mathcal{L}_{td}. \tag{10}$$

where α_1 and α_2 are tunable hyper-parameters.

For the spatial difference term, we introduce an uncertainty constraint (Ning et al. 2021) to facilitate the learning of difference representations in non-smooth regions:

$$\mathcal{L}_{sd} = \sum_{q\in\mathbb{Q}} (\boldsymbol{\sigma}^q - min(\boldsymbol{\sigma}^q)) ||\boldsymbol{D}^q_{GT} - \boldsymbol{D}^q_{HR}||_1, \tag{11}$$

where $||\cdot||_1$ represents the L_1 norm.

The temporal difference term consists of two components: adjacent frame and cross frame losses:

$$\begin{aligned}\mathcal{L}_{td} = &\underbrace{\sum_{q\in\mathbb{Q}} ||\mathcal{R}_{df}(\boldsymbol{\varphi}^q) - \Phi(\boldsymbol{D}^q_{GT})||_1}_{\text{adjacent frame}} \\ &+ \underbrace{\sum_{q\in\mathbb{Q}} ||\mathcal{R}_{df}(\widehat{\boldsymbol{\varphi}}^q) - \Phi(\boldsymbol{D}^q_{GT})||_1}_{\text{cross frame}},\end{aligned} \tag{12}$$

where $\mathcal{R}_{df}$ represents the difference reconstruction, composed of bicubic upsampling and convolutional layers. Φ is temporal difference computation, as defined in Eq. (5).

The total loss $\mathcal{L}_{total}$ integrates both reconstruction loss $\mathcal{L}_{rec}$ and difference regularization $\mathcal{L}_{diff}$, formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \beta \mathcal{L}_{diff}. \tag{13}$$

where β is a tunable hyper-parameter.

Experiments

Experimental Setups

Datasets. Following previous methods (Sun et al. 2023; Zhu et al. 2025), we evaluate STDNet on TarTanAir (Wang et al. 2020), DyDToF (Sun et al. 2023), and DynamicReplica (Karaev et al. 2023) datasets. Since dataset preprocessing details from prior approaches are unavailable, we redefine the training and test sets. Specifically, we utilize the hard scenes from TarTanAir for training, consisting of $6,164$ RGB-D frames in the train set and $1,228$ frames in the test set. Then, the pre-trained model on TarTanAir is evaluated on DyDToF (576 frames) and DynamicReplica (500 frames) without any fine-tuning. Besides, DynamicReplica is center-cropped to match the size of TarTanAir (640×480). All LR depth are generated from GT and RGB using the same syn-

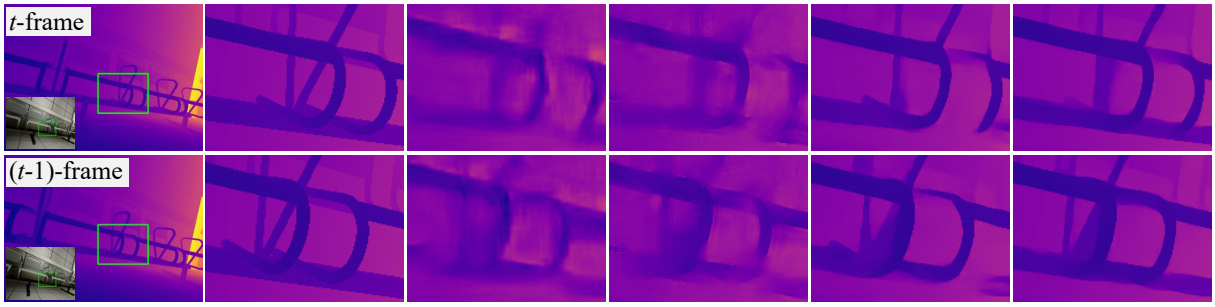


Figure 6: Visual results of consecutive frames on the TarTanAir at $\times 16$ upscaling.

Figure 7: Visual results on the DyDToF at $\times 8$ upscaling.

thesis pipeline as DVSR (Sun et al. 2023). We retrain the existing methods from scratch using their released code.

Implementation Details. Consistent with prior works (Sun et al. 2023; Zhu et al. 2025), we adopt root mean square error (RMSE), mean absolute error (MAE), and temporal end-point error (TEPE) as evaluation metrics (all measured in centimeters). During training, we randomly crop RGB and HR depth frames to 256×256. Besides, we employ the Adam optimizer (Kingma 2014) with an initial learning rate of 1×10^{-4} to train our STDNet. Our method is implemented using two NVIDIA RTX 4090 GPUs. The hyper-parameters are set as α_1=α_2=0.5 and β=0.01.

Comparison with the State-of-the-Art

We compare our STDNet with existing state-of-the-art approaches, including DJFR (Li et al. 2019), CUNet (Deng and Dragotti 2020), DKN (Kim, Ponce, and Ham 2021), FDSR (He et al. 2021), SUFT (Shi, Ye, and Du 2022), SGNet (Wang, Yan, and Yang 2024), C2PD (Kang et al. 2025), DORNet (Wang et al. 2025), and DVSR (Sun et al. 2023).

Quantitative Comparison. Tabs. 1-3 list quantitative comparisons across multiple datasets, demonstrating that our STDNet outperforms existing state-of-the-art approaches at $\times 4$, $\times 8$, and $\times 16$ scaling factors. Specifically, Tab. 1 shows that our method is superior to both the existing multi-frame DVSR and advanced single-frame approaches on TarTanAir. For example, compared to the second-best method, our STDNet reduces RMSE by$15.24cm$, MAE by $2.38cm$, and TEPE by $2.16cm$ on $\times 16$ VDSR. Additionally, Tabs. 2 and 3 further validate the generalization capability of our method on DyDToF and DynamicReplica. We observe that STDNet achieves outstanding performance, surpassing the suboptimal VDSR ($\times 16$) by $4.31cm$ in RMSE on DyDToF and by $0.15cm$ in RMSE on DynamicReplica.

Methods	$\times 4$		$\times 8$		$\times 16$	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
DJFR	22.39	4.72	32.26	6.90	44.76	11.73
CUNet	26.46	6.19	37.74	14.45	54.73	32.72
DKN	23.32	4.81	32.20	6.91	48.55	11.53
FDKN	22.61	4.65	32.17	7.10	43.82	11.57
FDSR	22.77	5.20	31.71	7.75	47.86	14.96
SUFT	59.06	17.37	63.00	20.35	80.35	36.97
SGNet	40.62	11.24	51.24	13.93	64.51	21.74
C2PD	26.05	6.35	31.19	11.68	49.12	29.03
DORNet	30.50	7.07	54.13	16.03	60.27	22.86
DVSR	<u>19.53</u>	<u>3.16</u>	<u>27.63</u>	<u>4.37</u>	<u>43.55</u>	<u>9.80</u>
STDNet	**18.23**	**3.04**	**26.87**	**4.09**	**39.24**	**8.72**

Table 2: Quantitative comparisons on the DyDToF.

Visual Comparison. Figs. 6 and 7 provide visual comparisons, clearly indicating that our method achieves more accurate depth recovery. For example, compared to previous approaches, the structure and shape of chair in Fig. 6 predicted by STDNet align more closely with the GT depth while exhibiting superior temporal stability. Additionally, Fig. 7 shows that our method results in more precise reconstruction of dynamic objects (e.g., dog) than others.

Model Complexity Analysis. Fig. 8 demonstrates that our STDNet maintains a comparable balance between parameters, performance, and inference time. Compared to single-frame approaches (DKN, SUFT, SGNet, C2PD, and DORNet), although our method exhibits higher time cost, it achieves a significant average reduction of $9.23M$ parameters and $35.82cm$ RMSE. Furthermore, STDNet outperforms the multi-frame DVSR with a $47.35ms$ improvement in inference speed and a $4.93cm$ gain in performance, while

Methods	×4		×8		×16	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
FDSR	0.42	0.08	0.80	0.31	1.51	0.66
SUFT	0.47	0.17	0.79	0.27	0.56	0.81
SGNet	0.44	0.08	0.72	0.25	1.50	0.75
C2PD	0.42	0.11	0.67	0.24	1.40	0.65
DORNet	0.46	0.10	0.60	0.17	1.39	0.79
DVSR	0.37	0.07	0.58	0.13	1.25	**0.48**
STDNet	**0.32**	**0.05**	**0.53**	**0.10**	**1.10**	0.51

Table 3: Quantitative comparisons on the DynamicReplica.

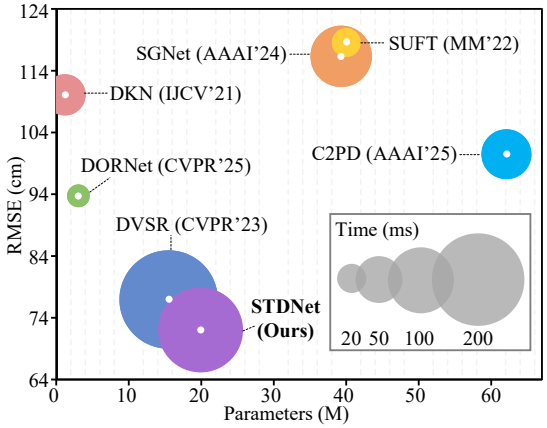


Figure 8: Complexity on the TarTanAir (×8) tested by a 4090 GPU. A larger circle area indicates larger time.

introducing only a modest parameter increase of $4.4M$.

Ablation Study

Spatial Difference and Temporal Difference. Fig. 9 illustrates an ablation study of spatial difference (SD) and temporal difference (TD). For the baseline, we replace all SD and TD in STDNet with concatenation operations, while keeping other architecture unchanged. As shown in Fig. 9(a), both SD and TD contribute to performance improvements over the baseline. For example, SD and TD reduce RMSE by $3.56cm$ and $14.02cm$ respectively on the DyDToF. When deployed together, STDNet achieves the best performance, surpassing the baseline by $17.94cm$ on the DyDToF.

Furthermore, Fig. 9(b) visualizes intermediate depth features using principal component analysis (PCA). It is clearly evident that both SD and TD facilitate more accurate depth structure. When SD and TD are combined, our STDNet produces sharper and clearer depth predictions. In summary, both quantitative and visual results demonstrate that our method effectively enhances VDSR performance, reconstructing high-quality depth videos.

Different Numbers of Adjacent Frames. Fig. 10(a) reveals the impact of different numbers of neighboring frames in TD. Compared to using only one RGB-D frames, incorporating additional neighboring frames significantly reduces the RMSE. However, the performance gains diminish when

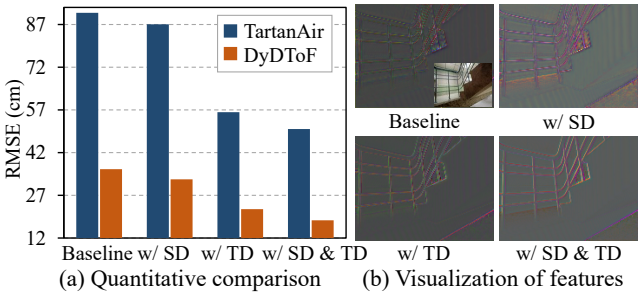


Figure 9: Ablation study of spatial difference (SD) and temporal difference (TD) on the TarTanAir and DyDToF (×4).

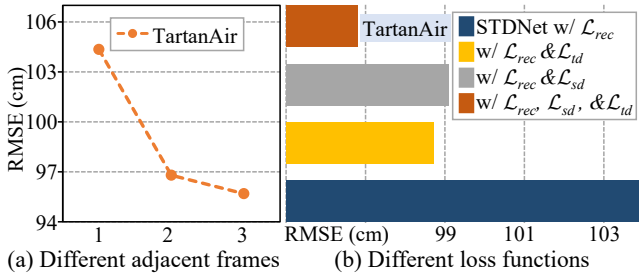


Figure 10: Ablation study of STDNet with (a) different numbers of adjacent frames and different loss functions (×16).

the number exceeds 2. Consequently, to balance computational cost and performance, our STDNet selects 2 frames (a adjacent frame and a cross frame) as the default setting.

Different Loss Functions. Fig. 10(b) presents the ablation study of different loss functions. The baseline is STDNet with the reconstruction loss $\mathcal{L}_{rec}$. These quantitative results demonstrate that both the spatial difference loss and temporal difference loss significantly improve performance. When $\mathcal{L}_{sd}$ and $\mathcal{L}_{td}$ are jointly employed, our method learns accurate spatiotemporal difference representations, thereby achieving high-quality depth restoration. Finally, STDNet surpasses the baseline by $7.08cm$ in RMSE on TarTanAir.

Conclusion

In this paper, we propose the spatiotemporal difference network, a novel framework that models spatiotemporal difference representations to address the inherent long-tailed distribution problems in VDSR. Specifically, we develop a spatial difference branch that incorporates a spatial difference mechanism to selectively transfer intra-frame RGB information to depth, effectively mitigating the long-tailed effects in spatial non-smooth regions. To enhance the temporal stability of predicted depth videos, our temporal difference branch implements the proposed temporal difference strategy, prioritizing the aggregation of multi-frame RGB-D features in temporal variation areas. Furthermore, a difference regularization is introduced to facilitate accurate difference representation learning. Extensive experiments demonstrate the effectiveness and superiority of our STDNet.

Acknowledgments

This work was supported by the NSFC under Grant Nos. U24A20330 and 62361166670.

References

Bi, J.; Wu, Q.; Qian, J.; Luo, L.; and Yang, J. 2025a. Dual Manifold Regularization Steered Robust Representation Learning for Point Cloud Analysis. In *AAAI*, 1844–1852.

Bi, J.; Wu, Q.; Qian, J.; Luo, L.; and Yang, J. 2025b. Structure-Aware Spherical Density Steered Cross-Domain Learning for Effective Point Cloud Understanding. *Pattern Recognition*, 112527.

Cao, J.; Li, Y.; Zhang, K.; and Van Gool, L. 2021. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*.

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 *CVPR*, 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 *CVPR*, 5972–5981.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, 168–172.

Chen, X.; Wang, H.; Chen, J.; Feng, K.; Liu, J.; Wang, X.; Zhang, W.; and Ni, B. 2024. Intrinsic Phase-Preserving Networks for Depth Super Resolution. In *AAAI*, 1210–1218.

De Lutio, R.; Becker, A.; D'Aronco, S.; Russo, S.; Wegner, J. D.; and Schindler, K. 2022. Learning graph regularisation for guided super-resolution. In *CVPR*, 1979–1988.

Deng, X.; and Dragotti, P. L. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3333–3348.

Dong, G.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024. Exploiting Dual-Correlation for Multi-frame Time-of-Flight Denoising. In *ECCV*, 473–489.

Gao, G.; Wang, Z.; Li, J.; Li, W.; Yu, Y.; and Zeng, T. 2022. Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer. *arXiv preprint arXiv:2204.13286*.

Guo, C.; Li, C.; Guo, J.; Cong, R.; Fu, H.; and Han, P. 2018. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, 28(5): 2545–2557.

He, L.; Zhu, H.; Li, F.; Bai, H.; Cong, R.; Zhang, C.; Lin, C.; Liu, M.; and Zhao, Y. 2021. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, 9229–9238.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025. Exploiting multimodal spatial-temporal patterns for video object tracking. In *AAAI*, 3581–3589.

Im, S.; Ha, H.; Choe, G.; Jeon, H.-G.; Joo, K.; and Kweon, I. S. 2018. Accurate 3d reconstruction from small motion clip for rolling shutter cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4): 775–787.

Isobe, T.; Jia, X.; Tao, X.; Li, C.; Li, R.; Shi, Y.; Mu, J.; Lu, H.; and Tai, Y.-W. 2022. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *CVPR*, 17411–17420.

Kang, J.; Cai, Q.; Tan, R.; Liu, Y.; and Liu, Z. 2025. C2pd: Continuity-constrained pixelwise deformation for guided depth super-resolution. In *AAAI*, 4212–4220.

Karaev, N.; Rocco, I.; Graham, B.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2023. Dynamicstereo: Consistent dynamic depth from stereo videos. In *CVPR*, 13229–13239.

Kim, B.; Ponce, J.; and Ham, B. 2021. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2): 579–600.

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

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, 335–351.

Li, Y.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2019. Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1909–1923.

Lian, J.; Du, X.; Liu, J.; Hui, L.; and Yang, J. 2025. Cross-Modal Driven Object Restoration for 3D Point Cloud Backdoor Defense. *IEEE Transactions on Information Forensics and Security*.

Lian, J.; Wang, D.-H.; Wu, Y.; and Zhu, S. 2023. Multi-branch enhanced discriminative network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 1263–1274.

Metzger, N.; Daudt, R. C.; and Schindler, K. 2023. Guided depth super-resolution by deep anisotropic diffusion. In *CVPR*, 18237–18246.

Ning, Q.; Dong, W.; Li, X.; Wu, J.; and Shi, G. 2021. Uncertainty-driven loss for single image super-resolution. *NeurIPS*, 34: 16398–16409.

Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; and Dong, C. 2022. Rethinking alignment in video super-resolution transformers. *NeurIPS*, 36081–36093.

Shi, W.; Ye, M.; and Du, B. 2022. Symmetric uncertainty-aware feature transmission for depth super-resolution. In *ACMMM*, 3867–3876.

Song, X.; Dai, Y.; Zhou, D.; Liu, L.; Li, W.; Li, H.; and Yang, R. 2020. Channel attention based iterative residual learning for depth map super-resolution. In *CVPR*, 5631–5640.

Sun, B.; Ye, X.; Li, B.; Li, H.; Wang, Z.; and Xu, R. 2021. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *CVPR*, 7792–7801.

Sun, Z.; Ye, W.; Xiong, J.; Choe, G.; Wang, J.; Su, S.; and Ranjan, R. 2023. Consistent direct time-of-flight video depth super-resolution. In *CVPR*, 5075–5085.

Tang, Q.; Cong, R.; Sheng, R.; He, L.; Zhang, D.; Zhao, Y.; and Kwong, S. 2021. Bridgenet: A joint learning network of depth map super-resolution and monocular depth estimation. In *ACMMM*, 2148–2157.

Wang, H.; Yang, M.; Lan, X.; Zhu, C.; and Zheng, N. 2022. Depth map recovery based on a unified depth boundary distortion model. *IEEE transactions on image processing*, 31: 7020–7035.

Wang, H.; Yang, M.; Zhu, C.; and Zheng, N. 2023a. RGB-guided depth map recovery by two-stage coarse-to-fine dense CRF models. *IEEE Transactions on Image Processing*, 32: 1315–1328.

Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; and Scherer, S. 2020. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 4909–4916.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 0–0.

Wang, X.; Chen, X.; Ni, B.; Tong, Z.; and Wang, H. 2023b. Learning continuous depth representation via geometric spatial aggregator. In *AAAI*, 2698–2706.

Wang, Z.; Yan, Z.; Pan, J.; Gao, G.; Zhang, K.; and Yang, J. 2025. DORNet: A Degradation Oriented and Regularized Network for Blind Depth Super-Resolution. In *CVPR*, 15813–15822.

Wang, Z.; Yan, Z.; and Yang, J. 2024. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *AAAI*, 5823–5831.

Wang, Z.; Yan, Z.; Yang, M.-H.; Pan, J.; Yang, J.; Tai, Y.; and Gao, G. 2024. Scene Prior Filtering for Depth Map Super-Resolution. *arXiv preprint arXiv:2402.13876*.

Yan, Z.; Jiao, J.; Wang, Z.; and Lee, G. H. 2025a. Event-Driven Dynamic Scene Depth Completion. *arXiv preprint arXiv:2505.13279*.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, G.; Li, J.; and Yang, J. 2022a. Learning complementary correlations for depth super-resolution with incomplete data in real world. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4): 5616–5626.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2022b. RigNet: Repetitive image guided network for depth completion. In *ECCV*, 214–230.

Yan, Z.; Wang, Z.; Dong, H.; Li, J.; Yang, J.; and Lee, G. H. 2025b. DuCos: Duality Constrained Depth Super-Resolution via Foundation Model. *arXiv preprint arXiv:2503.04171*.

Yan, Z.; Wang, Z.; Wang, K.; Li, J.; and Yang, J. 2025c. Completion as enhancement: A degradation-aware selective image guided network for depth completion. In *CVPR*, 26943–26953.

Yan, Z.; Zheng, Y.; Fan, D.-P.; Li, X.; Li, J.; and Yang, J. 2024. Learnable differencing center for nighttime depth perception. *Visual Intelligence*, 2(1): 15.

Ye, X.; Sun, B.; Wang, Z.; Yang, J.; Xu, R.; Li, H.; and Li, B. 2020. PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing*, 29: 7427–7442.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 9043–9053.

Yuan, J.; Jiang, H.; Li, X.; Qian, J.; Li, J.; and Yang, J. 2023. Recurrent structure attention guidance for depth super-resolution. In *AAAI*, 3331–3339.

Zhao, Z.; Zhang, J.; Gu, X.; Tan, C.; Xu, S.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2023. Spherical space feature decomposition for guided depth map super-resolution. In *ICCV*, 12547–12558.

Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; and Pfister, H. 2022. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, 5697–5707.

Zheng, H.; Han, W.; and Shen, J. 2025. Decoupling fine detail and global geometry for compressed depth map super-resolution. In *CVPR*, 951–960.

Zhong, Z.; Liu, X.; Jiang, J.; Zhao, D.; Chen, Z.; and Ji, X. 2021. High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing*, 31: 648–663.

Zhong, Z.; Liu, X.; Jiang, J.; Zhao, D.; and Ji, X. 2023. Deep attentional guided image filtering. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhou, M.; Yan, K.; Pan, J.; Ren, W.; Xie, Q.; and Cao, X. 2023. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision*, 131(1): 215–242.

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 *CVPR*, 25399–25408.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *CVPR*, 9308–9316.

Zhu, X.; Xiang, J.; Wang, X.; Liu, L.; Wang, Y.; Zhang, H.; Guo, F.; and Yang, X. 2025. Svdc: Consistent direct time-of-flight video depth completion with frequency selective fusion. In *CVPR*, 16619–16628.