SUPERF: NEURAL IMPLICIT FIELDS FOR MULTI-IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-resolution imagery is often hindered by limitations in sensor technology, atmospheric conditions, and costs. Such challenges occur in satellite remote sensing, but also with handheld cameras, such as our smartphones. Hence, super-resolution aims to enhance the image resolution algorithmically. Since single-image super-resolution requires solving an inverse problem, such methods must exploit strong priors, e.g. learned from high-resolution training data, or be constrained by auxiliary data, e.g. by a high-resolution guide from another modality. While qualitatively pleasing, such approaches often lead to "hallucinated" structures that do not match reality. In contrast, multi-image super-resolution (MISR) aims to improve the (optical) resolution by constraining the super-resolution process with multiple views taken with sub-pixel shifts. Here, we propose *SuperF*, a test-time optimization approach for MISR that leverages coordinate-based neural networks, also called neural fields. Their ability to represent continuous signals with an implicit neural representation (INR) makes them an ideal fit for the MISR task. The key characteristic of our approach is to share an INR for multiple shifted low-resolution frames and to jointly optimize the frame alignment with the INR. Our approach advances related INR baselines, adopted from burst fusion for layer separation, by directly parameterizing the sub-pixel alignment as optimizable affine transformation parameters and by optimizing via a super-sampled coordinate grid that corresponds to the output resolution. Our experiments yield compelling results on simulated bursts of satellite imagery and ground-level images from handheld cameras, with upsampling factors of up to 8. A key advantage of SuperF is that this approach does not rely on any high-resolution training data.

1 INTRODUCTION

The spatial resolution of imaging is often limited by sensor capabilities, atmospheric interference, and acquisition costs, affecting various domains including satellite remote sensing, smartphone photography, and medical imaging. Super-resolution (SR) aims to overcome such physical constraints algorithmically. Single-image super-resolution (SISR) methods tackle this inverse problem by relying on strong priors, typically learned from extensive high-resolution (HR) datasets (Ledig et al., 2017; Zhang et al., 2023), or through auxiliary guidance from complementary modalities (De Lutio et al., 2019; 2022; Metzger et al., 2023; Mei et al., 2025). Although SISR methods can produce visually appealing results, their reliance on learned priors often leads to *hallucinated* structures that diverge from the true underlying scene (Cohen et al., 2024). This may be tolerable for smartphone applications, but not for applications in medicine and science.

To mitigate some of these issues, multi-image super-resolution (MISR) has emerged as a special case of super-resolution by incorporating additional information from multiple low-resolution (LR) images captured with slight sub-pixel shifts (Tsai & Huang, 1984; Irani & Peleg, 1991; Elad & Feuer, 1997). As sub-pixel shifts vary across the repeated LR frames, the discretization introduces different aliasing artifacts in each frame. While these artifacts seem to be noise in the LR data, they can be leveraged as complementary information to compute the shared underlying high-resolution image (Wronski et al., 2019).

While MISR can be approached with *supervised learning-based* methods (Bhat et al., 2021a;b) when large training datasets with paired LR and HR data are available, MISR can also be achieved

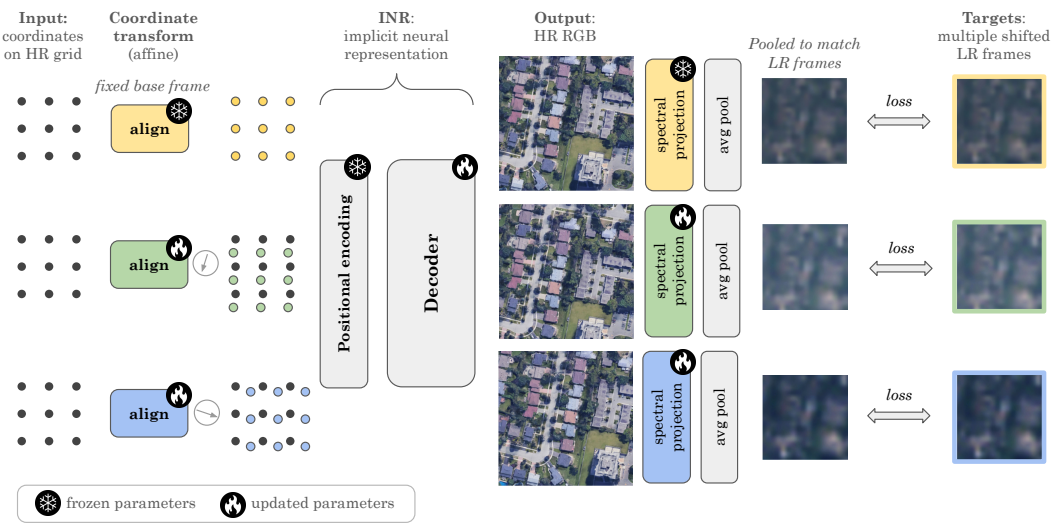


Figure 1: **Illustration of the proposed method.** *SuperF* achieves multi-image super-resolution by sharing an implicit neural representation (INR) across multiple low-resolution (LR) frames with sub-pixel shifts. The LR frames are aligned by jointly optimizing an affine coordinate transformation for each LR frame, together with the parameters of a coordinate-based multi-layer perceptron (MLP) that decodes the input coordinates to RGB values. Hence, leveraging the continuous characteristics of INRs for both the sub-pixel alignment in the pixel coordinate space *and* for representing the underlying high-resolution (HR) signal.

by *test-time optimization (TTO)* approaches that do not require offline training (Wronski et al., 2019; Lafenetre et al., 2023). The latter are particularly interesting, since HR data acquisition is expensive and the creation of large training datasets by pairing of LR images and HR data is non-trivial (Bhat et al., 2021a). Typically, MISR is associated with bursts of images captured in rapid succession. Repeated observations in satellite remote sensing also provide a multi-frame scenario with longer time intervals.[1]

In this work, we introduce *SuperF*, a test-time optimization approach for MISR leveraging the continuous field of implicit neural representations (INR). SuperF shares an INR across multiple shifted LR frames, while jointly estimating the frame-specific alignment. Iteratively refining both the alignment and the shared neural representation effectively reconstructs the underlying high-resolution image on a continuous field (see Fig. 1 for an illustration of the proposed method).

INRs are coordinate-based neural networks, also called neural fields[2], typically parameterized by multi-layer perceptrons (MLPs) that map continuous input coordinates (such as 2D image locations) directly to signals like RGB pixel intensities. Optimizing the parameters of such an MLP on an image implicitly encodes the image within its weights. Beyond image representation, INRs have been successfully adopted for data compression (Strümpler et al., 2022; Kwan et al., 2024), 3D shape modeling (Park et al., 2019; Mescheder et al., 2019), novel-view synthesis with neural radiance fields (NeRF) (Mildenhall et al., 2020), and burst fusion for denoising (Pearl et al., 2022) or layer separation of obstructions and background scenes (Nam et al., 2022; Chugunov et al., 2024).

The common unsupervised way to solve the MISR problem is to map the series of LR frames to a HR image, for example using steerable kernel regression (Wronski et al., 2019; Lafenetre et al., 2023). Instead of using the LR frames as an *input* to our model, we draw inspiration from the guided super-resolution work by De Lutio et al. (2019) and turn the problem formulation up-side down and treat the LR frames as reconstruction targets. While Nam et al. (2022) have explored such directions for burst fusion and layer separation tasks, their method was not designed to accurately solve sub-pixel frame alignment, which we show is crucial for MISR. Here, we build on these great ideas and design INRs dedicated for the MISR task. By directly parameterizing the affine transformations

[1]Like prior work we use the terms burst and multi-frame interchangeably (Wronski et al., 2019).

[2]We note that the term neural fields is used differently in computational neuroscience (Amari, 1977).

for the frame alignment and by introducing a supersampling strategy, we improve the sub-pixel alignment and consequently the MISR performance.

We empirically validate the proposed SuperF algorithm on bursts obtained from satellite imagery as well as ground-level images from handheld cameras. In both cases, SuperF gave compelling results. A key aspect of our approach is that it is a TTO method that avoids the need for large amounts of high-resolution training data. This also minimizes the risk of hallucinating high-resolution structures, as opposed to supervised learning based approaches. Our contributions are summarized as follows:

1. We propose *SuperF*, a test-time optimization method for MISR based on implicit neural representations. We demonstrate that jointly optimizing the sub-pixel frame alignment with an MLP shared across frames is both simple and key to adopt INRs to the MISR task.
2. Our method yields an improved sub-pixel alignment and continuous representation of the high-resolution signal, by directly parameterizing the affine transformations and by optimizing the INR with a supersampling strategy.
3. We introduce SatSynthBurst, a synthetic satellite burst dataset for MISR research and demonstrate that our proposed approach generalizes to different domains including ground-level bursts from handheld cameras and satellite image bursts.

2 Related work

2.1 Multi-Image Super-Resolution (MISR)

Existing MISR approaches contain both learning-free and learning-based methods. They are less prone to hallucinate structures compared to SISR. Test-time optimization (TTO) is applied to exploit natural hand tremors in handheld smartphone photography to capture bursts of slightly shifted raw images. Wronski et al. (2019) propose a steerable kernel regression that enables the direct RGB reconstruction without explicit demosaicing and improving resolution and signal-to-noise ratio, a technology built into the 'pixel' phone. This approach was later reimplemented and adapted for satellite burst applications (Lafenetre et al., 2023).

Like SISR approaches that aim to learn priors from large training datasets, the MISR problem has also been approached with both supervised (Bhat et al., 2021a;b; Cornebise et al., 2022) and self-supervised (Nguyen et al., 2022) learning. Deep neural network architectures for burst super-resolution were proposed to learn the alignment of multiple noisy RAW inputs in latent space via optical flow and the fusion with attention-based modules (Bhat et al., 2021a). Bhat et al. (2021b) proposed a deep reparameterization of the MISR problem and formulated the reconstruction objective in a learned latent space.

In this work, we propose a TTO approach leveraging the continuous nature of INR by jointly optimizing the alignment of low-resolution frames in a continuous coordinate space. Hence, our approach does neither require any high-resolution training data, nor a pre-processing step to register the LR frames. We compare our results with the approach proposed by Lafenetre et al. (2023), which is the closest state-of-the-art TTO approach and, at the same time, the only publicly available implementation. Since their high-resolution test data is not publicly available, we compare results on two different datasets. A handheld burst dataset (Bhat et al., 2021a) and a new synthetic satellite image burst dataset based on open high resolution images (Cornebise et al., 2022). As opposed to prior work that focuses on one domain, we demonstrate that our approach generalizes to different domains including satellite and ground-level image bursts.

2.2 Implicit Neural Representations (INR)

Recent advances in INRs have demonstrated the strength of representing continuous signals across various tasks (Essakine et al., 2025), but the development of INR has mainly been driven by 3D shape modeling (Park et al., 2019; Mescheder et al., 2019) and novel-view synthesis (Mildenhall et al., 2020). These techniques have been adopted for multi-view satellite data (Derksen & Izzo, 2021; Xiangli et al., 2022), species distribution modelling (Cole et al., 2023), and medical imaging to e.g. model 3D MIR volumes (Wu et al., 2021).

Recently INRs have been studied for single-image SR at arbitrary-scales (Chen et al., 2021; Cao et al., 2023; Chen et al., 2023; Zhu et al., 2025). Notably, Chen et al. (2021) propose the Local Implicit Image Function (LIIF) that models RGB values at arbitrary scaling factors. They devised a supervised learning approach that combines the explicit representation from a learned embedding with a local INR that is anchored in the nearest neighbor embeddings. Extending along this line of work, Becker et al. (2025) propose Thera, a neural heat field that explicitly models the point spread function (PSF) to enable analytically correct anti-aliasing at any resolution.

Here we bring forward an approach to leverage the continuous characteristics of INRs for *multi-image* SR. As opposed to prior work on INR for SISR, we do not investigate a supervised approach. Since it is challenging to build training datasets for multi-image super-resolution, we propose a TTO-based solution that does not require any high-resolution training data, but optimizes an INR on multiple shifted LR frames. Key to our proposed approach is the joint optimization of the INR with the alignment of the LR frames, which allows us to share an INR across frames leading to a continuous representation of the underlying high-resolution signal.

Closest to our work is the *NIR* approach presented by Nam et al. (2022). Although originally developed to fuse bursts for layer separation, NIR also serves as an INR baseline for the MISR task. Our method differs in three components. First, while NIR estimates transformation matrices $T_t = g(t)$ using a separate ReLU MLP g conditioned on the frame index t, we directly parametrize the transformation matrices as part of the model. Second, to improve sub-pixel alignment, we introduce a supersampling strategy to optimize the model on a high-resolution coordinate grid that is subsequently downsampled for supervision with the LR frames (similar strategies have been proposed to enhance details in novel view synthesis (Wang et al., 2022)). Finally, following prior MISR work (Wronski et al., 2019), rather than estimating transformations for all frames, we use the base frame as the reference coordinate system to relatively align all other LR frames. This reduces the degrees of freedom and facilitates evaluation with high-resolution reference data by avoiding misalignment.

3 METHODOLOGY

We describe images by functions $[0,1)^d \to \mathbb{R}^{n_c}$ mapping coordinates to intensities. In our application, we consider two-dimensional RGB frames in homogeneous coordinates, i.e., $d = 3$ and $n_c = 3$. Our input are T low-resolution frames $\mathbf{y}_{\text{LR}}^{(1)}, \ldots, \mathbf{y}_{\text{LR}}^{(T)}$ in discretized form, i.e., we are given the values at a finite discrete set of points $\mathcal{W} \subset [0,1)^d$. Our goal is to find an approximation $\hat{\mathbf{y}}_{\text{HR}}$ of the underlying high-resolution signal $\mathbf{y}_{\text{HR}}$ at points $\mathcal{V} \subset [0,1)^d$. Typically, $\mathcal{V}$ and $\mathcal{W}$ are grid points and $|\mathcal{V}| > |\mathcal{W}|$ because the $\mathbf{y}_{\text{LR}}^{(t)}$ are sampled with a lower resolution than the target resolution defined by $\mathcal{V}$.

Our approach is based on the assumption that $\mathbf{y}_{\text{LR}}^{(t)}(\boldsymbol{v}) \approx \varphi * \mathbf{y}_{\text{HR}}(\mathbf{A}^{(t)}\boldsymbol{v})$, where $\mathbf{A}^{(t)}$ is an affine transformation matrix and φ is a boxcar filter. Convolution with the boxcar filter implements a spatial average pooling. The affine transformation matrix models misalignments by rotation and translation in the homogeneous coordinate system. In contrast to standard registration methods, our goal is to also exploit misalignments by sub-pixel shifts, i.e., smaller than $\|\mathbf{w}_i - \mathbf{w}_j\|_\infty$ for any $\mathbf{w}_i, \mathbf{w}_j \in \mathcal{W}$.

3.1 IMPLICIT NEURAL REPRESENTATION (INR) SHARED ACROSS FRAMES

To optimize an implicit representation of an image, we make use of a coordinate-based multi-layer perceptron (MLP). The MLP model is denoted by $f_{\boldsymbol{\theta}}$ with learnable parameters $\boldsymbol{\theta}$. It is optimized to output the intensities $\hat{\mathbf{y}}$ (e.g., RGB pixel values) for the corresponding input coordinate $\mathbf{v} \in [0,1)^d$.

To share $f_{\boldsymbol{\theta}}$ for T shifted low-resolution frames, we need to *align* them on a sub-pixel scale. To achieve this, we make use of the continuous nature of INRs and optimize the parameters of affine transformation matrices $\hat{\mathbf{A}}^{(t)}$ that are applied to transform the input coordinates for each frame t. Following prior work (Wronski et al., 2019), we use the base frame as the reference coordinate system and set $\hat{\mathbf{A}}^{(1)} = I$, where I is the identity matrix (see Fig. 1). The coordinates $\mathbf{v}$ correspond to the high-resolution grid of the base frame:

$$\hat{\mathbf{y}}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{v}) = \hat{\rho}^{(t)}(f_{\boldsymbol{\theta}}(\hat{\mathbf{A}}^{(t)}\mathbf{v})) \tag{1}$$

The transformation matrices $\hat{\mathbf{A}}^{(t)}$ are directly parameterized by two translation parameters $\Delta x^{(t)}$ and $\Delta y^{(t)}$ as well as one rotation angle $\alpha^{(t)}$ for each frame. In contrast, Nam et al. (2022) proposed to estimate transformation matrices with another MLP for burst fusion. Since we only assume an approximate relationship between LR frames and expect some variation in brightness and contrast, our model optimizes a frame specific spectral projection $\rho^{(t)}$ with a scale and shift parameter per spectral band. For the base frame this projection ρ^1 is also fixed (scale 1 and shift 0).

3.2 Optimization with low-resolution frames

We propose a *supersampling* strategy to improve the sub-pixel alignment and consequently the implicit neural representation of the HR signal. During optimization, we run the INR at the high-resolution grid $\boldsymbol{v}$ corresponding to the resolution of the super-resolved output. Since we only have the $\mathbf{y}_{\text{LR}}^{(t)}$ available for the optimization, we need to match the output of the INR to the low-resolution frames. That is, we want to find $\boldsymbol{\theta}$ and $\hat{\mathbf{A}}^{(t)}$ such that

$$\hat{\mathbf{y}}_{\text{LR},\,\boldsymbol{\theta}}^{(t)}(\boldsymbol{v}) = \varphi * \hat{\rho}^{(t)}(f_{\boldsymbol{\theta}}(\hat{\mathbf{A}}^{(t)}\boldsymbol{v})) \tag{2}$$

equals $\mathbf{y}_{\text{LR}}^{(t)}(\boldsymbol{v})$ on $\boldsymbol{v} \in \mathcal{W}$. We fix the boxcar filter φ, which is implied by different resolutions of the discretized HR output and the given LR images. Ultimately, we optimize for multiple low-resolution frames by averaging a point-wise loss ℓ across the T frames:

$$\arg\min \mathcal{L}(\boldsymbol{\theta}, \hat{\mathbf{A}}^{(1)}, \ldots, \hat{\mathbf{A}}^{(T)}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{\boldsymbol{v}\in\mathcal{W}} \ell\left(\hat{\mathbf{y}}_{\text{LR},\,\boldsymbol{\theta}}^{(t)}(\boldsymbol{v}), \mathbf{y}_{\text{LR}}^{(t)}(\boldsymbol{v})\right) \tag{3}$$

In practice, the convolution with the boxcar filter and the sampling at grid points $\mathcal{W}$ is simply implemented by an average pooling. We use MLPs with ReLU activation functions and stochastic gradient descent with mini batches of frames.

3.3 Input transforms for high-resolution representations

Since we can only optimize the model on low-resolution 'views' of the underlying high-resolution signal, the MLP is prone to output only the low-frequencies of the signal. Hence, to recover high-frequencies captured by the multiple LR frames, we need to steer the MLP to output high-frequency details. In general, coordinate-based MLPs exhibit a spectral bias. The networks prioritize the reconstruction of low-frequency components of the target signal, whereas high-frequency details emerge only slowly during the convergence of optimization. Several approaches have been proposed to overcome this spectral bias (Sitzmann et al., 2020; Saragadam et al., 2023). We rely on the commonly used Fourier features (Tancik et al., 2020) as a positional encoding. The feature map $\gamma : [0,1)^d \to \mathbb{R}^{2m}$ is based on a random set of sine and cosine basis functions:

$$\gamma(\mathbf{v}) = \left[\cos(2\pi\mathbf{b}_1^{\mathrm{T}}\mathbf{v}), \ldots, \cos(2\pi\mathbf{b}_m^{\mathrm{T}}\mathbf{v}), \sin(2\pi\mathbf{b}_1^{\mathrm{T}}\mathbf{v}), \ldots, \sin(2\pi\mathbf{b}_m^{\mathrm{T}}\mathbf{v})\right]^{\mathrm{T}} \tag{4}$$

Each $\mathbf{b}_i \in \mathbb{R}^d$, $i = 1, \ldots, m$, is sampled from an isotropic multi-variate Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$, where the scale σ is a hyperparameter controlling the range of the sampled frequencies. We show that this hyperparameter is sensitive to the domain (e.g., satellite images vs. ground-level burst images), but the same value performs well across all samples within a domain.[3]

4 Experimental results and discussion

4.1 Datasets

Our experiments are based on datasets from two domains: remote sensing and handheld cameras (see examples in Fig. 2). First, we create a synthetic burst dataset from high-resolution satellite images to study various characteristics and the sensitivity of our proposed approach. Second, we demonstrate that our SuperF approach also generalizes to ground-level bursts from handheld cameras.

[3] Tancik et al. (2020) establish the relation between σ^2 and the bandwidth of the neural tangent kernel modelling the resulting MLP. They argue that a wider kernel supports the learning of high frequency components, but that a too wide kernel can lead to aliasing artifacts. They conclude that the parameter is problem dependent and has to be tuned.

SatSynthBurst (satellite imagery). To study MISR for satellite imagery bursts we constructed a synthetic burst dataset derived from 20 open high-resolution satellite images selected from the WorldStrat dataset (Cornebise et al., 2022) (see examples in Appendix). For each high-resolution sample, we generate 16 low-resolution frames with scale factors 2, 4, and 8, by randomly sampling sub-pixel shifts. Our dataset provides one LR *base frame* that is spatially aligned with the HR test images, a missing feature of existing datasets. This framework allows to study the influence of different upsampling factors keeping the HR resolution fixed. Furthermore, it gives us control over the sub-pixel shifts, and noise intensity. Although, our approach does not require the true parameters of the misalignment, they allow us to monitor the optimization dynamics when estimating the alignment. To realistically simulate the image formation process, we use spectral variations and additive Gaussian noise in all experiments (if not further specified). We follow the best practices for generating synthetic super-resolution data to mimic the modulation transfer function (MTF) of the Sentinel-2 sensor described by Lanaras et al. (2018). Details in the Appendix section A.1.

SyntheticBurst (ground-level imagery). To evaluate on handheld bursts of ground-level scenes, we make use of the SyntheticBurst data provided by Bhat et al. (2021a) and e.g. used by Bhat et al. (2021b). We select 50 out of the 300 provided ground level bursts that provide interesting high-resolution structures, i.e., we remove for example bursts that are crops of homogeneous areas such as building walls or skies. Each burst consists of 14 LR frames, originally at a scale factor of $\times 8$. To study different upsampling factors, we vary the HR output resolution by downsampling the HR reference images, but we keep the LR frames as provided to avoid changing the underlying noise model. Since this dataset does not provide a base LR frame aligned with the the HR test image, we run a brute force postprocessing to improve the alignment of the predictions before computing the error metrics (see Appendix section A.3).

4.2 Experimental setup

We follow standard practices in super-resolution and report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (using AlexNet), computed using the implementation by Bhat et al. (2021a). All experiments are implemented in PyTorch and executed on a single NVIDIA H100 GPU with 80 GB of VRAM (note, our experiments typically need around 1GB of VRAM). If not further specified, all experiments use the AdamW optimizer with a base learning rate of 2×10^{-3}, which is decayed to 1×10^{-6} over 2000 iterations using a cosine annealing schedule and a batch size of 1 frame.

During evaluation, a 16-pixel boundary is cropped from all sides to reduce edge artifacts. We additionally apply color matching as a post-processing step, following Bhat et al. (2021a), to correct for global color and intensity shifts between the reconstruction and the ground truth. The scale hyperparameter of the Fourier feature positional encoding is set to 10 for the SatSynthBurst and to 3 for the SyntheticBurst dataset.

4.3 Comparison to existing test-time optimization approaches

Baseline approaches. As we propose a TTO approach, we compare to a state-of-the-art TTO approach for MISR, a steerable kernel regression method by Lafenetre et al. (2023), which is an adapted version of the approach described by Wronski et al. (2019). We also compare to a burst fusion approach by Nam et al. (2022) (named NIR) and adapt it as a MISR baseline. Although developed for burst fusion for layer separation tasks, it is related to our method as it uses an INR with a built-in frame alignment (as introduced in section 2.2). To study the effect of each proposed methodological component, we integrate the NIR approach in our framework to keep all other components, that are design choices, the same. Thus, for both our SuperF and NIR, we use i) the same INR encoder (i.e., Fourier features with a ReLU MLP instead of Siren), ii) an affine matrix (instead of a homography), iii) the same batch optimization, and iv) the same frame-specific spectral projection. As proposed by Nam et al. (2022), we run NIR for up to 5k iterations. For reference, we also report the performance of a bilinear upsampling of the LR base frame.

Comparison results. We present quantitative results in Table 1 and Appendix Table 5 and show qualitative comparison in Fig. 2 (and Appendix Fig. 10–12). We see that our approach outperforms

Table 1: **Comparison with TTO baselines.** PSNR (↑) for different upscaling factors. Note that the SatSynthBurst fixes the HR output resolution while SyntheticBurst fixes the resolution of the LR frames. Hence, as the upsampling factor increases, we only expect lower performance metrics for SatSynthBurst. *Experimental setup*: upsampling factors ×2, ×4, ×8, 16 LR frames. Standard deviation across samples is given in parentheses and the number of iterations in square brackets.

	SatSynthBurst			SyntheticBurst		
	×2	×4	×8	×2	×4	×8
Bilinear	34.69 (3.50)	29.71 (3.64)	26.62 (3.68)	27.66 (3.50)	26.12 (3.72)	25.44 (3.82)
Lafenetre et al. (2023)	33.46 (3.62)	27.70 (3.79)	24.88 (3.71)	27.02 (3.29)	26.46 (3.05)	25.19 (2.97)
Nam et al. (2022) [2k]	26.26 (3.91)	24.63 (4.41)	23.85 (3.79)	23.62 (4.43)	22.69 (4.41)	22.28 (4.40)
Nam et al. (2022) [5k]	25.65 (5.82)	24.99 (4.12)	23.61 (2.97)	24.46 (4.31)	23.39 (4.32)	22.93 (4.33)
SuperF (ours) [2k]	36.73 (1.66)	32.94 (1.83)	28.87 (2.32)	29.38 (3.43)	27.90 (3.94)	27.08 (3.97)

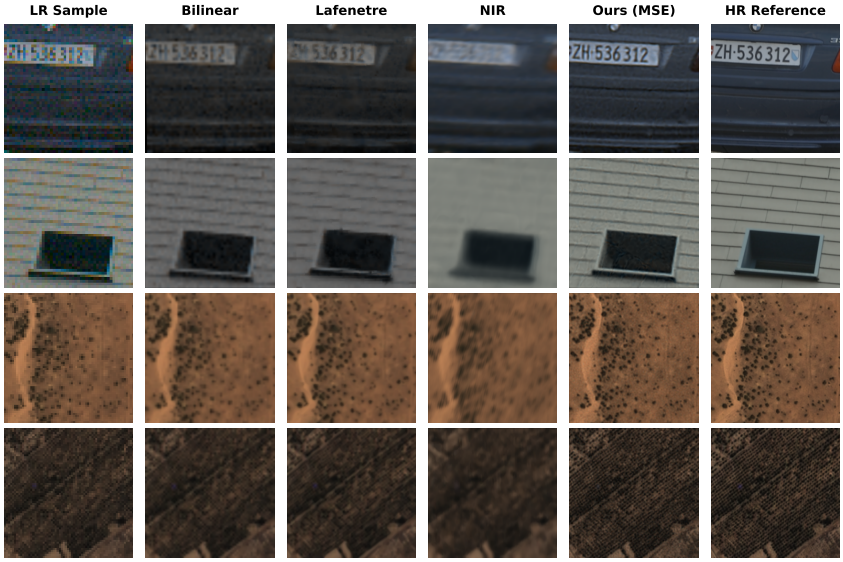


Figure 2: **Qualitative comparison with upsampling factor ×4.** From left to right, we show: one low-resolution (LR) frame, bilinear upsampling, steerable kernel regression (Lafenetre et al., 2023), NIR (Nam et al., 2022), our *SuperF* approach, and the high-resolution (HR) reference.

both baselines. While the approach by Lafenetre et al. (2023) improves compared to the bilinear baseline in terms of LPIPS, it does not yield better PSNR nor SSIM.

Qualitatively, the methods by both Lafenetre et al. (2023) and Nam et al. (2022) seem to be able to smooth and hence denoise the ground-level bursts, but lead to overly smooth results (Fig. 2). Furthermore, we observed that for some satellite scenes, NIR produces a constant output and collapses at the beginning of the optimization. Our proposed approach yields pleasing results that can deal with the high noise-level in the ground-level bursts and represent the high-resolution signal in satellite scenes.

4.4 Ablation studies and sensitivity analyses

To understand the individual components of our proposed methodology, we evaluate the performance gradually turning on each component in Table 2. We start with using our implementation of the INR with just a single LR frame (first row), but compare the resulting high-resolution reconstruction. This base experiment corresponds to an INR *without* using i) the Fourier feature positional encoding (FF); ii) the multiple LR frames (multi-frame); iii) the optimization of the alignment of the LR frames (align);

We confirm that the positional encoding is a crucial aspect for INR also in the MISR setup. However, optimization on a single LR frame is not able to recover any high-resolution details and, in

Table 2: **Ablation studies.** We study the importance of the individual components of our proposed approach. The base experiment (first row) corresponds to an INR *without*: i) the Fourier feature positional encoding (FF); ii) using multiple LR frames, i.e. a single frame (multi-frame); iii) optimizing the alignment the LR frames (align); *Experimental setup*: upsampling factor $\times 4$, 16 LR frames. Standard deviation across samples shown in parentheses.

			SatSynthBurst			**SyntheticBurst**		
FF	multi-frame	align	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
✗	✗	✗	20.33 (2.16)	0.337 (0.160)	0.661 (0.050)	16.63 (2.91)	0.225 (0.138)	0.720 (0.042)
✓	✗	✗	30.42 (3.24)	0.774 (0.073)	0.361 (0.047)	24.69 (3.08)	0.546 (0.107)	0.573 (0.047)
✓	✓	✗	28.11 (3.32)	0.663 (0.120)	0.458 (0.051)	22.83 (3.81)	0.523 (0.153)	0.592 (0.050)
✓	✓	✓	32.94 (1.83)	0.853 (0.025)	0.287 (0.035)	27.87 (3.92)	0.774 (0.102)	0.383 (0.070)

Table 3: **Effect of the proposed components** to advance the NIR baseline (Nam et al., 2022) on the SatSynthBurst dataset. The first row is the NIR baseline without any of our contributions. Variants incrementally add or remove: i) direct parameterization of T ("Direct T"), ii) training with supersampling ("SS"), and iii) using a fixed base frame ("FBF").

Method	Direct T	SS	FBF	PSNR ↑	SSIM ↑	LPIPS ↓	Align. Err. ↓	Iter.
NIR (Nam et al., 2022)	✗	✗	✗	24.63	0.539	0.595	0.650	2000
	✓	✗	✗	26.14	0.580	0.479	0.012	2000
	✗	✓	✗	24.76	0.482	0.593	0.079	2000
	✗	✗	✓	26.39	0.621	0.483	0.319	2000
	✗	✓	✓	24.76	0.482	0.593	1.324	2000
	✓	✗	✓	26.20	0.578	0.476	0.012	2000
	✓	✓	✗	31.30	0.818	0.295	0.012	2000
SuperF (ours)	✓	✓	✓	32.94	0.853	0.287	0.012	2000

fact, performs similar to a bilinear upsampling when used with the FF encoding (comparing Table 2 with Appendix Table 5). Next, we evaluate the performance of optimizing a shared INR on multiple LR frames without optimizing their alignment (third row). Compared to optimizing with a single LR frame, this leads to an even lower performance – a result of the sub-pixel shifts which blurs the signal. Only by optimizing the sub-pixel alignment together with the shared INR, our proposed approach is able to leverage the information in multiple LR frames, which leads to substantial improvement in performance (fourth row). These results hold systematically for both domains, satellite images (SatSynthBurst) and ground-level bursts (SyntheticBurst).

Effect of the proposed components. To understand which of the proposed components lead to an advancement over the closest method by Nam et al. (2022), we investigate the effect by turning each component on and off individually in Table 3. We find that using a direct paremetrization of the affine transformation parameters, instead of using another MLP to estimate the transformation is most crucial, followed by the supersampling strategy, to reduce the sub-pixel alignment error and MISR performance. Fixing the base frame is also complementary, but only when using our direct parametrization. Combined, all three components substantially improve super-resolution performance and reduce the sub-pixel alignment error (euclidean distance).

Sensitivity analyses Our method mainly depends on one key hyperparameter, the scale of the Fourier features σ^2, as described in section 3.3. We show that our algorithm is sensitive to this parameter in Appendix Fig. 7 and 8 and that the best Fourier feature scale depends on the domain. For the satellite image bursts an optimal scale is 10 and for the ground-level bursts it is 3. However, the optimal setting does not depend on the loss and the same setting generalizes across samples in the same domain. Furthermore, we investigate the sensitivity to the number of LR frames in the Appendix section C.2.

4.5 Results on real Sentinel-2 satellite images

We demonstrate that our method can be applied to real-world satellite images from Sentinel-2. We use available Sentinel-2 images from an AWS STAC endpoint, and use the cloud-free samples for super-resolution. These are real-world examples and therefore affected by noise due to lighting variation, changing landcover (e.g. crops), or seasonal variations like snow cover. In scenarios where the noise is dominating, our assumption of repeated observations of the same scene does not hold and further development is needed to account for such high noise levels.

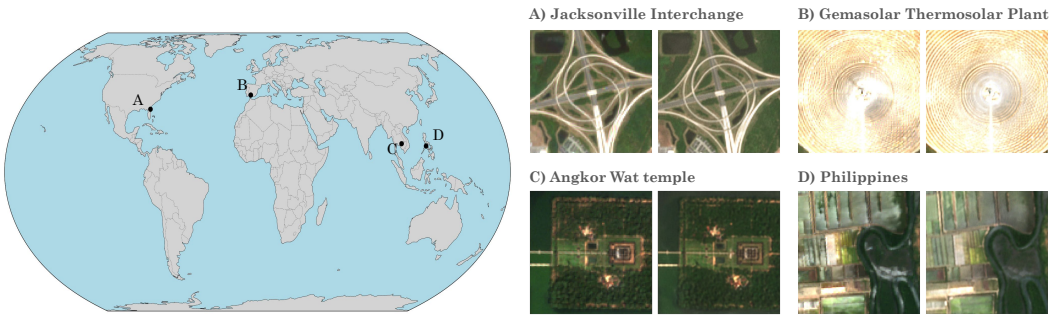


Figure 3: **Qualitative examples using real satellite images.** We demonstrate that our method can align and super-resolve real satellite images from the Sentinel-2 mission by an upsampling factor of 5 using a filtered time series from a Sentinel 2 STAC endpoint. Depending on the cloud cover this leads to a varying number of LR images retrieved within 3–5 months (A: 25, B:15, C:9, D:7).

4.6 Discussion

Our proposed MISR approach, which jointly optimizes the alignment of LR frames and a shared INR, exhibits several advantageous characteristics: i) While existing approaches require a pre-alignment step, SuperF directly works on large shifts by optimizing the alignment in continuous coordinate space (see Appendix section C.4). ii) As a TTO approach, there is no need for any high-resolution training data. This allows SuperF to be applied to new domains without any pretraining. However, some limitations exist.

Limitations. *Run time* may limit certain applications. Although our compact MLP is fairly memory-efficient, the iterative optimization process takes several seconds in our experiments (running non-optimized code). This may pose limitations for mobile device applications, but is less critical for remote sensing scenarios and other scientific and medical applications. A possible way to reduce the number of iterations needed may be to learn the initialization of the INR as shown by Tancik et al. (2021). *Real-world data* can be highly noisy. For instance, satellite imagery may also partially be affected by cloud cover and handheld ground-level bursts may depict changing scenes. We assume that the observations capture the same scene. Occlusions and other drastic changes between frames introduce noise, which requires further analyses. *Risk of overfitting* increases when setting the Fourier features scale hyperparameter too high. While this parameter depends on the domain, it is rather robust across samples within a domain. INR decoders that avoid such a positional embedding such as SIREN (Sitzmann et al., 2020) or WIRE (Saragadam et al., 2023) might be possible alternatives.

Impact on society. The ability to super-resolve publicly available real data like Sentinel-2 (see Fig. 3) enables a vast range of applications anywhere on Earth. This approach can support efforts to address critical societal challenges such as climate adaptation, biodiversity conservation, and food security, for example, by facilitating environmental monitoring of deforestation, tree cover, tree counting, and mapping agricultural fields. However, these technological advances also carry the potential for misuse, including in the context of geopolitical conflicts or resource exploitation.

5 Conclusion

We bring forward an approach to leverage the continuous characteristics of implicit neural representations for multi-image super-resolution. The key characteristic of *SuperF* is to jointly optimize the sub-pixel alignment of multiple low-resolution frames while sharing an INR across all frames. SuperF improves upon existing INR-based burst fusion approaches by optimizing INRs with a direct parameterization of the affine transformations and using a supersampling strategy, which leads to improved sub-pixel alignment and thus MISR performance. As a TTO method, SuperF does not require any high-resolution training data, which facilitates the applicability to new domains and minimizes the risk of hallucinating high-resolution structures.

Acknowledgments

This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1 and by the Global Wetland Center (grant number NNF23OC0081089) from Novo Nordisk Foundation.

References

Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 1977.

Alexander Becker, Rodrigo Caye Daudt, Dominik Narnhofer, Torben Peters, Nando Metzger, Jan Dirk Wegner, and Konrad Schindler. Thera: Aliasing-free arbitrary-scale super-resolution with neural heat fields. *arXiv preprint arXiv:2311.17643*, 2025.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021a.

Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *ICCV*, 2021b.

Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *CVPR*, 2023.

Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *CVPR*, 2023.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021.

Ilya Chugunov, David Shustin, Ruyu Yan, Chenyang Lei, and Felix Heide. Neural spline fields for burst image fusion and layer separation. In *CVPR*, 2024.

Regev Cohen, Idan Kligvasser, Ehud Rivlin, and Daniel Freedman. Looks too good to be true: An information-theoretic analysis of hallucinations in generative restoration models. *NeurIPS*, 2024.

Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *ICML*, 2023.

Julien Cornebise, Ivan Orsolic, and Freddie Kalaitzis. Open high-resolution satellite imagery: The WorldStrat dataset – with application to super-resolution. In *NeurIPS*, 2022.

Riccardo De Lutio, Stefano D'aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *ICCV*, 2019.

Riccardo De Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *CVPR*, 2022.

Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *CVPRW*, 2021.

Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 1997.

Amer Essakine, Yanqi Cheng, Chun-Wun Cheng, Lipei Zhang, Zhongying Deng, Lei Zhu, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Where do we stand with implicit neural representations? A technical and performance survey. *Transactions on Machine Learning Research*, 2025.

Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 1991.

Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. NVRC: Neural video representation compression. *NeurIPS*, 2024.

Jamy Lafenetre, Ngoc Long Nguyen, Gabriele Facciolo, and Thomas Eboli. Handheld burst super-resolution meets multi-exposure satellite imagery. In *CVPR Workshops*, 2023.

Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Kangfu Mei, Hossein Talebi, Mojtaba Ardakani, Vishal M Patel, Peyman Milanfar, and Mauricio Delbracio. The power of context: How multimodality improves image super-resolution. In *CVPR*, 2025.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *CVPR*, 2023.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Seonghyeon Nam, Marcus A Brubaker, and Michael S Brown. Neural image representations for multi-image fusion and layer separation. In *ECCV*. Springer, 2022.

Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised super-resolution for multi-exposure push-frame satellites. In *CVPR*, 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

Naama Pearl, Tali Treibitz, and Simon Korman. Nan: Noise-aware nerfs for burst-denoising. In *CVPR*, 2022.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *CVPR*, 2023.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020.

Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit neural representations for image compression. In *ECCV*. Springer, 2022.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *CVPR*, 2021.

Roger Y. Tsai and Thomas S. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, 1984.

Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. NeRF-SR: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (ToG)*, 2019.

Qing Wu, Yuwei Li, Lan Xu, Ruiming Feng, Hongjiang Wei, Qing Yang, Boliang Yu, Xiaozhao Liu, Jingyi Yu, and Yuyao Zhang. IREM: High-resolution magnetic resonance image reconstruction via implicit neural representation. In *MICCAI*, 2021.

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. In *ECCV*, 2022.

Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, Junpei Zhang, Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In *CVPRW*, 2023.

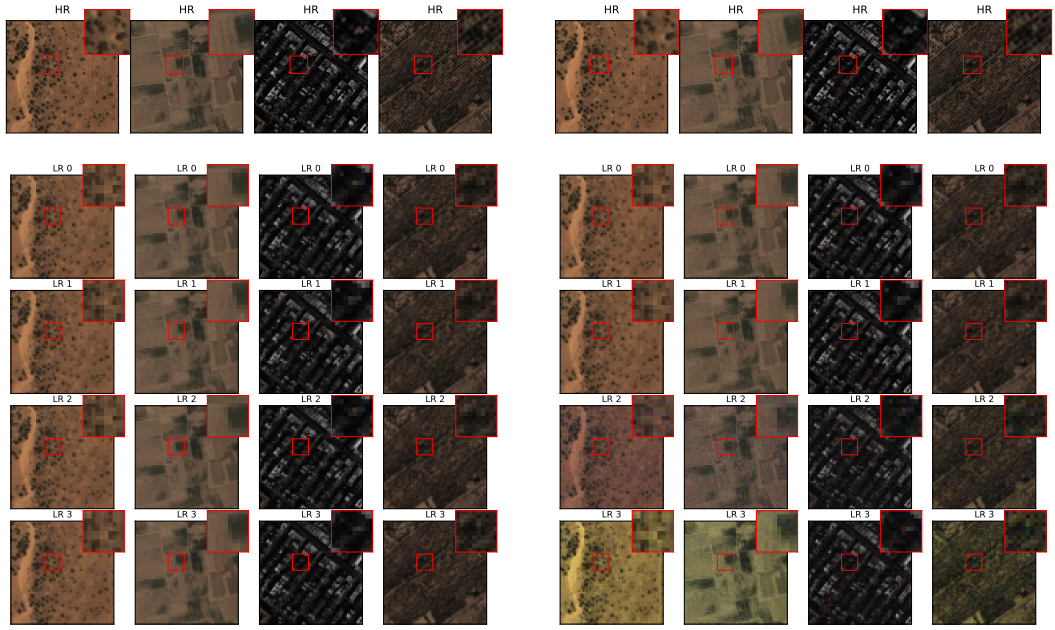
Jinchen Zhu, Mingjian Zhang, Ling Zheng, and Shizhuang Weng. Multi-scale implicit transformer with re-parameterization for arbitrary-scale super-resolution. *Pattern Recognition*, 2025.

A Dataset creation and evaluation procedure

In this section we provide details on i) the downsampling of high-resolution satellite images to create synthetic bursts of slightly shifted low-resolution images and ii) the postprocessing needed for evaluating the predicted high-resolution images.

A.1 Creation of the SatSynthBurst dataset (satellite imagery)

We constructed a synthetic burst dataset derived from 20 open high-resolution satellite images selected from the WorldStrat dataset (Cornebise et al., 2022) (see examples in Fig. 4). The high-resolution images from Airbus SPOT 6/7 satellite with a ground sampling distance (GSD) of up to 1.5 m are published under a CC BY-NC 4.0 license[4], which allows us to publicly redistribute our SatSynthBurst datasets under the same license for non-commercial purposes. We aim to simulate low-resolution images comparable to the Sentinel-2 mission, but at varying spatial resolutions allowing to study downsampling factors s of 2, 4, and 8. To simulate variation in the imaging conditions that could occur between images captured over several weeks, we incorporate spectral augmentations and additive Gaussian noise. Additionally, we follow the work by Lanaras et al. (2018) for generating synthetic super-resolution data using the modulation transfer function (*mtf*) of the Sentinel-2 sensor. Hence, before downsampling, we blur the high-resolution images with a Gaussian filter of standard deviation $u = 1/s$ pixels, which emulates the *mtf* of Sentinel-2 and, thus, the effective point spread function (*psf*) which is described as $psf = \sqrt{-2\log(mtf)/\pi^2}$. This is followed by an average pooling with a window size of $s \times s$.



(a) Without spectral augmentation (b) With spectral augmentation

Figure 4: **Examples of the SatSynthBurst dataset (factor $\times$4).** The top row shows the underlying high-resolution (HR) image. Below we show four slightly misaligned low-resolution (LR) frames.

A.2 Postprocessing for evaluation

We follow common practice in evaluating MISR results and use a spectral alignment proposed by Bhat et al. (2021a) to correct any spectral mismatch between the high-resolution prediction and the test image. Metrics like the PSNR and SSIM are rather sensitive to small misalignments, whereas LPIPS is more robust. Furthermore, we follow the evaluation protocol of Bhat et al. (2021a) and

[4] `https://creativecommons.org/licenses/by-nc/4.0/` (accessed: 2025-05-20)

Figure 5: **Examples of the SyntheticBurst dataset (factor $\times$8).** The top row shows the underlying high-resolution (HR) image. Below we show four slightly misaligned low-resolution (LR) frames.

mask out a buffer of 16 boundary pixels to avoid the effect of any boundary artifacts in the dataset (specifically, the SyntheticBurst dataset).

A.3 Evaluation on SyntheticBurst (ground-level imagery)

Unlike our SatSynthBurst dataset, the SyntheticBurst dataset (Bhat et al., 2021a) does not provide a base frame which is spatially aligned with the high-resolution test image (see examples in Fig. 5). Thus, an additional postprocessing step is needed, before a predicted HR image can be evaluated on the given HR test image. Therefore, we employ a brute force spatial alignment strategy to align the predicted image with the test image using an affine transformation consisting of rotation and a spatial translation. Our strategy selects the optimal translation within a 4×4 pixel neighborhood and the optimal rotation angle within a range of $[0, 4]$ degrees.

We have experimented with the alignment strategy presented by Bhat et al. (2021a), that uses a trained PWC-Net (Sun et al., 2018) to estimate the optical flow from the prediction to the reference test image. However, this led to artifacts in the warped prediction, which is why we chose the brute force postprocessing.

B Implementation details

We summarize the hyperparameter settings for both datasets in Table 4. The only hyperparameter that is differs between datasets is the Fourier feature scale.

C Additional results

C.1 Comparison of baselines with additional metrics

We provide additional evaluation metrics including PSNR, SSIM, and LPIPS for the baseline comparison in Table 5.

Table 4: **Hyperparameter settings.**

Hyperparameters	SatSynthBurst	SyntheticBurst
LR resolution	128 / 64 / 32	48
HR resolution	256	96 / 192 / 384
Optimizer	AdamW	
Learning rate sched.	Cosine annealing	
Learning rate base	2×10^{-3}	
Learning rate min	1×10^{-6}	
Weight decay	0.05	
Adam β	(0.9, 0.999)	
Batch size	1 LR frame per iteration	
Training iterations	2000	
Fourier feature scale	10	3
positional encoding dim	256	
INR decoder	MLP (4 layers, ReLU, dim=256)	

Table 5: **Comparison of test-time optimization methods.** Ours uses Fourier feature with scale 10 for SatSynthBurst (satellite) and scale 3 for SyntheticBurst (ground-level). *Experimental setup*: upsampling factor $\times 4$, 16 LR frames. Standard deviation across samples is given in parentheses and the number of iterations in square brackets.

	SatSynthBurst			SyntheticBurst		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bilinear	29.71 (3.64)	0.746 (0.104)	0.382 (0.043)	26.12 (3.72)	0.703 (0.121)	0.455 (0.077)
Lafenetre et al. (2023)	27.70 (3.79)	0.680 (0.130)	0.261 (0.055)	26.46 (3.05)	0.664 (0.121)	0.384 (0.118)
Nam et al. (2022) [2k]	24.63 (4.42)	0.539 (0.175)	0.595 (0.076)	22.69 (4.41)	0.576 (0.171)	0.616 (0.089)
Nam et al. (2022) [5k]	24.99 (4.13)	0.544 (0.167)	0.587 (0.082)	23.39 (4.32)	0.606 (0.165)	0.574 (0.090)
SuperF (ours) [2k]	32.94 (1.83)	0.853 (0.035)	0.287 (0.054)	27.90 (3.95)	0.774 (0.102)	0.385 (0.070)

C.2 Sensitivity to the number of LR frames

We study the sensitivity to the number of available LR frames in Fig. 6. This is a critical aspect for both application domains. Handheld bursts might be limited in the number of frames since the scene might change for long overall exposure times. Satellite imagery like Sentinel-2 are captured with a revisit period of ≈5 days. We thus need to consider longer time windows to obtain multi-frame satellite images. However, longer time windows may lead to changing appearance of the scene due to activity on the ground or seasonality, which will hinder MISR. Furthermore, cloud-free images may be scarce, depending on the geographic region.

For the SatSynthBurst dataset, we observe that the PSNR saturates with 8 samples for the factor $\times 2$, but keeps increasing slightly when using 16 samples for the larger upsampling factors. In contrast, the PSNR for the ground-level bursts keeps improving with more frames even for the factor $\times 2$. Additional frames may help to reduce the high noise level in the ground-level bursts.

C.3 Sensitivity to Fourier feature scale

As shown in the main paper, our method is sensitive to the Fourier feature scale, and the optimal hyperparameter depends on the domain, i.e. satellite imagery and ground-level bursts. We show the qualitative effect of the different Fourier features scales in Fig. 8. Setting the scale too low leads to over-smoothing, whereas setting it too high leads to grainy artifacts. However, we find that a single parameter setting performs well across samples within a domain. We use the optimal setting for upsampling factor 4 for all experiments including factor 2 and 8 (see hyperparameter setting in Table. 4).

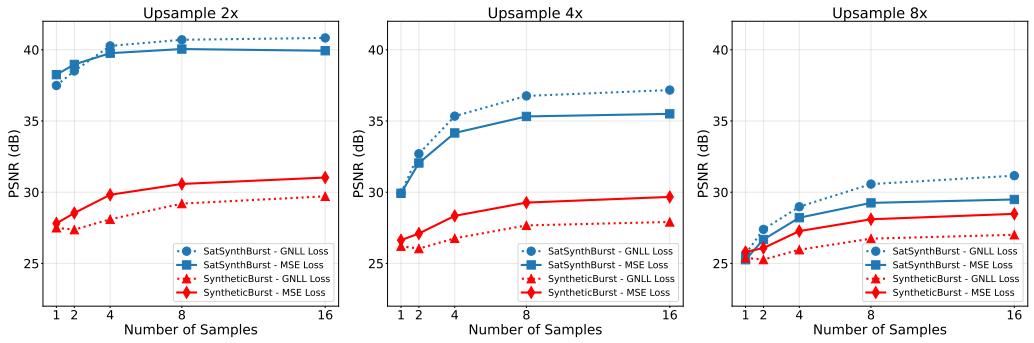
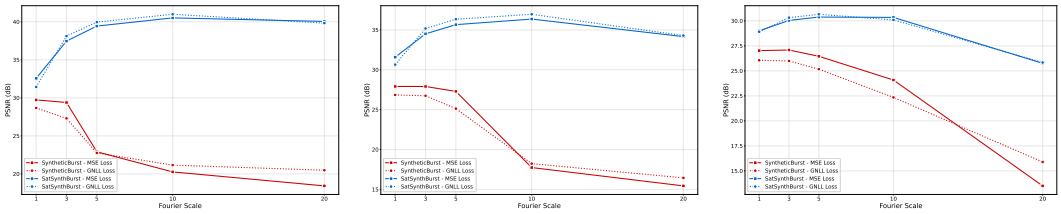


Figure 6: **Sensitivity to the number of LR frames.** From left to right, we report PSNR for upsampling factors 2, 4, and 8 by varying the number of LR frames on the horizontal axis.



(a) Upsample 2× (b) Upsample 4× (c) Upsample 8×

Figure 7: **Sensitivity analysis of the Fourier feature scale.** The optimal hyperparameter depends on the domain, i.e. satellite imagery (SatSynthBurst in blue) and ground-level bursts (SyntheticBurst in red) require different settings. However, the optimal setting is invariant to the loss. For SyntheticBurst we see a small difference between the upsamling factor experiments. However, we note that the two datasets differ in the strategy of creating different upsampling factors. The SyntheticBurst varies the HR output resolution with a fixed resolution of the LR frames. Hence, the absolute output resolution may affect the optimal Fourier feature scale. In contrast, the SatSynthBurst dataset fixes the HR output resolution and varies the resolution of the LR frames.

C.4 Ability to align large shifts without pre-processing

While existing methods (Wronski et al., 2019; Lafenetre et al., 2023) rely on a pre-alignment procedure to first reduce the misalignment to sub-pixel shifts, our method can directly work on multi-pixel shifts. Although the PSNR drops consistently with increased shifts, when comparing the results of bursts with sub-pixel shifts against bursts with shifts up to 4 LR pixels (see Table 6), our method keeps outperforming the baseline.

C.5 Experimental setup

We follow standard practices in super-resolution and report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (using AlexNet), computed using the implementation by Bhat et al. (2021a). All experiments are implemented in PyTorch and executed on a single NVIDIA H100 GPU with 80 GB of VRAM (note, our experiments typically need around 1GB of VRAM). If not further specified, all experiments use the AdamW optimizer with a base learning rate of 2×10^{-3}, which is decayed to 1×10^{-6} over 2000 iterations using a cosine annealing schedule and a batch size of 1 frame.

During evaluation, a 16-pixel boundary is cropped from all sides to reduce edge artifacts. We additionally apply color matching as a post-processing step, following Bhat et al. (2021a), to correct for global color and intensity shifts between the reconstruction and the ground truth. The scale hyperparameter of the Fourier feature positional encoding is set to 10 for the SatSynthBurst and to 3 for the SyntheticBurst dataset. However, our method breaks in the extreme case with upsampling factor 8, i.e. the shifts of 4 LR pixels correspond to 32 pixels in the high-resolution image of size 256×256

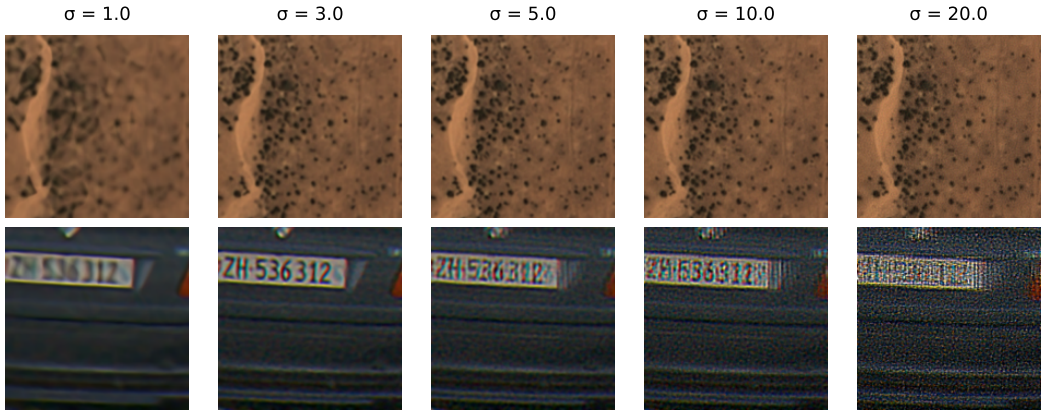


Figure 8: **Effect of the Fourier feature scale σ for MSE (upsampling factor $\times$4).** The optimal hyperparameter depends on the domain, i.e. satellite imagery and ground-level bursts require different settings. Setting the scale too low leads to over-smoothing, whereas setting it too high leads to grainy artifacts.

Table 6: **Comparing sub-pixel with large misalignments (PSNR).** We compare our SuperF results on bursts with sub-pixel shifts, i.e. max shifts of 1.0 LR pixels, vs. bursts with large shifts up to 4.0 LR pixels. *Experimental setup*: upsampling factor $\times$2, $\times$4, $\times$8; 16 LR frames. Standard deviation across samples shown in parentheses.

		SatSynthBurst		
		$\times$2	$\times$4	$\times$8
	Bilinear	34.73 (3.65)	29.81 (3.88)	26.83 (4.03)
max shift: 1.0 LR pixels	SuperF (ours)	39.93 (2.49)	35.50 (2.39)	29.49 (2.71)
max shift: 4.0 LR pixels	SuperF (ours)	38.24 (2.56)	31.49 (4.13)	18.28 (6.37)

pixels (i.e., >12% relative shift). Further investigation is needed to study if this issue could be resolved with a different hyperparameter setting, e.g. by increasing the number of iterations or the learning rate.

C.6 Qualitative results for real Sentinel-2 satellite image data

We demonstrate that our method can be applied to real world satellite images from the publicly available Sentinel-2 satellite images. We use available Sentinel-2 images from a AWS STAC endpoint, and use the cloud-free samples for super-resolution. These are real-world examples and therefore affected by noise due to lighting variation, changing landcover (e.g. crops), or seasonal variations like snow cover. In scenarios where the noise is dominating, our assumption of repeated observations of the same scene does not hold and further development is needed to account for such high noise levels.

C.7 Qualitative results for different upsamling factors

We provide additional qualitative comparisons for both satellite image and ground-level image bursts at upsampling factor $\times$2 (Fig. 10), $\times$4 (Fig. 11), and $\times$8 (Fig. 12). We note that the two datasets differ in the strategy of creating versions with different upsampling factors. The SatSynthBurst dataset fixes the HR output resolution and varies the resolution of the LR frames. Hence, we expect lower performance metrics for SatSynthBurst as the upsampling factor increases. In contrast, the SyntheticBurst varies the HR output resolution with a fixed resolution of the LR frames. Thus, the metrics are not comparable across upsampling factors.

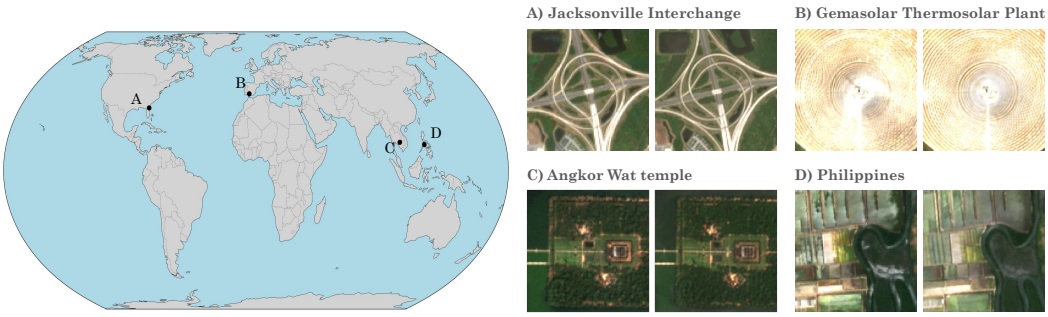


Figure 9: **Qualitative examples using real satellite images.** We demonstrate that our method can align and super-resolve real satellite images from the Sentinel-2 mission by an upsampling factor of 5 using ”clean” time series from a Sentinel 2 STAC endpoint. *Top right*: Gemasolar Thermosolar Plant (15 LR images). *Middle right*: Phillipines (7 LR images). *Bottom right*: Angkor Wat (9 LR images). *Left*: Jacksonville Interchange (25 LR images).

We find that for the satellite imagery, our results for the upsampling factor $\times 8$ start to become grainy. This may be a result of overfitting with a suboptimal hyperparameter setting for the Fourier feature scale.

D Use of Large Language Models

We used LLMs as search engines to support literature research and as coding assistants.

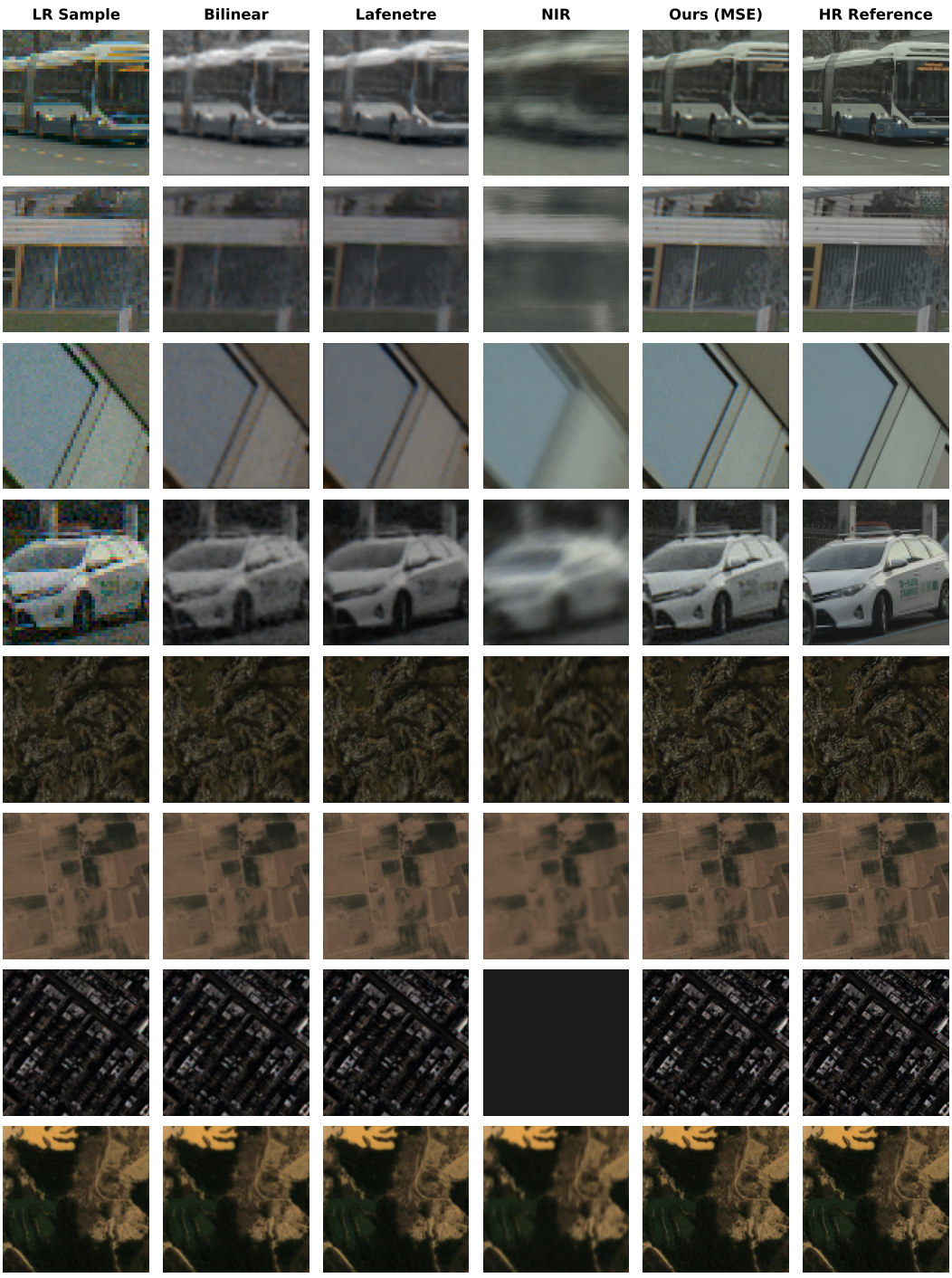


Figure 10: **Qualitative comparison with upsampling factor $\times 2$.** From left to right, we show: one low-resolution (LR) frame, bilinear upsampling, steerable kernel regression (Lafenetre et al., 2023), NIR (Nam et al., 2022), our *SuperF* approach, and the high-resolution (HR) reference.

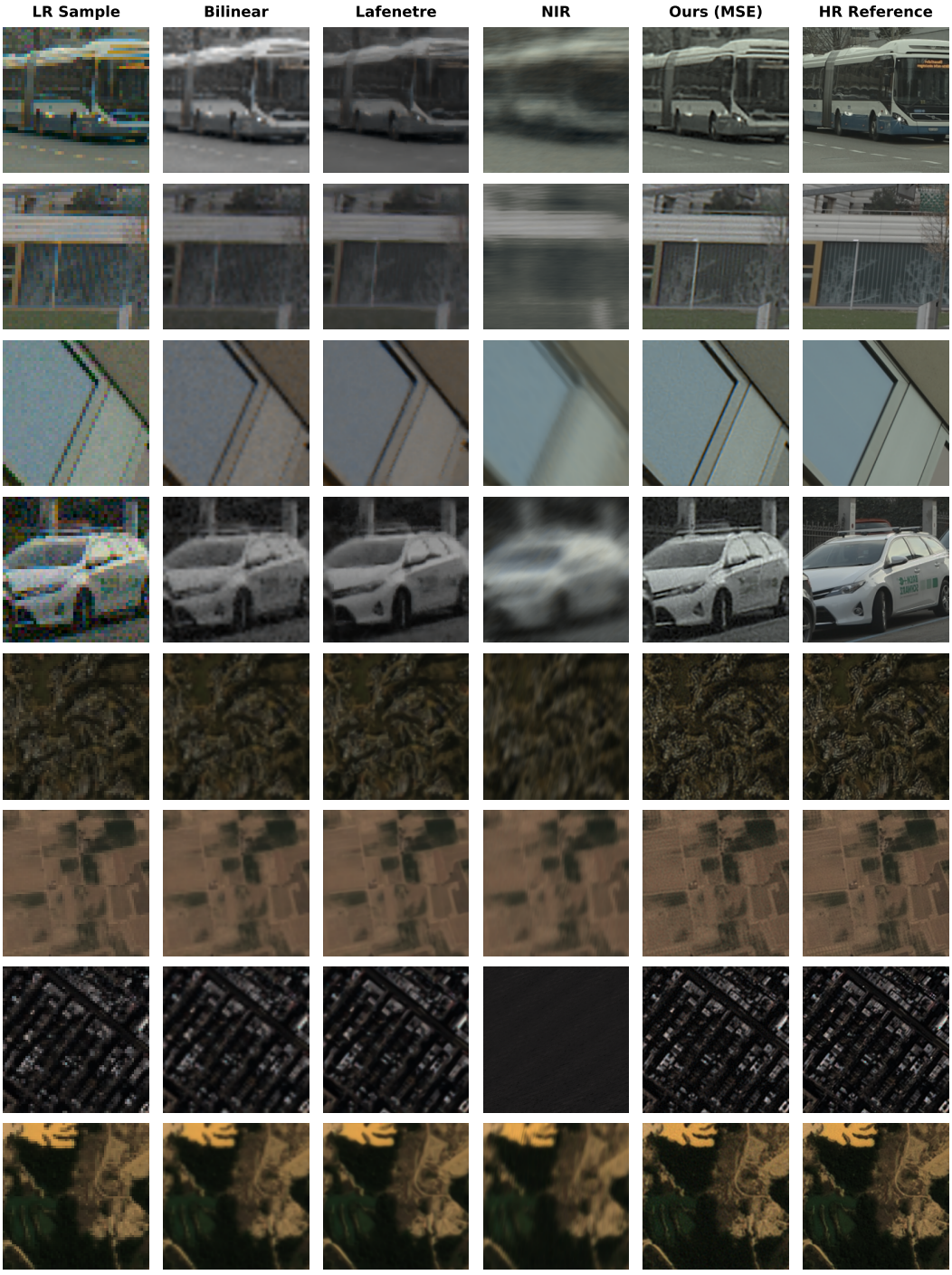


Figure 11: **Qualitative comparison with upsampling factor $\times$4.** From left to right, we show: one low-resolution (LR) frame, bilinear upsampling, steerable kernel regression (Lafenetre et al., 2023), NIR (Nam et al., 2022), our *SuperF* approach, and the high-resolution (HR) reference.

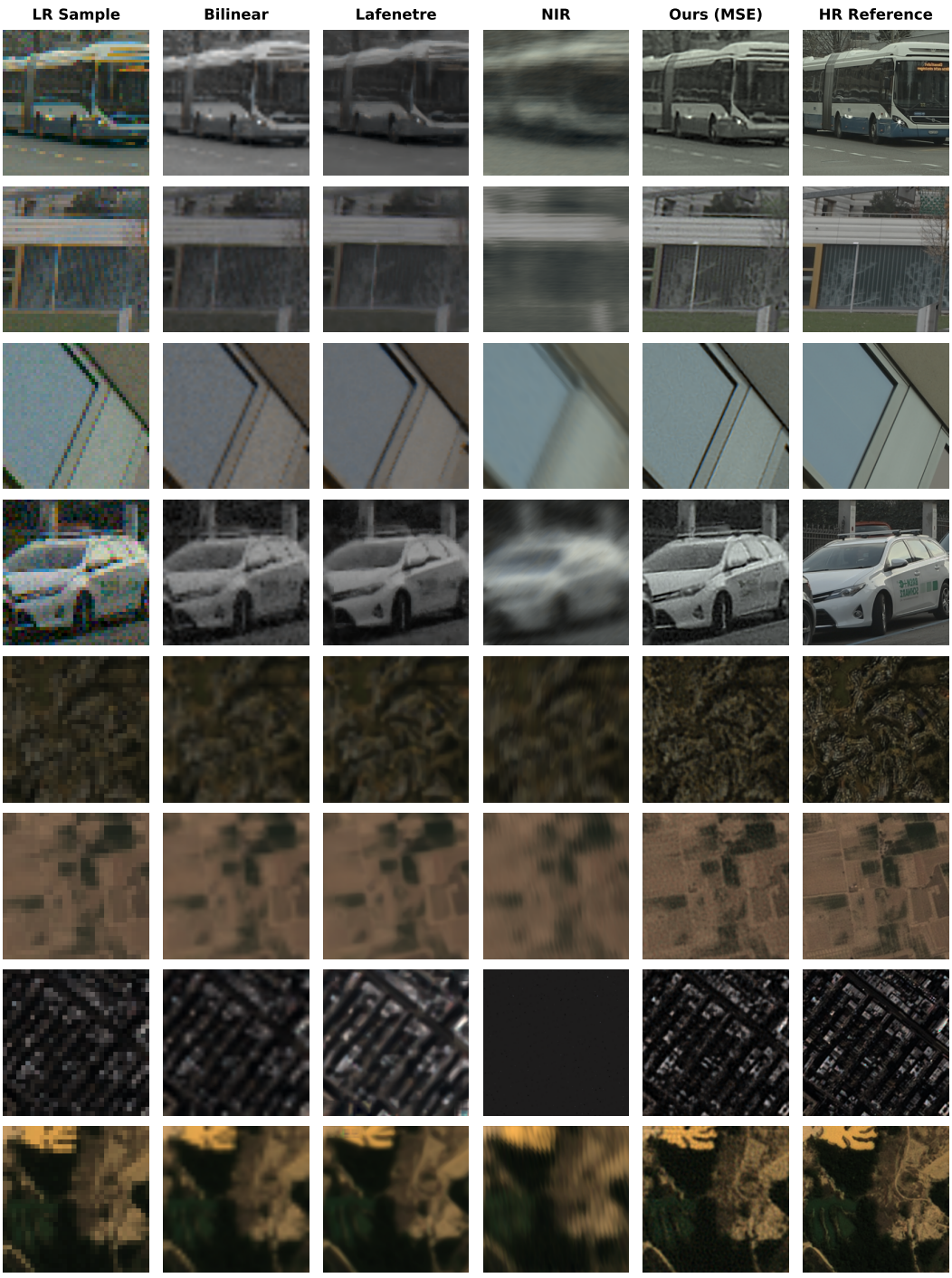


Figure 12: **Qualitative comparison with upsampling factor $\times$8.** From left to right, we show: one low-resolution (LR) frame, bilinear upsampling, steerable kernel regression (Lafenetre et al., 2023), NIR (Nam et al., 2022), our *SuperF* approach, and the high-resolution (HR) reference.