Deep 3D World Models for Multi-Image Super-Resolution Beyond Optical Flow

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 November 2024, accepted 5 December 2024, date of publication 9 December 2024, date of current version 20 December 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3514188*

RESEARCH ARTICLE

# Deep 3D World Models for Multi-Image Super-Resolution Beyond Optical Flow

**LUCA SAVANT AIRA[1], (Member, IEEE), DIEGO VALSESIA[1], (Member, IEEE), ANDREA BORDONE MOLINI[2], GIULIA FRACASTORO[1], (Member, IEEE), ENRICO MAGLI[1], (Fellow, IEEE), AND ANDREA MIRABILE[2]**

[1]Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
[2]Zebra Technologies, EC2N 4AW London, U.K.

Corresponding author: Luca Savant Aira (luca.savant@polito.it)

This work was supported in part by the Project PNRR-NGEU under Grant MUR-DM 352/2022, and in part by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (Program "RESTART") under Grant PE00000001.

**ABSTRACT** Multi-image super-resolution (MISR) allows to increase the spatial resolution of a low-resolution (LR) acquisition by combining multiple images carrying complementary information in the form of sub-pixel offsets in the scene sampling, and can be significantly more effective than its single-image counterpart. Its main difficulty lies in accurately registering and fusing the multi-image information. Currently studied settings, such as burst photography, typically involve assumptions of small geometric disparity between the LR images and rely on optical flow for image registration. We propose EpiMISR, a novel MISR method that can increase the resolution of sets of images acquired with arbitrary, and potentially wildly different, camera positions and orientations, generalizing the currently studied MISR settings. Our proposed model moves away from optical flow and explicitly uses the epipolar geometry of the acquisition process, together with transformer-based processing of radiance feature fields to substantially improve over state-of-the-art MISR methods in presence of large disparities in the LR images.

**INDEX TERMS** Deep learning, super-resolution (SR).

## I. INTRODUCTION

Image super-resolution (SR) is the task of recovering a high-resolution (HR) version of an image from degraded low-resolution (LR) observations. It is a longstanding inverse problem in the imaging field and has numerous practical applications due to camera limitations and image acquisition conditions. Most of the literature focuses on estimating the HR image from a single input image (SISR). While recent deep learning approaches [1], [2], [3] have tremendously advanced the state of the art, SISR remains highly ill-posed due to the limited high-frequency information available in a single image. Multi-image SR methods (MISR), on the other hand, are presented with multiple samplings of a given scene, carrying complementary information at a sub-pixel level. MISR techniques seek to accurately fuse the multiple LR images to obtain SR images with significantly higher quality than what is achievable by SISR methods. Only recently the deep learning literature has started exploring the multi-image setting due to increased difficulty in creating benchmark datasets as well as developing effective methods that can handle accurate image registration.

MISR can be seen as a generalization of the classic Stereo-SR setting [4], in which a pair of images is captured, often with a tightly controlled geometry to simplify the fusion process. At the moment, the most studied MISR settings are in the context of video [5] where successive frames provide the multiple images, remote sensing images [6] where satellite revisits of the same scene are exploited, and burst photography, where a set of photos is acquired in rapid succession such in [7], [8], or [9]. All these settings present a common denominator in that variations in the acquisition

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

geometry among the multiple images are relatively small, resulting in relatively small disparities in the image pixels. For example, in burst SR, geometric variations are mostly due to natural hand shaking. This is desirable because the SR process requires subpixel shifts in the sampling grid, and obtaining them with minimal overall movement only simplifies the fusion process. For this reason, works in this field resort on using forms of optical flow estimation between LR images to accurately register them. Optical flow estimates a translation vector for each pixel of an image in order to warp it to a target image. Such a transformation between flat camera planes may struggle in presence of complex 3D transformations.

It is thus clear that the aforementioned small-parallax settings that have been currently studied are restrictive and do not allow to account for many interesting scenarios for super-resolution where the LR images come from cameras with wildly different positions and orientations. As examples, one can think of sets of security cameras which image a scene from significantly different vantage points, or sets of images of a scene collected in the wild with no control over the acquisition process.

In this paper, we present EpiMISR, a new method designed for the general MISR setting, where a set of LR images are acquired by cameras with arbitrary positions and orientations, and our task is to super-resolve one (or more) of them. We move away from the optical flow based models, in favour of an explicit use of epipolar geometry with techniques inspired by recent works in the NeRF literature [10]. However, contrary to the NeRF literature, we are not concerned with novel view synthesis, but rather follow the standard SR approach of restoring one of the observed LR images. Our proposed method, called EpiMISR, leverages strong spatial priors necessary for the SR task and transfomer-based processing of radiance feature fields to achieve effective fusion of images with large discrepancies in acquisition geometries. We show that EpiMISR substantially improves over the state-of-the-art SR techniques developed for the more restrictive scenarios.

## II. RELATED WORK

### A. SINGLE-IMAGE SUPER-RESOLUTION

Single image super-resolution (SISR) is a long-standing problem in the field of computer vision, aiming at recovering a high-resolution (HR) image $I^{\mathrm{HR}}$ given its degraded version $I^{\mathrm{LR}}$. In its simplest form, the forward model of the problem is:

$$I^{\mathrm{LR}} = (K * I^{\mathrm{HR}}) \downarrow_s \tag{1}$$

where $\downarrow_s$ denotes decimation by a factor $s$ and $*$ denotes a convolution with degradation kernel $K$.

Note that this problem is ill-posed as the degradation process is non-injective. To overcome this challenge, two main families of approaches have been proposed: regularization methods and data-driven methods. Regularizers such as total variation impose handcrafted a-priori knowledge to establish a criterion in order to choose a plausible SR image, as done by [11] and [12]. Data-driven approaches, instead, extract this knowledge directly from data. Modern deep learning approaches to SISR [13], [14], [15] descend from the pioneer works of [16] and [17]. A recent state-of-the-art neural network design is SwinIR [1] which leverages a windows-attention-based architecture. It is also worth mentioning that some works [18] tackle the blind SISR problem, i.e., when the degradation process is not known and hence should be estimated. Finally, a branch of the literature is concerned with lightweight architectures, such as the one by [2].

### B. MULTI-IMAGE SUPER-RESOLUTION

The ill-posedness of SISR is intuitively reduced if extra images of the same scene are available. This MISR approach can be further specialized in the multiframe-SR if these extra images comes from adjacent frames of a video, burst-SR if they comes from a photo-burst, stereo-SR if the single extra image is the stereo companion of the target one.

Multiframe-SR and burst-SR assume small geometric disparity as there are small camera movements between successive acquisitions. Exploiting this fact, the first step in algorithms for these settings is typically to register the images to each other using optical flow models [19]. Recent works in the context of the burst-SR challenge by [20], such as [7], [8], and [9] follow this approach, relying on neural networks modules estimating optical flow. However, optical flow models geometric relations as locally translational on the camera plane, and, as such, is limited in its expressive power. This is fine when the geometric disparity is small, but a general setting may benefit for a more accurate account of the 3D geometry.

Similarly, lightfield SR [21] employs a familiar grid-like arrangement of multiple cameras with minimal disparities. Consequently, it facilitates simpler image fusion techniques and does not impose as stringent robustness requirements as a setting with large disparities. For instance, our scenario necessitates addressing potential occlusions and non-Lambertian surfaces. Unlike light field SR, which can comfortably rely on Lambertian approximations due to its small disparities, this approach does not exhibit clear generalizability to the large-disparity setting explored in our study.

The stereo-SR setting, instead, assumes only the presence of two cameras (i.e., just one extra image) and the acquisition setting is typically controlled so that camera poses only differ by an horizontal shift. Recently, [4], [22], [23] developed methods for stereo-SR that utilizes an attention mechanism to perform image alignment implicitly. Finally, a branch of the literature is concerned with SISR or stereo-SR of omnidirectional images [24], [25].

To the best of our knowledge, this is the first work tackling the problem of generic multi-image super-resolution, i.e., there are no assumptions about the number of images or the relative poses of the cameras. Hence, we move away from

2D image alignment processes and leverage a full deep 3D world model.

### C. NeRF AND IMAGE FUSION

NeRF architectures are neural world models, as they encode information from posed images in the weights of a neural network in a 3D-geometrically consistent way. In their original formulation by [10], a multilayer perceptron encodes the 5D radiance field of a given scene. An alternative 3D scene representation, based on Gaussian-shaped primitives, is proposed in [26]. Further NeRF evolutions, such as [27] and [28] aim to avoid per-scene training, learning general priors by introducing a feature extractor and exploiting constraints from epipolar geometry in an explicit way. References [29], [30], [31], and [32] improve the ray casting procedure with cone casting and more advanced space sampling mechanisms. Some works move away from the physically-grounded volumetric rendering integral by replacing it with transformers acting on a feature space, and address the novel view synthesis task using both per-scene training [33], [34] or using an inductive approach [35]. Reference [36] uses a similar architecture to perform 3D human joints localization and [37] to perform point cloud reconstruction. Also other works, such as [38], are concerned with multi-image fusion leveraging transformers in their pipelines. However, they differ from our work in that they do not deal with a super-resolution problem and are often limited by processing images in pairs and then aggregating the results with non-parametric processes. Recently, NeRF-like models have also been used to address inverse problems in imaging, of which super-resolution is an example. References [39] and [40] address the case where the input views are noisy, discovering outstanding denoising performance. References [41] and [42], instead, tackle the problem of superresolving the NeRF 3D geometry model, hence being capable of generating novel-views at a higher resolution. Our work differs significantly from them in that we are concerned with super-resolution of existing views only and we do not optimize on a per-scene basis, but rather leverage a training set to train an image fusion model that can be then used for an arbitrary scene with an arbitrary number of views with an arbitrary geometry.

## III. PROPOSED METHOD

We address the setting in which a number of images of a given scene are acquired from arbitrary vantage points, possibly with large geometric disparity. These images have low resolution and we seek to super-resolve one of them by suitably combining the complementary information carried by the other images. Our proposed method, called **EpiMISR**, is a MISR neural network which explicitly accounts for the epipolar geometry by exploiting camera poses and processing 3D feature fields in a NeRF-like manner. Given $V + 1$ LR views of a static scene, and the corresponding intrinsic and extrinsic camera parameters, our task is to obtain a HR version of one of them, which we will call the *target view*, by also leveraging information from the $V$ extra views. In the parlance of NeRF models, this is referred to as *not-novel* view synthesis.

EpiMISR is not optimized on a per-scene basis, but rather uses a training set to learn the function needed to perform image fusion with an arbitrary geometry for the SR task in a supervised way. As shown in the high-level overview in Fig. 1, EpiMISR consists of three main modules, named SISR-FE, CAP and MIFF, which create SR features, sample them along epipolar lines and fuse them, and will be detailed in the following sections. Notice that EpiMISR also computes a super-resolved image from only the target view, called $I^{\mathrm{SISR}}$. We found that a loss function optimizing the fidelity of both the SISR and MISR outputs with respect to the HR ground truth, such as

$$L = \mathcal{L}\left(I^{\mathrm{MISR}}, I^{\mathrm{HR}}\right) + \alpha\mathcal{L}\left(I^{\mathrm{SISR}}, I^{\mathrm{HR}}\right) \tag{2}$$

provided more stable performance over a variable range of available views and ensured that the degenerate case of a single view ($V = 0$) recovers the performance of the SISR backbone. In our experiments, we used the L1 loss as $\mathcal{L}$.

### A. SISR-FE MODULE

The SISR-FE (Single Image Super-Resolution Feature Extractor) module is shared across views and its purpose is to capture strong spatial priors (local correlation and, possibly, non-local self-similarity) to extract features supported on a super-resolved image grid. Each pixel in this super-resolved grid is geometrically positioned on the camera plane associated to each particular view, but its feature vector captures the information of a neighborhood. The increased resolution with respect to the original allows finer processing by the other modules. Being part of a modular approach, SISR-FE can leverage any state-of-the-art SISR architecture by truncating the final projection to RGB space. More formally, let $I_v^{\mathrm{LR}}$ be the $v$-th view as input of the module, its output will be a set of $C$ feature maps at $s$ times the resolution:

$$\text{SISR-FE} : I_v^{\mathrm{LR}} \in \mathbb{R}^{H,W,3} \rightarrow F_v \in \mathbb{R}^{sH,sW,C} \tag{3}$$

where $v = 0$ denotes the target view. We also remark that a SISR image prediction $I^{\mathrm{SISR}}$ is obtained from $F_0$ via projection of features to RGB values, and it is used as a basis for the multi-image residual correction estimated by the other modules.

### B. CAP MODULE

In order to handle potentially large geometric disparities in camera poses, epipolar geometry is employed instead of the optical flow modules commonly used in the burst SR literature. A deterministic, non-learnable module called CAP (CastAndProject) is used to implement epipolar geometry with an approximate pinhole camera model. Given a pixel on the SR target view grid, there exists an associated

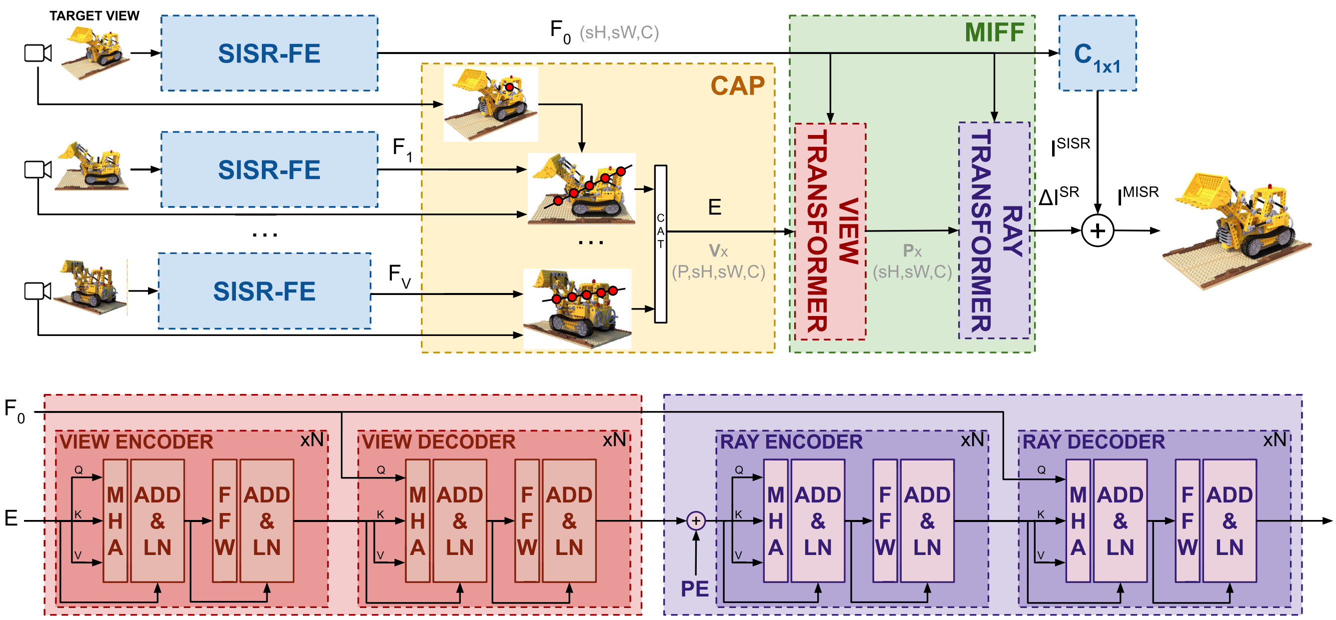



**FIGURE 1.** EpiMISR Architecture. From the LR target view and the extra views super-resolved features are obtained by any single-image SR network (SISR-FE), sampled along epipolar lines associated to pixels in the target view (CAP) and fused (MIFF) to produce a residual correction to single-image SR.

straight line, called the epipolar line, for each of the extra views, such that the line will intersect with the object imaged by the target pixel. The CAP module is shared across the extra views, and receives as input the camera parameters of the target view $\mathcal{P}_0$, the camera parameters of the $v$-th view $\mathcal{P}_v$ and the super-resolved feature map of the $v$-th view $F_v$ to compute the epipolar features $E_v$.

$$\mathrm{CAP}_{\mathcal{P}_0} : (F_v, \mathcal{P}_v) \rightarrow E_v \in \mathbb{R}^{P,sH,sW,C} \tag{4}$$

The epipolar features tensor $E_v$ denotes the epipolar lines for view $v$ sampled at $P$ locations, for each pixel and feature in the target view.

The purpose of this module is to build the tensor $E_v$ so that the following MIFF module can efficiently scan the epipolar line in search of features in the extra views that match the feature in the target view at each target pixel position, thus effectively exploiting inter-view information. For each pixel in the target view, CAP casts a ray in the 3D space passing through the center of the target camera and the selected pixel (using $\mathcal{P}_0$). Along this ray, $P$ points are sampled. For each sampled point, the module computes the projection point onto the image plane of the extra view (using $\mathcal{P}_v$). As the obtained coordinates can be non-integer, the module bicubically resamples the super-resolved feature maps $F_v$ at the correct coordinates. This also highlights the importance of having features $F_v$ on a super-resolved grid to properly account for fine details. The module also generates a boolean mask to flag invalid projected points that are outside the feature map or behind the extra camera. We also note that CAP samples points hyperbolically along the ray, so that the points are equally spaced when projected on the image planes.

### C. MIFF MODULE

The MIFF (Multi Image Feature Fusion) module receives as input the epipolar feature tensors $E_1, \ldots, E_V$ returned by the CAP module, containing features from the extra views, warped and aligned to the target view. Its task is to aggregate them to return a residual correction to the SISR image of the target view that accounts for the information of the other views.

$$\mathrm{MIFF} : (F_0, \{E_1, \ldots, E_V\}) \rightarrow \Delta I^{\mathrm{SR}} \in \mathbb{R}^{sH,sW,3} \tag{5}$$

The final super-resolved version of the target view is then obtained by:

$$I^{\mathrm{MISR}} = I^{\mathrm{SISR}} + \Delta I^{\mathrm{SR}}. \tag{6}$$

Similarly to [33], we drop the classical physics-based volume integral formulation, replacing it with two transformers that aggregate the information from the extra views directly in a feature space. The two transformers work in a cascade fashion, with the first transformer aggregating the views (*view transformer*) and the second transformer aggregating the points along the ray (*ray transformer*). Using the notation from [43], each transformer is formed by an encoder and a decoder module. We refer the reader to Fig. 1 for a detailed block diagram of the following explanation.

The encoder for the view transformer considers the sequence of $V$ epipolar feature tensors $E_v$ as input and derives joint features by means of a stack of several multihead self-attention layers, feed-forward layers and LayerNorm layers [44]. This operation is crucial as it allows for the fusion of independently computed features $E_v$ from each view. By leveraging self-attention layers we enable the network to derive more intricate and integrated joint features.

Also notice that this operation is equivariant to the ordering of the views and does not depend on the specific number of views $V$ available. The output of the view transformer encoder is a sequence of length $V$ of joint features. This is provided as input to the decoder together with the super-resolved features $F_0$ of the target view. The decoder uses multiple cross-attention layers to correlate the features of the target view with those extracted from the other views. Its output summarizes the content of the views in a feature field, equivalent to the radiance field in the physics-based approach of NeRF.

Next, the ray transformer replaces the physics-backed volumetric integral to integrate the feature field over the ray. Again an encoder-decoder structure is used. The encoder performs self-attention over the sequence of $P$ ray points to mix the ray features. Then the decoder uses cross-attention between the super-resolved features $F_0$ of the target view and the output of the encoder to estimate the RGB residual image correction $\Delta I^{SR}$ that is added to the SISR image.

Notice that performing the aggregation along the ray and then along the views is not optimal. However, performing both aggregations together in a single step is too computationally demanding, hence we perform first the aggregation along the views and then along the ray.

## IV. EXPERIMENTAL RESULTS

### A. EXPERIMENTAL SETTING

In this work, we address the MISR task with a supervised learning paradigm. In order to properly characterize the proposed method from an experimental standpoint, we need a setting with multiple images having relatively large disparity compared to the more conventionally studied burst SR setting. Consequently, we use the DTU dataset [45], which is already known in the NeRF literature, for this new SR setting. In particular, we utilize the rectified DTU dataset,[1] comprising 124 different scenes, with 49 posed views per scene, each view having 1600 × 1200 pixels. For reasons of computational efficiency, we first bicubically downsample the original images by a factor of 4 obtaining the 400 × 300 HR images from which degraded LR images are derived. We split the dataset into train, validation and test. Validation set is formed by only scene 47 while the test set is formed by scenes 3, 10, 13, 18, 30, 63, 77, 99, 103. All the other 114 scenes form the train split. From each scene, multiple input sets are extracted by selecting as the target view a random image among the 49 and then choosing the nearest $V$ images as extra views, with respect to camera centers. The number of extra views during training is $V = 7$ and, unless otherwise stated,the same number is also used for testing. The angle between the target view and the other views ranges between 11 and 33 degrees, averaging around 15 degrees, which is in line with our large disparity setting.

In our experiments, the SISR-FE module is based on the SwinIR architecture [1] in order to be comparable with recent methods in the burst SR literature. We also present some ablations with simpler designs for SISR-FE in Sec. IV-E. The number of points sampled by the CAP module along the ray during training is $P = 256$, and, unless otherwise stated, the same number is used during testing. Finally, regarding the MIFF module, we set the number of encoder and decoder layers to 4 for both transformers.

The training pipeline of EpiMISR for the following experiments consists of two steps. First, we pretrain the SISR-FE module and its RGB projection as a SISR neural network on the DIV2K dataset from [46], and finetune it on the DTU dataset. Then the whole EpiMISR architecture is trained end-to-end for the MISR task, using the loss in Eq. 2 with $\alpha = 1$.

We employ the Adam optimizer for the end-to-end optimization of EpiMISR. The SISR-FE module is frozen to the pretrained weights for the first 350 iterations to train the sole MIFF module and stabilize the training, followed by an additional 150 iterations to finetune the whole network. The learning rate is linearly warmed up for the first 60 epochs starting with $10^{-6}$ up to $10^{-4}$. A multi-step scheduler halves it at epochs 150, 250. For the final 150 epochs, the learning rate is set to $10^{-5}$ and further halved at epochs 80, 120. We train on four A100 GPUs for about 7 days.

We compare the proposed technique to a number of state-of-the-art approaches for multi-image super-resolution in the literature. However, we remark that our setting with relatively large parallax and free camera positions is new and different from existing settings in the super-resolution literature. The closest match is the burst SR literature, which however only considers small disparities and does not use camera poses. We consider BSRT [9] as the state-of-the-art for the burst SR literature, and DBSR [7] as additional baseline. The NeRF literature has recently published the NeRF-SR method by [41]. We consider this method as an interesting additional point of reference which follows the NeRF methodologies and explicitly uses camera poses. However, NeRF-SR follows a different settings as it is concerned with novel view synthesis at a higher resolution rather than not-novel view enhancement and it does not follow the supervised learning paradigm. A recent preprint by [42] proposes Super-NeRF, but it has not been tested due to the lack of publicly available code. Besides, its setting is also different because, similarly to NeRF-SR, it does not follow the supervised learning paradigm, it focuses on novel view synthesis and, moreover, it optimizes for perception metrics and not for distortion. All methods in our comparisons have been retrained using the authors' code and following the same pretraining procedure of EpiMISR. The number of epochs for their training has been chosen to maximize their performance on a validation set. A minor modification has been made to the burst methods to use RGB images instead of RAW mosaiced images.

Code for EpiMISR is available at https://github.com/mezzelfo/EpiMISR.

[1] Third light setting, as it is the most uniform.

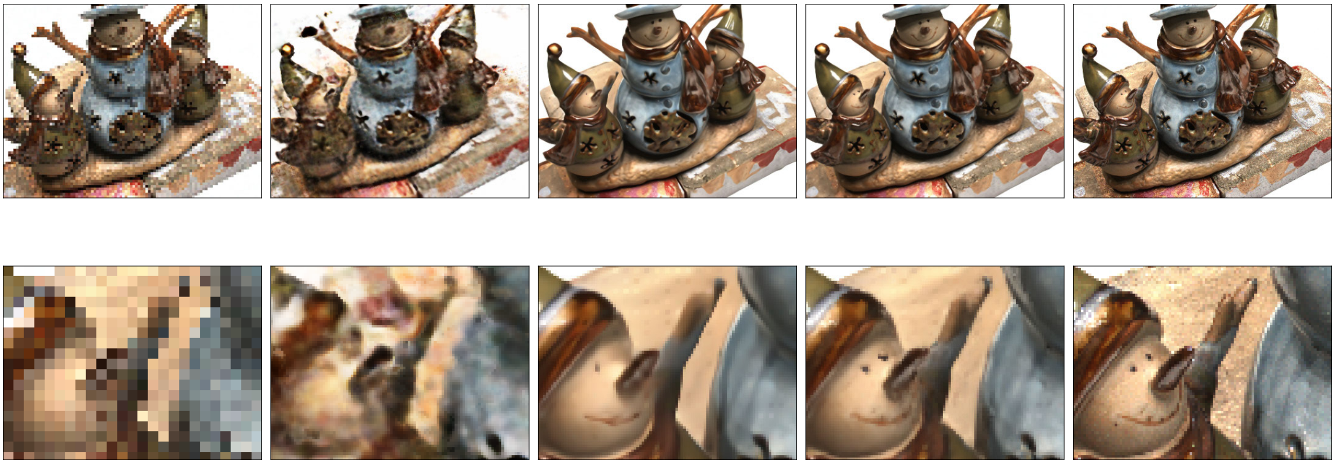


FIGURE 2. DTU scene 3 with 4× scale factor. From left to right: LR nearest neighbours interpolation (19.31 dB), NeRF-SR (19.75 dB), BSRT (23.60 dB), EpiMISR (24.43 dB), HR ground truth.

TABLE 1. Quantitative results for MISR on DTU dataset.

| | | No. Params | PSNR ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|---|
| 4× | **EpiMISR** | 23.30M | **28.60** | **0.11** | **0.87** |
| | BSRT ( [9]) | 20.56M | 27.84 | 0.13 | 0.85 |
| | Burstormer ( [47]) | 3.27M | 26.76 | 0.16 | 0.84 |
| | DBSR ( [7]) | 12.91M | 26.36 | 0.20 | 0.80 |
| | NeRF-SR ( [41]) | 1.19M | 23.17 | 0.32 | 0.64 |
| | SwinIR ( [1]) | 14.70M | 26.87 | 0.17 | 0.82 |
| | 3DGS ( [26]) | ≈ 1.88M | 22.73 | 0.30 | 0.73 |

TABLE 2. Quantitative results for MISR on the Google Scanned Objects and LLFF datasets.

| | GSO | | | LLFF | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | LPIPS ↓ | SSIM ↑ | PSNR ↑ | LPIPS ↓ | SSIM ↑ |
| **EpiMISR** | **31.50** | **0.04** | **0.96** | **23.07** | **0.20** | **0.74** |
| BSRT ( [9]) | 30.09 | 0.05 | 0.95 | 22.85 | 0.24 | 0.72 |
| SwinIR ( [1]) | 29.29 | 0.07 | 0.95 | 22.27 | 0.29 | 0.69 |

### B. MAIN EXPERIMENT

Table 1 reports our main results on the DTU dataset for a 4× SR factor. For quantitative evaluation, we use PSNR as quality metric and LPIPS and SSIM as perceptual metrics.[2] Metrics are computed after cropping 16 pixels on each side to avoid border effects. It can be noticed that some multi-image methods with weak geometric priors struggle to improve over the SISR result of SwinIR. As a sanity check, we tested but not reported in the table the SISR performance of EpiMISR after all the finetuning procedures, and saw that it is just marginally above the reference SwinIR results (26.96 dB), confirming that improvements actually come from the use of multiple images. Instead, the state-of-the-art from the burst SR literature (Burstormer) is not able to handle the high disparities present in the DTU dataset, highlighting the difference between the burst SR task and the generic multi-image SR task. We found that a previous burst SR method (BSRT) performs better than Burstormer but still shows a significantly lower PSNR of about 0.8 dB compared to EpiMISR, highlighting the importance of explicitly modeling the problem geometry at the core of our model rather than relying on optical flow. NeRF-SR does not show competitive performance, which is expected for several reasons: i) it targets the novel view synthesis setting; ii) it is optimized on a per-scene basis, thus not being able to learn powerful image priors from training data; iii) it is a much smaller model. We also compare EpiMISR with the 3DGS [26] method, which, in this task, does not show competitive performance as it shares the same drawbacks of NeRF-SR. Fig. 2 shows a qualitative comparison between the proposed method and the other baselines. It can be noticed that EpiMISR provides more accurate details.

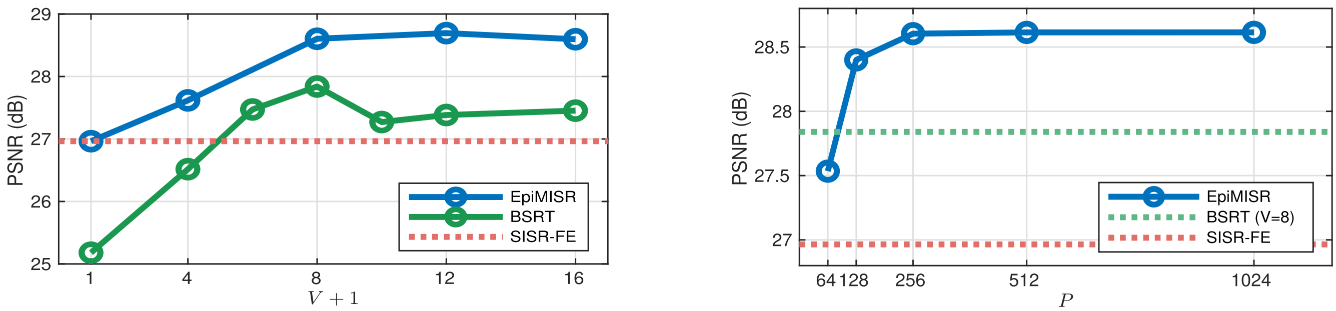
[2]We remark that all methods, except NeRF-SR, optimize for distortion rather than perception, see [48] for distortion vs. perception tradeoff.

### C. EXPERIMENTS ON GSO DATASET AND LLFF DATASET

In this section we report our results on the 1023 scenes from the Google Scanned Objects dataset [28] (512 × 512 pixels) and on the 8 scenes from LLFF dataset [49] (≈ 2100 × 1600) pixels), for a 4× SR factor. Table 2 reports the evaluation results of EpiMISR, BSRT and SwinIR methods on the Google Scanned Objects dataset and on the LLFF dataset. All the methods are trained only on DTU dataset as previously described and are not finetuned on the GSO dataset nor on the LLFF dataset, hence these results shows that EpiMISR outperforms baselines even on an unseen data distribution.

### D. NUMBER OF VIEWS AND NUMBER OF POINTS ALONG THE RAYS

In this section, we study the impact of two important parameters of the proposed method, namely $V$, the number of extra views, and $P$ the number of points along the ray.




(a) Comparison of PSNR for different $V$.

(b) Comparison of PSNR for different $P$.

**FIGURE 3. PSNR with respect to $V$ and $P$.**

**TABLE 3. Comparison of different SISR-FE modules in terms of MISR and SISR performance on the DTU dataset.**

| SISR-FE Module | No. Params | PSNR (SISR) ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|
| **SwinIR** | 14.85M | 28.60 (26.96) | 0.11 | 0.87 |
| RLFN ( [2]) | 0.86M | 28.05 (26.38) | 0.12 | 0.86 |
| Bicubic + conv3×3 | 7.94k | 25.73 (24.13) | 0.23 | 0.79 |
| Bicubic + conv1×1 | 1.80k | 24.56 (24.04) | 0.27 | 0.76 |

**TABLE 4. Challenging geometry setting.**

| | | No. Params | PSNR ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|---|
| | **EpiMISR** | 23.30M | **27.00** | **0.15** | 0.82 |
| 4× | BSRT ( [9]) | 20.56M | 26.82 | 0.16 | **0.83** |
| | SwinIR ( [1]) | 14.70M | 26.87 | 0.17 | 0.82 |

It can be expected that increasing the number of views $V$ allows to integrate extra information and increase the quality of the SR image. However, diminishing returns are expected, especially for extra views with very large disparity. Fig. 3 reports the PSNR of the SR image for different number of views used by the super-resolution process. Images are added by expanding the neighborhood of available views around the target, so they are progressively farther or more angled with respect to the target. We notice that only a marginal improvement is obtained increasing from 8 to 16 views. Regarding views, we also remark that EpiMISR can process an arbitrary number of input views with an arbitrary ordering, as its operations are invariant in that dimension.

The number of ray points $P$ determines the density of the feature field that takes the place of the radiance field in our model. This parameter is strictly tied to the resolution of the images and the scene characteristics, and its sampling should be fine enough to capture the fine details of the scene. Fig. 3b shows that a too small value of $P$ has a significant impact on SR quality, while performance saturates beyond the chosen value of $P = 256$.

### E. SISR-FE ABLATION

The EpiMISR modular design allows to decouple the fusion of multiple images using the 3D geometry from the super-resolved feature extraction, which can leverage advances in SISR methods or be tuned for the desired complexity. In this section, we present some MISR results using different SISR-FE modules in order to study its impact on overall performance. Results are shown in Table 3. Unsurprisingly, the SwinIR architecture used in the main experiment provides the best performance but it is also a relatively large model. However, it is interesting to notice that the RLFN architecture by [2] from the NTIRE 2022 challenge on Efficient Super-Resolution is able to still improve over BSRT with a fraction of the parameters. We also notice that bicubic upsampling followed by $1 \times 1$ RGB-to-features convolution is not sufficient to provide reasonable performance, highlighting the need for operations that capture a local context larger than 1 pixel. In fact, when bicubic upsampling is followed by $3 \times 3$ convolution the subsequent MIFF module is able to successfully exploit the local context as the overall performance increases by 1.17dB while the SISR performance stays almost the same. We also notice that the PSNR difference between the single-image and multi-image results is stable around 1.6 dB, proving that the MIFF module is relatively robust to the single-image processing.

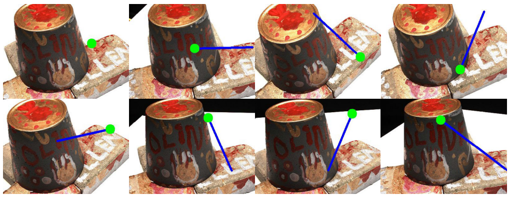
### F. WIDER-BASELINE EXPERIMENT

In this section we present an experiment where views are taken very far apart and asymmetrically with respect to the target view in order to challenge the method and the state-of-the-art BSRT. Table 4 reports the PSNR obtained by BSRT and EpiMISR when compared to the SISR PSNR. It can be noticed that in this challenging setting, BSRT degrades to the SISR performance, while EpiMISR still provides an improvement. This more challenging geometry is created by taking the $V-1$ extra views that are at median distance (out of all the views available in the dataset) with respect to the distance to the target view camera center.

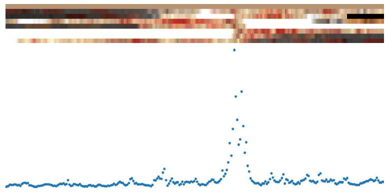
### G. ANALYSIS OF RAY ATTENTION

In this section, we present an interpretation of the attention map generated by the ray transformer within the MIFF module as a depth map. Fig. 4a illustrates a typical input

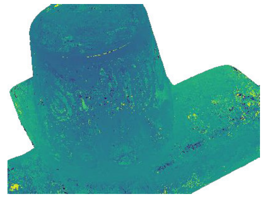
IEEE Access



(a) Input image set with epipolar lines.



(b) Strip aligned against Ray Transformer attention.



(c) Depth map.

**FIGURE 4. An example of depth map generation.**

image set. The first image is the target view, while the subsequent $V = 7$ images are the extra views. Let us fix the pixel to be superresolved in the target image. The CAP modules casts a ray through this pixel and projects it onto the other views. This process yields samples along the epipolar lines, which are collected to form a "strip" of dimensions $P \times (V + 1)$, depicted in Fig. 4b (depiction is in RGB space instead of feature vectors). There are $P$ columns because the CAP module samples $P$ points along the epipolar lines, and there are $V+1$ rows because there are $V+1$ epipolar lines. It is worth noting that the first row comprises repeated instances of the same pixel, as the epipolar line collapses to a single point in the target view. Thanks to the property of epipolar geometry, there is a region along the strip, which we will call "strip alignment region", where all the views are imaging the same 3D point, hence the sampled feature map should report similar information. The attention weights generated by the ray transformer are also visualized in Figure 4b and we can see they reach their maximum in the alignment region, meaning that the MIFF module has identified the correspondences across all extra images. Moreover, the position of the maximum attention weight provides an estimate of the depth of the object imaged by the selected pixel in the target view. A noisy depth map for all pixels can be extracted in this unsupervised way and is visualized in Fig. 4c.

### H. FAILURE CASES AND MORE QUALITATIVE RESULTS

Fig. 7 shows the Empirical Cumulative Distribution Function (ECDF) of the PSNR improvements of EpiMISR with respect to BSRT on all the DTU dataset test split. The failure cases, that are the instances in the DTU test dataset where BSRT outperforms EpiMISR, are rare, as the ECDF(0) $\approx$ 2.04%. Fig. 6 shows an example of such rare cases while Fig. 5 reports some DTU scenes results where the proposed method outperforms the baselines.

### I. SENSITIVITY ANALYSIS TO CAMERA PARAMETER ESTIMATION

In this section we present two experiments that show EpiMISR performance when camera parameters are not available (first experiment) or noisy (second experiment).

Camera parameters in the DTU dataset are highly accurate, as they have been obtained from a calibration procedure. The availability of calibrated poses is not guaranteed in practical application. Hence, in this experiment we use the state-of-the-art HLOC algorithm [50] to infer poses from the LR images alone. We then input these inferred poses to EpiMISR and obtain the predicted HR images. Comparing with the ground truth HR images, we report a PSNR of 28.10 dB, which is degraded from the result with accurate poses but still superior to BSRT which does not need that information, confirming that, while improvements in 3D geometry modeling lead to improved performance, poses retrieved from a practical algorithm are sufficient to obtain state-of-the-art super-resolution performance.

As a second experiment, a sensitivity analysis to perturbations of the extrinsic camera parameters is shown in Fig. 8. It shows the PSNR achieved when the 6-D DTU pose is perturbed to simulate uncertainty. A diagonal zero-mean Gaussian with parameter $\sigma_{\text{translation}}$ is used to perturb the translational components. A simple symmetric distribution over $SO(3)$ with parameter $\sigma_{\text{rotation}}$ is used to perturb the rotational component. As Fig. 8 shows, the performance of EpiMISR degrades in higher noise poses regime, but it is still superior to BSRT in a lower noise regime and, overall, it exhibits a stable trend.

Finally, we remark that camera parameter estimation from LR images performed disjointly from the SR process is clearly suboptimal. Future work may significantly improve the results by designing joint methods that correct an initial pose estimation while performing super-resolution, similarly to what is done by NeRF methods for in-the-wild images [51].

### J. VIEW CONSISTENCY

In this section we present an experiment where the view consistency is assessed. As the setting we study is that of not-novel view synthesis, we are only concerned with generating details that are consistent with the LR observations of the target view we want to super-resolve, and it is outside the scope of the method to enable novel view synthesis. The transformers used as building blocks of our method implicitly ensure that only consistent information is borrowed from the other views via the attention mechanism. Table 5 reports an additional result about the PSNR between the LR target image and the SR target image when degraded to LR.

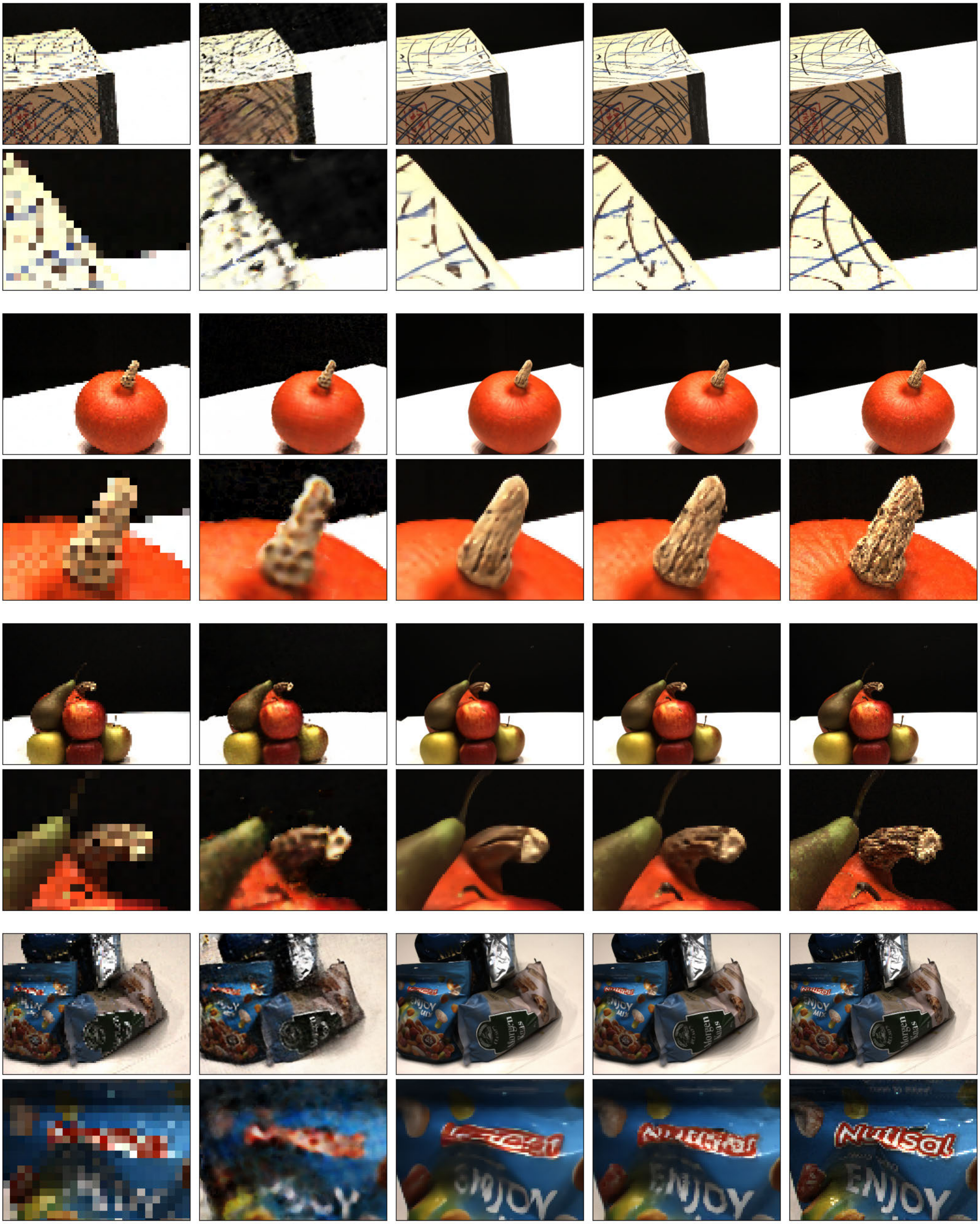


**FIGURE 5.** Qualitative results of some DTU test scenes with 4× scale factor. From left to right: LR nearest neighbours interpolation, NeRF-SR, BSRT, EpiMISR, HR ground truth.

**FIGURE 6.** A qualitative example of a failure case (DTU dataset, scan 63). This is an example where BSRT outperforms EpiMISR. From left to right: LR nearest neighbours interpolation, NeRF-SR, BSRT, EpiMISR, HR ground truth.

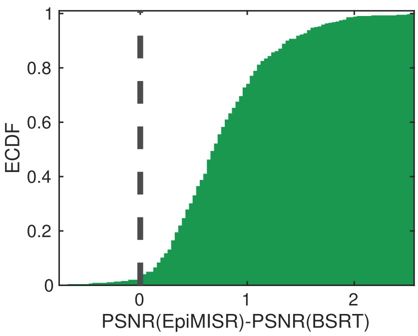



**FIGURE 7.** ECDF of the PSNR improvements of EpiMISR with respect to BSRT on the test split of the DTU dataset.

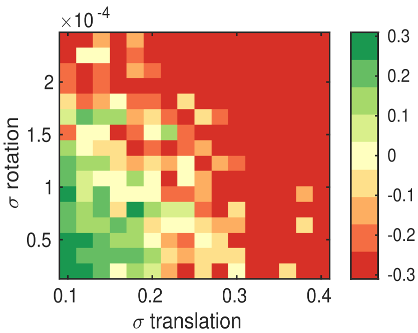



**FIGURE 8.** EpiMISR PSNR gain (dB) over BSRT for different noise regimes on camera poses for a single test image.

**TABLE 5.** View consistency. PSNR between the degraded SR images and the LR images.

| | LR - PSNR (dB) ↑ |
|---|---|
| **EpiMISR** | **30.71** |
| BSRT ( [9]) | 30.08 |
| SwinIR ( [1]) | 29.14 |

## V. CONCLUSION AND FUTURE WORKS

We presented a novel setting for multi-image super-resolution that addresses the case of image sets with arbitrary camera placements, potentially involving large disparities. The explicit use of epipolar geometry in our approach enables notable improvements over existing methods that rely on optical flow. However, the method faces some limitations, particularly with very noisy camera parameters, which can hinder reconstruction quality. Additionally, it assumes a pinhole camera model, and a generalization of the CAP module is needed to handle more complex projections, such as fish-eye or wide-angle lenses. Future work could also explore incorporating adaptive mechanisms that correct pose estimation errors during the super-resolution process to increase the robustness and applicability of the method.

## REFERENCES

[1] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, ''SwinIR: Image restoration using swin transformer,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[2] F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, F. Chen, and L. Fu, ''Residual local feature network for efficient super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 765–775.

[3] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, ''Activating more pixels in image super-resolution transformer,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.

[4] X. Chu, L. Chen, and W. Yu, ''NAFSSR: Stereo image super-resolution using NAFNet,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1238–1247.

[5] Z. Wang, J. Chen, and S. C. H. Hoi, ''Deep learning for image super-resolution: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[6] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, ''DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3644–3656, May 2020.

[7] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, ''Deep burst super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9205–9214.

[8] B. Lecouat, J. Ponce, and J. Mairal, ''Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2350–2359.

[9] Z. Luo, Y. Li, S. Cheng, L. Yu, Q. Wu, Z. Wen, H. Fan, J. Sun, and S. Liu, ''BSRT: Improving burst super-resolution with swin transformer and flow-guided deformable alignment,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 997–1007.

[10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, ''NeRF: Representing scenes as neural radiance fields for view synthesis,'' *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[11] S. D. Babacan, R. Molina, and A. K. Katsaggelos, ''Total variation super resolution using a variational approach,'' in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 641–644.

[12] Y.-R. Li, D.-Q. Dai, and L. Shen, "Multiframe super-resolution reconstruction using sparse directional regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 7, pp. 945–956, Jul. 2010.

[13] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.

[14] H. Huang, L. Shen, C. He, W. Dong, and W. Liu, "Differentiable neural architecture search for extremely lightweight image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2672–2682, Dec. 2023.

[15] D. Huang, X. Zhu, X. Li, and H. Zeng, "CLSR: Cross-layer interaction pyramid super-resolution network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6273–6287, Nov. 2023.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[18] Z. Luo, Y. Huang, L. Shang, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 5632–5643.

[19] S. Baker and T. Kanade, "Super resolution optical flow," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. CMU-RI-TR-99-36, Oct. 1999.

[20] G. Bhat et al., "NTIRE 2022 burst super-resolution challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1041–1061.

[21] S. Zhang, S. Chang, and Y. Lin, "End-to-end light field spatial super-resolution network using multiple epipolar geometry," *IEEE Trans. Image Process.*, vol. 30, pp. 5956–5968, 2021.

[22] Z. Zhang, B. Peng, J. Lei, H. Shen, and Q. Huang, "Recurrent interaction network for stereoscopic image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2048–2060, May 2023.

[23] J. Tang, C. Lu, Z. Liu, J. Li, H. Dai, and Y. Ding, "CTVSR: Collaborative spatial–temporal transformer for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 5018–5032, Jun. 2024.

[24] X. Chai, F. Shao, Q. Jiang, and H. Ying, "TCCL-Net: Transformer-convolution collaborative learning network for omnidirectional image super-resolution," *Knowl.-Based Syst.*, vol. 274, Aug. 2023, Art. no. 110625.

[25] X. Chai, F. Shao, H. Chen, B. Mu, and Y.-S. Ho, "Super-resolution reconstruction for stereoscopic omnidirectional display systems via dynamic convolutions and cross-view transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[26] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, Jul. 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[27] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4576–4585.

[28] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "IBRNet: Learning multi-view image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4688–4697.

[29] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5835–5844.

[30] B. K. S. Isaac-Medina, C. G. Willcocks, and T. P. Breckon, "Exact-NeRF: An exploration of a precise volumetric parameterization for neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 66–75.

[31] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-NeRF: Anti-aliased grid-based neural radiance fields," 2023, *arXiv:2304.06706*.

[32] X. Huang, Q. Zhang, Y. Feng, X. Li, X. Wang, and Q. Wang, "Local implicit ray function for generalizable radiance field representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 97–107.

[33] M. Varma, P. Wang, X. Chen, T. Chen, S. Venugopalan, and Z. Wang, "Is attention all that NeRF needs?" in *Proc. 11th Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–22.

[34] M. Suhail, C. Esteves, L. Sigal, and A. Makadia, "Generalizable patch-based neural rendering," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland. Springer, Jan. 2022, pp. 156–174.

[35] M. Suhail, C. Esteves, L. Sigal, and A. Makadia, "Light field neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8259–8269.

[36] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7776–7785.

[37] X. Wang, Z. Zhu, G. Huang, F. Qin, Y. Ye, Y. He, X. Chi, and X. Wang, "MVSTER: Epipolar transformer for efficient multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 573–591.

[38] Q. Huang, M. Hu, and D. Jones Brady, "Array camera image fusion using physics-aware transformers," 2022, *arXiv:2207.02250*.

[39] N. Pearl, T. Treibitz, and S. Korman, "NAN: Noise-aware NeRFs for burst-denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12662–12671.

[40] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "NeRF in the dark: High dynamic range view synthesis from noisy raw images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16169–16178.

[41] C. Wang, X. Wu, Y.-C. Guo, S.-H. Zhang, Y.-W. Tai, and S.-M. Hu, "NeRF-SR: High quality neural radiance fields using super-sampling," in *Proc. 30th ACM Int. Conf. Multimediax*, Oct. 2022, pp. 6445–6454.

[42] Y. Han, T. Yu, X. Yu, Y. Wang, and Q. Dai, "Super-NeRF: View-consistent detail generation for NeRF super-resolution," 2023, *arXiv:2304.13518*.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 1–11.

[44] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[45] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 406–413.

[46] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[47] A. Dudhane, S. Waqas Zamir, S. Khan, F. Shahbaz Khan, and M.-H. Yang, "Burstormer: Burst image restoration and enhancement transformer," 2023, *arXiv:2304.01194*.

[48] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[49] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–14, Aug. 2019.

[50] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12708–12717.

[51] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the wild: Neural radiance fields for unconstrained photo collections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7210–7219.

**LUCA SAVANT AIRA** (Member, IEEE) received the B.Sc. degree in mathematics for engineering and the M.Sc. degree in mathematical engineering from the Politecnico di Torino, in 2019 and 2022, respectively, where he is currently pursuing the Ph.D. degree in artificial intelligence. His research interests include 3D world representation, multi-image super-resolution, and uncertainty estimation.

**DIEGO VALSESIA** (Member, IEEE) received the Ph.D. degree in electronic and communication engineering from the Politecnico di Torino, in 2016. He is currently an Assistant Professor with the Department of Electronics and Telecommunications (DET), Politecnico di Torino. His main research interests include processing of remote sensing images, and deep learning for inverse problems in imaging. He is a member of the EURASIP Technical Area Committee for Signal and Data Analytics for Machine Learning and a member of the ELLIS Society. He was a recipient of the IEEE ICIP 2019 Best Paper Award and the IEEE Multimedia 2019 Best Paper Award. He is a Senior Area Editor of IEEE Transactions on Image Processing, for which he received the 2023 Outstanding Editorial Board Member Award.

**ANDREA BORDONE MOLINI** received the M.Sc. degree in computer networks engineering from the Politecnico di Torino, Turin, Italy, and the M.Sc. degree in communications systems security from Telecom ParisTech, Paris, France, as part of the double degree program, in 2016, and the Ph.D. degree in electronic and communication engineering from the Politecnico di Torino, in 2021. He is currently a Senior Computer Vision Algorithm Researcher with Zebra Technologies, London.

**GIULIA FRACASTORO** (Member, IEEE) received the Ph.D. degree in electronic and telecommunications engineering from the Politecnico di Torino, Turin, Italy, in 2017. She is currently an Assistant Professor with the Department of Electronics and Telecommunications (DET), Politecnico di Torino. Her research interests include graph signal processing, system identification, and deep learning. She was a recipient of the IEEE International Conference on Image Processing (ICIP) 2019 Best Paper Award, the IEEE Picture Coding Symposium (PCS) 2016 Best Paper Award, and the IEEE International Conference on Image Processing (ICIP) 2015 Best Student Paper Award. She is an Associate Editor of IEEE Transactions on Image Processing and IEEE Transactions on Signal and Information Processing Over Networks.

**ENRICO MAGLI** (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the Politecnico di Torino, Italy, in 1997 and 2001, respectively. He is currently a Full Professor with the Politecnico di Torino, where he leads the Image Processing and Learning Group, performing research in the fields of deep learning for image and video processing, image compression, and image forensic for multimedia and remote sensing applications. He is a fellow of the ELLIS Society for the advancement of artificial intelligence in Europe. He was a recipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award, the IEEE ICIP 2015 Best Student Paper Award (as the Senior Author), the IEEE ICIP 2019 Best Paper Award, the IEEE Multimedia 2019 Best Paper Award, and the 2010 and 2014 Best Associate Editor Award of IEEE Transactions on Circuits and Systems for Video Technology. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and the *EURASIP Journal on Image and Video Processing*. He was an IEEE Distinguished Lecturer from 2015 to 2016.

**ANDREA MIRABILE** received the B.Sc. degree in electronic engineering from the Universitã di Catania, Italy, and the M.Sc. degree in bioengineering with neurotechnology from Imperial College London, U.K., in 2013. He is currently the Global Director of Artificial Intelligence with Zebra Technologies, where he has been leading AI initiatives, since 2021. With over a decade of experience, his expertise spans deep learning, computer vision, natural language processing (including generative AI), and MLOps.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement