# SuperF: Neural Implicit Fields for Multi-Image Super-Resolution

**Sander Riisøen Jyhne**[*]
University of Agder

**Christian Igel**
University of Copenhagen

**Morten Goodwin**
University of Agder

**Per-Arne Andersen**
University of Agder

**Serge Belongie**
University of Copenhagen

**Nico Lang**[*]
University of Copenhagen

## 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. Since single-image super-resolution requires solving an inverse problem, such methods must exploit strong priors, e.g. learned from high-resolution training data. While qualitatively pleasing, such approaches are prone to "hallucinated" structures. In contrast, multi-image super-resolution (MISR) aims to improve the resolution by constraining the process with multiple views. We propose *SuperF*, a test-time optimization MISR approach that leverages the continuous characteristics of implicit neural representations (INR). It shares an INR for multiple shifted low-resolution frames while jointly optimizing the frame alignment. Our experiments yield compelling results on simulated and real bursts of satellite imagery as well as on ground-level images. As SuperF does not rely on high-resolution training data, it can be applied to any place on Earth.

## 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 [14, 28], or through auxiliary guidance from complementary modalities [7, 8, 17, 15]. 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 [4]. This may be tolerable for smartphone applications, but not for applications in 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 [26, 10, 9]. 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 [27]. While MISR can be approached with *supervised learning-based* methods [1, 2] when large training datasets with paired LR and HR data are available, MISR can also be achieved by *test-time optimization (TTO)* approaches that do not require offline training [27, 13]. 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 [1]. Typically, MISR is associated with bursts of images captured in rapid

[*] Corresponding authors: sander.jyhne@kartverket.no, nila@di.ku.dk

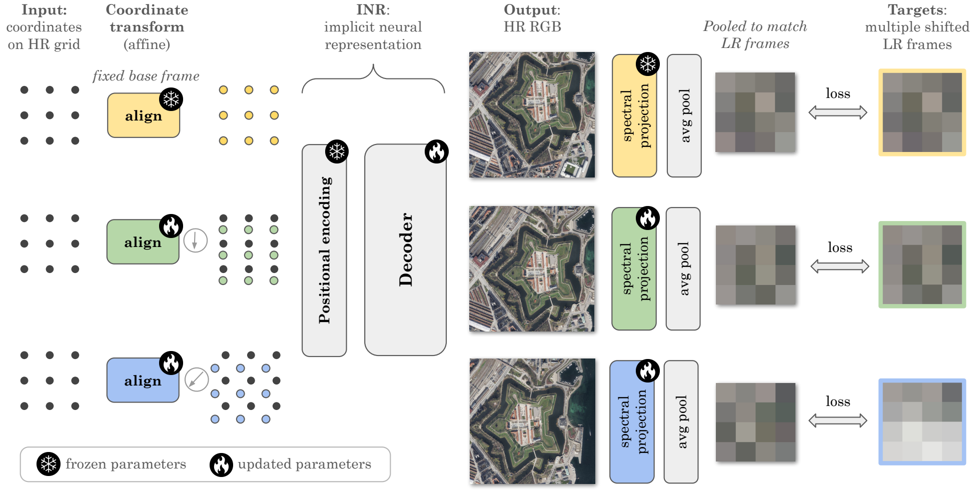

Figure 1: **Illustration of the proposed method.** *SuperF* achieves multi-image super-resolution by sharing an implicit neural representation (INR) across multiple low-resolution (LR) frames with sub-pixel shifts. The LR frames are aligned by jointly optimizing an affine coordinate transformation for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

succession, but repeated observations in satellite remote sensing also provide a multi-frame scenario with longer time intervals.

In this work, we introduce *SuperF*, a TTO 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 , 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 [24, 12], 3D shape modeling [20, 16], novel-view synthesis with neural radiance fields (NeRF) [18], and burst fusion for denoising [21] or layer separation of obstructions and background scenes [19, 3]. INRs have also found their application in Earth observation tasks to encode geospatial data based on geographic coordinates [5, 11, 22].

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 [27, 13]. 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. [7] and turn the problem formulation up-side down and treat the LR frames as reconstruction targets. While Nam et al. [19] 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 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.

## 2 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 $\mathbf{y}_{\text{LR}}^{(1)}, \dots, \mathbf{y}_{\text{LR}}^{(T)}$ in discretized form, i.e., we are given the values at a finite 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 $\varphi$ is a boxcar filter. 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 $\mathbf{w}_i, \mathbf{w}_j \in \mathcal{W}$.

**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 [27], 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, $\hat{\mathbf{y}}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{v}) = \hat{\rho}^{(t)}(f_{\boldsymbol{\theta}}(\hat{\mathbf{A}}^{(t)}\mathbf{v}))$. 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, [19] 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 $\rho^1$ is also fixed (scale 1 and shift 0).

**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

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

equals $\mathbf{y}_{\text{LR}}^{(t)}(\boldsymbol{v})$ on $\boldsymbol{v} \in \mathcal{W}$. We fix the boxcar filter $\varphi$, 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 $\ell(\hat{\mathbf{y}}_{\text{LR},\,\boldsymbol{\theta}}^{(t)}(\boldsymbol{v}), \mathbf{y}_{\text{LR}}^{(t)}(\boldsymbol{v}))$ for all $\boldsymbol{v} \in \mathcal{W}$ for all $T$ frames. The convolution with the boxcar filter and the sampling at grid points $\mathcal{W}$ is implemented by an average pooling. We use MLPs with ReLU activation functions and stochastic gradient descent with mini batches of frames.

## 3 Experimental results and discussion

**Datasets and baselines.** Our experiments are based on datasets from two domains: remote sensing and handheld cameras. First, we create the SatSynthBurst dataset from 20 open high-resolution satellite images selected from the WorldStrat dataset [6]. Second, we demonstrate that our SuperF approach also generalizes to ground-level bursts from handheld cameras using the SyntheticBurst dataset [1].

As SuperF uses TTO, we compare to a state-of-the-art TTO method for MISR, a steerable kernel regression method by [13], which is based on [27]. We also compare to a burst fusion approach by [19] (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. 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 [25] with a ReLU MLP instead of

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 $\times 2$, $\times 4$, $\times 8$, 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. [13] | 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. [19] [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. [19] [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) |
| Ours MSE [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) |

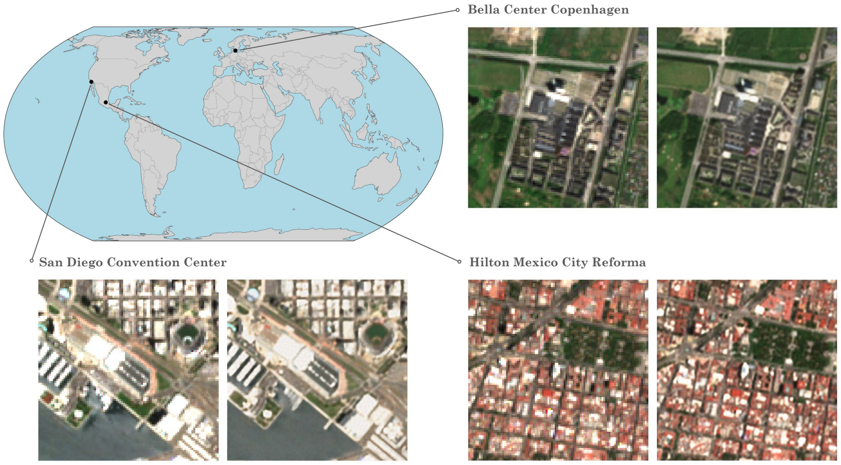

Figure 2: **Qualitative examples.** Real Sentinel-2 (left) and SuperF ×5 super-resolution (right).

Siren [23]), 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. [19], we run NIR for up to 5k iterations. The experimental setup is given in Appendix A.

**Quantitative results.** We present quantitative results in Table 1. Interestingly, the TTO baselines cannot outperform a simple bilinear baseline on PSNR. While [13] has been designed for MISR for satellite imagery, [19] is not explicitly designed for MISR, but for layer-separation at the original resolution. Our approach outperforms both baselines across several upsampling factors (see Table 1).

**Super-resolving any place on Earth.** Applied to real satellite bursts from Sentinel-2 images, SuperF yields qualitatively compelling results and can deal with the high noise-levels (see Fig. 2). Given a location, we retrieve Sentinel-2 images and apply a strict cloud-filter to select the valid frames within three months, leading to 10–25 images per burst.

**Discussion.** Our proposed MISR approach, which jointly optimizes the alignment of LR frames and a shared INR, exhibits several advantageous characteristics: 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 and reduces the risk of hallucinating structures. However, some limitations exist. Our method depends on one key hyperparameter, the scale of the Fourier features. 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. Although our compact MLP is fairly memory-efficient, the iterative optimization process takes several seconds in our experiments. This may pose limitations for mobile device applications, but is less critical for remote sensing applications. 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 repeated observations capture the same scene. Occlusions and other drastic changes between frames introduce noise, which requires further analyses.

## 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

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

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

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

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

[11] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *AAAI*, 2025.

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

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

[14] 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.

[15] 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.

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

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

[18] 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.

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

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

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

[22] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *ICLR*, 2024.

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

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

[25] 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.

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

[27] 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*, 2019.

[28] 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.

## A Experimental setup

We follow standard practices in super-resolution and report Peak Signal-to-Noise Ratio (PSNR). 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 $2 \times 10^{-3}$, which is decayed to $1 \times 10^{-6}$ 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. [1], 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.