Machine Vision and Applications (2025) 36:69
https://doi.org/10.1007/s00138-025-01685-w

RESEARCH

Hstr-net: reference based video super-resolution with dual cameras

H. Umut Suluhan[1] · Abdullah Enes Doruk[2] · Hasan F. Ates[2] · Bahadir K. Gunturk[3]

Received: 11 October 2024 / Revised: 3 March 2025 / Accepted: 21 March 2025 / Published online: 23 April 2025
© The Author(s) 2025

Abstract

High-spatio-temporal resolution (HSTR) video recording plays a crucial role in enhancing various imagery tasks that require fine-detailed information. State-of-the-art cameras provide this required high frame-rate and high spatial resolution together, albeit at a high cost. To alleviate this issue, this paper proposes a dual camera system for the generation of HSTR video using reference-based super-resolution (RefSR). One camera captures high spatial resolution low frame rate (HSLF) video while the other captures low spatial resolution high frame rate (LSHF) video simultaneously for the same scene. A novel deep learning architecture is proposed to fuse HSLF and LSHF video feeds and synthesize HSTR video frames. The proposed model combines optical flow estimation and (channel-wise and spatial) attention mechanisms to capture the fine motion and complex dependencies between frames of the two video feeds. Simulations show that the proposed model provides significant improvement over existing reference-based SR techniques in terms of PSNR and SSIM metrics. The method also exhibits sufficient frames per second (FPS) for aerial monitoring when deployed on a power-constrained drone equipped with dual cameras. The source code is publicly available at https://github.com/umutsuluhan/HSTRNet.

Keywords Super-resolution · Deep learning · Attention · Wide-area monitoring

Abbreviations

HSTR	High-spatio-temporal resolution
RefSR	Reference-based-super-resolution
HSLF	High spatial resolution low frame rate
LSHF	Low spatial resolution high frame rate
SR	Super resolution
HR	High resolution
LR	Low resolution
TFI	Temporal frame interpolation
SISR	Single image super resolution
REF	High resolution reference
HSR	High spatial resolution
LSR	Low spatial resolution
WAS	Wide-area surveillance
VFI	Video frame interpolation
Q	Query
K	Key
V	Value
PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity index measure
LPIPS	Learned perceptual image patch similarity
C	Contextual representation
PM	Patch correspondences

✉ H. Umut Suluhan
suluhan@arizona.edu

✉ Hasan F. Ates
hasan.ates@ozyegin.edu.tr

Abdullah Enes Doruk
enes.doruk@ozu.edu.tr

Bahadir K. Gunturk
bkgunturk@medipol.edu.tr

[1] Electrical and Computer Engineering Department, University of Arizona, Tucson, AZ 85721, USA

[2] Faculty of Engineering, Ozyegin University, 34794 Istanbul, Turkey

[3] School of Engineering and Natural Sciences, Istanbul Medipol University, 34810 Istanbul, Turkey

1 Introduction

High spatio-temporal resolution (HSTR) cameras provide detailed information in both spatial and temporal dimensions, benefiting many video-processing tasks. While these cameras record at high frame rates, commercial models often compromise on spatial quality [1]. For instance, some

cellphone cameras can record slow-motion videos at 1080p with a frame rate of 120 or 240 fps, but this drops to 60 fps when shooting in 4K. Professional high-speed cameras capture recordings with high spatial and temporal quality, but they come at a high retailing price [1]. Also, ultra-wide angle cameras can provide high-resolution outputs of wide-area scenes, albeit at a lower frame rate. To address these limitations and achieve HSTR video feeds, video super-resolution has emerged as a promising approach that produces high-quality video from one or more low-resolution feeds of the same scene.

Video super-resolution can be leveraged to produce high-quality video feeds from dual camera sources, suitable for a variety of scenarios including wide-area monitoring, mobile phone video generation, etc. Wide-area monitoring over large regions necessitates the identification and monitoring of numerous relatively small and distant objects. It is employed for the purpose of identifying vehicles, pedestrians, anomalies, and security-related occurrences across extensive regions, utilizing either aerial vehicles flying at high altitudes or fixed camera poles. Nevertheless, as the spatial resolution of these objects decrease, the task of precise detection becomes increasingly challenging, relying more heavily on the object's motion characteristics. Consequently, the camera's frame rate becomes a critical factor in achieving accurate motion estimation and analyzing complex dynamic scenarios. Additionally, object detection often relies on visible light, which is highly affected by the background and the lighting conditions of the environment [2, 3]. Enhancing low-light [3] and low-resolution images is crucial for video surveillance [2]. Therefore, high-contrast and high-resolution imagery is vital for object detection. Wide-area monitoring demands wide-angle, high-resolution cameras capable of continuously monitoring expansive areas, preferably at a high frame rate. However, wide-angle, high-resolution cameras and those with high frame rates are both costly and bulky for deployment on small scale drones. Another application scenario is for mobile phones with multiple cameras featuring both wide-angle (low-detail) and telephoto (high-detail) lenses. The widespread use of multiple cameras mounted on resource-constrained devices aligns well with the concept of generating a high-quality high frame rate wide-angle video by utilizing the telephoto camera feed as HR reference and wide-angle camera feed as LR frame.

Comprehensive studies have been conducted to improve and assess [4–6] the quality of video footage captured by commercial cameras. The pursuit of enhanced resolution has been investigated in two main aspects: spatial super-resolution (SR), which involves synthesizing high-resolution (*HR*) frames from corresponding low-resolution (*LR*) frames, and temporal frame interpolation (TFI), aimed at enhancing the temporal resolution of videos with low frame rates. However, the challenge lies in estimating the missing spatial and temporal information for these tasks. TFI tends to introduce motion blur and artifacts due to its assumption of linear motion between adjacent frames, while spatial interpolation often results in overly smoothed frames, thereby missing essential high-frequency components [7]. Single image super-resolution (SISR) methods [8–15] also often yields blurred results due to its inability to accurately recover high-frequency details that are absent in the *LR* image [16–22].

An alternative approach involves the application of computational imaging techniques, which leverage multiple cameras with varying resolutions and frame rates to produce high spatio-temporal resolution (HSTR) videos [1, 23]. The dual camera systems offer cost-effectiveness and adaptability to various scenarios. We adopt a similar strategy and put forth the idea of merging two camera feeds: one offering high spatial resolution but low frame rate (HSLF), and the other providing low spatial resolution but high frame rate (LSHF). This fusion enables generation of a high spatio-temporal resolution (HSTR) video, addressing the limitations associated with single-camera resolution enhancement methods. To realize the dual camera setup, we follow the reference-based super-resolution (RefSR) approach that utilizes a high-resolution reference (*REF*) image to perform texture transfer between high-quality spatial components and low-quality counterparts in the *LR* image. RefSR consistently produces quantitatively and visually better results when compared to SISR, mainly due to the effective transfer of relevant textures from the *REF* image to the *LR* image [1, 16–21, 23–27]. Our method further enhances the texture transfer process by employing a novel patch-correspondence algorithm that focuses on local correspondences of subregions between frames.

RefSR-based networks typically operate in the spatial domain, employing a correspondence matching strategy between all possible pairs of *REF* and *LR* patches for subsequent texture transfer. However, exhaustive patch-matching strategy struggles to establish pixel-level correspondences in the temporal domain, a challenge frequently encountered in processing multiple video feeds, despite the high computational demand. In the realm of RefSR, attention-based mechanisms play a crucial role in determining the correspondences between patches in *LR* and *REF* images [18–20, 22]. These mechanisms encompass both hard and soft-attention techniques, facilitating the matching of similar patches while assigning weights to these correspondences to balance their influence based on patch similarity. Despite its high-quality outcome, this approach requires searching and matching of all possible patch pairs, which results in quadratic computational complexity [26]. To address this

limitation, we introduce a lightweight attention mechanism that leverages local similarities within a window of patches between neighboring frames. This method ensures precise texture transfer by exploiting natural similarities within the same subregion of multiple frames, while also reducing the computational complexity, memory usage, and execution time. In addition, optical flow is employed to extract warped features that are fused with patch-based attention features for more accurate and detailed high-frequency texture transfer.

Considering the aforementioned studies, there is a clear need for RefSR networks that are reliable, low in complexity and highly accurate, while also accommodating the energy and memory constraints of embedded systems. In this work, we present a configuration involving a camera recording high spatial resolution (HSR) - low frame rate (LFR) video feed, alongside a secondary camera capturing low spatial resolution (LSR) - high frame rate (HFR) video feed. Our HSTR-Net model combines these two video feeds to generate high spatial resolution high frame rate (HSTR) video content. HSTR-Net is a deep neural network designed to be lightweight and capable of almost real-time processing. It leverages both a reference (*REF*) frame and a low-resolution (*LR*) frame inspired by reference-based super-resolution techniques. This approach enables HSTR-Net to synthesize missing frames in the HSLF video using the corresponding *LR* frame from the LSHF video feed. To achieve this, HSTR-Net employs motion estimation/compensation and also incorporates an attention-based patch-matching mechanism. Subsequently, a fusion process is applied to generate the missing *HR* frame. A distinctive feature of our approach is the novel patch-matching strategy that transforms the self-attention mechanism from the Swin transformer [28] into a patch correspondence solution for RefSR. This innovative approach allows for high processing speeds on embedded systems, as it conducts patch matching through a lightweight attention mechanism, as opposed to previous RefSR methods that involve comparing all possible pairs of patches from the two input frames.

HSTR-Net deviates from our previous work [29] in terms of both complexity and quantitative performance. Suluhan et al. [29] requires a total of 5 input frames to construct the corresponding high-resolution frame, which in turn demands increased computational power, thereby resulting in a higher energy consumption footprint. In addition, the model in Suluhan et al. [29] is only applicable in scenarios where the *LR* and *REF* frames are temporally close to each other, thereby limiting its applicability for high frame rate increases. On the other hand, HSTR-Net offers a more generalized solution which can be successfully utilized for low frame rate video sequences with fast and complex motion, even when it is hard to achieve a precise registration of features between the *LR* and *REF* frames.

HSTR-Net stands out as a lightweight RefSR network that demonstrates competitive performance both quantitatively and qualitatively, making it suitable for deployment on embedded systems. HSTR-Net outperforms state-of-the-art in RefSR by 0.28 dB on the Vimeo90K [7] septuplet dataset while significantly reducing processing time. When evaluated on the higher-resolution VisDrone [30] aerial dataset, HSTR-Net excels with a remarkable 7.14 dB improvement and an average run-time reduction of $\%54$ over the next fastest RefSR model. Furthermore, HSTR-Net's superior performance is also assessed on the MAMI [31] dataset, a wide-area surveillance (WAS) dataset recorded using a multi-camera setup. Benchmark results on the MAMI dataset underscore HSTR-Net's adaptability and generalization capability, showcasing its ability to outperform prior approaches by 1.1 to 4.76 dB while maintaining high visual quality. Additionally, HSTR-Net is deployed on Nvidia Jetson AGX Xavier Developer Kit [32], achieving a processing rate of 4.62 frames per second.

The novel contributions of this paper can be summarized as follows:

- HSTR-Net enables high-quality video feed generation for aiding tasks such as object detection or mobile video processing on resource constrained devices thanks to its lightweight architecture.
- HSTR-Net architecture introduces novel and low-complexity patch correspondence using cross-frame attention for processing fast, complex and rapidly changing motion between frames.
- Optical flow-based warped features and patch-based attention features between the *REF* and *LR* frames are fused in HSTR-Net architecture for optimal high-frequency detail and texture transfer.
- HSTR-Net achieves state-of-the-art results, both quantitatively and visually, for Vimeo, VisDrone and MAMI dual-camera datasets.

Section 2 discusses the previous work in SR, VFI, and RefSR. Section 3 gives the architectural details of HSTR-Net. Section 4 provides details of training, simulations, and evaluation of results. Section 5 concludes the paper with a discussion of future work.

2 Related work

2.1 Reference-based super resolution (RefSR)

Single Image Super-Resolution (SISR) encounters challenges in incorporating high-frequency details into the super-resolution process, primarily because it relies solely on *LR* images. In response to this limitation, Reference-based Super-Resolution (RefSR) is an innovative approach that introduces an additional image or frame as reference. This reference frame serves as a means to infuse high-frequency details, complementing the low-resolution features and enhancing texture transfer and fusion within the super-resolution process.

The literature presents various RefSR methods that rely on patch matching between the reference and low-resolution images, followed by the transfer of textures from the *REF* to the *LR* image [16, 18–20, 24–27]. Despite achieving competitive results, these methods face challenges in capturing fine-grained information between successive frames in a video feed. Other studies adopt an approach involving optical flow estimation and compensation on neighboring frames, followed by a fusion process to obtain the final result [1, 23]. However, optical flow alone struggles to handle the rapid movement and frequent high rotation characteristics often associated with drones.

Zhang et al. [16], Yang et al. [18] rely on exhaustive patch matching between the *REF* and *LR* images, which can be computationally intensive and thereby unsuitable for deployment on resource-constrained platforms. Jiang et al. [24] propose two separate methodologies, including a correspondence network that learns robust transformation matching for addressing scale and rotation misalignments, and a teacher-student correlation approach that incorporates distilled knowledge for texture transfer. Nevertheless, a performance degradation occurs when dealing with substantial scale misalignments between the *REF* and *LR* images.

Lu et al. [19], Xia et al. [26] offer an alternative approach to the problem by introducing coarse-to-fine correspondence matching instead of exhaustively iterating through all potential matches, which reduces the model complexity while achieving satisfying results. Huang et al. [21] divides the RefSR task into two distinct stages. Initially, they apply single image super-resolution (SISR) and then leverage the generated SR image to facilitate texture transfer. This division helps address the issues of reference-underuse and reference-misuse, which pertain to insufficient and irrelevant texture transfers, respectively. Guo et al. [33] introduces a dual-search strategy, initially filtering out irrelevant textures from the *REF* data during the first pass to narrow down the search space. In the subsequent stage, refined search space allows for accurate texture transfer between the *REF* and *LR* images. Mei et al. [34] incorporates perceptual and adversarial loss into their approach, utilizing a feature reuse network to overcome the limitations associated with relying solely on reconstruction loss. This adaptation helps prevent the generation of overly smooth SR images. Furthermore, they introduce dynamic filters and a spatial attention mechanism to mitigate the issues of irrelevant texture transfer and misalignment between the *LR* and *REF* images. Shim et al. [35] employs deformable convolutions to achieve adaptive alignment between *LR* and *REF* frames, effectively learning this alignment implicitly.

Mobile phone manufacturers have introduced a new dual-camera setup, featuring multiple lenses such as telephoto and wide-angle cameras. These cameras capture scenes from varying angles, enhancing the overall user experience. Wang et al. [25], Lee et al. [27] leverage telephoto and wide-angle frames from mobile phones and address content disparities between the frames by transferring missing texture information from the wide-angle frames. However, this sometimes leads to blurriness at the edges of the resulting images. In contrast, Zou et al. [36] proposes an alternative approach. They employ a geometric matching method in combination with patch-matching to achieve more precise alignment between the *LR* and *REF* frames, enhancing the subsequent texture transfer process.

The application of attention-based networks in RefSR is quite natural, as it can facilitate patch matching through self-attention mechanisms. In many instances, attention-based RefSR approaches have focused on conducting global patch matching using a combination of hard and soft attention mechanisms [18, 20, 22]. Hard attention is typically employed to identify the most similar patches between the *REF* and *LR* images. This involves an exhaustive iteration over all possible patch pairs to determine the best matches. Subsequently, soft attention comes into play, assigning weights to each patch matching in the final result. The purpose of soft attention is to ensure that the contribution of each match is fairly considered and eliminating the risk of treating all patch pairs as equally significant. Jiang et al. [24] integrates attention mechanisms into the fusion process. An attention map is generated based on extracted features and patch correspondences, with the goal of achieving a focused and enhanced fusion. Cao et al. [37] integrates the Swin transformer to replace the traditional ResBlock for feature aggregation, significantly enhancing model performance, yet also increasing model complexity and runtime. Aslahishahri [38] uses hierarchical transformer model and double attention block for superior results in benchmark refSR datasets. While these attention-based strategies are valuable, they are often computationally demanding and they may not be sufficient to address all the complexities of RefSR for video sequences exhibiting complex motion.

Recent literature also contains domain-specific or task-specific applications of RefSR models. Dong et al. [39] proposes a change-aware diffusion model (Ref-Diff) for RefSR, utilizing land cover change priors to guide the denoising process. The method injects these priors to handle reference utilization in unchanged areas and content reconstruction in changed regions, decoupling semantics-guided and reference texture-guided denoising. Min et al. [40] introduces a Domain Matching (DM) module designed to mitigate domain gaps between reference and low-resolution images in RefSR. The module is integrated in a plug-and-play manner to enhance texture transfer in remote sensing scenarios, where domain shifts often occur due to variations between satellite sources. Zheng et al. [41] proposes a multi-correspondence matching framework for RefSR, addressing reference underuse and misuse by identifying multiple similar correspondences within the reference image. The framework aligns and aggregates multiple reference features with the LR feature, determining their contributions based on similarity, and integrates seamlessly with existing RefSR models. Tu et al. [42] introduces a gradient-assisted texture-enhancement GAN (RGTGAN) for RefSR in remote sensing. The method incorporates a gradient-assisted module for feature extraction, a dense-intern deformable convolution for feature alignment, and a dense-restore-residual module for feature transfer, addressing challenges in texture reconstruction and dataset limitations in remote sensing RefSR.

2.2 Video frame interpolation (VFI)

Video frame interpolation (VFI) is employed to recover missing frames within a video sequence by leveraging information from neighboring frames. Typically, optical flow estimation between adjacent frames is utilized to create a warped reference frame, which is subsequently fused with additional features to generate the final output frame. Recent work [43, 44] proposes architectures that estimate optical flow between consecutive frames and utilizes deformable convolution for the motion compensation process. Deformable convolution offers enhanced adaptability to unseen data compared to traditional warping methods since it is trainable. Jin et al. [45], Lin et al. [46], Li et al. [47] predict bi-directional flows between consecutive frames, which are subsequently used for the warping process. They employ frame synthesis networks to refine the generated warped frames. Hu et al. [48] separates frame interpolation into two parallel processes, which are a temporal prediction network that focuses on motion-related tasks such as motion estimation and compensation, and a spatial prediction network that extracts visual content and features from neighboring frames. Plack et al. [49] initially extracts features at various resolutions and uses them as inputs for each level of the network. Each level comprises two transformer blocks and a flow estimation block, and they utilize the output from the previous layer alongside the feature inputs. Zhou et al. [50] enhances motion estimation by gathering correlation information between frames and refining motion fields to generate more accurate motion information. This work also generates occlusion maps and synthesize the final frame using occlusion maps and estimated motions. Li et al. [51] proposes extracting dense correlations before motion estimation to eliminate occlusions and blurs caused by large motions. This work generates multiple occlusion maps and motion fields to synthesize the final frame accurately, without occlusions or blurs. Liu et al. [52] employs two input frames and context features to synthesize an intermediate frame by estimating trajectories, reducing the impact of blur and distortion resulting from optical flow-based warping. Zhang et al. [53] performs space-time video super-resolution and uses two *LR* frames for motion estimation, followed by forward and backward warping operations to produce five *HR* frames. Luo et al. [54] leverages the information from four *HR* frames to combine the extraction of multi-level spatio-temporal dependencies with the task of motion estimation and compensation to generate the final output. Li et al. [55] utilizes four high-resolution (*HR*) input frames. These frames are employed for extracting spatio-temporal features and creating a multi-scale contextual representation, which serves as input to a fusion module based on deformable convolution.

In the context of wide-area monitoring, where video frame rates are typically low, conventional VFI methods struggle to cope with the sudden and rapid motions often encountered in wide-angle video feeds. This limitation arises from their heavy reliance on assumptions of constant and linear motion. Therefore, RefSR is a more viable approach for spatio-temporal resolution enhancement in wide-angle videos, since it can employ the *LR* frame information at the generated output when accurate motion estimation or *REF-LR* registration is not possible. In other words, RefSR combines the advantages of SISR and VFI to generate high-quality artifact-free HSTR videos.

3 Methodology

We propose a lightweight reference-based frame interpolation technique, namely HSTR-Net, to generate high spatio-temporal video feed from HSR-LFR and LSR-HFR video footages recorded by a dual camera setup. The architecture is inspired by RIFE [56] and illustrated in Fig. 1. The missing HR_t frame is synthesized by using an LR_t frame from the LSR-HFR video and a neighboring HR_{t-1} or HR_{t+1} frame from the HSR-LFR video as a reference. Motion estimation

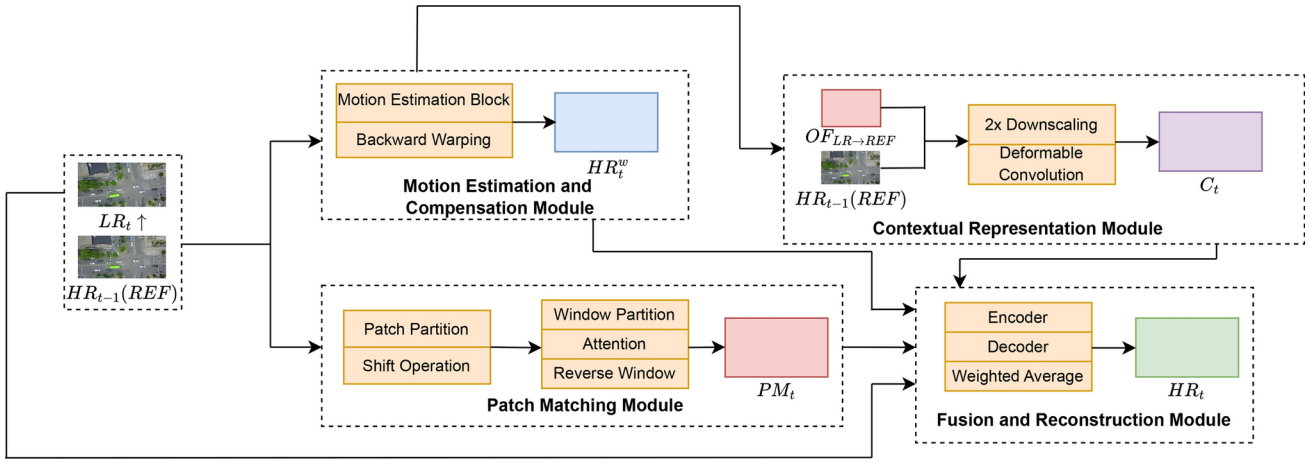


Fig. 1 Overall architecture of HSTR-Net. Motion estimation and compensation are applied to *LR* and *REF* frames to generate a warped reference frame. An attention-based patch-matching scheme is applied to input frames to make use of correspondences between similar textures. A contextual representation module extracts multi-scale features by applying deformable convolution at various resolutions. Lastly, a fusion and reconstruction module is used to generate the final *HR* output frame

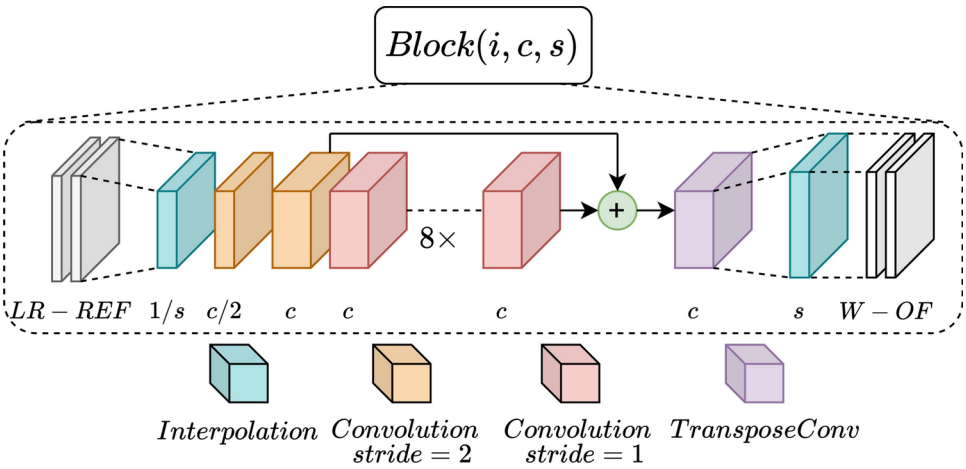


Fig. 2 Motion estimation block architecture (*i, c, s* correspond to input channel size of a block, output channel size for convolution layers and scale factor, respectively)

and compensation techniques are applied as the initial block for generating a warped reference frame. Contextual representation extraction phase produces features and information at various scales using estimated optical flow and warped frame. Attention-based patch-matching module performs correspondence estimation between patches of *REF* and *LR* in a local manner to exploit similarities occurring between frames in two video feeds. Lastly, a fusion module synthesizes HR_t by weighting the features gathered from prior stages. The HSTR-Net architecture generates a high spatial quality *HR* output through the coupled motion compensation and accurate texture transfer from *REF* to *LR* frame while considering execution time and power constrains. The local patch-matching mechanism provides necessary features for reliable and accurate texture transfer, since frames of time-synchronized video feeds have similar spatial textures. Furthermore, the reduced model complexity of single reference-based HSTR-Net enables deployment onto an embedded system composed of dual cameras. HSTR-Net achieves 4.62 fps on Jetson Xavier board with an input resolution of 672×380, which is sufficient for wide-area monitoring on drones.

3.1 Motion estimation and compensation module

HSTR-Net relies on accurate motion estimation and compensation since optical flow accuracy is important for the quality of the generated image in optical flow-based models. HSTR-Net motion estimation predicts optical flow from the *HR* reference frame to the *LR* frame, which is denoted as $OF_{LR\to REF}$. The motion estimation module consists of three consecutive blocks ($Block_i, 0 \leq i \leq 2$) that extract motion at different scales; each block consists of downsampling and multiple convolutional layers as illustrated in Fig. 2. Each block has *i, c, s* parameters that are input channel size, output channel size for convolution layers, and scale factor, respectively. Blocks 0, 1, and 2 have 6, 7, and 7 input channels and inputs are downsampled by scale

factors of 4, 2, and 1, respectively for each block. $Block_0$ has 6 input channels and produces a warped reference image with 3 channels and optical flow with 4 channels (representing both $OF_{LR\rightarrow REF}$ and $OF_{REF\rightarrow LR}$ with each having 2 channels) that are concatenated channel-wise and sent to the next block as input. $Block_1$ has 7 input channels and generates similar outputs as $Block_0$. $Block_2$ is the last block in the motion estimation module that has 7 input channels and calculates the final optical flows of $OF_{LR\rightarrow REF}$ and $OF_{REF\rightarrow LR}$. Inside each block, the first two convolutional layers are used to match the output channel size and further downsample the input with a stride of 2. Then, 8 convolutional layers with the same input and output channel size are employed. Outputs of these two sub-blocks are added and fed to the transpose convolution layer, which has an output channel size of 4 and stride of 2. Interpolation is employed at the end to increase the resolution of the estimated optical flow field to that of the input frame.

For motion compensation, a fixed warping operator is used to generate the warped HR_t^W from the *REF* frame. The warping module applies sub-pixel motion compensation to accurately obtain the target frame using vertical and horizontal flow components of $OF_{LR\rightarrow REF}$. HR_t^W constitutes an initial prediction for HR_t; however it requires further refinement with the necessary features in order to obtain artifact-free high-quality results.

3.2 Contextual representation module

The contextual representation module is illustrated in Fig. 3. This module takes the *REF* image and $OF_{LR\rightarrow REF}$ as input and generates contextual feature maps C_0, C_1, C_2, C_3 at four different scales to support the subsequent fusion process. The module gradually lowers the resolution of the input image by a factor of 2 using convolutional layers. Optical flow is also downsampled to match the feature map size. These convolutional layers have a stride of 2 and output channel sizes as (16, 32, 64, 128). Then, this module applies deformable convolution at each scale, treating optical flow as an offset in convolution for producing warped reference frame features at various resolutions. Deformable convolution is a trainable feature registration operator that provides more precise alignment for the features of the reference frame when compared to the traditional fixed warping. A learned warping operator can provide better alignment of extracted features with the synthesized frame. The deformable convolution module is a built-in *PyTorch* function, which is efficient in terms of memory and computational complexity.

3.3 Attention-based patch matching module

In this section, we introduce our novel attention-based architecture inspired by Liu et al. [28], which transforms self-attention for single input frame into a patch correspondence module for reference-based super-resolution. Texture transfer from *REF* to *LR* frame is an important phase that requires accurate detection of similar spatial textures. Prior work focuses on iterating over all possible pairs of patches in two images leading to inefficient models in terms of execution time and limiting the ability to deploy on power-constrained systems. However, our module is a lightweight attention mechanism that performs patch-matching operations in various resolutions. In addition, the preceding methods rely on global patch-matching which is suitable for images with large texture differences. However, it is unable to exploit the local similarities between neighboring frames. Our patch-matching module is solely based on local correspondences with a window structure for detecting and transferring similar textures.

In this module, *LR* and *REF* frames are initially partitioned into 4×4 patches by a convolution layer with a stride of 4 and output channel size of 48 and these patches are given as input to the sequential patch-matching blocks. Firstly, a shift operation is applied on these patches to extract features at different locations and scales, and 3×3 windowed groups are formed from shifted patches. Then, a novel self-attention mechanism is applied on these windows

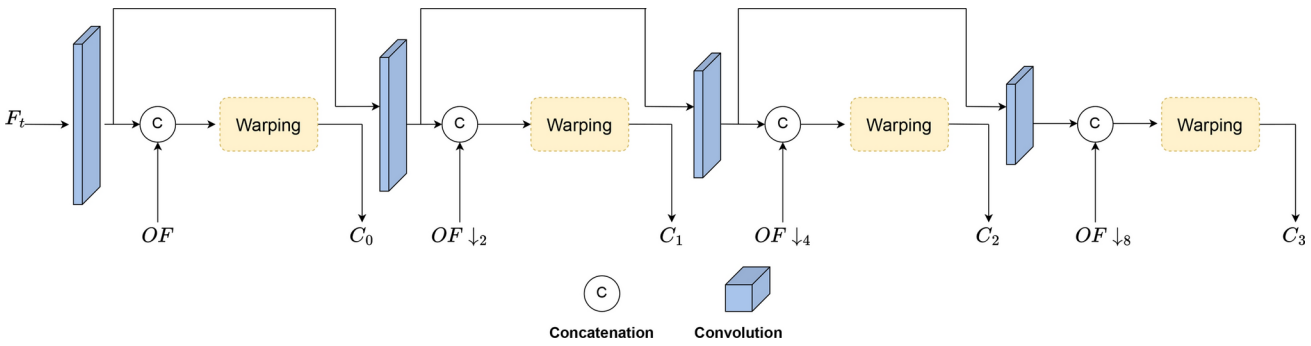


Fig. 3 High-level diagram of contextual representation module. F_t and *OF* denote input frame and optical flow respectively. In each level, the input feature map and *OF* is downscaled by a factor of 2. C_0, C_1, C_2 and C_3 denote differently sized outputs of ContextNet

as detailed below. A pipeline of 6 patch matching blocks is created by forming 3 groups with each having 2 sub-blocks. Additionally, a patch merging mechanism is employed on the last two groups for increasing channel size and reducing the resolution of input feature maps. Consecutive patch matching groups have output channel sizes of (48, 96, 192), respectively. This approach offers patch-matching in various resolutions and localizations that enable a better adjusted and robust fusion process. A high-level overview of this module is illustrated in Fig. 4.

Patch Matching Block:

This paper proposes a lightweight patch-matching block that performs self-attention on two given inputs. This block is the backbone of our attention mechanism and it applies attention at different scales by processing prior outputs of itself.

The first sub-block in the patch matching block combines patches into windows to calculate correspondence between patches within the same window. In the second sub-block, a shifting operation is applied on partitioned windows. The two successive sub-blocks are connected in order to introduce cross-patch correspondence and exploit the local information at various positions. This approach enables focusing on the local similarity between neighboring patches and is a suitable method for video super-resolution, since similar textures lie in adjacent subareas of the neighboring patches. Then, the self-attention operation of Liu et al. [28] is transformed into *LR-REF* patch-matching as presented in Eq. 1. The query (Q), key (K), and value (V) properties of both *LR* and *REF* frames are calculated for attention using linear layers. During attention calculation, two matrix multiplications are performed, which are $Q_{LR} \times K_{REF}$ and $Q_{REF} \times K_{LR}$, to obtain correspondences between *LR* and *REF* frames in both directions. The result of first matrix multiplication, $Q_{LR} \times K_{REF}$, is multiplied with V_{REF} to obtain intermediate attention AT_i, since we are interested in transferring textures from *REF*. The other result, AT_{LR}, is multiplied with the AT_i to incorporate the *LR* attention information and get the final attention output of AT:

$$\begin{aligned} AT_{REF} &= Q_{LR} \times K_{REF} \\ AT_{LR} &= Q_{REF} \times K_{LR} \\ AT &= AT_{LR} \times (AT_{REF} \times V_{REF}) \end{aligned} \tag{1}$$

AT carries information on the patch-by-patch similarity measure between the two images, which provides a powerful reference for the texture transfer process. Unlike prior work, this approach is able to leverage the attention mechanism's advantages in an efficient manner resulting in faster and low-memory forward propagation. Following the setting of [28], a patch merging mechanism is used to create a hierarchical representation of feature maps at different resolutions. These feature maps are represented as PM_0, PM_1, PM_2 in Fig. 4 and they will be input to the fusion module to provide the matching high-frequency texture. Our patch-matching block is able to detect similar and relevant patches between *LR* and *REF* frames for the texture transfer process in an efficient way, leading to an optimized and high-performance network.

3.4 Fusion module

The fusion module is the final block to reconstruct the HR_t frame. Our fusion block, $\mathcal{F}$, is a modified UNet[57] auto-encoder structure that processes contextual information in both finer and coarser resolutions, which is illustrated in Fig. 5. In addition, it has residual connections between corresponding encoder and decoder layers, which exploit feature information from the earlier layers in the latter phases of the decoding process. The inputs of the auto-encoder block are HR_t^W (warped reference frame), LR_t, PM (patch correspondences), and C (contextual feature maps). The outputs of the block are residual high-frequency components (R) and a weight mask M:

$$R, M = \mathcal{F}(HR_t^W, LR, PM, C) \tag{2}$$

A subsequent reconstruction phase is employed to weigh the final contributions of *LR* and the warped reference frame, HR_t^W, utilizing the residual high-frequency component R

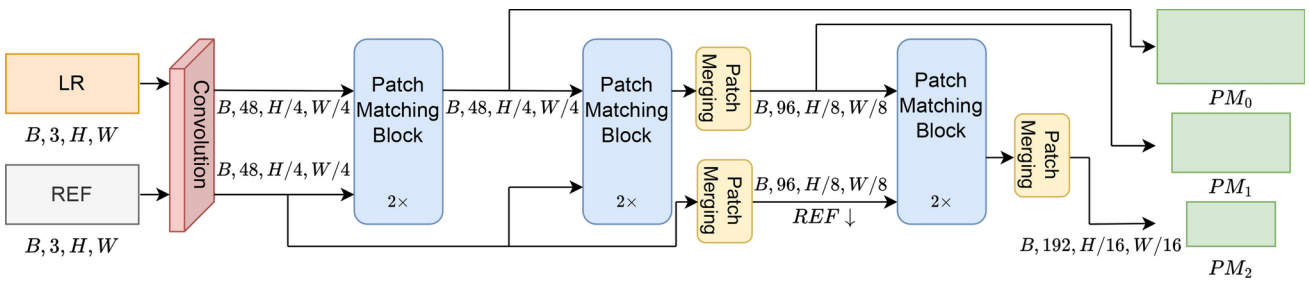


Fig. 4 High-level diagram of patch matching module. Given *LR* and *REF* frames, patch-partition is applied by a convolution layer. Later, patch-matching groups are applied to produce patch-matching outputs. Each output is used as an intermediate input for the next group and preserved for the final output. The last two groups apply patch-merging to decrease resolution by a factor of 2

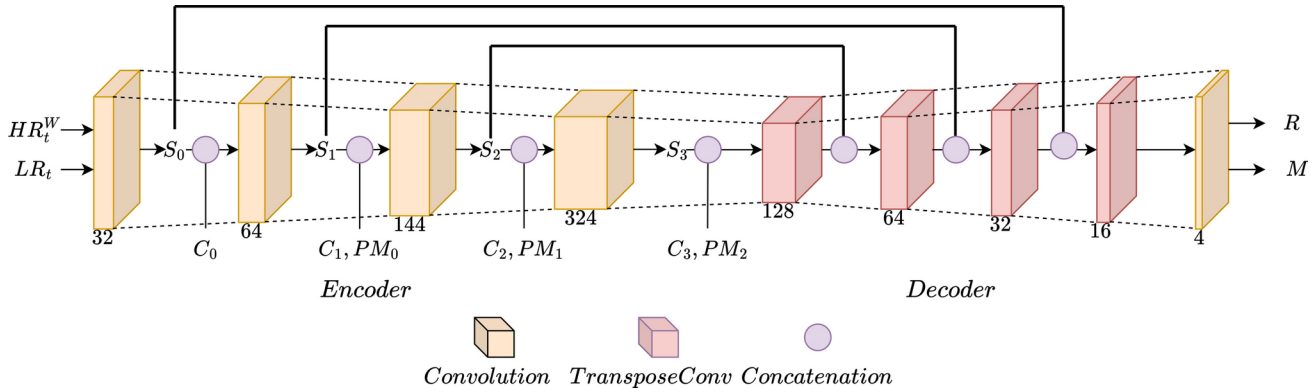


Fig. 5 Auto-encoder architecture of fusion module. The contextual feature maps C_i and patch correspondences PM_{i-1} are concatenated with appropriate encoder layer outputs S_i with matching resolutions and provided as input to the next encoder layer

and mask M generated by auto-encoder architecture. The weighted reconstruction is defined as below:

$$\hat{HR}_t = M \cdot HR_t^W + (1 - M) \cdot LR + R \tag{3}$$

The auto-encoder block has four downsampling and four upsampling layers with a final convolutional layer. Each downsampling layer consists of convolution followed by a ReLU operation where each upsampling layer is a combination of transpose convolution followed by a ReLU. All convolution and transpose convolution layers including the final layer have a kernel size of 3. Output channel sizes of consecutive downsampling and upsampling layers are (32, 64, 144, 304) and (128, 64, 32, 16) respectively. The last layer has 16 as the input channel size and 4 as the output channel size. Three channels of the output give residual components R for each color band, and the other output is the weighting mask. Frame $\hat{HR}_t$ is synthesized by using R and mask M as illustrated in Eq. 3. The weighted average of the reference warped frame and *LR* image is calculated using the mask, which balances the contribution of each input frame to the synthesized output. The residual R compensates for any missing high-frequency texture that may not be matched with or transferred from the *REF* frame.

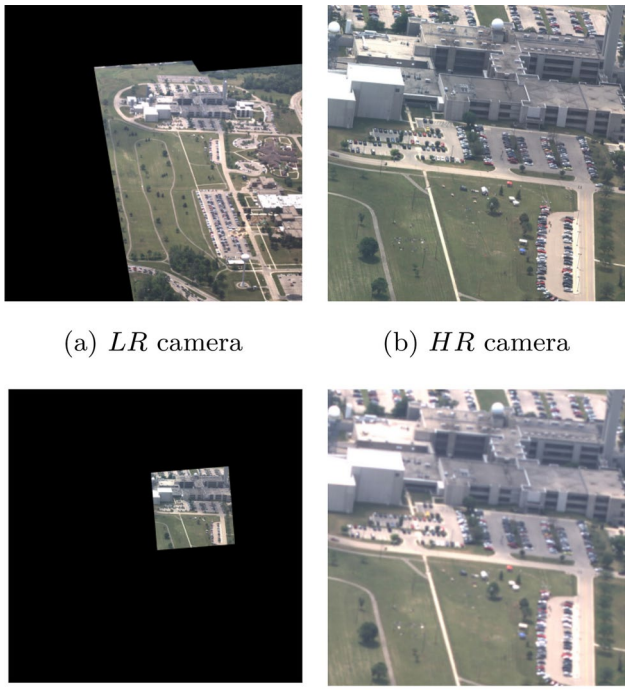
4 Experiments

4.1 Datasets

HSTR-Net is trained and tested on *HR-LR* image pairs from Vimeo90K [7], VisDrone [30] and MAMI [31] datasets. Vimeo90K is a benchmark dataset for SR, VFI and RefSR applications. Vimeo90K septuplet dataset has 91,701 sequences arranged as 7 consecutive frames with a fixed resolution of 256×448; HSTR-Net is trained and tested on the septuplet dataset. The triplet dataset of Vimeo90k has 73,171 3-frame sequences with resolution of 256×448. Vimeo90K triplet dataset is only used for testing purposes to demonstrate the robustness of HSTR-Net across different content. The Vimeo90K dataset comprises low-motion scenes with large objects from everyday environments, and its extensive size allows for training HSTR-Net across a diverse range of scenes. Furthermore, VisDrone is an aerial imagery dataset consisting of drone-captured footages; this dataset is widely used in aerial object detection and tracking challenges. High-resolution VisDrone videos are downsampled to 672×380 resolution in order to simulate typical resolutions for wide-angle frame sequences. This dataset presents new challenges because it includes fast drone motion/rotation, small moving objects, etc. Since both datasets only have *HR* frames, whereas RefSR requires corresponding *LR* frames, 4x bicubic downsampling followed by 4x bicubic upsampling is applied to synthetically generate the *LR* frames.

Cross-dataset generalization is vital to evaluate the robustness of deep learning models in real-world applications. A model adapting across multiple datasets without fine-tuning proves its ability to handle diverse motion patterns, object sizes, and scene complexities, where training data may not always align with real-world conditions. To assess HSTR-Net's robustness across multiple datasets, we test HSTR-Net on the VisDrone dataset after training on Vimeo90K.

In addition to these two datasets, a true dual-camera setup is employed using the Minor Area Motion Imagery (MAMI) [31] dataset. MAMI contains four *HR* and one wide-angle *LR* camera capturing the aerial video of a military campus. Figure 6a and b show sample frames of the *LR* and *HR* cameras, respectively. *LR* camera frames are matched and registered with *HR* camera frames and cropped to 512×512 sized outputs in order to generate the MAMI dataset for RefSR purposes (see Fig. 6d). Our MAMI RefSR dataset consists of 8, 018 train and 1, 983 test samples with a



(a) LR camera (b) HR camera

(c) registered HR image (d) registered LR image

Fig. 6 Visuals from MAMI dataset

resolution of 512×512. MAMI is a challenging dual-camera setup, since it includes camera misalignment, varying color and contrast, different orientations of the cameras, etc. Therefore, diverse augmentation techniques are necessary to train a generalized network, which is adapted to distortions and misalignments of the dual-camera setup. To the best of our knowledge, MAMI is the first real-life dual-camera dataset that specifically targets reference-based super-resolution in WAS.

4.2 Training and experimental setup

HSTR-Net is trained on the Vimeo90K septuplet training set by using HR_{t-1} or HR_{t+1} frame as *REF* for LR_t with %50 probability. Random cropping, rotation and flipping are applied for data augmentation during training. Each frame is cropped randomly to 128×128 resolution. Random rotation varying from -10 to 10 degrees and flipping with %20 probability are applied as additional data augmentation techniques. L_1 loss is used for training to synthesize sharp and qualitatively better-looking frames while slightly losing quantitative accuracy. HSTR-Net is trained for 62 epochs with a learning rate of 10^{-5}.

Peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [58] are the metrics used for evaluation. LPIPS calculates the perceptual similarity between two images based on learned feature representations. We adapt the input set configuration from [1] and designate the fourth frame of the septuplet test sequence as the *LR* frame and the fifth frame as *REF* for the evaluation of RefSR models. For VFI models, the third and fifth frames are used as input to synthesize the fourth frame. Luo et al. [54], Li et al. [55] models already provide quantitative results for the fourth frame utilizing the first, third, fifth, and seventh frames for inference. On the triplet dataset, the second frame is downsampled and used as *LR* and the third frame becomes *REF* for RefSR models, while the first and third frames are provided as input to the VFI models. Lastly, LR_t and HR_{t+1} frames are used on VisDrone test sequences for RefSR model evaluations.

Table 1 Quantitative comparison on Vimeo90k septuplet dataset (the best result is highlighted in bold, and the second best result in italics)

Method	Model	PSNR	SSIM	Time(ms)	# of Parameters (M)
RefSR 1LR + 1 HR	TTSR[18]	33.27	0.940	72.13	6.7
	SSEN [35]	34.45	0.958	84.32	5.0
	SRNTT [16]	35.76	0.975	1326.28	5.7
	MASA-SR [19]	36.57	0.980	68.72	4.0
	C2-Matching [24]	38.40	0.983	90.82	–
	CrossNet [23]	39.17	0.985	–	–
	DATSR [37]	39.26	0.983	156.43	19.4
	FRFSR [34]	39.55	0.985	123.65	16.8
	AWNet [1]	*39.66*	**0.986**	*56.02*	27.3
	HSTRNet	**39.94**	**0.986**	**29.56**	11.1
VFI 2 HR	RIFE [56]	35.78	0.972	–	–
VFI 4 HR	BP3D [54]	35.76	0.959	–	–
	ESTF [55]	35.75	0.961	–	–

4.3 Results and discussions

We first present quantitative comparison results of RefSR techniques on Vimeo90K and VisDrone datasets. We employ state-of-the-art RefSR methods to compare against HSTR-Net. VFI results are also provided for reference. In Table 1, measurements regarding RefSR models are collected from septuplet dataset of our benchmark except [23]. The result for CrossNet [23] is taken from [1]. VFI methods [54, 55] provide PSNR/SSIM results on their respective papers; [56] is re-evaluated on the benchmark. In Table 2, for the triplet dataset, the results of the tested methods are taken from their respective papers if they are available; otherwise the trained models are tested on the Vimeo90K triplet benchmark. In Table 3, available RefSR models are compared on the VisDrone test set. To accurately assess the cross-dataset generalization of each model to an aerial video

Table 2 Quantitative comparison on Vimeo90k triplet dataset (the best result is highlighted in bold, and the second best result in italics)

Method	Model	PSNR	SSIM
RefSR 1LR + 1 HR	TTSR [18]	31.61	0.927
	SSEN [35]	32.12	0.944
	MASA-SR [19]	35.08	0.970
	SRNTT [16]	35.14	0.972
	C2-Matching [24]	37.07	0.980
	FRFSR [34]	37.43	0.981
	DATSR [37]	37.96	0.980
	AWNet [1]	*38.36*	**0.983**
	HSTR-Net	**38.45**	**0.983**
VFI 2 HR	RIFE [56]	35.61	0.978
	[49]	36.08	0.979
	UPR-Net [45]	36.42	0.981
	FGDCN [43]	36.46	0.981
	AMT [51]	36.53	0.982
	TTVFI [52]	36.54	0.981
	DQBC [50]	36.57	0.981
VFI 4 HR	ESTF [55]	35.14	0.976

Table 3 Quantitative comparison on VisDrone dataset (the best result is highlighted in bold, and the second best result in italics)

Method	Model	PSNR	SSIM	LPIPS	Time(ms)
RefSR 1LR + 1 HR	SSEN [35]	22.76	0.796	0.244	231.16
	C2-Matching [24]	25.25	0.868	0.198	260.08
	AWNet [1]	26.16	0.852	–	*116.53*
	TTSR [18]	27.57	0.883	0.238	214.40
	DATSR [37]	29.76	0.926	0.109	326.87
	MASA-SR [19]	30.65	0.940	0.110	160.30
	FRFSR [34]	31.12	0.953	0.071	248.32
	SRNTT [16]	*32.17*	*0.959*	**0.065**	3527.22
	HSTR-Net	**33.30**	**0.965**	*0.072*	**53.35**
	HSTR-Net-*f*	34.13	0.969	0.067	53.35

dataset, we perform the simulations without any fine-tuning. In the tables, the best RefSR method is highlighted in bold, and the second best result is highlighted in italics.

As seen from Tables 1 and 2, HSTR-Net generates superior results when compared to state-of-the-art RefSR and VFI methods. HSTR-Net achieves 0.28 dB and 0.09 dB enhancement in PSNR metric with respect to second best RefSR result (i.e. AWNet) for septuplet and triplet datasets respectively, while reducing execution time by 47% for Vimeo90K video resolution. Despite only using an *HR* and an *LR* frame, HSTR-Net surpasses models involving multi-frame setup such as [54, 55] as well as VFI models that use 2 *HR* frames, since HSTR-Net's unique patch-matching module efficiently and accurately transfer high-frequency textures, which are obtained from *REF* to *LR* frame. Our unique patch-matching mechanism successfully contributes to the HSTR-Net, since it is designed specifically to capture local patch similarities between frames rather than global pixel or patch matching as in previous RefSR methodologies. Since adjacent patches are spatially similar for neighboring frames, the method provides accurate matching results at low complexity. Additionally, coupling motion estimation/compensation with the patch-matching method increases the model's robustness and accuracy by capturing dynamic movements between adjacent frames.

The drone footages of VisDrone often include blurs, distortions caused by high-speed rotation and complex motion of the drone and moving objects at the scene. Therefore, accurate and reliable motion estimation and compensation complemented with necessary features and textures for the fusion process is essential for acquiring both quantitatively and visually pleasant RefSR results. HSTR-Net synthesizes better results compared to other RefSR models quantitatively on the VisDrone dataset, since it successfully combines motion compensation and texture transfer stages, as opposed to other tested models with limited capabilities for modeling complex motions. Additionally, local similarity embedding gives an edge to HSTR-Net, which is not present in global matching-based methodologies. For instance, AWNet achieves comparable performance to HSTR-Net on Vimeo90K; however the model fails to compete with HSTR-Net on the VisDrone dataset and gives more than 7 dB lower quality. HSTR-Net surpasses the second best approach (i.e. SRNTT) by 1.13 dB in terms of PSNR and by 0.006 in SSIM, while SRNTT achieves lowest LPIPS score on the VisDrone dataset; however SRNTT is not suitable for real-time processing due to its high computational complexity. In terms of perceptual quality, HSTR-Net provides high-quality predictions as quantified by LPIPS metric in Table 3 as well as demonstrated visually in Fig. 7. Despite being trained on the simple and low motion sequences of Vimeo90K, HSTR-Net model is capable of capturing the fast and complex motions of VisDrone sequences. This outcome demonstrates HSTR-Net's generalization capability and the model's ability to capture fine details and maintain structural consistency in complex aerial video sequences, effectively transferring knowledge across datasets with minimal loss of quality. Table 3 also includes a model named HSTR-Net-*f*, which is fine-tuned HSTR-Net on the VisDrone training dataset. Fine-tuning HSTR-Net increases PNSR by less than 1 dB, which shows that HSTR-Net is a well-generalized model. In fact, when the results from Tables 1, 2 and 3 are analyzed, it is clear that HSTR-Net is by far a more generalizable model than its competitors.

Figure 7 provides visual comparison between tested methods for selected frames of VisDrone dataset. As observed in Fig. 7, HSTR-Net is able to capture fine details and generates visually sharp images. Other techniques are unable to cope with high rotations and increased speed of the drone leading to blurry and distorted outputs, since most existing architectures heavily rely on correspondence matching and

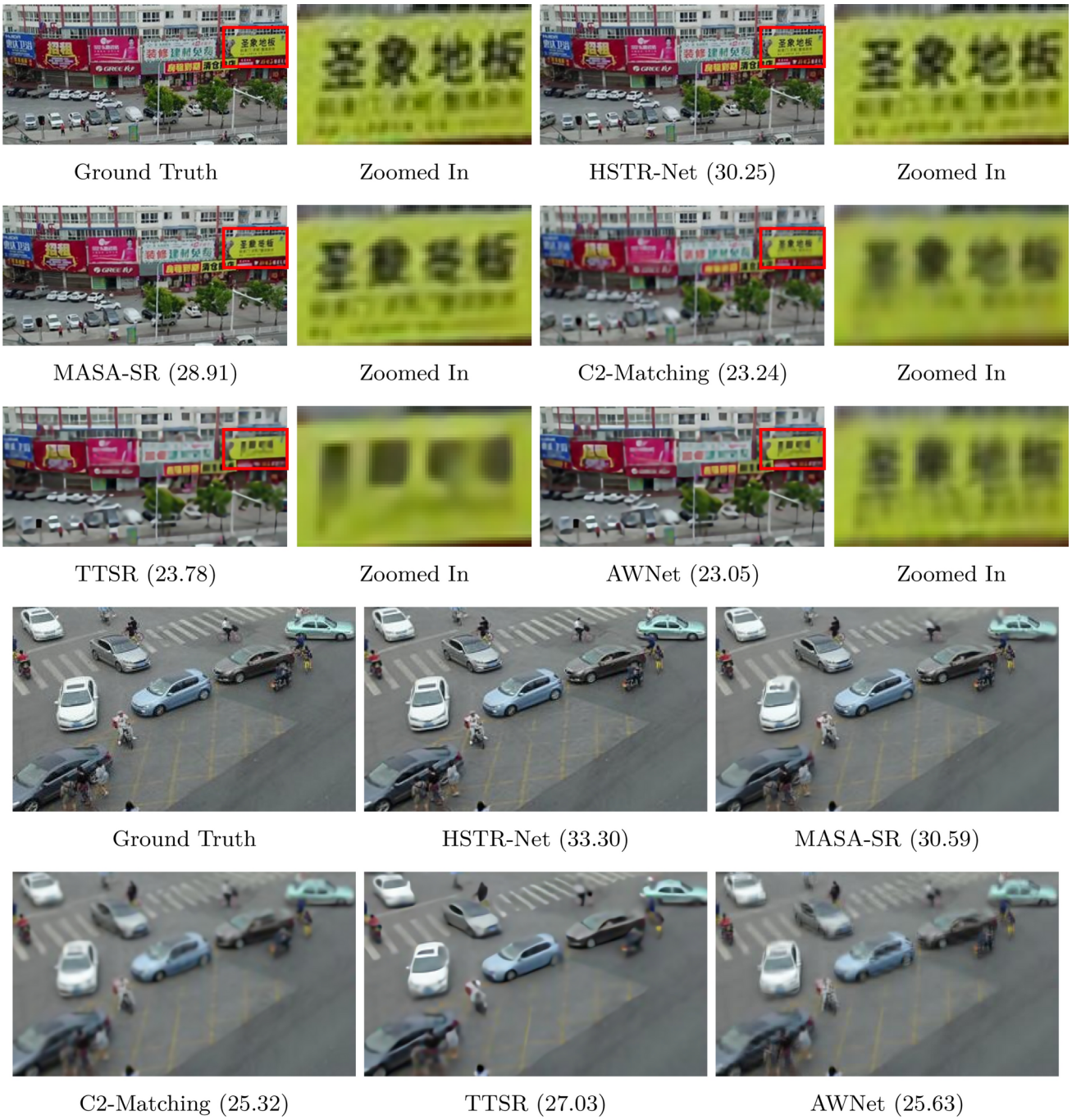


Fig. 7 Qualitative comparison of different models on VisDrone dataset (PSNR in dB)

do not include separate modules for motion estimation and compensation. Our approach is capable of synthesizing finer details, since it combines optical flow estimation and patch-matching schemes in an optimal way. To effectively showcase HSTR-Net's ability to capture fine details, we present a zoomed-in view of the red-labeled regions in each prediction in Fig. 7. The zoomed-in area focuses on a text region that is particularly challenging for video super-resolution models. Text reconstruction poses a significant challenge for many video enhancement techniques due to its small size, intricate details, and the impact of motion blur or compression artifacts in *LR* frames. Conventional approaches that rely solely on optical flow estimation or patch matching often struggle to preserve such fine-grained details. In contrast, HSTR-Net integrates both optical flow estimation and patch-matching-based texture transfer, enabling it to effectively address these challenges. By combining these techniques, HSTR-Net not only models complex motion patterns accurately

Table 4 Comparison of RefSR networks on MAMI dataset (the best result is highlighted in bold, and the second best result in italics)

Model	PSNR	SSIM	LPIPS
TTSR [18]	20.58	0.678	0.552
SSEN [35]	22.32	0.719	0.510
MASA-SR [19]	22.90	0.728	0.498
DATSR [37]	24.05	0.772	0.373
FRFSR [34]	*24.24*	*0.785*	*0.349*
HSTRNet	**25.34**	**0.804**	**0.326**

but also preserves and reconstructs fine spatial details, such as text, with greater precision. This capability highlights HSTR-Net's advantage over traditional methods, as it can produce high-quality, readable text even in challenging conditions, such as low-resolution inputs and complex motion.

Next, HSTR-Net model is tested on the MAMI dataset to demonstrate that HSTR-Net is capable of adapting to real-life scenarios, thereby constitutes a suitable solution for WAS. For comparison, we fine-tune TTSR [18], SSEN [35], FRFSR [34], DATSR[37], and MASA-SR [19] along with HSTR-Net on the MAMI training set. Since Vimeo90K is a synthetically generated training set, this fine-tuning is necessary in order for the models to adjust to the dual-camera issues, such as registration errors and differences in point spread function, color and contrast between the two camera frames. As observed in Table 4, PSNR values are comparably lower due to the challenging characteristics of the MAMI dataset. However, HSTR-Net is superior to its counterparts by 1.11 to 4.76 dB, since the patch-matching scheme accurately complements the motion compensation module. Besides having higher PSNR and SSIM metrics, HSTR-Net generates visually sharp and pleasant images compared to its counterparts as illustrated in Fig. 8 and also by the LPIPS metric, thanks in part to the augmentation techniques employed during the training process to reduce the impact of distinct characteristics of *LR* and *REF* images. However, other methods are unable to compensate for the misalignments, contrast differences, etc., which eventually leads to blurry, color distorted and/or contrast distorted frames as presented in Fig. 8.

In Tables 1 and 3 computational complexity of HSTR-Net is also compared against existing methods. Timing results in Table 1 provide inference times for 256×448-sized frames of Vimeo90K septuplet and triplet datasets. As seen from Table 1, HSTR-Net successfully balances accuracy and execution time, when compared to the second-best result provided by AWNet [1]. In addition to execution times as direct measure of computational complexity, we also report the parameter count for each model to provide insights into network size. Conventional RefSR methods [16, 18, 19, 35] have smaller network sizes, reducing memory requirements. However, they are computationally intensive, resulting in slower execution times, rendering them less suitable for real-time processing on resource-constrained devices. In contrast, advanced models [34, 37], while offering improved

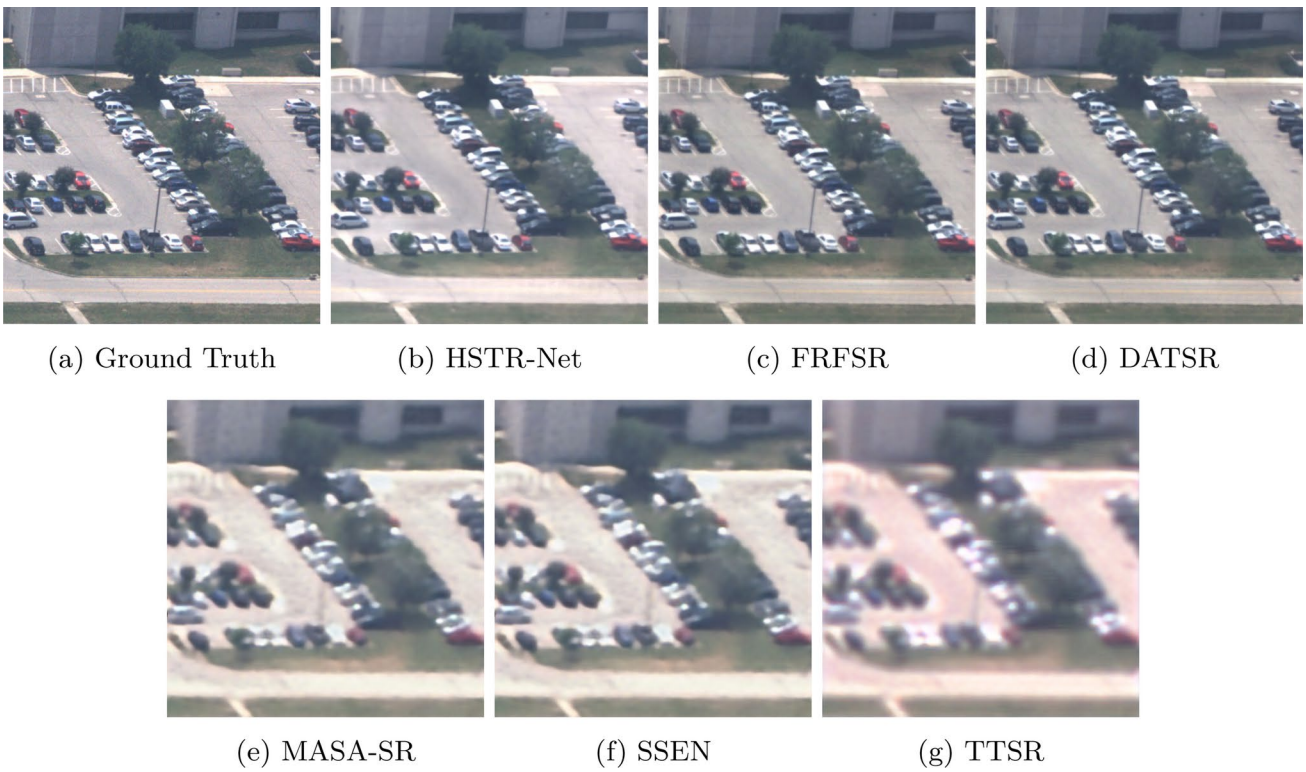


Fig. 8 Qualitative comparison of different models on MAMI dataset

Table 5 Execution times (in ms) for VisDrone dataset on Xavier platform

	HSTR-Net	MASA-SR	TTSR
Time (ms)	**216**	1012	8955

performance, have larger network sizes that lead to increased computational complexity and longer execution times, limiting their practical use on devices with limited processing power. AWNet [1] has the largest model size among the approaches tested, posing challenges for real-time execution. On the other hand, HSTR-Net strikes an optimal balance between network complexity and size, providing a 2.77x increase in network size compared to MASA-SR [19], but a 2.32x improvement in execution time. This makes HSTR-Net particularly suitable for real-time applications, such as UAV surveillance, where both performance and computational efficiency are essential. The efficiency of HSTR-Net is more observable when data resolution is increased. State-of-the-art RefSR methodologies cannot cope with the high resolution (672×380) of the VisDrone dataset and fail to provide acceptable inference speeds for real-time applications. On the other hand, HSTR-net model complexity is able to linearly scale with the increase in resolution, which promises satisfactory frame rates for embedded deployment on a drone. HSTR-Net model achieves a $\%54$ speedup over second fastest RefSR model (i.e. AWNet) on the VisDrone dataset. We also deployed our model and two RefSR networks on Nvidia Jetson AGX Xavier Developer Kit [32] to showcase HSTR-Net's ability of low-overhead execution on an embedded platform. As presented in Table 5, HSTR-Net outperforms other tested methods, outputting 4.62 fps on average, while MASA-SR and TTSR cannot even process a frame per second. This result indicates that HSTR-Net can be deployed in a drone setup involving dual cameras and can still achieve sufficient frame rates for detection, tracking or other surveillance tasks in wide-area monitoring.

Figure 9 showcases the ground truth, the corresponding *LR* frame, and the HSTR-Net prediction from a real-world dataset captured by drone cameras. The *HR* video has been temporally downsampled to simulate a HSLF camera scenario. The scene captures a variety of vehicles, such as minivans and cars, moving along a highway. Given the diversity of vehicle types and the frequent overlap in their appearances, object detection in the *LR* frame (Fig. 9b) proves to be particularly challenging due to motion blur and the lack of fine-grained details. Despite these difficulties, HSTR-Net successfully generates a frame where the vehicles and background are more distinct and detailed. This improvement is particularly evident in the enhanced textures of the vehicles, where HSTR-Net applies its patch-matching-based texture transfer technique to reconstruct fine details that are often lost in LR frames. By leveraging both motion estimation and attention mechanisms, HSTR-Net effectively restores spatial details, making objects more recognizable and improving overall visual clarity. This demonstrates the model's ability to overcome the challenges posed by occluding and fast-moving objects, highlighting its potential for real-world applications.

4.4 Ablation study

In this section, several versions of HSTR-Net are compared to demonstrate the improvement in terms of quantitative metrics with proposed modifications. Table 6 presents



(a) Ground Truth (b) Corresponding *LR* frame (c) HSTR-Net

Fig. 9 The real-time operating performance of HSTR-Net on a surveillance drone with two cameras. Ground truth *HR* image is provided for visual comparison purposes. HSTR-Net output displays rich high-frequency details in vehicles and the background, when compared to the corresponding *LR* frame

Table 6 Comparison of different HSTR-Net models on Vimeo septuplet and VisDrone datasets

Method	Vimeo	VisDrone
	PSNR–SSIM	PSNR–SSIM
HSTRNet-*b*	36.86–0.978	32.02–0.957
HSTRNet-*i*	38.56–0.982	33.42–0.965
HSTRNet-*d*	39.34–0.984	33.93–0.968
HSTR-Net	39.94–0.986	34.13–0.969

results for four different HSTR-Net models. Each model is separately trained on both Vimeo and VisDrone training sets. HSTRNet-*b* is the baseline model performing motion estimation/compensation using IFNet from RIFE [56], followed by contextual representation extraction and fusion processes. HSTRNet-*i* is the model in which the motion estimation/compensation module is modified for RefSR purposes. Since single RefSR requires only the optical flow $OF_{LR\rightarrow REF}$, the original architecture is modified to produce a single warped reference frame, which in turn provides a simplified module with lower channel sizes and also with more accurate motion estimation. The model HSTRNet-*d* incorporates deformable convolution, which uses a memory-efficient built-in PyTorch module, rather than non-trainable warping for the contextual feature extraction process. HSTR-Net is the original model presented in the paper that includes the unique patch-matching scheme as well.

As observed in Table 6, HSTRNet-*b* model with general purpose optical flow estimation is limited by its ability to accurately match similar spatial textures between *REF* and *LR* frames, due to solely relying on motion compensation. Therefore, designing a task-specific motion estimation and compensation module alongside a powerful yet efficient deformable convolution stage enhances the performance of the architecture. Additionally, video frames often include similar spatial content between neighboring frames that can be exploited for better transfer of high-frequency features from *REF* to *LR* frame. Our Swin-based attention and patch-matching module proposes a new way of local patch-matching between the two frames in a memory-efficient manner, increasing the PSNR and SSIM metrics on both Vimeo septuplet and VisDrone datasets.

Figure 10 illustrates the complementary roles of each module in HSTR-Net. Figure 10a presents a cropped ground truth from a basketball game, depicting a ball descending through the net. Figure 10b shows frame HR_t^w, which is the result of applying optical flow-based warping to the *REF* frame HR_{t+1}. While optical flow is capable of capturing local motion at a sub-pixel level, it struggles with fast motion and cannot synthesize the ball at its correct position. Figure 10c presents the output of HSTR-Net incorporating deformable convolution and patch-matching modules. Here, deformable convolution enhances optical flow in capturing complex motion but still remains effective only for small, local displacements. On the other hand, patch matching establishes long-range relationships between *REF* and *LR* frames to address fast motion, but is limited to pixel-level accuracy due to its low-complexity implementation. By integrating optical flow, deformable convolution and patch matching, HSTR-Net becomes capable of handling complex motion in diverse scenes, including both fast and slow motion with sub-pixel accuracy. Figure 10c shows the precise ball positioning in the synthesized output image of HSTR-Net.

HSTR-Net model is also applicable for increasing the frame rate by a factor more than 2. For 4x frame rate increase, HSTR-Net can be used to generate multiple *HR* frames, given LR_τ, $\tau = t-1, t, t+1$ and REF_τ, $\tau = t-2, t+2$. Initially the model predicts HR_{t-1} and HR_{t+1}. Then, HR_t is predicted by using $(\hat{HR}_{t-1}, LR_t)$ and $(\hat{HR}_{t+1}, LR_t)$ frame pairs as input to HSTR-Net, for incorporating high-frequency features from both directions and increasing the accuracy. Lastly, the average of the two predictions is taken to synthesize the final $\hat{HR}_t$ prediction. The process is shown in Eq. 4:



(a) Ground truth (b) OF-based warped image (c) HSTR-Net output

Fig. 10 Visual comparison of OF-based and HSTR-Net based synthesized output images. Due to the complementary use of optical flow, deformable convolution and patch matching, HSTR-Net is capable of capturing fast and complex motion in the scene

Table 7 Quantitative results of intermediate HSTR-Net on Vimeo and VisDrone datasets, where 2nd, 3rd and 4th represents the respective frames in a five-frame group in which frames are ordered from 1 to 5

Frame	Vimeo	VisDrone
	PSNR–SSIM	PSNR–SSIM
2nd	39.95–0.986	34.3–0.971
3rd	39.23–0.984	33.80–0.968
4th	39.94–0.986	34.32–0.971
Avg	39.71–0.985	34.16–0.970

Table 8 Comparison of Different HSTRNet models

Method	1 HR + 1 LR	2 HR + 3 LR
	PSNR–SSIM–Time	PSNR–SSIM–Time
Vimeo Triplet	38.45–0.983–29.56	40.62–0.988–64.11
Vimeo Septuplet	39.94–0.986–29.56	41.42–0.988–64.11
VisDrone	33.30–0.965–53.35	35.06–0.973–81.79

$$\begin{aligned} \hat{HR}_{t-1} &= \mathcal{M}(LR_{t-1}, REF_{t-2}) \\ \hat{HR}_{t+1} &= \mathcal{M}(LR_{t+1}, REF_{t+2}) \\ \hat{HR}_t^1 &= \mathcal{M}(LR_t, \hat{HR}_{t-1}) \\ \hat{HR}_t^2 &= \mathcal{M}(LR_t, \hat{HR}_{t+1}) \\ \hat{HR}_t &= 0.5\hat{HR}_t^1 + 0.5\hat{HR}_t^2 \end{aligned} \tag{4}$$

where $\mathcal{M}$ denotes HSTR-Net. For the evaluation of 4x frame rate upsampling, the second and fifth *HR* frames of the septuplet dataset are used for generating the third and fifth *HR* frames. Then, the fourth frame is generated using these two intermediate reconstructed frames. Table 7 shows that competitive and highly accurate results are achieved even though intermediate *REF* frames are synthesized by the model. On the VisDrone dataset, a consecutive group of five frames are used to simulate 4x upsampling. Results in Table 7 show that the middle frame of the five-frame tuple yields average PSNR/SSIM results that are close to the other two generated frames, despite using lower quality *REF* frames at the input.

Table 8 makes a comparison with our previous work called HSTR-Net-interp [29]. HSTR-Net-interp utilizes HR_τ, $(\tau = t-1, t+1)$ and LR_τ, $(\tau = t-1, t, t+1)$ frames to generate HR_t frame. This model has a similar architecture as the original HSTR-Net involving motion compensation, contextual representation extraction, and fusion process performed sequentially, but without any patch matching module. The model synthesizes warped reference frames, $HR_\tau^W(\tau = t-1, t+1)$ and $LR_\tau^W(\tau = t-1, t+1)$, using both *HR* and *LR* reference images during motion compensation process. Then, contextual representations, $(HR_\tau^C, LR_\tau^C, \tau = t-1, t+1)$ are calculated by gradually decreasing the input resolution for extracting valuable information and textures to aid the latter fusion process. Lastly, the fusion process takes $(HR_\tau^W, LR_\tau^W, HR_\tau^C, LR_\tau^C, LR_t)$ tuple as input, generating a mask and high-frequency residual channel at the output. Considering all these, HSTR-Net-interp is bulky due to the use of 5 input frames, even though it has a simpler architecture when compared to HSTR-Net. As seen from Table 8, HSTR-Net-interp offers better results in terms of both PSNR and SSIM metrics. However, HSTR-Net is a better generalized model that can be applied at different frame rates and for different content with complex motion. HSTR-Net-interp requires fine-tuning when the motion complexity and the frame rate of the tested scenario changes, thereby providing a less generalized solution. Also, HSTR-Net provides highly scalable performance in terms of time complexity and memory usage.

5 Conclusion

This paper addresses the problem of generating high spatio-temporal resolution (HSTR) video from multiple lower-resolution cameras using computational imaging techniques. The proposed solution involves a dual camera system that simultaneously captures high spatial resolution low frame rate (HSLF) video and low spatial resolution high frame rate (LSHF) video for the same scene. A novel deep learning architecture is introduced to fuse the two video feeds and generate high spatio-temporal resolution (HSTR) video frames using reference-based super-resolution (RefSR). The proposed model incorporates optical flow estimation, channel-wise attention, and spatial attention mechanisms to capture complex dependencies and fine motion between frames. Through experiments, it is demonstrated that the proposed model outperforms existing reference-based SR techniques in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics. Moreover, the model is suitable for deployment on drones equipped with dual cameras, making it a promising solution for aerial surveillance applications.

Acknowledgements This work is supported in part by TUBITAK Grant - Project No: 118E891

Author contributions H.U.S. wrote the code, performed the simulations, wrote the main manuscript text. A.E.D. performed some of the simulations and tests, wrote part of the manuscript and prepared Fig. 8. H.F.A. developed the research idea and solution, wrote parts of the manuscript and corrected the English of the manuscript. B.K.G. developed the research concept, proposed solutions to the problem. All authors reviewed the manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Cheng, M., Ma, Z., Asif, M.S., Xu, Y., Liu, H., Bao, W., Sun, J.: A dual camera system for high spatiotemporal resolution video acquisition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3275–3291 (2020)
2. Zhou, M., Zhao, X., Luo, F., Luo, J., Pu, H., Xiang, T.: Robust rgb-t tracking via adaptive modality weight correlation filters and cross-modality learning. ACM Trans. Multimed. Comput. Commun. Appl. **20**(4), 1–20 (2023)
3. Zhou, M., Wu, X., Wei, X., Xiang, T., Fang, B., Kwong, S.: Low-light enhancement method based on a retinex model for structure preservation. IEEE Trans. Multimedia **26**, 650–662 (2023)
4. Zhou, M., Lan, X., Wei, X., Liao, X., Mao, Q., Li, Y., Wu, C., Xiang, T., Fang, B.: An end-to-end blind image quality assessment method using a recurrent network and self-attention. IEEE Trans. Broadcast. **69**(2), 369–377 (2022)
5. Zhou, M., Chen, L., Wei, X., Liao, X., Mao, Q., Wang, H., Pu, H., Luo, J., Xiang, T., Fang, B.: Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment. IEEE Trans. Broadcast. **69**(2), 396–405 (2023)
6. Zhou, M., Shen, W., Wei, X., Luo, J., Jia, F., Zhuang, X., Jia, W.: Blind image quality assessment Exploring content fidelity perceptibility via quality adversarial learning. Int. J. Comput. Vision (2025). https://doi.org/10.1007/s11263-024-02338-7
7. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vision (IJCV) **127**(8), 1106–1125 (2019)
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
9. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
10. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
11. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
12. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. Advances in neural information processing systems **28** (2015)
13. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, (2018)
14. Zhang, K., Gao, X., Tao, D., Li, X.: Single image super-resolution with non-local means and steering kernel regression. IEEE Trans. Image Process. **21**(11), 4544–4556 (2012). https://doi.org/10.1109/TIP.2012.2208977
15. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Comput. Graphics Appl. **22**(2), 56–65 (2002)
16. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7982–7991 (2019)
17. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8425–8434 (2020)
18. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5791–5800 (2020)
19. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6368–6377 (2021)
20. Lin, R., Xiao, N.: Residual channel attention connection network for reference-based image super-resolution. In: 2021 8th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS). IEEE, pp. 307–313 (2021)
21. Huang, Y., Zhang, X., Fu, Y., Chen, S., Zhang, Y., Wang, Y.-F., He, D.: Task decoupled framework for reference-based super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5931–5940 (2022)
22. Liu, X., Li, J., Duan, T., Li, J., Wang, Y.: Dsma: Reference-based image super-resolution method based on dual-view supervised learning and multi-attention mechanism. IEEE Access **10**, 54649–54659 (2022)
23. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 88–104 (2018)
24. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2103–2112 (2021)
25. Wang, T., Xie, J., Sun, W., Yan, Q., Chen, Q.: Dual-camera super-resolution with aligned attention modules. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2001–2010 (2021)
26. Xia, B., Tian, Y., Hang, Y., Yang, W., Liao, Q., Zhou, J.: Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. arXiv preprint arXiv:2201.04358 (2022)
27. Lee, J., Lee, M., Cho, S., Lee, S.: Reference-based video super-resolution using multi-camera video triplets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17824–17833 (2022)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

29. Suluhan, H.U., Ates, H.F., Gunturk, B.K.: Dual camera based high spatio-temporal resolution video generation for wide area surveillance. In: 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 1–8 (2022)
30. Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q., Ling, H.: Vision meets drones: past, present and future. **1**(2), 8, arXiv preprint arXiv:2001.06303 (2020)
31. Minor Area Motion Imagery (2013). https://www.sdms.afrl.af.mil/index.php?collection=mami2013
32. Jetson AGX Xavier Evaluation Board. Accessed: 2023-07-03 (2023). www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-xavier/
33. Guo, K., Chen, L., Zhu, X., Kui, X., Zhang, J., Shi, H.: Double-layer search and adaptive pooling fusion for reference-based image super-resolution. ACM Trans. Multimedia Comput. Commun. Appl. (2023). https://doi.org/10.1145/3604937
34. Mei, X., Yang, Y., Li, M., Huang, C., Zhang, K., Lió, P.: A feature reuse framework with texture-adaptive aggregation for reference-based super-resolution. arXiv preprint arXiv:2306.01500 (2023)
35. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8425–8434 (2020)
36. Zou, H., Xu, L., Okatani, T.: Geometry enhanced reference-based image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6123–6132 (2023)
37. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer. In: European Conference on Computer Vision. Springer, pp. 325–342 (2022)
38. Aslahishahri, M., Ubbens, J., Stavness, I.: HiTSR: A hierarchical transformer for reference-based super-resolution. arXiv preprint arXiv:2408.16959 (2024)
39. Dong, R., Yuan, S., Luo, B., Chen, M., Zhang, J., Zhang, L., Li, W., Zheng, J., Fu, H.: Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27684–27694 (2024)
40. Min, J., Lee, Y., Kim, D., Yoo, J.: Bridging the domain gap: A simple domain matching method for reference-based image super-resolution in remote sensing. IEEE Geosci. Remote Sens. Lett. **21**, 1–5 (2023)
41. Zheng, H., Heng, H., Yan, Z., Zeng, K., Fang, J., Qiang, B.: A generic multi-correspondence matching framework for reference-based image super-resolution. IEEE Trans. Instr. Measur. (2024). https://doi.org/10.1109/TIM.2024.3428600
42. Tu, Z., Yang, X., He, X., Yan, J., Xu, T.: RGTGAN: Reference-based gradient-assisted texture-enhancement GAN for remote sensing super-resolution. IEEE Trans. Geosci. Remote Sens. **62**, 1–21 (2024)
43. Lei, P., Fang, F., Zhang, G.: Flow guidance deformable compensation network for video frame interpolation. arXiv preprint arXiv:2211.12117 (2022)
44. Li, C., Wu, G., Sun, Y., Tao, X., Tang, C.-K., Tai, Y.-W.: H-vfi: Hierarchical frame interpolation for videos with large motions. arXiv preprint arXiv:2211.11309 (2022)
45. Jin, X., Wu, L., Chen, J., Chen, Y., Koo, J., Hahm, C.-h.: A unified pyramid recurrent network for video frame interpolation. arXiv preprint arXiv:2211.03456 (2022)
46. Lin, X., Zhao, L., Liu, X., Chen, J.: Mvfi-net: Motion-aware video frame interpolation network. In: Proceedings of the Asian Conference on Computer Vision, pp. 3690–3706 (2022)
47. Li, Y., Zhu, Y., Li, R., Wang, X., Luo, Y., Shan, Y.: Hybrid warping fusion for video frame interpolation. Int. J. Comput. Vision **130**(12), 2980–2993 (2022)
48. Hu, M., Jiang, K., Liao, L., Nie, Z., Xiao, J., Wang, Z.: Progressive spatial-temporal collaborative network for video frame interpolation. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2145–2153 (2022)
49. Plack, M., Briedis, K.M., Djelouah, A., Hullin, M.B., Gross, M., Schroers, C.: Frame interpolation transformer and uncertainty guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9811–9821 (2023)
50. Zhou, C., Liu, J., Tang, J., Wu, G.: Video frame interpolation with densely queried bilateral correlation. arXiv preprint arXiv:2304.13596 (2023)
51. Li, Z., Zhu, Z.-L., Han, L.-H., Hou, Q., Guo, C.-L., Cheng, M.-M.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9801–9810 (2023)
52. Liu, C., Yang, H., Fu, J., Qian, X.: Ttvfi: Learning trajectory-aware transformer for video frame interpolation. arXiv preprint arXiv:2207.09048 (2022)
53. Zhang, Y., Wang, H., Zhu, H., Chen, Z.: Optical flow reusing for high-efficiency space-time video super resolution. IEEE Trans. Circ. Syst. Video Technol. **33**(5), 2116–2128 (2022)
54. Luo, Y., Pan, J., Tang, J.: Bi-directional pseudo-three-dimensional network for video frame interpolation. IEEE Trans. Image Process. **31**, 6773–6788 (2022)
55. Li, H.-D., Yin, H., Liu, Z.-H., Huang, H.: Enhanced spatial-temporal freedom for video frame interpolation. Appl. Intell. **53**(9), 10535–10547 (2022)
56. Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Real-time intermediate flow estimation for video frame interpolation. In: European Conference on Computer Vision, pp. 624–642 (2022). Springer
57. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 234–241 (2015)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

H. Umut Suluhan is a PhD student in the Department of Electrical & Computer Engineering at The University of Arizona. He received his BS degree in Computer Science from Ozyegin University in 2023. His research interests are broadly in high-performance computing and reconfigurable architectures with a focus on runtime software design and resource management algorithm development for heterogeneous computing systems.

Abdullah Enes Doruk received his Bachelor's degree in Electrical and Electronics Engineering from Bursa Technical University and is currently pursuing a Master's degree in Artificial Intelligence Engineering at Özye?in University, where he is also a researcher at the Deep-VIP Lab. His work focuses on computer vision and deep learning, particularly object detection, semantic segmentation, and domain adaptation using real and synthetic data. He has industry experience in developing AI systems for autonomous vehicles and robotics, and has authored multiple publications in visual perception and adaptation.

Hasan F. Ates received his Ph.D. degree from the Department of Electrical Engineering, Princeton University in 2004. He was a Research Associate at Sabanci University, between 2004 and 2005. He held positions of Assistant, Associate and Full Professorship at Isik University between 2005 and 2018. He was at Istanbul Medipol University between 2018 and 2022. He is currently a Professor in the Department of Computer Science, Ozyegin University, which he joined in September 2022. He is the author/co-author of more than 50 peer-reviewed publications in the areas of image/video processing/coding and computer vision.

Bahadir K. Gunturk received the B.S. degree in Electrical Engineering from Bilkent University, Turkey, in 1999, and the Ph.D. degree in Electrical Engineering from the Georgia Institute of Technology, in 2003. From 2003 to 2014, he was with the Department of Electrical and Computer Engineering, Louisiana State University. Since 2014, he has been with Istanbul Medipol University, where he is currently a Professor. He has published more than 50 peer-reviewed journals/conference papers in the areas of image processing and computer vision.