Accelerating Stereo Image Simulation for Automotive Applications using Neural Stereo Super Resolution

Hamed Haghighi, Mehrdad Dianati, Valentina Donzella, Kurt Debattista

***Abstract*—Camera image simulation is integral to the virtual validation of autonomous vehicles and robots that use visual perception to understand their environment. It also has applications in creating image datasets for training learning-based vision models. As camera image simulation takes into account a wide variety of external and internal parameters, achieving a high-fidelity simulation is a computationally expensive process. Recently, several neural network-based techniques have been proposed to reduce the computational complexity of image rendering, a critical element of the camera simulation pipeline. However, the existing methods are tailored for monocular camera images and are not optimised for stereo images, which are widely used in autonomous driving applications. To address this, we propose a technique based on Stereo Super Resolution (SSR) to speed up the simulation of stereo images. The proposed method first simulates stereo images at a lower resolution, then super-resolves them to their original resolution using our introduced SSR model, ETSSR. We evaluated the performance of our technique using the CARLA driving simulator and created our own synthetic dataset for training ETSSR. The evaluations indicate that our approach can speed up stereo image simulation by a factor of up to 2.57 over various resolutions. Moreover, it shows that our ETSSR achieves on-par or superior performance compared to the state-of-the-art models, using significantly fewer parameters and FLOPs. We have made our source code and dataset available at https://github.com/hamedhaghighi/ETSSR.**

I. Introduction

Cameras are widely used in Autonomous Vehicles (AVs) and other robotic applications to aid the perception of the operational environment. Camera image simulation refers to the process of producing synthetic images based on simulated scene descriptions [1]. It is commonly used in virtual system validation [2] and has recently gained popularity as a method for creating/augmenting data for training learning-based vision models [3]. Many real-world applications, including AVs, use stereo cameras to provide depth information for 3D vision. As a result, simulation-based development and validation of such systems require fast and accurate stereo image simulation techniques such as the one proposed in this paper.

Camera image simulation is inherently a complex process due to the large number of parameters that need to be taken into account. Consequently, it demands a high level of computational resources. While this problem may be tolerated in some applications, it poses a significant challenge in the simulation-based validation of safety-critical technologies such as AVs. This stems from the fact that billions of miles of virtual driving are often needed to prove safety and reliability of such systems [4]. Hence, the development of faster-than-real-time simulation setups becomes crucial for system developers. In light of this necessity, our study aims to accelerate the simulation of stereo images that are widely utilised in AVs and related technologies.

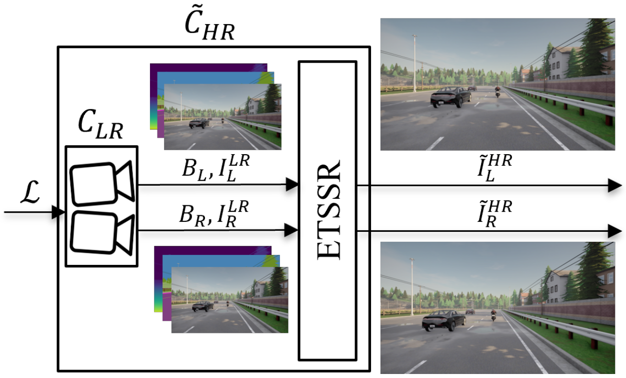


Fig. 1. Using Stereo Super Resolution (SSR) to accelerate stereo image simulation. First, a Low Resolution (LR) stereo camera model C_{LR} produces LR stereo images I_L^{LR}, I_R^{LR} and the respective auxiliary buffers B_L, B_R based on the description of the scene $\mathcal{L}$. Second, our proposed SSR model, ETSSR, produces High Resolution (HR) stereo images $\tilde{I}_L^{HR}, \tilde{I}_R^{HR}$ feeding in I_L^{LR}, I_R^{LR} and B_L, B_R. The accelerated HR camera model $\tilde{C}_{HR}$ is built by integrating the C_{LR} and the SSR model.

There are two generic approaches to camera image simulation in the literature: physics-based and data-driven techniques [2]. Physics-based simulation techniques are computationally complex as they rely on detailed calculations of complicated physical phenomena. To avoid the complexity of such models, a wide range of data-driven methods have been proposed in recent years. In the context of AVs and robotics, data-driven methods often leverage generative adversarial networks (GANs) [5] to enhance the realism of simulated images. Although these methods demonstrate promising results for offline applications, their adequacy for real-time and faster-than-real-time simulation remains a major technical challenge. In the gaming application, data-driven techniques use neural image enhancement models to reduce the computational costs of image rendering [6]. Although these techniques have

This research is supported in part by the University of Warwick's Centre for Doctoral Training in Future Mobility Technologies and in part by the Hi-Drive Project through the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 101006664.

The authors are with the Warwick Manufacturing Group (WMG), University of Warwick, Coventry, U.K.

* Corresponding author: Hamed.Haghighi@warwick.ac.uk

enabled real-time image rendering at high resolution, the deployed neural networks are not specifically optimised for stereo images in terms of computational complexity and image quality.

This paper proposes a technique based on Stereo Super Resolution (SSR) to accelerate the simulation of stereo images. As depicted in Figure 1, our approach initially models stereo camera images in Lower Resolution (LR) and then uses our proposed SSR model, ETSSR, to reconstruct them to their original (High) Resolution (HR). Our ETSSR is inspired by the success of Swin Transformer [7] models in the field of image restoration [8]. Specifically, these models have proven to offer superior PSNR/FLOPs or PSNR/parameters trade-offs when compared to other state-of-the-art models. Swin Transformer models have been employed for single-view feature extraction in image restoration tasks. However, we have taken this notion further by introducing Disparity-aware Swin Cross Attention Module (DSCAM) for stereo image scenarios. Utilising simulation buffers such as disparity maps, our DSCAM can extract cross-view features more efficiently and globally than the ubiquitous Parallax Attention Module (PAM) [9] used in the existing SSR models.

We carried out several experiments in the CARLA [10] simulation framework to assess the performance of our proposed technique. We also created a dataset of synthetic stereo images at different resolutions to train and evaluate super-resolution models. Our results imply that incorporating cross-view information in ETSSR can lead to improved performance compared to using two independent single-view models. Additionally, we show that our ETSSR model is considerably more computationally efficient than the state-of-the-art super-resolution models, while achieving comparable or superior performance in terms of image quality. The contributions of this paper can be summarised as follows:

- The introduction of a technique for accelerating stereo image simulation for autonomous driving applications: The proposed method permits stereo images to be simulated up to 2.57 times faster in CARLA, while preserving the output quality.
- The design of a novel transformer-based SSR architecture, ETSSR, achieving comparable or superior performance to state-of-the-art models, while also being more efficient and lightweight.
- Provision of a public dataset of synthetic images for super-resolution tasks: The dataset was created using CARLA, covering a diverse set of driving scenes.

The rest of this paper is structured as follows. Section II reviews relevant works on camera image simulation for AVs, image super-resolution, and neural image reconstruction techniques. Section III describes the proposed speed up technique, as well as the different components of our ETSSR network. Section IV elaborates on the experimental settings, speed up analysis, ablation study, and comparison to the state-of-the-art methods. Finally, concluding remarks and the potential future work are given in Section V.

II. Related work

In this section, we review the related work in each relevant field and explain how our method differs from those already existing. We present studies concerning camera image simulation for AVs in Section II-A, image super-resolution models in Section II-B, and neural image reconstruction for image rendering in Section II-C.

A. Camera Image Simulation for Autonomous Driving

Camera image simulation involves modelling external elements relating to the simulated environment, *e.g.* light reflection, as well as internal factors relating to the sensor specifications, such as camera lenses. The degree to which each factor is precisely modelled determines the fidelity of the image simulation. This is implicitly specified by the simulation objective in autonomous driving applications. For instance, low-latency simulation frameworks [10]–[12] were developed at the cost of fidelity for real-time testing scenarios, *e.g.* hardware-in-the-loop-testing. In contrast, high-fidelity approaches [13]–[15] were used to produce synthetic datasets for offline applications. In this case, exclusive simulation pipelines were designed using physics-based simulation techniques, and toolboxes [16].

Recently, data-driven simulation techniques have demonstrated success in synthesising realistic images for driving applications [5]. In this regard, a wide range of schemes has been utilised in the literature for incorporating generative models [17] into the image simulation pipeline. This includes modelling different parts of the pipeline, such as the image sensor [18], optical system [19], and weather disturbances [20]–[22], or the entire pipeline [23]. The main objective of the works in this area is to provide photo-realistic images for offline training of machine learning-based perception models. Our proposed method is also data-driven as we leverage a machine learning model for camera image simulation. However, in contrast to the existing methods, we investigate the effectiveness of our approach for real-time or faster-than-real-time settings by measuring our method's computational complexity and runtime performance.

B. Image Super Resolution

Image Super-Resolution is the technique of recovering High Resolution (HR) single or multiple image(s) from the corresponding single or multiple Low Resolution (LR) image(s). In the following, we review recent methods in the literature for single and stereo image super-resolution tasks, focusing on Deep Learning (DL)-based methods as they constitute state-of-the-art.

1) Single Image Super Resolution: Single Image Super Resolution (SISR) is a category of image super-resolution whereby the input is restricted to a single image. Regarding the first use of DL models in image super-resolution, Dong et al. [24] designed a shallow Convolutional Neural Network (CNN) to learn a mapping between LR and HR images. Following that, Kim et al. [25] proposed a very deep CNN (VDSR) inspired by VGG-Net [26] to enhance super-resolution. Along

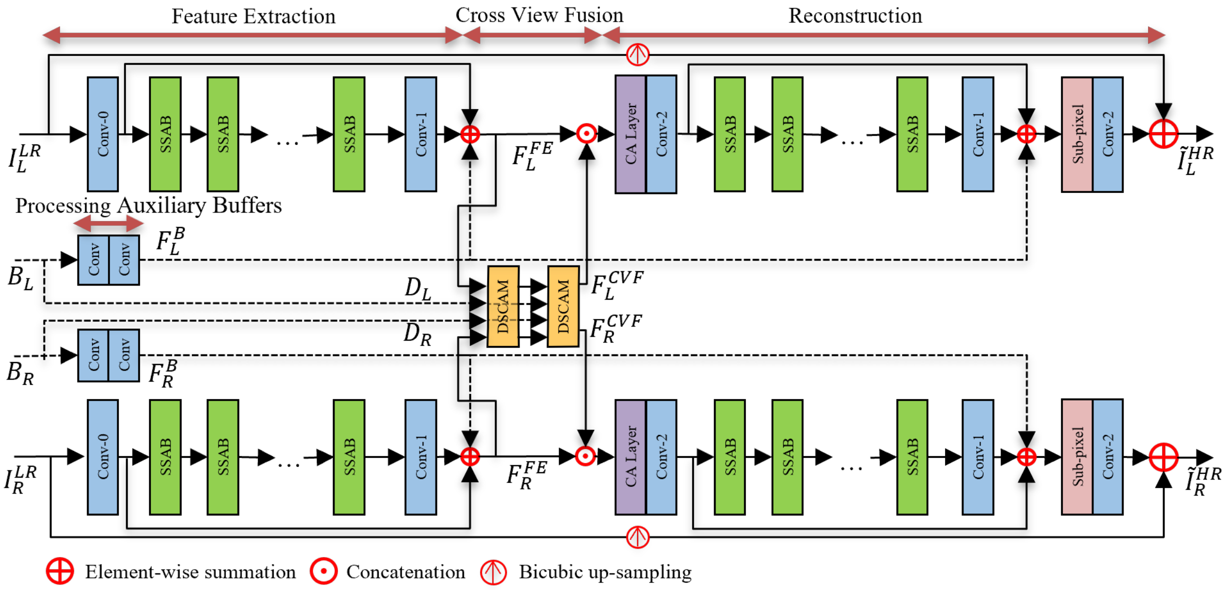


Fig. 2. An overview of our ETSSR's architecture. Our ETSSR inputs LR images of the left and right views ($I_L^{LR}, I_R^{LR} \in \mathbb{R}^{H\times W\times 3}$) as well as their respective auxiliary buffers ($B_L, B_R \in \mathbb{R}^{H\times W\times 4}$), and outputs the super-resolved images ($\tilde{I}_L^{HR}, \tilde{I}_R^{HR} \in \mathbb{R}^{sH\times sW\times 3}$).

with the advancement of DL-based architectures in image recognition, SISR models became more deep and complex. In this regard, residual learning was used by Lim et al. [27], residual dense blocks and attention mechanism including channel and spatial attention was leveraged by Zhang et al. [28], [29]. In the most recent work, Liang et al. [8] proposed a fully transformer-based model which outperforms previous CNN-based models.

2) Stereo Super Resolution: Stereo Super Resolution (SSR) is another type of image super-resolution which takes advantage of inter-view information provided by stereo images. In terms of pioneering DL-based models, Jeon et al. [30] developed a two-stage CNN, namely StereoSR, that exploits parallax prior as cross-view information. To overcome large variations in disparity maps, Wang et al. [9] proposed Parallax Attention Module (PAM) for finding correspondence between views in an unsupervised manner. Ying et al. [31] suggested a stereo attention module to exploit inter-view information at various points of pre-trained SISR models. To address the occlusion issue in SSR, Wang et al. [32] designed a Bi-directional PAM (bi-PAM) to use inter-view symmetric cues. Several i-lateral losses have also been incorporated into the model to enforce stereo consistency and robustness to illumination changes. Dai et al. [33] proposed a unified framework, SSRDEFNet, to simultaneously perform SSR and disparity estimation. The structure of SSRDEFNet is recursive, with the disparity estimator assisting SSR and vice versa. Lately, Chu et al. [34] developed NAFSSR, which is inspired by NAFNET [35] architecture for single-view feature extraction and expanded by a set of cross attention modules for the multi-view scenarios.

Compared to DL-based SSR models, our ETSSR model stands out due to some key distinctions. First, our ETSSR is a novel transformer-based SSR network that uses swin attention [7] as the backbone. This helps the network aggregate intra /inter-view features more computationally efficient than the state-of-the-art models. Second, our ETSSR exploits auxiliary simulation buffers to further boost its performance.

C. Neural Image Enhancement for Image Rendering

Image rendering, also known as scene rendering, is the most computationally intensive element of camera image simulation. There is a growing body of research that use neural networks to reduce the costs of image rendering, particularly for gaming application. The proposed techniques in this area can be separated into two categories: (1) denoising sparsely ray-traced images and (2) super-sampling as an anti-aliasing approach.

Regarding the first category, neural networks are used to denoise monte-carlo renderings with low samples per pixel. For instance, Kalantari et al. [36] used multi-layer perceptron in addition to fixed filters as a denoising model. Following that, CNNs were used in the work by Bako et al. [37] to predict the filters needed for noise removal. The method improved earlier results due to the implicit learning of complex filters by the CNN model. In a similar study, Chaitanya et al. [38] proposed a U-Net-based [39] network to predict the de-noised input directly. Recurrent connections were also added to this model to increase temporal stability. Other contributions in this line include the works accomplished by Vogels et al. [40] and Hasselgren et al. [41], which further enhanced the reconstruction quality by predicting filter kernels at different scales.

In the second category, neural super-sampling is used to avoid under-sampling artefacts primarily jagged edges in the spatial domain and flickering in the temporal domain. Deep-learning Super-Sampling (DLSS) [42] was the first method to employ neural networks for up-sampling rendered images in

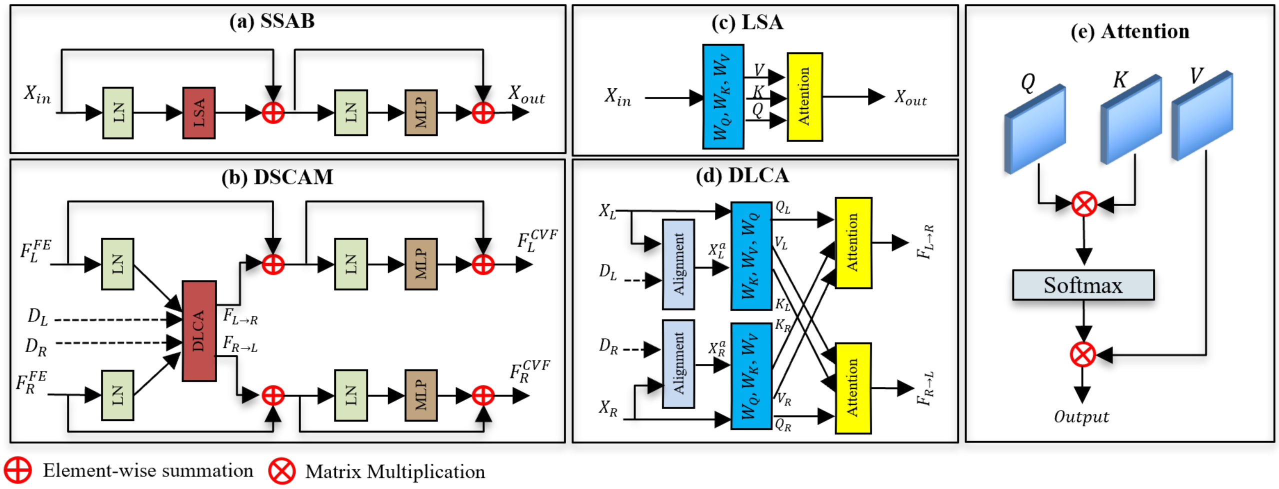


Fig. 3. Details of our ETSSR's components. (a) Swin Self Attention Block (SSAB). (b) Disparity-aware Swin Cross Attention Module (DSCAM). (c) Local Self Attention (LSA). (d) Disparity-aware Local Cross Attention (DLCA). (e) Attention mechanism. The layers LN and MLP stand for Layern-Norm and Multi-Layer Perceptron, respectively.

real-time. In the subsequent work, Xiao et al. [43] devised a U-Net-based network that uses temporal information based on motion vectors to increase output quality and stability of super-sampling.

Our work is inline with the above methods since we use neural networks to reconstruct high quality synthetic images from low quality ones. However, it differs from the existing methods as ETSSR is optimised for stereo images and outperforms SISR models (see Section IV-E); also, it focuses on the spatial domain for super-resolution.

III. Proposed Method

In this section, we elaborate on our proposed technique. We first provide the rationales for using SSR to accelerate stereo image simulation in Section III-A. We then describe the overall architecture of our ETSSR in Section III-B and its two major components in Section, III-C, and III-D, respectively.

A. Speeding up Stereo Image Simulation

One of the factors that adds to the complexity of camera image simulation is the resolution of the image sensor. Our acceleration technique uses this insight to speed up the simulation by applying super-resolution techniques. Specifically, for stereo image simulation, we leverage inter-view information to effectively super-resolve stereo images. As shown in Figure 1, our approach is comprised of two stages. In the first stage, LR stereo camera model C_{LR} maps the description of scene $\mathcal{L}$ to LR stereo images $I_L^{LR}, I_R^{LR} \in \mathbb{R}^{H\times W\times 3}$ (with $H \times W$ image resolution) and the respective auxiliary buffers $B_L, B_R \in \mathbb{R}^{H\times W\times 4}$ as:

$$I_L^{LR}, I_R^{LR}, B_L, B_R = C_{LR}(\mathcal{L}). \tag{1}$$

The description of scene $\mathcal{L}$ includes all the information required for producing an image, including object properties, light sources, and weather conditions. The auxiliary buffers B_L, B_R consist of disparity maps D_L, $D_R \in \mathbb{R}^{H\times W\times 1}$ and semantic segmentation layouts $S_L, S_R \in \mathbb{R}^{H\times W\times 3}$, which can be obtained from intermediary buffers during the simulation. In the second stage, our SSR model, ETSSR, reconstructs HR images $\tilde{I}_L^{HR}, \tilde{I}_R^{HR} \in \mathbb{R}^{sH\times sW\times 3}$ (s is the up-scaling factor) feeding in I_L^{LR}, I_R^{LR} and B_L, B_R as:

$$\tilde{I}_L^{HR}, \tilde{I}_R^{HR} = ETSSR(I_L^{LR}, I_R^{LR}, B_L, B_R). \tag{2}$$

Sequentially, the two steps approximate the HR stereo camera $\tilde{C}_{HR}$ which maps $\mathcal{L}$ into the reconstructed HR stereo images $\tilde{I}_L^{HR}, \tilde{I}_R^{HR}$ as:

$$\tilde{I}_L^{HR}, \tilde{I}_R^{HR} = \tilde{C}_{HR}(\mathcal{L}). \tag{3}$$

Assuming that the sum of the ETSSR's runtime and simulation time of C_{LR} is less than the simulation time of the original C_{HR}, the proposed technique can speed up the simulation as we will explore in Section IV-B.

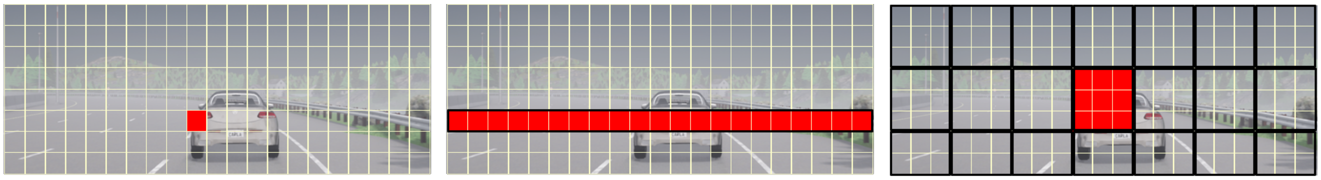
B. Network Architecture

As shown in Figure 2, our ETSSR network's architecture contains four stages: processing auxiliary buffers, feature extraction, cross-view fusion, and reconstruction. All the stages except cross-view fusion share the same operation and weights for both views. The details of each stage are presented in the following sections.

1) Processing Auxiliary Buffers: Auxiliary buffers B_L, B_R consist of disparity and semantic segmentation maps. The former provides the distance between corresponding pixels in two views, and the latter contains the pixel-level object classes. We feed this additional information into our ETSSR to further improve the performance. To process the auxiliary buffers, firstly, buffer features $F^B \in \mathbb{R}^{H\times W\times 64}$ are extracted using two 3×3 convolutions as :

$$F^B = Conv(Conv(B)). \tag{4}$$

Secondly, F^B is added to the upper layers of the model as we shall see in the rest of this section.



(a) Query pixel

(b) Cross attention context in PAM

(c) Cross attention context in DSCAM

Fig. 4. Comparison between cross attention context in PAM and our DSCAM. (a) A query pixel in an image view attends to a local window in the opposite view. (b) Cross attention context in PAM that corresponds to the parallax line of the query pixel in the opposite view. (c) Cross-attention context in DSCAM which corresponds to the patch enclosing the query pixel in the opposite view.

2) Feature Extraction: Feature extraction inputs I^{LR} and provides informative features $F^{FE} \in \mathbb{R}^{H \times W \times 64}$ for subsequent components. Feature extraction consists of a 3×3 convolution at the beginning and a series of Swin Self-Attention Blocks (SSABs) with residual learning (more details in Section III-C). Considering K SSAB layers for feature extraction, the process can be formulated as:

$$\begin{aligned} &F^0 = Conv(I^{LR}) \\ &F^i = SSAB_i(F^{i-1}) \quad i = 1, 2, 3, ..., k \\ &F^{FE} = F^0 + F^B + Conv(F^K). \end{aligned} \tag{5}$$

3) Cross-View Fusion: In cross-view fusion, the F^{FE} from two views are fused by the proposed Disparity-aware Swin Cross Attention Module (DSCAM). The DSCAM, firstly, registers the images using disparity maps D_L, D_R and then performs local Swin cross attention (more details in Section III-D). Considering F_L^{FE}, F_R^{FE} and D_L, D_R as inputs, DSCAM outputs $F_L^{CVF}, F_R^{CVF} \in \mathbb{R}^{H \times W \times 64}$ as:

$$F_L^{CVF}, F_R^{CVF} = DSCAM(F_L^{FE}, F_R^{FE}, D_L, D_R). \tag{6}$$

our ETSSR contains two subsequent layers of DSCAM to leverage a larger attention context.

4) Reconstruction: Following the feature fusion, F_L^{CVF}, F_R^{CVF} are further processed to reconstruct the HR images. In the first step of reconstruction, features go through a channel attention layer, followed by a convolution layer. Afterwards, the resulting features are fed to a series of SSABs similar to the feature extraction stage. Finally, a pixel-shuffle layer followed by a convolution layer produces the super-resolved image $\tilde{I}^{HR}$. Taking the left view as an example and considering K SSAB layers, the reconstruction process can be formulated as:

$$\begin{aligned} &F_L^{Rec_0} = Conv(CA(Concat(F_L^{CVF}, F^{FE}))) \\ &F_L^{Rec_i} = SSAB_i(F_L^{Rec_{i-1}}) \\ &F_L^{Rec} = F_L^{Rec_0} + F_L^{Rec_K} + F_L^B \\ &\tilde{I}_L^{HR} = Conv(PixelShuffle(F_L^{Rec})), \end{aligned} \tag{7}$$

where $Concat()$ and $CA()$ functions are concatenation and channel attention functions respectively.

C. Swin Self Attention Block

Swin Self Attention Block (SSAB) is the main building block of our ETSSR. SSAB is based on the Swin Transformer [7] architecture, which has recently demonstrated remarkable PSNR/parameters and PSNR/FLOPs trade-offs in image restoration tasks [8]. To achieve this, Swin Transformer-based models calculate self-attention on small image windows, making the process more efficient, and periodically shift the windows to prevent the loss of global context, i.e., a drop in performance. For simplicity, shifting and splitting operations are typically omitted from illustrations, but they are placed at the beginning of each SSAB layer. As shown in Figure 3-(a), the SSAB consists of Layer-Norm (LN), Local Self Attention (LSA, details are shown in Figure 3-(c)), and Multi-Layer Perceptron (MLP). Considering $X_{in} \in \mathbb{R}^{L \times 64}$ ($L = W_c \times W_c$, where W_c is the patch size) as a patch of an input image, SSAB outputs $X_{out} \in \mathbb{R}^{L \times 64}$ following the below equations:

$$\begin{aligned} &X_{mid} = LN(LSA(X_{in})) + X_{in} \\ &X_{out} = MLP(LN(X_{mid})) + X_{mid}. \end{aligned} \tag{8}$$

To perform LSA, first, query, value, and key matrices $Q, K, V \in \mathbb{R}^{L \times 64}$ are computed as:

$$Q,\ K,\ V = X_{in}W_Q,\ X_{in}W_K,\ X_{in}W_V. \tag{9}$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{64 \times 64}$ are the weights of fully-connected network. Then, the attention function (shown in Figure 3-(e)) is calculated as:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{64})V. \tag{10}$$

D. Disparity-aware Swin Cross Attention Module

We propose Disparity-aware Swin Cross Attention Module (DSCAM) to mitigate the issue of large computation in the ubiquitous Parallax Attention Module (PAM) [9]. As shown in Figure 4-(b), in PAM, each pixel of an image view attends to its parallax line, equivalent to an image row of length W in the counter-part view. In applications where cameras have a wide horizontal field of view, *e.g.* autonomous driving, PAM demands a huge and unnecessary calculation for cross attention. On the other hand, the DSCAM performs Swin cross attention on corresponding image patches with size w_c, where $w_c^2 < W$ (Figure 4-(c)). We also feed in disparity maps to align the patches in left and right views so that the relevant context is available for cross attention. As shown in Figure 3-(b), the architecture of DSCAM is similar to SSAB with Disparity-aware Local Cross Attention (DLCA, details are shown in Figure 3-(d)) replacing LSA. In DLCA, firstly, image features

TABLE I

Speed up analysis of our ETSSR's variants for different image resolutions. The average simulation time of the CARLA's stereo camera in the original resolution is referred to as t_{HR}, while that of the corresponding LR stereo camera plus our ETSSR's runtime is referred to as $t_{LR} + t_{SR}$ (equivalent to the time proposed by our approach). The PSNR/SSIM column shows the average output quality of the super-resolved images, which is indicated separately for the left and right view.

Image Resolution	t_{HR} (ms)	Model Variants	$t_{LR} + t_{SR}$ (ms)	Speed Up	PSNR / SSIM (left)	PSNR / SSIM (right)
1280 × 720	131	ETSSR_S	58	2.26	33.03 / 0.8999	33.04 / 0.9002
		ETSSR_B	62	2.11	33.12 / 0.9017	33.14 / 0.9021
		ETSSR_L	65	2.02	33.18 / 0.9023	33.20 / 0.9021
1920 × 1080	173	ETSSR_S	68	2.54	34.93 / 0.9245	34.94 / 0.9247
		ETSSR_B	75	2.31	34.96 / 0.9247	34.98 / 0.9250
		ETSSR_L	82	2.11	34.96 / 0.9251	34.99 / 0.9253
2560 × 1440	231	ETSSR_S	90	2.57	38.20 / 0.9569	38.22 / 0.9566
		ETSSR_B	103	2.24	38.45 / 0.9581	38.46 / 0.9579
		ETSSR_L	117	1.97	38.54 / 0.9592	38.56 / 0.9588

TABLE II

Ablation Study of our ETSSR. We incrementally add our proposed modules to the baseline model to observe the effect on the efficiency and output quality. The best result is bolded.

Models	Params (M)	FLOPs (T)	PSNR/SSIM (left)	PSNR/SSIM (right)
Baseline	0.90	0.607	38.16/0.9560	38.19/0.9558
Baseline + SCAM	0.58	0.514	38.14/0.9546	38.17/0.9544
Baseline + DSCAM (SCAM + Disparity)	0.58	0.514	38.39/0.9577	38.40/0.9573
ETSSR	0.62	0.533	**38.44/0.9579**	**38.46/0.9576**

from two views X_L, X_R get aligned using disparity maps D_L, D_R as:

$$\begin{aligned} X_L^a &= \mathit{Index\_Select}(X_L, X_L^j + D_L) \\ X_R^a &= \mathit{Index\_Select}(X_R, X_R^j - D_R), \end{aligned} \tag{11}$$

where X_L^j, X_R^j are grid maps containing the column index of each pixel, and $\mathit{Index\_Select}()$ selects the input indexes from the specified array. Secondly, aligned features constitute keys K and values V for cross attention as:

$$\begin{aligned} Q_L, K_L, V_L &= X_R W_Q, X_R^a W_K, X_R^a W_V \\ Q_R, K_R, V_R &= X_L W_Q, X_L^a W_K, X_L^a W_V \\ F_{R \to L} &= \mathit{Attention}(Q_L, K_R, V_R) \\ F_{L \to R} &= \mathit{Attention}(Q_R, K_L, V_L), \end{aligned} \tag{12}$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{64 \times 64}$ are the weights of fully-connected network.

E. Loss

The L1 norm between super-resolved image $\tilde{I}^{HR}$ and ground-truth I^{HR} is chosen as an objective function for training our ETSSR:

$$\mathit{Loss} = |\tilde{I}_L^{HR} - I_L^{HR}| + |\tilde{I}_R^{HR} - I_R^{HR}|. \tag{13}$$

IV. Performance Evaluation

In this section, we evaluate the performance of our technique. In Section IV-A, we first describe our experimental settings, including the provided dataset, training parameters, evaluation metrics, and our ETSSR's variants. We then discuss the speed up achieved by our approach in Section IV-B. In Section IV-C and IV-D, the effectiveness of the proposed ETSSR's modules and the cross-view fusion component is analysed. Finally, we present a quantitative and qualitative comparison to the state-of-the-art super-resolution models in Section IV-E.

TABLE III

Analysis of cross-view fusion. We train our ETSSR with five different strategies to investigate the effect of cross-view fusion and its optimal spot. The best result is bolded.

Models	PSNR/SSIM (left)	PSNR/SSIM (right)
ETSSR with No Fusion	37.93/0.9529	37.94/0.9527
ETSSR with Input Fusion	38.02/0.9544	38.04/0.9541
ETSSR with Late Fusion	38.17/0.9552	38.20/0.9550
ETSSR with Early Fusion	38.40/0.9573	38.41/0.9571
ETSSR	**38.44/0.9579**	**38.46/0.9576**

A. Experimental Settings

1) Dataset: In order to train our ETSSR, we need access to a considerable number of synthetic stereo images and simulation buffers captured at different resolutions. Although there are off-the-shelf datasets of synthetic images, *e.g.* Virtual Kitti [44], the images are only rendered at a single resolution. Moreover, there is a lack of information regarding the simulation time required for simulating images at different resolutions and rendering quality. Due to these limitations, we built our own dataset, namely CARLA's Multi-Resolution Stereo Images (CMRSI), using the CARLA [10] driving simulator. We modelled an RGB stereo camera with an image resolution of 2560 × 1440 (following 16:9 aspect ratio of standard automotive image sensors [45]) and a baseline of 0.5 metres. We placed the stereo camera model on a vehicle model controlled by CARLA's auto-pilot algorithm. The vehicle was driven across all eight maps of CARLA to record a total of 800 frames, with approximately 100 frames per map. In addition to RGB images, we captured the scene's disparity map and semantic segmentation layouts as auxiliary buffers. Notably, all the data (including RGB images and auxiliary buffers) were also recorded at a resolution of 640 × 360 in order to produce LR frames. Throughout our evaluations, we maintain the super-resolution up-scaling factor to be 4 × 4.

2) Training: We split the CMRSI into training and test sets following the 80:20 ratio. During the training phase, we

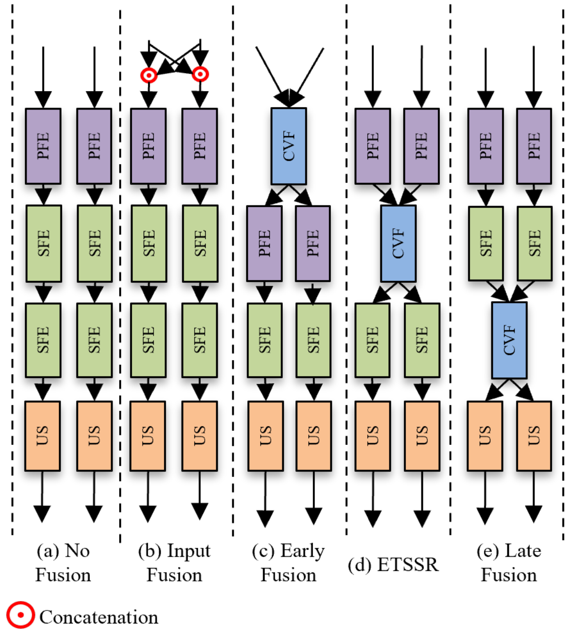
TABLE IV
QUANTITATIVE COMPARISON TO THE STATE-OF-THE-ART METHODS. THE BEST RESULT IS IN RED, AND THE SECOND BEST RESULT IS IN BLUE.

Models	Params (M)	FLOPs (T)	PSNR/SSIM (left)	PSNR/SSIM (right)
Bicubic	-	-	35.64/0.9343	35.65/0.9341
VDSR	0.67	4.917	37.48/0.9498	37.50/0.9493
RCAN	15.36	7.020	38.08/0.9542	38.09/0.9539
EDSR	38.90	18.282	38.10/0.9556	38.09/0.9553
RDN	22.04	10.168	38.31/0.9559	38.33/0.9556
StereoSR	1.15	8.461	36.19/0.9383	36.17/0.9381
SSRDEFNet	2.26	5.004	36.62/0.9544	36.62/0.9550
PASSRnet	1.42	1.745	37.89/0.9541	37.92/0.9543
iPASSR	1.43	1.102	38.11/0.9557	38.13/0.9549
NAFSSR	0.46	15.327	38.14/0.9555	38.16/0.9553
ETSSR_S (***Ours***)	0.47	0.362	38.20/0.9569	38.22/0.9566
ETSSR_B (***Ours***)	0.62	0.533	38.45/0.9581	38.46/0.9579
ETSSR_L (***Ours***)	0.76	0.703	38.54/0.9592	38.56/0.9588

cropped LR images to 30 × 90 overlapping patches with stride 20, similar to the training settings of other super-resolution networks. We train our model with a batch size of 32 for 80 epochs and use the Adam optimiser with a learning rate of $2e^{-4}$ halving every 30 epochs. We implement our model with Pytorch library and run the entire experiments on Nvidia Quadro RTX 5000.

3) Evaluation metrics: PSNR and SSIM [46] are used to assess the quality of super-resolved images. The former measures pixel-level error, while the latter focuses on structural similarity. We calculate the metrics for the left and right image views independently. The total number of parameters in millions (M) and floating point operations (FLOPs) in trillions are also used to measure the capacity and efficiency of the super-resolution networks.

4) Model Variants: To study the trade-off between the complexity and performance of our ETSSR, we designed three model variants: ETSSR_S (small), ETSSR_B (base), and ETSSR_L (large). The difference between the variants is in the number of SSABs used in the feature extraction and reconstruction components of our ETSSR. For ETSSR_S ,ETSSR_B and ETSSR_L, we consider two, four, and six SSAB layers respectively. By increasing the number of SSABs, more informative features are extracted, and the ability of image reconstruction is raised. Unless specified, the ETTSR model refers to the ETSSR_B throughout the paper.



(a) No Fusion (b) Input Fusion (c) Early Fusion (d) ETSSR (e) Late Fusion

Fig. 5. Arrangement of our ETSSR's components in different fusion schemes. PFE, CVF, SFE and US stand for primary feature extraction, cross-view fusion, secondary feature extraction, and up-sampling. PFE is equivalent to our ETSSR's feature extraction component and the SFE refers to a series of SSABs in our ETSSR's reconstruction component.

B. Speed up Analysis

In this section, we investigate the effectiveness of our acceleration technique. As shown in Table I, we compare the simulation time of stereo images in CARLA with the time required by our approach for different image resolutions and ETSSR variants. We further report the output quality of images super-resolved by our ETSSR variants in terms of the evaluation metrics. The time required for the stereo image simulation in the original resolution is referred to as t_{HR}, while that of the LR stereo image ($\frac{1}{4} \times \frac{1}{4}$ of the original resolution) and the runtime of our ETSSR is referred to as t_{LR} and t_{SR} respectively. It should be noted that the simulation time of stereo images is averaged over scenarios of varying complexity. Moreover, our ETSSR model is optimised with PyTorch's Tensor-RT library at 16-bit precision. As shown, our method can speed up the camera image simulation by factors between 1.97 and 2.57, depending on the ETSSR variant and image resolution. The different variations of ETSSR allow a trade-off between output quality and execution time. The quantitative metrics also show that our ETSSR can reconstruct HR images with high quality. This will be analysed further in Section IV-E.

C. Ablation Study

To assess the impact of each proposed module in our ETSSR, we progressively add the modules to a baseline model and evaluate their performance. The baseline model uses SSABs for feature extraction and reconstruction, similar to our ETSSR. However, it does not take in auxiliary buffers and uses PAM instead of DSCAM for cross-view fusion. Our ablation study involves three stages: firstly, we replace PAM in the baseline with a Cross Swin Attention Module (CSAM, which is DSCAM without disparity) to analyse the impact of

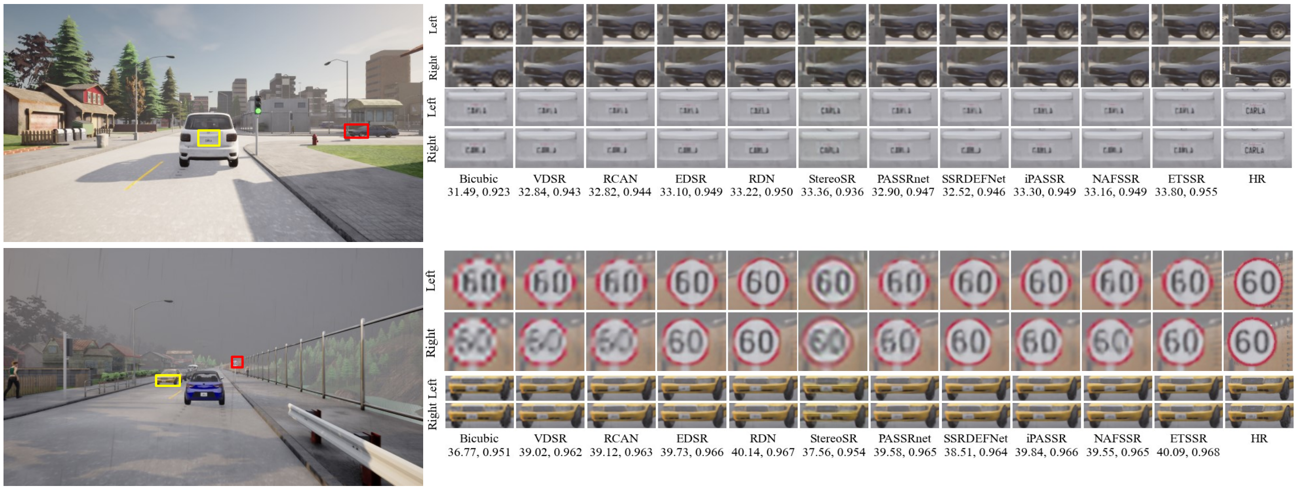


Fig. 6. Qualitative comparison to the state-of-the-art methods. The values below each image represent the image's PSNR/SSIM averaged over the left and right views.

cross Swin attention. Secondly, we examine the effectiveness of disparity alignment by feeding the disparity maps into CSAM (DSCAM). Thirdly, we input auxiliary buffers into the model (which constructs our ETSSR model) to check whether extra information is advantageous. As shown in Table II, replacing PAM with CSAM reduces the parameters and FLOPs by 36% and 15%, respectively, while PSNR/SSIM remains relatively similar. This indicates the effectiveness of swin cross attention in boosting the model's efficiency. Using DSCAM in the second stage raises the PSNR by 0.24 dB averaged over two views, which shows the importance of disparity maps in assisting cross attention. In the last stage, feeding auxiliary buffers elevates the PSNR/SSIM at the expense of a slight increase in the computation, which proves to be negligible in practice.

D. Analysis of Cross-View Fusion

In this section, we investigate the effect of cross-view fusion as well as its optimal spot in our ETSSR's architecture. To analyse the effect of cross-view fusion, we replace the DSCAM with two layers of SSABs in each view and train the separated models independently (no fusion). To find the optimal spot for cross-view fusion, we train the model with four different fusion schemes: (1) concatenating the left and right images at the input and training the separated models independently (Input Fusion), (2) fusing the features before the primary feature extraction component (Early Fusion), (3) fusing the features after the primary feature extraction component (ETSSR), and (4) fusing the features after the secondary feature extraction (Late Fusion). Figure 5 depicts the arrangement of the components for different fusion schemes. As shown in Table III, excluding the cross-view fusion causes a 0.51 dB decrease in PSNR comparing to our ETSSR. This highlights the significance of cross-view information and SSR's superiority over two separate SISR models. It is also evident that the original ETSSR model outperforms the input, late, and early fusion schemes. This implies that the ideal spot to fuse information is within the middle of the network, where intra-view features are rich enough and can also assist the super-resolution in the counterpart view.

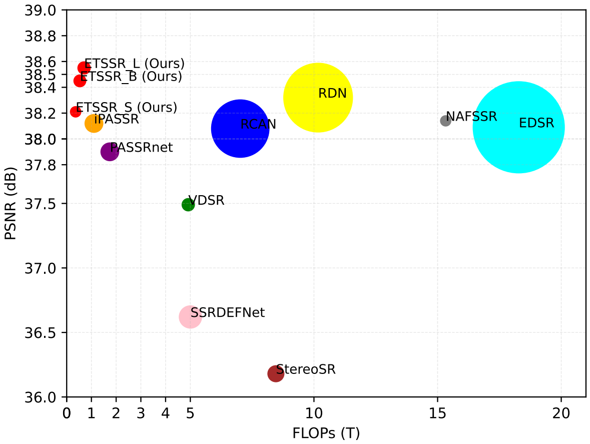


Fig. 7. PSNR vs FLOPs for state-of-the-art super-resolution methods. The size of each point is proportional to the number of parameters in the model. PSNR is averaged over the left and right views.

E. Comparison to State-Of-The-Arts

In this section, we compare our ETSSR to several state-of-the-art SSR and SISR methods. We select StereoSR [30], PSSRNet [9], iPASSR [32], SSRDEFNet [33], and NAFSSR [34] (the most efficient version) from SSR models and VDSR [25], EDSR [27], RCAN [29], and RDN [28] from SISR networks. We train all the models from scratch on the CMRSI training set and evaluate them on the test set.

1) Quantitative Comparison: We quantitatively compare our ETSSR to the state-of-the-art super-resolution models in Table IV. As shown, the ETSSR_L achieves the maximum SSIM/PSNR while using less than 4% of parameters and 8% of the FLOPs of the closest method, RDN. Compared to the efficient models in terms of FLOPs, the closest method is iPASSR which consumes more than three times as many FLOPs of our ETSSR_S while falling -0.09 dB short in PSNR. Figure 7 depicts the scatter plot of PSNR against FLOPs for all methods, making the difference between the methods more apparent.

2) Qualitative Comparison: We qualitatively compare our ETSSR to state-of-the-art methods by visualising the model's output in Figure 6. We chose the scenes from two distinct CARLA maps and the frames where multiple objects are visible in the camera's field of view. As shown, our ETSSR can reconstruct the image with a high degree of detail, while other methods may suffer from blurriness or artefacts. Bear in mind that our goal in this research is not to visually exceed the state-of-the-art methods but rather to propose an efficient SSR network that preserves the HR image quality. Readers can view this video for supplementary results at full resolution in other CARLA maps.

V. Conclusion and Future Work

In this research, we proposed a technique based on SSR to speed up the simulation of stereo images, which are commonly used in autonomous driving applications. To efficiently super-resolve synthetic stereo images, we designed ETSSR, a novel SSR network. According to our experiments in CARLA, the proposed technique can speed up the stereo image simulation by a factor of up to 2.57 over different image resolutions. Moreover, our ablation study showed that each of the proposed ETSSR's components contributed to the performance improvement. We also highlighted the significance of cross-view fusion on the model's performance and realised that the optimal spot for feature fusion lies in the middle of the network. Our comparative study revealed that ETSSR is more efficient and lightweight than the state-of-the-art super-resolution models while also achieving comparable or superior results in terms of image quality.

We identify three prospective research directions for future work. Firstly, we propose that efficient incorporation of temporal information (e.g. previous data frames) and utilisation of spatio-temporal consistency losses could further improve the output quality of ETSSR. In this case, the acceleration technique can be reformulated as stereo video super-resolution [47], however, with careful consideration of the model's efficiency. Secondly, we anticipate that the ongoing research on the Transformer model's efficiency [48] may enable an even greater acceleration of our approach. Thirdly, we suggest investigating the feasibility of a similar technique for other sensors used in AVs, such as Lidar or radar. For this, the super resolution model needs to be customised to reconstruct the data in the representation provided by the sensor. We believe that this work lays the foundation for further research on the efficiency of camera image simulation and the customisation of modern rendering techniques for autonomous driving or robotics applications.

References

[1] J. Chen, K. Venkataraman, D. Bakin, B. Rodricks, R. Gravelle, P. Rao, and Y. Ni, "Digital camera imaging system simulation," *IEEE Transactions on Electron Devices*, vol. 56, pp. 2496–2505, 2009.

[2] A. Elmquist and D. Negrut, "Modeling cameras for autonomous vehicle and robot simulation: An overview," *IEEE Sensors Journal*, vol. 21, pp. 25 547–25 560, 11 2021.

[3] S. I. Nikolenko, *Synthetic Data for Deep Learning*, 1st ed. Springer Cham, 2021, vol. 174.

[4] N. Kalra and S. M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Santa Monica, CA: RAND Corporation, 2016.

[5] M. Uricar, P. Krizek, D. Hurych, I. Sobh, S. Yogamani, and P. Denny, "Yes, we gan: Applying adversarial techniques for autonomous driving," *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 2019, 2 2019.

[6] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik, "Advances in neural rendering," *Computer Graphics Forum*, vol. 41, no. 2, pp. 703–735, 2022.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10 002, 3 2021.

[8] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2021-October, pp. 1833–1844, 8 2021.

[9] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 12 242–12 251, 3 2019.

[10] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "Carla: An open urban driving simulator," pp. 1–16, 10 2017.

[11] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2149–2154, 2004.

[12] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles," *Springer Proceedings in Advanced Robotics*, vol. 5, pp. 621–635, may 2017.

[13] Z. Liu, M. Shen, J. Zhang, S. Liu, H. Blasinski, T. Lian, and B. Wandell, "A system for generating complex physically accurate sensor images for automotive applications," *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 2019, 2 2019.

[14] Z. Liu, J. E. Farrell, and B. A. Wandell, "Isetauto: Detecting vehicles with depth and radiance information," *IEEE Access*, vol. 9, pp. 41 799–41 808, 2021.

[15] K. Debattista, T. Bashford-Rogers, C. Harvey, B. Waterfield, and A. Chalmers, "Subjective evaluation of high-fidelity virtual environments for driving simulations," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 1, pp. 30–40, 2018.

[16] M. Pharr, W. Jakob, and G. Humphreys, "Physically based rendering: From theory to implementation: Third edition," *Physically Based Rendering: From Theory to Implementation: Third Edition*, pp. 1–1233, 11 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[18] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Sensor transfer: Learning optimal sensor effect image augmentation for sim-to-real domain adaptation," *IEEE Robotics and Automation Letters*, vol. 4, pp. 2431–2438, 9 2018.

[19] Q. Zheng and C. Zheng, "Neurolens: Data-driven camera lens simulation using neural networks," *Computer Graphics Forum*, vol. 36, pp. 390–401, 12 2017.

[20] M. Tremblay, S. S. Halder, R. de Charette, and J. F. Lalonde, "Rain rendering for evaluating and improving robustness to bad weather," *International Journal of Computer Vision*, pp. 1–20, 9 2020.

[21] R. Gong, D. Dai, Y. Chen, W. Li, and L. V. Gool, "Analogical image translation for fog generation," *arXiv*, 6 2020.

[22] H. Wang, Z. Yue, Q. Xie, Q. Zhao, Y. Zheng, and D. Meng, "From rain generation to rain removal," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 786–14 796.

[23] Z. Yang, Y. Chai, D. Anguelov, Y. Zhou, P. Sun, D. Erhan, S. Rafferty, and H. Kretzschmar, "Surfelgan: Synthesizing realistic sensor data for

autonomous driving." IEEE Computer Society, 2020, pp. 11 115–11 124.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 295–307, 12 2014.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 1646–1654, 11 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9 2014.

[27] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2017-July, pp. 1132–1140, 7 2017.

[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 12 2018.

[29] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. R. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision*, 2018.

[30] D. S. Jeon, S. H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1721–1730, 12 2018.

[31] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 496–500, 2020.

[32] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 766–775, 11 2020.

[33] Q. Dai, J. Li, Q. Yi, F. Fang, and G. Zhang, "Feedback network for mutually boosted stereo image super-resolution and disparity estimation," *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, vol. 18, pp. 1985–1993, 6 2021.

[34] X. Chu, L. Chen, and W. Yu, "Nafssr: Stereo image super-resolution using nafnet," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1238–1247, 2022.

[35] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 17–33.

[36] N. K. Kalantari, S. Bako, and P. Sen, "A machine learning approach for filtering monte carlo noise," *ACM Transactions on Graphics (TOG)*, vol. 34, 7 2015.

[37] S. Bako, T. Vogels, B. McWilliams, M. Meyer, J. Novák, A. Harvill, P. Sen, T. Derose, and F. Rousselle, "Kernel-predicting convolutional networks for denoising monte carlo renderings," *ACM Transactions on Graphics (TOG)*, vol. 36, 7 2017.

[38] C. R. Chaitanya, A. S. Kaplanyan, C. Schied, M. Salvi, A. Lefohn, D. Nowrouzezahrai, and T. Aila, "Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, 7 2017.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," vol. 9351. Springer Verlag, 2015, pp. 234–241.

[40] T. Vogels, F. Rousselle, B. McWilliams, G. Röthlin, A. Harvill, D. Adler, M. Meyer, and J. Novák, "Denoising with kernel prediction and asymmetric loss functions," *ACM Transactions on Graphics (TOG)*, vol. 37, 7 2018.

[41] J. Hasselgren, J. Munkberg, M. Salvi, A. Patney, and A. Lefohn, "Neural temporal adaptive sampling and denoising," *Computer Graphics Forum*, vol. 39, pp. 147–155, 5 2020.

[42] E. Andrew, P. Jukarainen, and A. Patney, "Truly next-gen: Adding deep learning to games and graphics," 2019.

[43] L. Xiao, S. Nouri, M. Chapman, A. Fix, D. Lanman, and A. Kaplanyan, "Neural supersampling for real-time rendering," *ACM Transactions on Graphics (TOG)*, vol. 39, 7 2020.

[44] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4340–4349.

[45] F. E. Sahin, "Long-range, high-resolution camera optical design for assisted and autonomous driving," *Photonics*, vol. 6, no. 2, 2019.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[47] R. Xu, Z. Xiao, M. Yao, Y. Zhang, and Z. Xiong, "Stereo video super-resolution via exploiting view-temporal correlations," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 460–468.

[48] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Comput. Surv.*, vol. 55, no. 6, dec 2022.

Hamed Haghighi is a PhD candidate with the Warwick Manufacturing Group (WMG) at University of Warwick, UK. He received a B.Sc. (2016) in Software Engineering from Isfahan University of Technology (Isfahan, Iran) and an M.Sc. (2019) in Artificial Intelligence from University of Tehran (Tehran, Iran). His research interests include machine learning, computer vision, computer graphics, and autonomous vehicles.

Mehrdad Dianati (Senior Member, IEEE) is the Head of Intelligent Vehicles Research Directorate and technically leads Networked Intelligent Systems (Cooperative Autonomy) research at Warwick Manufacturing Group (WMG), University of Warwick. He has over 28 years of combined industrial and academic experience, with 20 years in leadership roles in multi-disciplinary collaborative R&D projects, in close collaboration with the Automotive and ICT industries. He is also the Co-Director of Warwick's Centre for Doctoral Training on Future Mobility Technologies, training doctoral researchers in the areas of intelligent and electrified mobility systems in collaboration with the experts in the field of electrification from the Department of Engineering of the University of Warwick. In the past, he served as an Associate Editor for the IEEE Transactions on Vehicular Technology and several other international journals, including IET Communications. Currently, he is the Field Chief Editor of Frontiers in Future Transportation.

Valentina Donzella received her BSc (2003) and MSc (2005) in Electronics Engineering from University of Pisa and Sant'Anna School of Advanced Studies (Pisa, Italy), and her PhD (2010) in Innovative Technologies for Information, Communication and Perception Engineering from Sant'Anna School of Advanced Studies. In 2009, she was a visiting graduate student at McMaster University (Hamilton, ON, Canada) in the Engineering Physics department. She is currently Full Professor, and Head of the Intelligent Vehicles - sensors group at WMG, University of Warwick, UK; before this position, she was a MITACS and SiEPIC postdoctoral fellow at the University of British Columbia (Vancouver, BC, Canada), in the Silicon Photonics group. She is first author and co-author of several journal papers on top tier optics journals. Her research interests are: LiDAR, Intelligent Vehicles, integrated optical sensors, sensor fusion, and silicon photonics. Dr Donzella is Full College member of EPSRC and Senior Fellow of Higher Education Academy.

Kurt Debattista received a B.Sc. in mathematics and computer science, an M.Sc. in psychology, an M.Sc. degree in computer science, and a Ph.D. from the University of Bristol. He is currently a Professor with WMG, at the University of Warwick. His research interests are high-fidelity rendering, HDR imaging, machine learning, and applied perception.