

# Segmentation-Aware Latent Diffusion for Satellite Image Super-Resolution: Enabling Smallholder Farm Boundary Delineation

Aditi Agarwal<sup>1</sup> Anjali Jain<sup>2</sup> Nikita Saxena<sup>1</sup> Ishan Deshpande<sup>1</sup> Michal Kazmierski<sup>1</sup>  
 Abigail Annkah<sup>3</sup> Nadav Sherman<sup>3</sup> Karthikeyan Shanmugam<sup>1</sup> Alok Talekar<sup>1</sup>  
 Vaibhav Rajan<sup>1</sup>

<sup>1</sup>Google DeepMind, <sup>2</sup>Google, <sup>3</sup>Google Research  
 aditie@google.com

## Abstract

*Delineating farm boundaries through segmentation of satellite images is a fundamental step in many agricultural applications. The task is particularly challenging for smallholder farms, where accurate delineation requires the use of high resolution (HR) imagery which are available only at low revisit frequencies (e.g., annually). To support more frequent (sub-) seasonal monitoring, HR images could be combined as references (ref) with low resolution (LR) images – having higher revisit frequency (e.g., weekly) – using reference-based super-resolution (Ref-SR) methods. However, current Ref-SR methods optimize perceptual quality and smooth over crucial features needed for downstream tasks, and are unable to meet the large scale-factor requirements for this task. Further, previous two-step approaches of SR followed by segmentation do not effectively utilize diverse satellite sources as inputs. We address these problems through a new approach, **SEED-SR**, which uses a combination of conditional latent diffusion models and large-scale multi-spectral, multi-source geo-spatial foundation models. Our key innovation is to bypass the explicit SR task in the pixel space and instead perform SR in a segmentation-aware latent space. This unique approach enables us to generate segmentation maps at an unprecedented  $20\times$  scale factor, and rigorous experiments on two large, real datasets demonstrate up to 25.5% and 12.9% relative improvement in instance and semantic segmentation metrics respectively over approaches based on state-of-the-art Ref-SR methods.*

## 1. Introduction

Satellite remote sensing provides invaluable data at a global scale in many scientific domains. In agriculture and environmental studies, it can drive improvements in multiple applications such as land cover mapping [12], crop iden-

tification [1] agricultural monitoring [38] and crop yield estimation [46] and agricultural drought monitoring [52]. Effective solutions to these problems can have significant impact on global food security and environment [39]. A foundational step in several agricultural applications is *field boundary delineation*, the task of predicting the polygonal boundaries and constituent areas of crop fields from overhead satellite images [24, 35, 54]. The problem is especially challenging for *smallholder farms*, containing fields smaller than 2 hectares [15]; and is crucial for the Global South comprising about two-thirds of the developing world’s 3 billion rural people in about 475 million households [39].

The task is formulated as an image segmentation problem and several deep learning models, developed for natural images, have been adopted for satellite images [4, 12, 24, 35, 54, 57], which differ from natural (RGB) images in several aspects such as their geometry, atmospheric effects, spectral bands and radiometric information [31].

An important practical constraint is posed by the spatio-temporal tradeoff in this context – limitations in remote sensing technologies and high hardware costs prohibit us from obtaining simultaneously high temporal resolution and high spatial resolution images on a large scale [36, 69]. E.g., very high resolution (VHR) satellite data (e.g.,  $\leq 1$  m/px Ground Sample Distance (GSD), from Airbus, WorldView satellites) which contains detailed spatial information, essential for fine-grained segmentation of smallholder farms, is only available (bi-)annually; while low resolution (LR) satellite data (e.g., 10-30 m/px GSD from Sentinel-2, Landsat-8, PlanetScope satellites) are available every 1-8 days but lack details necessary for this task. Since agricultural processes are dynamic and seasonal, field boundaries may change at higher temporal frequency (up to bi-weekly [14], e.g., see Fig. 1) and delineating them from LR satellite images is nearly impossible for many smallholder farms.

To address this problem, Reference-based Super-Resolution (Ref-SR) techniques can be used which utilize

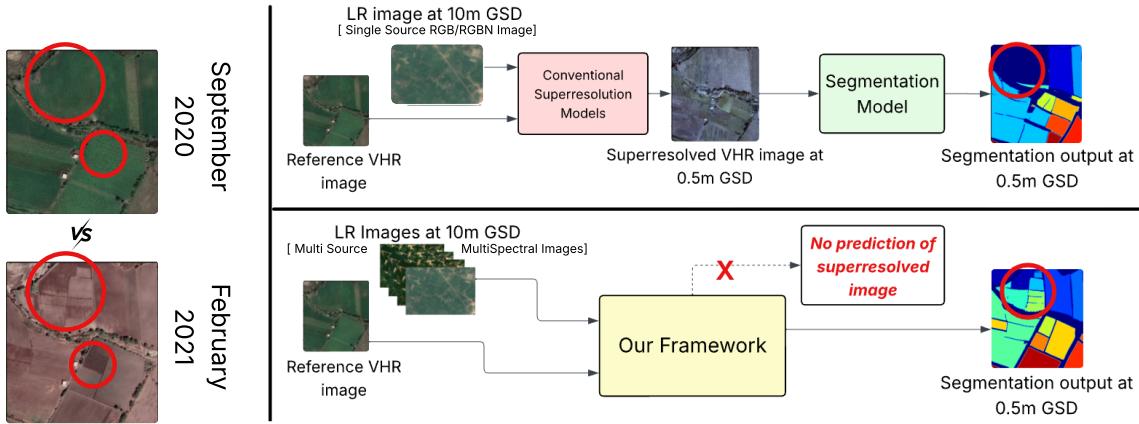


Figure 1. (Left) Example of subtle temporal changes in agricultural field boundaries. (Right) SEED-SR (below) directly generates high-resolution segmentation from an older VHR reference and a recent LR temporal stack, unlike current (above) 2-step super-resolution and segmentation. Best viewed in colour.

image super-resolution (SR) methods to obtain HR images from one or more LR images at a given time period, with guidance from previously available co-located HR image. Reconstructing semantically correct and perceptually plausible HR images via Ref-SR requires (i) adequate semantic and texture modeling of land cover changes between Ref and LR, and (ii) in the satellite context, modeling the differences in sensors used which make the LR and HR images aligned only at a semantic level and not at pixel level. These issues have been addressed in various deep learning models have been developed for Ref-SR [2, 8–10, 20, 22, 34, 60, 63, 67, 68]. The desired segmentation is obtained from the super-resolved image (see fig. 1).

However, three challenges remain, which hinder the effective use of such Ref-SR based approaches for smallholder farm boundary delineation. First, the scale factor of existing methods remain limited to less than  $16\times$  (see Sec. 2), which is insufficient to discern subtle shifts in field boundaries (see fig. 1). Since pixel-based SR methods are optimized for perceptual quality and not task-specific accuracy, high-frequency image features related to fine-scale and inter-field boundary integrity are lost. Second, most SR methods use only RGB inputs, with some using an additional NIR band, and do not effectively utilize multi-spectral, multi-temporal information available from diverse Earth Observation data sources. Finally, a limitation found in most previous works is due to the use of the downsampled versions of HR images as ‘proxies’ for real LR images. In reality, HR and LR images are rarely acquired simultaneously, leading to genuine temporal changes on the ground (e.g., vegetation phenology, new construction, transient objects) rather than resolution differences only. Furthermore, images from different sensors inherently possess distinct characteristics, creating a ‘style gap’ in terms of color rendition, texture, and radiometric properties. Models trained solely on such proxy LR images fail to generalize when confronted with the complex degradations present in

real-world satellite imagery [64].

We develop a new method, SEED-SR (Segmentation Embedding Enhancement via Diffusion - for Super Resolution) to address these problems by utilizing a combination of conditional latent diffusion models [44], and a large-scale multi-spectral, multi-source geo-spatial foundation models (FM), which are increasingly being developed [3, 5, 13, 21, 26, 42, 71]. The key idea of our approach is to bypass the explicit SR task in the pixel space and instead perform SR in a segmentation-aware latent space. Thus, in SEED-SR we ‘super-resolve’ an LR embedding to an HR embedding, using latent diffusion models, which offers the benefits of improved fidelity, and scale consistency [44]. We obtain the embeddings through FMs, thereby reducing our training time and effectively utilizing large-scale, diverse data that is used in pretraining the FMs. However, the FMs yield fixed-size embeddings which are very high-dimensional which makes diffusion challenging, a problem that we solve through architectural innovations within SEED-SR. Thus, we demonstrate how rich feature hierarchies learned from geo-spatial FMs can be effectively utilized to solve an open challenge in super resolution.

In summary, this work makes the following key contributions:

- We develop SEED-SR, a method to generate very high resolution segmentation maps (at 50cm GSD) for field boundary delineation, at an unprecedented  $20\times$  super-resolution, from multi-temporal, low-resolution multi-source, multi-spectral satellite image inputs (at 10m GSD) and a high resolution historical reference image.
- SEED-SR showcases a unique way to leverage multiple pre-trained large-scale geo-spatial foundation models, and to learn a mapping between their high-dimensional embeddings with latent diffusion models.
- We rigorously evaluate our method on real-world data comprising LR and HR images from different satellite sensors. Our experiments on the field boundary delin-

eration task, for smallholder farms across multiple countries in the Global South, show that our method outperforms state-of-the-art methods following the two-step pixel-level SR and segmentation approach.

## 2. Related Work

**Super-resolution (SR)** has been extensively studied in computer vision [32, 37, 53] and in remote sensing (RS) [50]. We summarize the closest related works in Table 1 along 6 dimensions – whether or not the method (1) was designed for RS images [RS], (2) was trained/optimized for segmentation [OPT], (3) was designed/evaluated with real LR satellite data [Real], (4) whether the multispectral LR input was used [MS] (5) considered Ref images as inputs [Ref], (6) their maximum scale factor reported [SF].

Table 1. Summary of Related Works

Method	RS	OPT	Real	MS	Ref	SF
[8, 41, 61]	x	x	NA	x	x	4x
[7, 27, 28, 62]	✓	x	x	x	x	6x
[6, 22, 30, 63, 65]	x	x	NA	x	x	8x
[2, 20, 34, 68]	x	x	NA	x	✓	8x
[9, 17, 67]	✓	x	x	x	✓	4x
[10, 29]	✓	x	x	x	✓	16x
[55]	✓	x	✓	x	x	16x
[11, 16, 18, 66, 70]	✓	x	x	✓	x	4x
[48, 59]	✓	✓	✓	x	x	3x-5x
<b>SEED-SR (Ours)</b>	✓	✓	✓	✓	✓	20x

Many SR techniques have been developed based on deep learning architectures for both general settings e.g., SRCNN [8], DInN [41], and SADN[61] and specifically for remote sensing (RS), e.g., FunSR[7], LGCNet [27], TransENet [28]. Generative Adversarial Networks (GAN) have been widely used in SR for remote sensing applications (e.g., TTSR [65], WTRN [30], C2-Matching [22], AMSA [63], and DATSR [6]). However, GANs have largely been outperformed by diffusion-based models [47] which offer more realistic reconstruction and stable training.

Very few SR techniques have utilized multi-spectral LR inputs, and they usually just use the NIR band in addition to RGB, e.g., in [11, 16, 18, 66, 70]. Experimental comparisons in [56] against TransENet [28] show that addition of NIR does not significantly improve performance over RGB. Most focus on the SR task only and not on downstream analysis after SR, notable exceptions include [48, 59] who use multi-task networks for joint segmentation and super-resolution. None of these use Ref images to guide SR, and most achieve low scale factors of  $\leq 8x$ .

**Reference-based SR:** Among SR methods, our work belongs to the category of Ref-SR models [2, 9, 20, 22, 34, 63, 67, 68]. Most recent Ref-SR models for remote sensing have been obtained through conditional diffusion, e.g.,

DCDMF [29] and HSR-Diff [60] and RefDiff [10], which incorporate various conditions derived from the LR and Ref images into the denoising process of diffusion. The state-of-the-art RefDiff model improves texture reconstruction through the use of land cover change priors (obtained via change detection algorithms) and the use of SFT layers [58] which enable semantics-guided and reference texture-guided denoising. However, in terms of addressing the problem of large (20x) spatial resolution gap, its performance remains limited, also seen in our experiments.

Further, there are two limitations found in most previous works on SR for remote sensing. First, as discussed earlier, proxy LR images from downsampled HR images lead to models that are not generalizable. Second, as shown in previous works (e.g., [55]) and our experiments, enhancing general visual quality through SR does not directly translate to the high accuracy required for downstream applications. Since image-centric SR typically optimizes for global perceptual quality or pixel-wise fidelity (e.g., PSNR/SSIM), they can inadvertently smooth over or misrepresent the subtle yet crucial features necessary for fine-grained segmentation. In our work, we address both these limitations.

**Farm boundary delineation:** Previous deep learning approaches developed specifically for smallholder field boundary segmentation, such as [12, 35, 57], have utilized HR images only and have not tackled the SR problem; and have also not used multi-spectral, multi-source inputs.

## 3. Problem Statement

Let  $X_{t,h}$  denote a HR, RGB image at time  $t$  and  $S_{t,h}$  denote a segmentation map obtained from  $X_{t,h}$  and let  $X_{t',r}$  denote a single-source HR, RGB Ref image from the same geographical area, captured at time  $t' < t$  with at least a 6-month interval between  $t'$  and  $t$ . We specifically select a 6 months interval to ensure that there are noticeable landscape differences between  $X_{t,h}$  and  $X_{t',r}$ . Let  $[X_{t_m,l}, \dots, X_{t,l}]$  represent a sequence of multi-temporal, multi-spectral, multi-source LR images captured within a temporal window of size  $t - t_m$  such that  $t' \leq t_m < t$ . Our aim is to develop a model  $\mathcal{M} : ([X_{t_m,l}, \dots, X_{t,l}], X_{t',r}) \rightarrow S_{t,h}$  which learns a segmentation map, for field boundary delineation, from the input LR and Ref images.

## 4. Our Approach: SEED-SR

We build our desired model  $\mathcal{M}$  using three key components: two distinct Geo-spatial Foundation Models (FM) and a conditional diffusion model. The first FM is the AlphaEarth Foundations model [5] a pre-trained model to obtain task-agnostic features from multi-temporal, multi-spectral, multi-source LR images. This allows us to obtain embeddings,  $e_l$ , of our LR input sequence  $[X_{t_m,l}, \dots, X_{t,l}]$ . The second FM is an encoder-decoder model [49] which

takes HR images as input and is specifically fine-tuned for agricultural field segmentation. We use the encoder derived from this FM to obtain embeddings,  $e_r, e_h$  of Ref ( $X_{t_m,r}$ ) and HR ( $X_{t,h}$ ) images respectively. Thus, through the use of both the FMs, our problem boils down to learning the distribution of HR embeddings ( $e_h$ ) from the LR embeddings ( $e_l$ ) and the Ref embeddings ( $e_r$ ). This is challenging because of the high dimensionality of the latent information-rich embeddings,  $e_r, e_h$ . We develop a novel adaptation of the conditional denoising diffusion model [19] to learn  $e_h$ , conditioned on  $e_r, e_l$ . With the learnt HR embeddings, the decoder of the second FM can be used to obtain the required segmentation map  $S_{t,h}$ . Fig. 2 shows the overall schematic.

#### 4.1. Foundation Models for Geospatial Reasoning

We utilize two distinct pre-trained foundation models:

- Low Resolution Foundation Model (LR-FM): We use the pre-trained AlphaEarth Foundations Model [5] for processing LR imagery. The model utilizes multi-spectral images from multiple sources for a given time period ( $[t_m \dots t]$ ) – Sentinel-2 L1C top-of-atmosphere (bands B2,3,4,8,11), Landsat 8/9 T1 TOA (bands B2,3,4,5,6,8,10) and Sentinel-1 GRD (bands HH, HV, VV, VH depending on availability), all resampled to 10m and summarizes them into embeddings of size (128,128,64) pixels spanning an area of  $1280 \times 1280 m^2$ . We generate embeddings that span the 4 weeks before time  $t$ . We denote the pre-trained encoder derived from the model by  $E_{LFM}$ .
- High Resolution Foundation Model (HR-FM): This encoder-decoder model [49], is pre-trained and fine-tuned for agricultural field segmentation specifically for small-holder farms. The model processes HR images (here, 0.5 m GSD covering  $320 \times 320 m^2$ , resized to  $480 \times 480$  pixels for the encoder). The encoder,  $E_{HSM}$ , produces (120,120,3840) dimensional feature maps. The decoder,  $D_{HSM}$ , is used to obtain the final HR segmentation map.

We use encoders of these FMs to obtain the embeddings:  $e_l = E_{LFM}([X_{t_m,l}, \dots, X_{t,l}])$ ,  $e_h = E_{HSM}(X_{t,h})$ ,  $e_r = E_{HSM}(X_{t',r})$ . We sample an area of  $320 \times 320 m^2$  from the model output of  $E_{LFM}$  corresponding to the area in the HR images to yield a (32,32,64) sized tensor. To integrate  $e_l$  with spatial feature maps  $e_r$  and  $e_h$  this 64D vector is spatially tiled and projected to match the dimensions of  $e_r$ .

#### 4.2. Latent Diffusion for Embedding Generation

We estimate the desired conditional distribution,  $p(e_h|e_l, e_r)$  by training a parameterized model,  $G_\theta$ . To this end, we adapt the denoising diffusion probabilistic model (DDPM) [19] to operate in the latent space of  $E_{HSM}$ , i.e., to predict the embedding  $e_h$  from its noise-perturbed version. Following standard DDPM, our diffusion model consists of a forward process which

gradually adds Gaussian noise to the HR embedding  $e_h$  over  $T$  discrete timesteps following a predefined noise variance schedule and a reverse process which learns to recover the original embedding  $e_{h,0}$  from its noised( $\epsilon$ ) versions  $e_{h,\tau}$  conditioned on  $e_l, e_r$  and the timestep  $\tau$  where  $\tau \in [1, T]$ . This is achieved by training a neural network  $G_\theta(e_{h,\tau}, \tau, c)$  to predict  $e_{h,0}$ , where  $c = (e_l, e_r)$  represents the combined conditioning information. The model is trained by minimizing the mean squared error between the true embedding and its prediction:  $L_{DDPM} = \mathbb{E}_{\tau, e_{h,0}, \epsilon} [\|e_{h,0} - G_\theta(e_{h,\tau}, \tau, c)\|_2^2]$ .

#### 4.3. Handling High-Dimensional Embeddings

Our network  $G_\theta(e_{h,\tau}, \tau, c)$  follows a U-Net architecture [45], with novel architectural elements, strategically designed to extract and utilize relevant information at each stage, from high-dimensional inputs. Appendix C has details of each block and dimensions of all the layers.

##### 4.3.1. Encoder Design

Our encoder consists of 4 encoder blocks, each comprising a convolution block (as detailed in Appendix C) and a max pooling layer. It processes a noised HR (120,120,3840)-size embedding as its primary input, along with a reference HR (120,120,3840)-size embedding, a spatially resized LR (32,32,64)-size embedding, and a diffusion timestep identifier. The encoder path progressively spatially downsamples this combined input, integrating the timestep embedding. However, the channel depth is progressively increased (320, 480, 560, 640, and a bottleneck of 784 channels). These substantial channel capacities are deliberately chosen to ensure that the rich information contained within the initial high-dimensional embeddings is effectively propagated and transformed without significant loss.

**Self Attention, Convolution and Cross Attention.** Initially, the input modalities (upsampled  $e_l, e_r$  and the noised high-resolution input  $e_h$ ) are concatenated, forming a high-dimensional tensor with 7744 channels. This concatenated input undergoes a self-attention mechanism. This crucial step allows the model to learn complex spatial correlations and inter-dependencies across the different input features. Following self-attention, the resulting feature map is a (120, 120, 64) dimensional tensor, effectively condensing the information while preserving key spatial relationships. To integrate this contextual information with the model’s encoder pathway, this (120, 120, 64) attention output is spatially downsampled (to 7x7). The downsampling reduces memory usage from 14 GB to 2.75 MB and FLOPs from 423 GFLOPs to 5 MFLOPs making diffusion feasible in this feature space. Additional details about memory fitting are in Sec. C.1. This downsampling is achieved using a convolutional block (comprising of 2D convolutions, LayerNorm(s), and ReLU activations (as detailed in Appendix C) to match the spatial dimensions (7x7) of the fea-

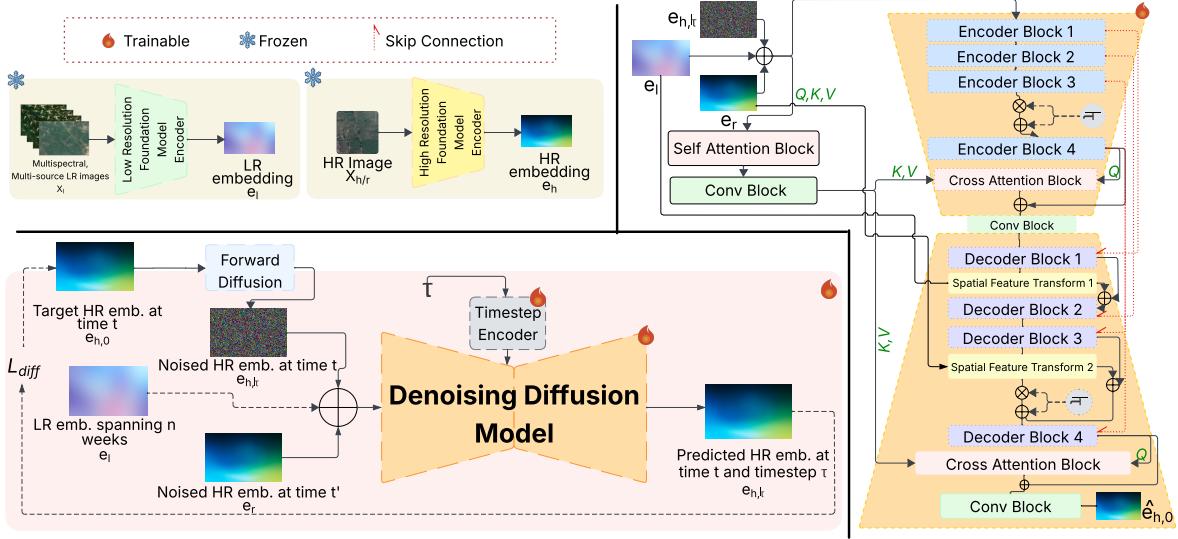


Figure 2. (Left) Training Diagram: A conditional latent diffusion U-Net predicts a clean HR embedding ( $e_{h,0}$ ) from a noised input ( $e_h$ ), conditioned on concatenated LR ( $e_l$ ) and reference HR ( $e_r$ ) features. (Right) Architecture of the conditional denoising diffusion model( $G_\theta$ ). Best viewed in colour.

tures at the deepest part of the encoder (Encoder Block 4). This downsampled attention map, rich in global context, is then fused via cross-attention with the output of the encoder block 4 (which has 640 channels). This allows the encoder’s features to be conditioned by the global context derived from all input modalities. Furthermore, the decoder also leverages this contextual understanding in its final stage via cross-attention.

#### 4.3.2. Decoder Design

Each of the 4 decoder blocks comprises transposed convolution and convolutional layers. The decoder path reconstructs the spatial resolution using skip connections from the encoder, and incorporates the timestep embedding. Inspired by the approach in [10], we strategically apply conditioning using Spatially-adaptive Feature Transform (SFT) blocks [58] in the decoder. Each SFT layer encodes the function  $F_{out} = \gamma \odot F_{in} + \beta$  where  $F_{out}, F_{in}$  are the output and input features of the layer and the scale ( $\gamma$ ) and shift ( $\beta$ ) parameters are adaptively learnt from the input conditions (see Appendix C for more details). The resized LR embedding modulates early-stage decoder features using a SFT block, while the reference HR embedding similarly conditions mid-stage decoder features through another SFT block. SFT layers play an important role in our decoder as they allow the  $e_l$  and  $e_r$  embeddings to provide detailed, spatially varying, and nuanced conditioning signals.

Before the final output, late-stage decoder features are further refined by a cross-attention mechanism employing the fused HR contextual representation established at the input stage, enabling the decoder to refine its predictions using this comprehensive, early-stage feature summary. The network culminates in a convolutional layer that outputs a

direct prediction of the input noised HR embedding.

#### 4.4. Segmentation Output Generation

During inference, the pretrained FM encoders ( $E_{LFM}, E_{HSM}$ ) are used to obtain the embeddings  $e_l, e_r$ , which, in turn, are used in the trained denoising diffusion model along with a sampled latent from an Isotropic Gaussian distribution  $\mathcal{N} \sim (0, I)$  to obtain the predicted HR embedding  $\hat{e}_{h,0}$ . This embedding is used in the pretrained decoder  $D_{HSM}$  of HR-FM to obtain a semantic segmentation map. In order to promote sample variation, we use four different seeds for different inference runs and compute the mean of semantic segmentation logits obtained from the runs. The mean of semantic segmentation logits are fed to a watershed algorithm [40] to obtain an instance segmentation map.

### 5. Experimental Evaluation

#### 5.1. Datasets and Preprocessing

Our experiments are on two datasets. The first comprises images from **Vietnam**, taken from a publicly available benchmark dataset, AI4SmallFarms [43], for agricultural field boundary delineation. The images are collected between 2021 and 2023, and the train/val/test splits contain 5296/432/753 images. The second is a larger proprietary dataset spanning the entirety of **India**, collected between 2019-22, with 32611/291/2615 images in train/val/test splits. In both cases, the satellite images encompass smallholder farms at national scales across diverse landscapes including farms on hilly terrains, arid regions, river plains and other areas. The high-resolution images are processed into (640,640) tiles at a 0.5m GSD, co-located

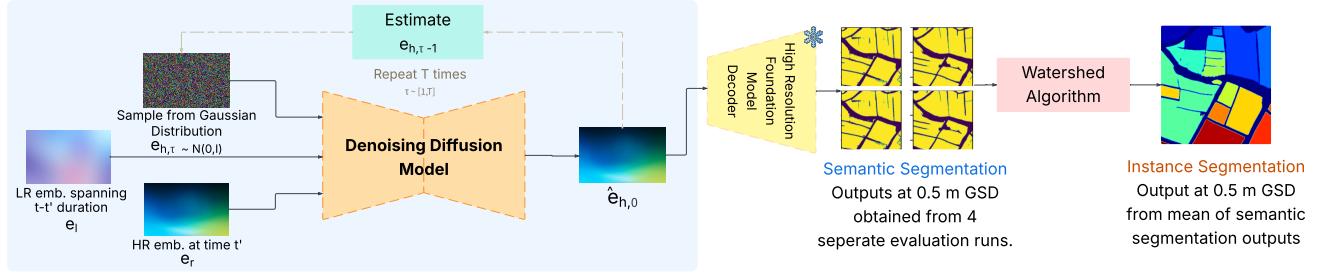


Figure 3. Inference uses a denoising diffusion framework: the model iteratively estimates the clean HR embedding ( $\hat{e}_{h,0}$ ) from a noised  $e_{h,\tau}$ , conditioned on LR ( $e_l$ ) and Ref ( $e_r$ ). This final  $\hat{e}_{h,0}$  yields a semantic segmentation map via a frozen  $D_{HSM}$  decoder. 4 samples are generated from distinct noise seeds; their semantic segmentation logits are averaged and then watershed-processed for instance segmentation. Best viewed in colour.

low-resolution embedding are obtained from  $E_{LFM}$  at 10m GSD. Embeddings are generated using the  $E_{LFM}$  encoder, initially producing (128,128,64) feature maps which are subsequently cropped to (32,32,64) to correspond with the high-resolution image areas. Thus both HR images and  $e_l$  are co-located and represent the same  $320 \times 320 \text{ m}^2$  area. Additional details about the dataset collection and labelling are present in Appendix B, with a summary of input details in Table 5.

## 5.2. Experiment Settings

To rigorously validate our approach, we perform evaluation against two complementary sets of reference annotations. The first, termed ‘**teacher predictions**’, represents the HR-FM’s inference on the target high-resolution imagery. The second comprises ‘**human labels**’, which serve as our manually curated ground truth. Against these human labels, the teacher predictions demonstrated significant agreement (see Appendix B). Also, results on teacher predictions allow us to compare with the case of segmentation directly on an HR image, if it were available at the same time. We use semantic and instance segmentation metrics ( $mIoU_S$ ,  $mIoU_I$ ) and pixel-level binary classification metrics to evaluate performance. Formal definitions are in Appendix B. Appendix C has hyperparameter details.

## 5.3. Baseline Methods

Our baselines follow the common two-step approach wherein SR, in the pixel space, is performed first and the super-resolved image is then segmented. We use 4 categories of SR methods for benchmarking (i) Bicubic upsampling, serving as a standard interpolation baseline; (ii) competitive SR methods on natural images: SRCNN [8], DNN [41]; (iii) SR algorithms specifically designed for remote sensing imagery: FunSR[7], LGCNet[27] and (iv) the state-of-the-art Ref-SR method RefDiff[10], which was shown to outperform GAN-based methods, TTSR [65], WTRN [30], C2-Matching [22], AMSA [63], and DATSR [6] as well as diffusion-based models, EDM [23] and HSR-Diff [60]. We feed changes in field boundary segmentation masks as

the change prior to RefDiff. Since baseline methods cannot utilize our multi-spectral multi-source inputs, we follow their settings closely, i.e., use downsampled HR images as LR ‘proxies’ and the recommended hyperparameters. We use HR-FM to obtain segmentation maps from the super-resolved images for all baselines, thus ensuring a fair comparison with our method which uses the decoder,  $D_{HSM}$ , of HR-FM.

## 5.4. Results: Field Boundary Segmentation

Table 6 shows the results of our method and all the baseline methods. B details the evaluation metrics utilized for comparision. In both the datasets we observe the same performance trend: Among the baselines, methods specifically designed for satellite images, FunSR, LGCNet and RefDiff, outperform the rest; and RefDiff which effectively utilizes a Ref image in its modeling performs the best. All the baselines, where pixel-based SR is followed by segmentation, are outperformed by our method across all the metrics. In particular, our model achieves 25.5% and 14.67% improvement in  $mIoU_I$  and 12.9% and 10.87% improvement in  $mIoU_S$  over RefDiff in the two datasets respectively.

## 5.5. Results: Qualitative Comparison

Fig. 4 illustrates the output masks obtained by our method and 3 baselines for a few sample inputs. The first two columns show the Ref and HR images and the their segmentation masks as predicted by HR-FM. The circled portions of the masks indicate the changes which have occurred in the field boundaries – in the first and second row we observe that multiple farms have merged into a larger farm. In both cases, our method delineates the boundaries more accurately compared to the baselines. In the last row we observe the noisy artifacts added during SR by other methods, which are absent in our reconstruction. More examples are discussed in Appendix B, including cases where our method does not perform well.

Table 2. Segmentation Performance on India (above) and Vietnam (below) datasets. Comparison with teacher predictions (left) and human labels (right). Metrics include – A: Accuracy, P: Precision, R: Recall, F1: F1 Score and mean  $IoU$  for instance ( $I$ ) and semantic ( $S$ ) segmentation.

Method	India Dataset											
	Teacher Predictions						Human Labels					
	$mIoU_I$	$mIoU_S$	A	P	R	F1	$mIoU_I$	$mIoU_S$	A	P	R	F1
Bicubic	10.9	15.25	59.41	16.21	8.07	8.84	10.3	15.22	60.7	15.66	8.66	9.08
SRCCNN	28.71	37.25	69.91	36.92	23.67	24.57	27.23	35.81	68.10	36.86	21.63	23.32
DINN	31.29	43.92	71.03	38.47	24.94	25.37	28.85	40.16	69.41	37.97	23.12	25.64
LGCNet	32.05	42.96	71.15	40.13	28.83	28.10	30.76	41.29	69.03	39.25	24.19	26.42
FunSR	37.41	48.27	74.59	44.32	32.00	32.40	35.15	45.07	71.78	43.62	28.82	30.21
RefDiff	47.96	65.73	81.48	58.02	54.77	54.05	45.10	62.94	79.18	57.11	50.59	51.59
SEED-SR(Ours)	<b>58.44</b>	<b>74.96</b>	<b>86.33</b>	<b>62.11</b>	<b>64.68</b>	<b>61.51</b>	<b>56.61</b>	<b>71.06</b>	<b>83.55</b>	<b>60.94</b>	<b>59.38</b>	<b>58.34</b>

Method	Vietnam Dataset (AI4SmallFarms)											
	$mIoU_I$	$mIoU_S$	A	P	R	F1	$mIoU_I$	$mIoU_S$	A	P	R	F1
	15.70	24.48	61.67	29.47	19.37	18.83	14.67	23.32	56.28	32.41	18.32	18.28
Bicubic	28.23	40.92	69.59	49.06	34.44	35.61	27.73	38.07	68.32	47.98	31.86	32.15
SRCCNN	32.71	45.64	71.09	53.91	37.18	39.14	29.92	44.45	69.55	52.86	35.31	39.08
DINN	35.96	48.61	71.28	58.17	37.61	43.58	32.59	47.03	68.19	57.40	34.78	41.24
LGCNet	37.16	56.24	72.85	62.06	41.18	47.96	35.59	54.23	69.52	59.02	40.76	55.44
FunSR	44.82	63.02	77.43	71.49	66.32	68.24	43.16	60.98	75.17	69.56	63.51	67.39
RefDiff	<b>50.91</b>	<b>71.01</b>	<b>81.34</b>	<b>76.76</b>	<b>74.97</b>	<b>74.43</b>	<b>49.49</b>	<b>67.61</b>	<b>78.72</b>	<b>72.94</b>	<b>75.65</b>	<b>71.89</b>

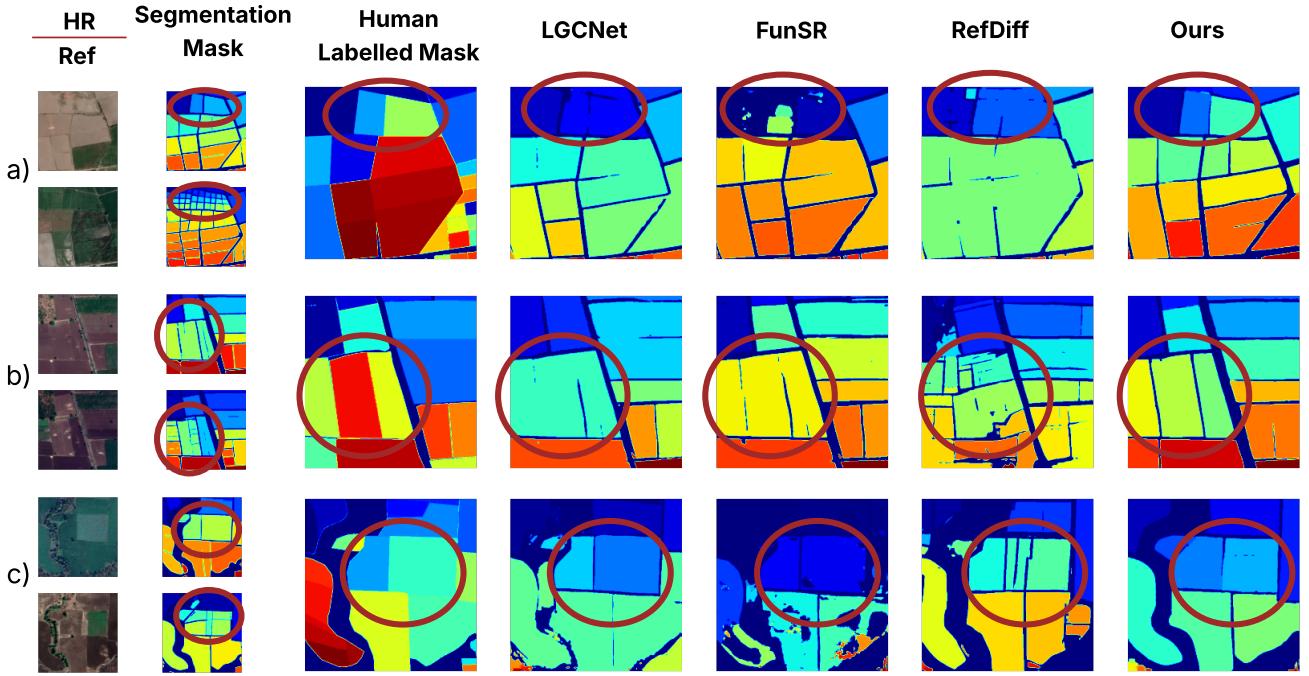


Figure 4. Results on 3 examples from the India Dataset. Best viewed in colour.

## 5.6. Results: Geographic Generalizability

To quantitatively demonstrate geographical generalization, we conducted zero-shot generalization experiments. We used our model, trained only on the India dataset Sec. 5.1, and evaluated its performance on two unseen regions from different continents: (i) **Vietnam** and (ii) **Kenya**. The former is the same AI4SmallFarms [43] test set described in Sec. 5.1, while the latter is from [24]. On Vietnam, our India-trained model achieves **46.24**  $mIoU_I$  against human labels and **47.62**  $mIoU_I$  against teacher predictions demon-

strating strong generalization across different smallholder farming systems in Asia. Notably, the India-trained model’s performance on Vietnam (46.24  $mIoU_I$ ) exceeds the best baseline (RefDiff 44.82 [10]) even when RefDiff was fully trained on the Vietnam data (see Tab. 6). On Kenya, the India-trained model achieves **51.49**  $mIoU_I$  against human labels and **53.42**  $mIoU_I$  against teacher predictions, demonstrating strong generalization across farming systems in a different continent.

Table 3. Zero shot performance ( $mIoU_I$ ) across geographies.

Dataset	Human Labels	Teacher Predictions
Vietnam	46.24	47.62
Kenya	51.49	53.42

Table 4. Ablation Studies

Method Config	Components		Model Inputs		Teacher Predictions	
	$E_{LFM}$	$E_{HSM}$	LR	Ref	$mIoU_I$	$mIoU_S$
A	x	x	✓	✓	37.41	41.78
B	x	x	✓	x	16.27	23.41
C	✓	x	✓	x	17.49	25.83
D	x	✓	x	✓	48.73	56.19
E	x	✓	✓	✓	52.19	65.28
SEED-SR	✓	✓	✓	✓	<b>58.44</b>	<b>74.96</b>

## 5.7. Results: Ablation Studies on SEED-SR

To systematically dissect the impact of  $E_{LFM}$ ,  $E_{HSM}$ , LR and Ref inputs on the final segmentation performance, as measured by  $mIoU_I$  and  $mIoU_S$  on teacher predictions, we conduct ablation studies on the India dataset.

Table 4 shows the results. In **Config A** LR and Ref image inputs are utilized without the specialized LFM and HSM encoders i.e., the diffusion model takes the multi-spectral, multi-temporal, Sentinel-2 LR images as input. Segmentation is performed by using the HR-FM model on the superresolved image (obtained by our diffusion model). This configuration yields significantly lower scores ( $mIoU_I = 37.41$ ,  $mIoU_S = 41.78$ ) compared to our full model. This corroborates our performance comparison in Table 6. To evaluate the scenario with minimal inputs and no specialized FM-based encoders, **Config B** utilized only the LR image input without any FM encoders or Ref. The LR input image here was a stack of 32 Sentinel-2 multi-spectral images corresponding to time  $\epsilon [t', t]$ . The target was the superresolved  $e_h$ . As expected, this configuration performed poorly further emphasizing the necessity of latent space (via FMs) and the guiding information from the ref HR imagery.

Next, we investigated the importance of each input modality used as conditions in our diffusion model. In **Config C**, we removed the HSM encoder for conditioning and the Ref image input, relying solely on the LFM encoder and the LR image sequence. This led to a drastic reduction in performance which clearly demonstrates the critical role of the HR Ref image and its HSM-derived embedding in providing high-fidelity prior information for the diffusion model. Similarly, **Config D** isolates the contribution of the Ref image by removing the LFM encoder and the LR image input. While this configuration performs better than relying only on the LR embeddings (Config C), it still performs significantly worse than the proposed method. This underscores that the temporal context and features extracted

by the LFM from the LR image sequence provide essential complementary information that the Ref image alone cannot supply. Finally, in **Config E**, a stack of 32 Sentinel-2 LR images corresponding to time  $\epsilon [t', t]$  are fed directly while reference image inputs are encoded with the HSM encoder. This suggests that the specialized feature extraction and representation provided by the LFM encoder for the LR data are vital for optimal performance, reinforcing the benefit of our dual-encoder FM strategy. The results from Configs C, D and E together strongly indicate that our model learns from both LR and Ref embeddings. Details of input dimensions and layer parameters across ablation configurations are in Appendix C. These results demonstrate the importance of each model component and validates that for combined SR and segmentation, a task-specific latent space powered by FMs outperforms pixel-space methods.

## 5.8. Results: Additional Experiments

Appendix B has additional results. To summarize, SEED-SR outperforms two-stage approaches when SAM [25] is used on super-resolved images from various baselines. **Sensitivity analysis** shows that (i) SEED-SR’s performance depends considerably on the underlying FM – while diffusion within SEED-SR learns accurate mappings between embeddings, it does not correct for errors in the input embeddings and (ii) clarifies our choice of 4 runs during inference. **Running time** analysis of SEED-SR with baselines shows that it is faster than other diffusion-based models (RefDiff), while being slower than non-generative alternatives (FunSR, LGCNet).

## 6. Conclusions, Limitations and Future Work

We develop SEED-SR, a novel latent diffusion-based method to generate HR segmentation maps and demonstrate its efficacy through experiments on smallholder field boundary delineation, where it distinguishes itself by effectively bridging a 20x resolution gap. To our knowledge, this work is the first to utilize a combination of latent diffusion model and two distinct FMs – one for processing multi-source, multi-spectral multi-temporal LR inputs and another, for defining a task-aware latent space and generating HR segmentation maps – for field boundary delineation.

While our novel framework shows substantial improvements in segmentation performance, there are several limitations which offer avenues for further research. Our inference is computationally intensive and new techniques to improve its running time could be investigated. Our method’s performance is inherently tied to the underlying FMs, with possibly dissimilar architectures, and achieving optimal semantic and pixel alignment between LR and HR embeddings remains challenging. Ways to improve the alignment, both through improved FM design and connector architectures such as ours, could be explored. Our framework can

easily be adapted to other tasks like object detection by using suitable task-specific FMs and could be evaluated in future work. More broadly, our work lays the foundation to build models which can derive fine-grained artifacts reliant on HR details using high-revisit-frequency LR satellite images to develop more accurate remote sensing applications.

## References

- [1] L. Agilandeswari, M. Prabukumar, V. Radhesyam, K. L. N. B. Phaneendra, and A. Farhan. Crop classification for agricultural applications in hyperspectral remote sensing images. *Applied Sciences*, 12(3):1670, 2022. [1](#)
- [2] Masoomeh Aslahishahri, Jordan Ubbens, and Ian Stavness. Darts: Double attention reference-based transformer for super-resolution. *arXiv preprint arXiv:2307.08837*, 2023. [2](#), [3](#)
- [3] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024. [2](#)
- [4] H. L. Aung, B. Uzkent, M. Burke, D. Lobell, and S. Ermon. Farm parcel delineation using spatio-temporal convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 76–77, 2020. [1](#)
- [5] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chen-hui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. [2](#), [3](#), [4](#)
- [6] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulin Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, 2022. [3](#), [6](#)
- [7] Keyan Chen, Wenyuan Li, Sen Lei, Jianqi Chen, Xiaolong Jiang, Zhengxia Zou, and Zhenwei Shi. Continuous remote sensing image super-resolution based on context interaction in implicit function space. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. [3](#), [6](#), [16](#), [17](#)
- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. [2](#), [3](#), [6](#)
- [9] Runmin Dong, Lixian Zhang, and Haohuan Fu. Rrsgan: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17, 2021. [3](#)
- [10] Runmin Dong, Shuai Yuan, Bin Luo, Mengxuan Chen, Jin-xiao Zhang, Lixian Zhang, Weijia Li, Juepeng Zheng, and Haohuan Fu. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27684–27694, 2024. [2](#), [3](#), [5](#), [6](#), [7](#), [12](#), [16](#), [17](#)
- [11] Simon Donike, Cesar Aybar, Luis Gómez-Chova, and Freddie Kalaitzis. Trustworthy super-resolution of multispectral sentinel-2 imagery with latent diffusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:6940–6952, 2025. [3](#)
- [12] Radhika Dua, Nikita Saxena, Aditi Agarwal, Alex Wilson, Gaurav Singh, Hoang Tran, Ishan Deshpande, Amandeep Kaur, Gaurav Aggarwal, Chandan Nath, Arnab Basu, Vishal Batchu, Sharath Holla, Bindiya Kurle, Olana Missura, Rahul Aggarwal, Shubhika Garg, Nishi Shah, Avneet Singh, Dinesh Tewari, Agata Dondzik, Bharat Adsul, Milind Sohoni, Asim Rama Praveen, Aaryan Dangi, Lisan Kadivar, E Abhishek, Niranjan Sudhansu, Kamlakar Hattekar, Sameer Datar, Musty Krishna Chaithanya, Anumas Ranjith Reddy, Aashish Kumar, Betala Laxmi Tirumala, and Alok Talekar. Agricultural landscape understanding at country-scale. *arxiv preprint 2411.05359*, 2024. [1](#), [3](#), [12](#)
- [13] Iris Dumeur, Silvia Valero, and Jordi Inglada. Paving the way toward foundation models for irregular and unaligned satellite image time series. *arXiv preprint arXiv:2407.08448*, 2024. [2](#)
- [14] Yetkin Özüm Durgun, Anne Gobin, Gregory Duveiller, and Bernard Tychon. A study on trade-offs between spatial resolution and temporal sampling density for wheat yield estimation using both thermal and calendar time. *International Journal of Applied Earth Observation and Geoinformation*, 86, 2019. [1](#)
- [15] Food and Agriculture Organization. The economic lives of smallholder farmers: An analysis based on household data from nine countries, 2015. [1](#)
- [16] Mikel Galar, Rubén Sesma, Christian Ayala, Lourdes Albizua, and Carlos Aranda. Super-Resolution of Sentinel-2 Images Using Convolutional Neural Networks and Real Ground Truth Data. *Remote Sensing*, 12(18):2941, 2020. [3](#)
- [17] Jingliang Guo, Mengke Yuan, Tong Wang, Zhifeng Li, Xiaohong Jia, and Dong-Ming Yan. Dtesr: Remote sensing imagery super-resolution with dynamic reference textures exploitation. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. [3](#)
- [18] Hao Han, Ziyi Feng, Wen Du, Sien Guo, Peng Wang, and Tongyu Xu. Remote sensing image classification based on multi-spectral cross-sensor super-resolution combined with texture features: A case study in the liaohe planting area. *IEEE Access*, 12:16830–16843, 2024. [3](#)
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [4](#), [18](#)
- [20] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *CVPR*, pages 5931–5940, 2022. [2](#), [3](#)
- [21] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. [2](#)

- [22] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, pages 2103–2112, 2021. 2, 3, 6
- [23] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, pages 26565–26577, 2022. 6
- [24] Hannah Kerner, Saketh Sundar, and Mathan Satish. Multi-region transfer learning for segmentation of crop field boundaries in satellite images with limited labels. In *Proceedings of the AAAI Workshop on AI to Accelerate Science and Engineering*, 2023. 1, 7
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 8, 13
- [26] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geobench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023. 2
- [27] Sen Lei, Zhenwei Shi, and Zhengxia Zou. Super-resolution for remote sensing images via local–global combined network. *IEEE Geoscience and Remote Sensing Letters*, 14(8): 1243–1247, 2017. 3, 6, 17
- [28] Sen Lei, Zhenwei Shi, and Wenjing Mo. Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 3
- [29] Shuangliang Li, Siwei Li, and Lihao Zhang. Hyperspectral and panchromatic images fusion based on the dual conditional diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 3
- [30] Zhen Li, Zeng-Sheng Kuang, Zuo-Liang Zhu, Hong-Peng Wang, and Xiu-Li Shao. Wavelet-based texture reformation network for image super-resolution. *IEEE Transactions on Image Processing*, 31:2647–2660, 2022. 3, 6
- [31] Thomas M. Lillesand, Ralph W. Kiefer, and Jonathan W. Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, 7th edition, 2015. 1
- [32] Hongying Liu, Zekun Li, Fanhua Shang, Yuanyuan Liu, Liang Wan, Wei Feng, and Radu Timofte. Arbitrary-scale super-resolution via deep learning: A comprehensive survey. *Information Fusion*, 102:102015, 2024. 3
- [33] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 17
- [34] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 2, 3
- [35] Wenting Mei, Han Wang, David Fouhey, Weichen Zhou, Iain Hinks, Justin M. Gray, Pinki Mondal, and Meha Jain. Using deep learning and very-high-resolution imagery to map smallholder field boundaries. *Remote Sensing*, 14(13):3046, 2022. 1, 3
- [36] Lingyao Meng, Bingfang Xu, Zhong Dong, Yongguang Zhang, Hong Chen, Kuang Zhao, and Jianlong Yang. Deep learning-based spatio-temporal fusion of remote sensing images: A review and new perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:262–281, 2022. 1
- [37] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 3
- [38] Deepak Murugan, Akanksha Garg, and Dharmendra Singh. Development of an adaptive approach for precision agriculture monitoring with drone and satellite data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5322–5328, 2017. 1
- [39] Isah Nabara and Norsida Man. *Enhancing Precision Farming Innovations for Global Food Security Through Agricultural Extension Services*, pages 119–142. 2025. 1
- [40] H.P. Ng, S.H. Ong, K.W.C. Foong, P.S. Goh, and W.L. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 61–65, 2006. 5
- [41] Quan H Nguyen and William J Beksi. Single image super-resolution via a dual interactive implicit neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4936–4945, 2023. 3, 6
- [42] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023. 2
- [43] Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. Ai4smallfarms: A dataset for crop field delineation in south-east asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 5, 7, 12, 13
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, pages 234–241. Springer, 2015. 4
- [46] Mitchell Roznik, Milton Boyd, and Lysa Porth. Improving crop yield estimation by applying higher resolution satellite ndvi imagery and high-resolution cropland masks. *Remote Sensing Applications: Society and Environment*, 25:100693, 2022. 1
- [47] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arxiv preprint 2104.07636*, 2021. 3
- [48] Luis Salgueiro, Javier Marcello, and Verónica Vilaplana. Seg-esrgan: A multi-task network for super-resolution and semantic segmentation of remote sensing images. *Remote Sensing*, 14(22), 2022. 3

- [49] David Schottlander and Tomer Shekel. Geospatial reasoning: Unlocking insights with generative ai and multiple foundation models, 2025. Accessed: 2025-04-29. [3](#), [4](#)
- [50] Maria Sdraka, Ioannis Papoutsis, Bill Psomas, Konstantinos Vlachos, Konstantinos Ioannidis, Konstantinos Karantzalos, Ilias Gialampoukidis, and Stefanos Vrochidis. Deep learning for downscaling remote sensing images: Fusion and super-resolution. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):202–255, 2022. [3](#)
- [51] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann N. Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cissé, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *CoRR*, abs/2107.12283, 2021. [12](#)
- [52] S. Skakun, N. Kussul, A. Shelestov, and O. Kussul. The use of satellite data for agriculture drought risk quantification in ukraine. *Geomatics, Natural Hazards and Risk*, 7(3):901–917, 2015. [1](#)
- [53] Hu Su, Ying Li, Yifan Xu, Xiang Fu, and Song Liu. A review of deep-learning-based super-resolution: From methods to applications. *Pattern Recognition*, page 110935, 2024. [3](#)
- [54] F. Waldner and F. I. Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote Sensing of Environment*, 245:111741, 2020. [1](#)
- [55] Ce Wang and Wanjie Sun. Semantic guided large scale factor remote sensing image super-resolution with generative diffusion prior. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:125–138, 2025. [3](#)
- [56] Chunyang Wang, Xian Zhang, Wei Yang, GaiGe Wang, Xingwang Li, Jianlong Wang, and Bibo Lu. Mswagan: Multispectral remote sensing image super-resolution based on multiscale window attention transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. [3](#)
- [57] S. Wang, F. Waldner, and D. B. Lobell. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *Remote Sensing*, 14(22):5738, 2022. [1](#), [3](#)
- [58] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. [3](#), [5](#)
- [59] Dawei Wen, Shihao Zhu, Yuan Tian, Xuehua Guan, and Yang Lu. Generating 10-meter resolution land use and land cover products using historical landsat archive based on super resolution guided semantic segmentation network. *Remote Sensing*, 16(12), 2024. [3](#)
- [60] Chanyue Wu, Dong Wang, Yunpeng Bai, Hanyu Mao, Ying Li, and Qiang Shen. Hsr-diff: Hyperspectral image super-resolution via conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2023. [2](#), [3](#), [6](#)
- [61] Hanlin Wu, Ning Ni, and Libao Zhang. Scale-aware dynamic network for continuous-scale super-resolution. *CoRR*, abs/2110.15655, 2021. [3](#)
- [62] Hanlin Wu, Jiangwei Mo, Xiaohui Sun, and Jie Ma. Latent diffusion, implicit amplification: Efficient continuous-scale super-resolution for remote sensing images. *arXiv preprint 2410.22830*, 2024. [3](#)
- [63] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *AAAI*, pages 2768–2776, 2022. [2](#), [3](#), [6](#)
- [64] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Yuan Wang, and Liangpei Zhang. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. *Information Fusion*, 96:297–311, 2023. [2](#)
- [65] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bainong Guo. Learning texture transformer network for image super-resolution. *CoRR*, abs/2006.04139, 2020. [3](#), [6](#)
- [66] Maialen Zabalza and Angela Bernardini. Super-Resolution of Sentinel-2 Images Using a Spectral Attention Mechanism. *Remote Sensing*, 14(12):2890, 2022. [3](#)
- [67] Jiayang Zhang, Wanxu Zhang, Bo Jiang, Xiaodan Tong, Keya Chai, Yanchao Yin, Lin Wang, Junhao Jia, and Xiaoxuan Chen. Reference-based super-resolution method for remote sensing images with feature compression module. *Remote Sensing*, 15(4):1103, 2023. [2](#), [3](#)
- [68] Lin Zhang, Xin Li, Dongliang He, Fu Li, Yili Wang, and Zhaoxiang Zhang. Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In *European Conference on Computer Vision*, pages 648–664. Springer, 2022. [2](#), [3](#)
- [69] Xiaolin Zhu, Jie Chen, Feng Gao, Xi Chen, Dongping Zhang, Yujie Guo, Huiting Ma, and Carsten Montzka. An updated review of spatial-temporal fusion in remote sensing. *Remote Sensing*, 10(12):1952, 2018. [1](#)
- [70] Xi Zhu, Yang Xu, and Zhihui Wei. Super-resolution of sentinel-2 images based on deep channel-attention residual network. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 628–631, 2019. [3](#)
- [71] Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth and climate foundation models. *arXiv preprint arXiv:2405.04285*, 2024. [2](#)

## A. Societal Impact

This research stands to make significant positive contributions, particularly in enhancing global food security and supporting sustainable agricultural practices. The developed super-resolution technology specifically addresses the challenge of monitoring agricultural land with high accuracy and timeliness, which is crucial for precision agriculture, improved yield estimation, and efficient resource management. This is especially impactful for smallholder farms, which characterize a vast portion of agricultural landscapes in the developing world and are essential to the livelihoods of billions. By enabling cost-effective and timely in-season monitoring, the technology can provide actionable insights to improve yields and mitigate losses, directly contributing to the economic well-being of these farmers and bolstering local and global food supplies.

## B. Additional Experimental Results

### B.1. Dataset Collection

The India Dataset is a proprietary collection of images acquired over the period 2019-2022, covering varied agricultural topographies within India. We follow established precedent from works such as [43], [51] in sourcing VHR imagery from Google Maps Satellite View. We take images sampled at 0.5 m GSD and from the following two satellite sources: Maxar Worldview, Airbus Pleiades.

Individual images represent a spatial extent of 320\*320  $m^2$ . Annotation was conducted by trained human personnel according to established protocols. An 'agricultural field' was operationally defined as a contiguous area meeting at least one of the following criteria: (1) cultivation of a single crop type, or (2) distinct visual boundaries separating it from neighboring parcels or non-agricultural features. The Vietnam Dataset was constructed using publicly available labels provided by [43]. As documented in [43], the high-resolution (HR) images at a given time  $t$ , denoted as  $X_t$  were originally collected for August 2021. However, our preliminary analysis revealed inconsistencies between these provided labels and the corresponding  $X_t$  imagery. Specifically, in some instances, reference images  $X_r$  predating  $X_t$  by six months or more exhibited greater concordance with the ground truth labels than the August 2021 images themselves. To ensure the integrity of our evaluation, the test set was subsequently filtered. This process retained only those image-label pairs where the August 2021 image ( $X_t$ ) demonstrated clear and strong agreement with its associated ground truth labels, thereby excluding cases where labels appeared to reflect a significantly different temporal state. Table 5 summarises the properties of the datasets used.

### B.2. Evaluation Metrics

We use standard pixel-level metrics to evaluate the performance of our binary classification task. Let  $TP_i, TN_i, FP_i, FN_i$  denote the number of true positive, true negative, false positive, false negative pixels respectively in a single test image  $i$ . For a test set with  $N$  images the semantic IoU  $IoU_S = \frac{1}{N} \sum_1^N \frac{TP_i}{TP_i + FP_i + FN_i}$ . At the instance level, we use the definition recommended for evaluating agricultural field instances (e.g., [12]), viz., Instance IOU  $IoU_I = \frac{1}{M} \sum_1^M (P_m^i \cap P^i) / (P_m^i \cup P^i)$ , where  $M$  is the number of ground truth instances in the test set,  $P_m^i$  is the total number of pixels in all the predicted  $i^{\text{th}}$  instances which have more than  $t\%$  overlap with the  $i^{\text{th}}$  ground truth instance and  $P^i$  is the total number of pixels in the ground truth  $i^{\text{th}}$  instance. Accuracy, Precision, Recall and F1 score are computed from  $TP_i, TN_i, FP_i, FN_i$  and averaged over the test set.

### B.3. Performance of Teacher and Human Labels

The Teacher model predictions i.e. the predictions from the HR-FM on the HR images yield an  $mIoU_S$  of 74.18 % and an  $mIoU_I$  of 71.17 % for the India Dataset and  $mIoU_S$  of 68.47 % and an  $mIoU_I$  of 63.29 % for the Vietnam subset of the AI4SmallFarms Dataset.

### B.4. Ablations with Inference Timesteps

Experimental results suggested that stopping the reverse diffusion process early produced better results. Table 6 displays the results for both India dataset and Vietnam dataset. Figure 5 shows some qualitative examples from the last 100 denoising steps during inference time.

### B.5. Qualitative results: Additional Examples

To provide a more nuanced understanding of our model's performance characteristics, Figure 6 presents a selection of additional qualitative examples. These examples highlight the efficacy of our proposed method under certain conditions and its failure modes.

- **Demonstrating Robustness to New Structures** Examples a,b showcase instances where our model successfully identifies and segments newly developed structures or significant emergent changes within the scene. This demonstrates our model's capability to detect objects or alterations not present in the reference image but from the LR inputs.
- **Superiority of Reference-Based Methods in Specific Contexts** In contrast, c,f,g illustrate specific scenarios where Ref-SR methods(RefDiff[10] and SEED-SR(Ours) exhibit superior performance compared to non-reference-based approaches. These cases typically involve subtle textural changes.
- **Failure Mode: Logit Inconsistency** Example d reveals a specific failure mode observed in our model. In this in-

Table 5. Summary of Input Data Characteristics

Category	Item	Detail / Specification
<b>LR Inputs</b>	Satellite Sources	Sentinel-2 L1C, Landsat 8/9, and Sentinel-1 GRD
	Spatial Resolution	10m GSD (Resampled)
	Temporal Span	4 weeks before the target time ( $t$ )
	Embedding Dimension from $E_{LFM}$	(128,128,64) – $\rightarrow$ (32,32,64) [After Cropping]
<b>HR Inputs</b>	Per Embedding Spatial Coverage	$320 \times 320 m^2$
	Spatial Resolution	0.5m GSD
	Embedding Dimension from $E_{HFM}$	(120,120,3840)
	Per Embedding Spatial Coverage	$320 \times 320 m^2$
	India Dataset Collection Time	2019–2022
	AI4SmallFarms Data Collection Time	August 2021 [43]

Table 6. Comparision of Performance on early stopping during Inference time for India and Vietnam Dataset. A: Accuracy, P: Precision, R: Recall

Method	India Dataset											
	Teacher Predictions						Human Labels					
	$mIoU_I$	$mIoU_S$	A	P	R	F1	$mIoU_I$	$mIoU_S$	A	P	R	F1
$T_{inf,500}$	49.16	72.44	84.89	60.49	61.27	58.67	48.21	70.71	83.48	59.80	57.28	56.58
$T_{inf,470}$	58.44	74.96	86.33	62.11	64.68	61.51	56.61	71.06	83.55	60.94	59.38	58.34
Vietnam Dataset (AI4SmallFarms)												
Method	$mIoU_I$	$mIoU_S$	A	P	R	F1	$mIoU_I$	$mIoU_S$	A	P	R	F1
$T_{inf,500}$	49.14	72.30	82.10	76.68	77.92	75.92	48.77	68.20	79.11	72.77	78.30	72.59
$T_{inf,470}$	50.91	71.01	81.34	76.76	74.97	74.43	49.49	67.61	78.72	72.94	75.65	71.89

stance, the introduction of noisy artifacts or erroneous detections in the output change map can be attributed to significant disagreement among the four distinct segmentation logits generated internally by our model prior to taking an average.

- **Comparative Advantage:** **Reduced Noise/Oversegmentation** Finally, e provides a comparative example where other existing methods exhibit common failure modes such as oversegmentation of changed regions or the introduction of spurious noisy artifacts in their predictions.

## B.6. Evaluation of Segmentation Performance using SAM for Baseline Methods

We present results in Figure 8 for zero-shot performance of SAM[25] on the superresolved images generated by baseline methods and provide a qualitative comparison against segmentation results produced by HR-FM and SEED-SR’s superresolved segmentation output. Visual results clearly demonstrate that the segmentation from HR-FM and SEED-SR are considerably more accurate and coherent for agricultural fields compared to zero-shot SAM, which largely struggles with delineating farm boundaries.

## B.7. Foundation Model Dependency and Failure Modes

SEED-SR’s performance is coupled with the underlying FMs, a limitation we discuss in the conclusions. To further investigate this dependency, we isolated a subset of the test data where the per example mIoU of the teacher predictions (HR-FM’s performance on ground truth HR Images) against the human labels is poor (instance IoU  $< 60\%$ ). This analysis confirms that SEED-SR’s performance is bounded by the capabilities of the HR-FM. On the subset where the HR-FM performs poorly ( $< 60\% \text{ IoU}$ ), HR-FM achieves 52.02 IoU, and SEED-SR achieves 37.56 IoU, as seen in Table 7. This expected behavior validates that the model successfully learns the mapping to the target latent space, but cannot correct for errors inherent in that space. Visual inspection shows that most of these images correspond to particularly difficult regions with few fields, or extremely small fields or dry/hilly terrains. As expected, the best baseline method, RefDiff also performs poorly at  $mIoU_I = 25.44$  in this set.

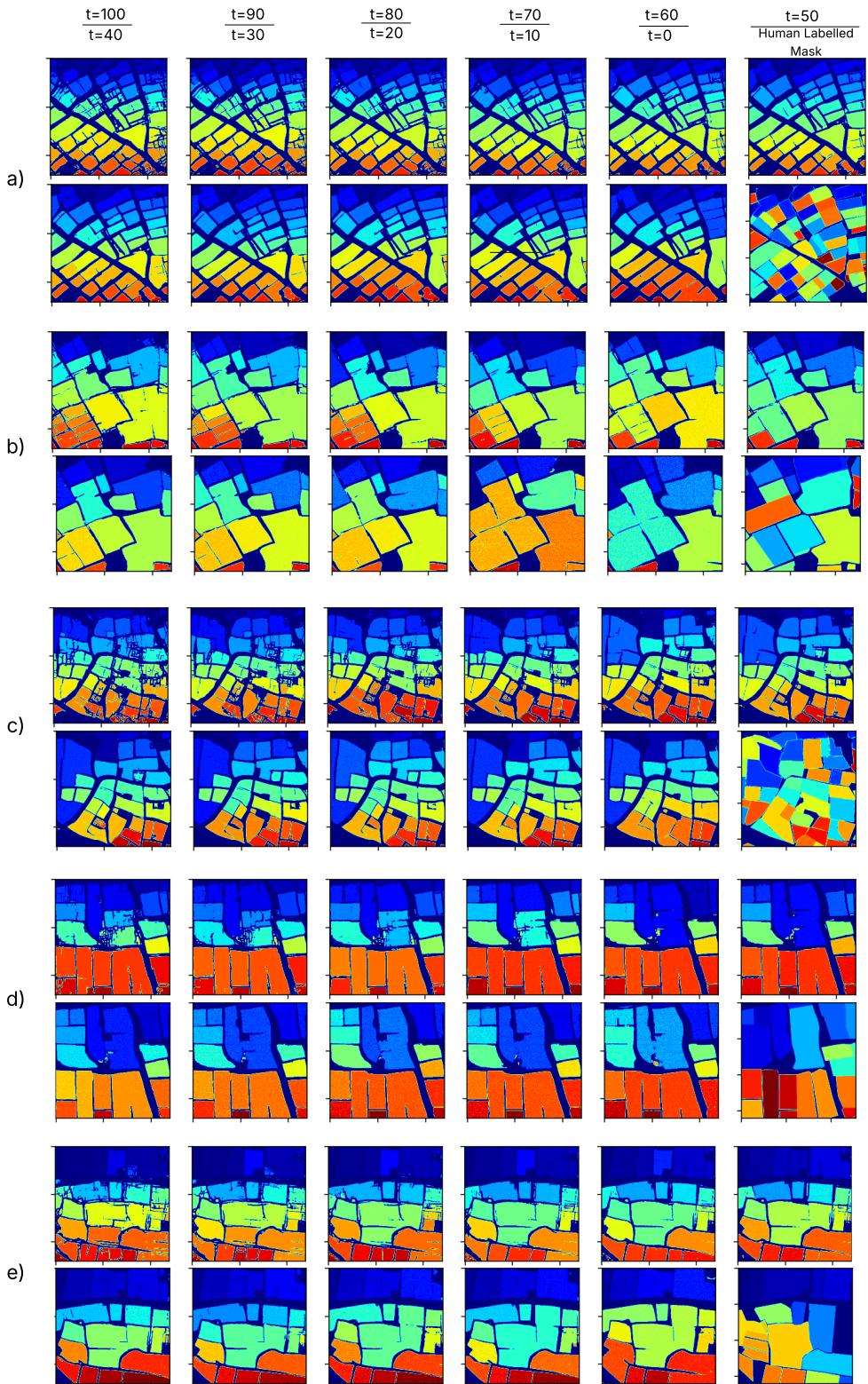


Figure 5. Qualitative results of last 100 denoising steps during inference for India Dataset.

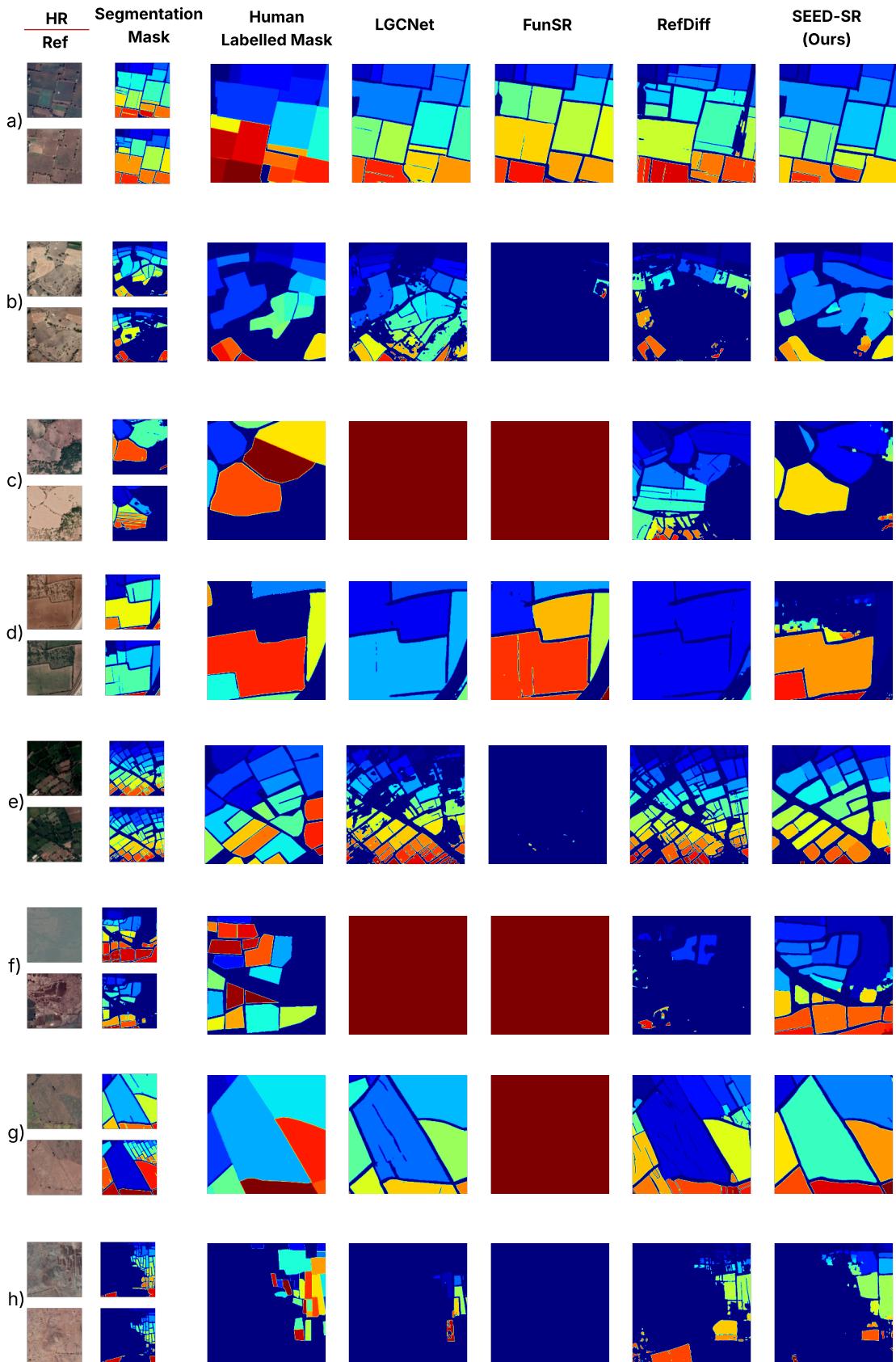


Figure 6. Additional Qualitative Examples from the India Dataset for Visual Comparison

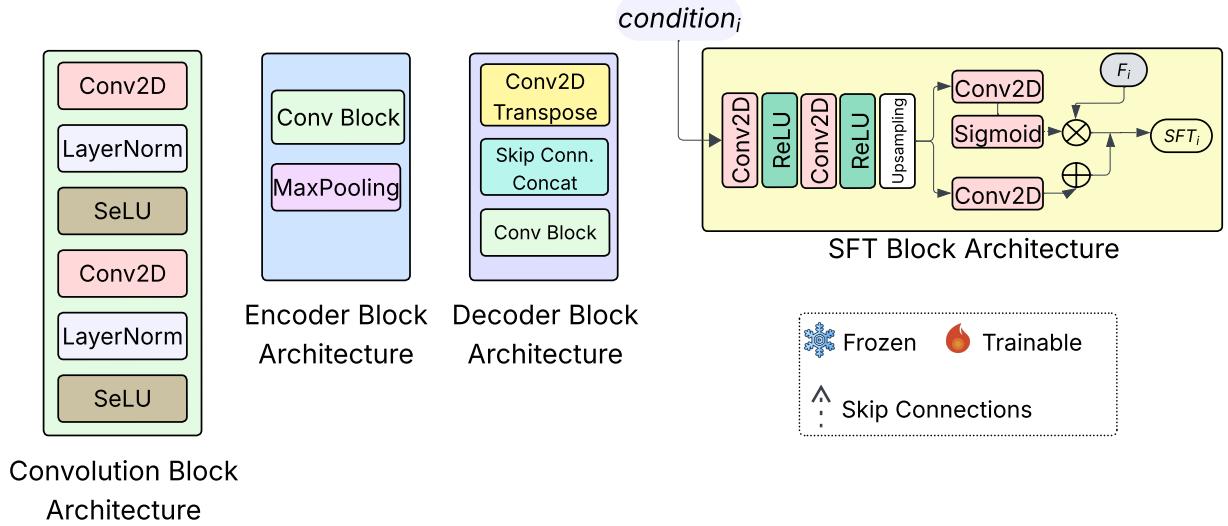


Figure 7. Architecture of ConvBlock, Encoder Block, Decoder Block and SFT as used in Figure 2

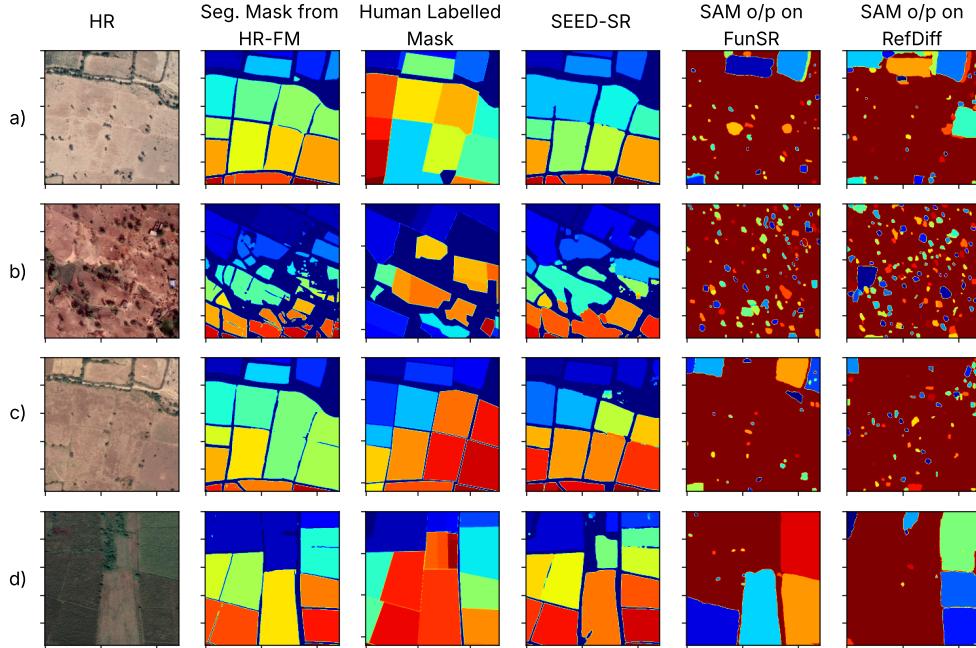


Figure 8. Zero Shot Instance Segmentation Outputs by SAM on Image Superresolution Baselines (RefDiff[10] and FunSR[7]) on the India Dataset

## B.8. Computational Cost

Our model runs for 200,000 training steps. Each training step for a batch size of 128 takes approximately 1 second over 8 Google TPUs. The complete training framework took 58.3 hours. The complete inference framework along

with metrics computation took approximately 3 hours. The inference step takes 2.64 seconds per denoising step for a batch size of 320.

Table 8 shows the average inference time per example at maximum batch size for SEED-SR and three base-

Method	$mIoU_I$ (vs. Human Labels)
HR-FM (Teacher prediction)	52.02
SEED-SR Prediction	37.56
RefDiff Prediction	25.44

Table 7. Performance on subset of the test data where the per example  $mIoU_I$  of the teacher predictions( HR-FM’s performance on ground truth) against the human labels is poor ( $mIoU_I < 60\%$ )

lines. While diffusion methods (SEED-SR, RefDiff) are slower than non-generative methods (LGCNet, FunSR), the substantial performance gain justifies this trade-off for high-accuracy applications. Notably, SEED-SR (3.968s) is slightly faster than the comparable SOTA diffusion baseline, RefDiff (4.12s). The average inference time per example at maximum batch size was calculated at maximum throughput to take advantage of the hardware’s parallelization capabilities across models.

Table 8. Model Inference Time Comparison

Method	Average Inference time per Example
SEED-SR(Ours)	3.968 s
RefDiff [10]	4.12 s
FunSR [7]	0.7 s
LGCNet [27]	0.64 s

### B.9. Choice of different number of masks for the production of the final segmentation map

Diffusion models are heavily reliant on the random seed for the variability in examples and often a mean of predictions is taken to reduce stochastic variance. The masks were computed for the validation set of the India dataset(291 images). Table 9 shows the performance for different number of masks. The masks were randomly chosen for each example. The results show clear gains up to N=4, with diminishing returns thereafter. We selected N=4 as the optimal balance between performance ( $\approx 2.3 mIoU_I$  gain over N=2) and inference cost.

## C. Additional Model Details

### C.1. Novelty in DDPM Architecture

Our contribution is the novel integration of two distinct geospatial FMs via a latent diffusion model for 20x task-specific SR. This is non-trivial due to the high-dimensional embeddings (120x120x3840) . The novelty in our DDPM architecture Sec. 4.3 includes the integration strategy for multi-modal conditioning (LR and Ref) using self-attention on concatenated inputs and introducing a convolution block

Table 9. Sensitivity to number of masks

No. of Masks	Teacher Predictions ( $mIoU_I$ )	Human Labels ( $mIoU_I$ )
1	49.9693	49.7525
2	55.2753	53.8581
3	55.5124	54.6811
<b>4</b>	<b>57.7448</b>	<b>56.0503</b>
5	57.7419	56.0247
6	57.6382	56.4666
7	57.8608	56.5775
8	58.3667	56.5203

to downsample the feature map spatially (to 7x7) before applying cross-attention at the bottleneck (L207-217); at extremely high dimensions. This innovation reduced memory usage from 14 GB to 2.75 MB and FLOPs from 423 GFLOPs to 5 MFLOPs (See Table), making diffusion feasible in this feature space.

Table 10. Computational Cost Comparison justifying the design choice of DDPM architecture

Metric	Case A (7x7)	Case B (120x120)
Tokens per sample	49	14,400
Attention memory	$\sim 0.15$ MB	$\sim 13.2 \times 10^3$ MB (float32)
Total memory	$\sim 2.75$ MB	$\sim 14 \times 10^3$ MB
FLOPs	$\sim 5$ MFLOPs	$\sim 423 \times 10^3$ MFLOPs

## C.2. Experimental Details

SEED-SR was trained using the AdamW optimizer [33] with a learning rate  $5 \times 10^{-4}$ , a batch size of 128, for a total of 200,000 steps. A total of  $T = 500$  timesteps were employed for the forward process and  $T = 470$  steps were employed for the reverse diffusion processes. The noise variance schedule for the diffusion process was linearly interpolated from  $\beta_1 = 10^{-4}$  to  $\beta_T = 0.02$ . The training was done on Google’s TPUv5.

## C.3. Diffusion Process

### C.3.1. Forward Process

The forward process gradually adds Gaussian noise to the HR embedding  $e_h$  over  $T$  discrete timesteps. Given  $e_{h,0}$  the noised embedding  $e_{h,\tau}$  at timestep  $\tau[0, T]$  is obtained by sampling from the distribution:

$$q(e_{h,\tau}|e_{h,0}) = \mathcal{N}(e_{h,\tau}; \sqrt{\bar{\alpha}_{h,0}}, (1 - \bar{\alpha}_{\tau})\mathbf{I})$$

where

$$e_{h,\tau} = \sqrt{\bar{\alpha}_{\tau}}e_{h,0} + \sqrt{1 - \bar{\alpha}_{\tau}}\varepsilon, \text{ for } \varepsilon \approx N(0, I)$$

The variables  $\alpha_{\tau} = 1 - \beta_{\tau}$  and  $\bar{\alpha}_{\tau} = \prod_{i=1}^{\tau} \alpha_i$ , where  $\beta_{\tau} \tau=1^T$  is a predefined noise variance schedule.

### C.3.2. Reverse Denoising Process

The reverse process learns to recover the original clean embedding  $e_{h,0}$  from its noised version  $e_{h,\tau}$  conditioned on  $e_l, e_r$  and the timestep  $\tau$ . This is achieved by training a neural network  $G_\theta(e_{h,\tau}, \tau, c)$  to predict an estimate of the clean embedding,  $\hat{e}_{h,0} = G_\theta(e_{h,\tau}, \tau, c)$ , where  $c = \text{condition}(e_l, e_r)$  represents the combined conditioning information. The model is trained by minimizing the mean squared error between the true clean embedding and its prediction:

$$L_{\text{diffusion}} = \mathbb{E}_{\tau, e_{h,0}, \varepsilon} [\|e_{h,0} - G_\theta(e_{h,\tau}, \tau, c)\|_2^2]$$

The sampling process proceeds as below:

We begin by sampling  $e_{h,\tau}$  from a standard isotropic Gaussian distribution,

$$e_{h,\tau} \sim \mathcal{N}(0, \mathcal{I})$$

For each timestep  $\tau$  from T down to 1, to obtain  $e_{h,\tau-1}$  from  $e_{h,\tau}$  we estimate  $\hat{e}_{h,0}$  from the model  $G_\theta$ ,

$$\hat{e}_{h,0} = G_\theta(e_{h,\tau}, \tau, c)$$

Using the predicted  $\hat{e}_{h,0}$  and the current  $e_{h,\tau}$ , the mean of the posterior distribution  $q(e_{h,\tau-1}|e_{h,\tau}, \hat{e}_{h,0})$  denoted by  $\tilde{\mu}_t(e_{h,\tau}, \hat{e}_{h,0})$  is calculated as:

$$\tilde{\mu}_t(e_{h,\tau}, e_{h,0}) = \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\hat{e}_{h,0} + \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}e_{h,\tau}$$

The sample for the previous timestep is then drawn by adding scaled Gaussian noise to the posterior mean:

$$e_{h,\tau-1} = \tilde{\mu}_t(e_{h,\tau}, e_{h,0}) + \sigma_t z$$

where  $z \sim \mathcal{N}(0, \mathcal{I})$  if  $\tau > 1$  and  $z=0$  if  $\tau = 1$ . The choice of noise variance  $\sigma_\tau^2 = \beta_\tau$  follows from [19]. So the sampling step becomes,

$$e_{h,\tau-1} = \tilde{\mu}_t(e_{h,\tau}, e_{h,0}) + \sqrt{\beta_\tau} z$$

Thus the  $\hat{e}_{h,0}^{final}$  is the generated high-resolution embedding, which is then passed to the frozen decoder  $D_{HSM}$  to produce the final segmentation map.

### C.3.3. Spatially Feature Transform (SFT) blocks

To specifically introduce and leverage the distinct characteristics of the low-resolution input ( $e_l$ ) and the reference input ( $e_r$ ), we employ Spatially Feature Transform (SFT) blocks at different stages within the decoder. Figure 7 shows a detailed architecture of the SFT Block, Convolution Block, Encoder Block and Decoder Block as shown in Figure 2. Architecturally, each SFT block first processes its conditioning input through a pair of 3x3 convolutional layers with ReLU activations (transforming channels to 128, then 64).

Let  $c_{in}$  be the input conditioning in the SFT block, the convolutional block followed by ReLU produces,

$$C_1 = \text{ReLU}(\text{Conv}_{128,3 \times 3}(c_{in}))$$

Progressively we apply another convolutional block to reduce the channels to 64,

$$C_2 = \text{ReLU}(\text{Conv}_{64,3 \times 3}(C_1))$$

The output  $C_2$  is then bilinearly upsampled to match the spatial dimensions (H,W) of the main feature map F that it will modulate, resulting in  $C_{up}$ . Subsequently, from the upsampled conditioning features  $C_{up}$ , two separate 3x3 convolutional branches, are used to predict spatial-specific modulation parameters: a scale factor ( $\gamma$ ) and a shift factor ( $\beta$ ). Both  $\gamma$  and  $\beta$  will have the same number of channels ( $n_F$ ) as the main feature map F.

$$\gamma = \sigma(\text{Conv}_{n_F,3 \times 3}(C_{up}))$$

$$\beta = \text{Conv}_{C_F,3 \times 3}(C_{up})$$

Finally, the learned scale ( $\gamma$ ) and shift ( $\beta$ ) parameters are applied to the main feature map F using an element-wise affine transformation. The feature map F is first scaled by  $\gamma$  and then the shift  $\beta$  is added such that,

$$F_{out} = \gamma \odot F + \beta$$

## C.4. Layerwise Dimensions in Our Model

Table 5 displays the layerwise dimensions of SEED-SR's denoising diffusion model.

## C.5. Changes in Ablation studies

Table 6 specifies the layerwise changes in model dimensions because of the different inputs of ablation studies.

Table 11. Layerwise Dimensions of our Model

Layer	Dimension
Input Noised HR	(120,120,3840)
Input Ref	(120,120,3840)
Input LR	(32,32,64)
Input Timestep	(1,)
Upsampled LR	(120,120,64)
Concatenated Encoder Input	(120,120,7744)
Encoder Block 1	(60,60,320)
Encoder Block 2	(30,30,480)
Encoder Block 3	(15,15,560)
Encoder Block 3 with Timestep Embedding	(15,15,560)
Encoder Block 4	(7,7,640)
Attention Block	(7,7,640)
Bottleneck Conv Block	(7,7,784)
Decoder Block 1	(15,15,640)
SFT Block 1	(15,15,640)
Decoder Block 2	(30,30,560)
Decoder Block 3	(60,60,480)
SFT Block 2	(60,60,480)
Decoder Block 3 with SFT and Timestep Embedding	(60,60,480)
Decoder Block 4	(120,120,320)
Attention Block	(120,120,320)
Convolution Block	(120,120,3840)

Table 12. Layerwise Input Dimension Changes for Ablation Studies

(a) Config A: S2 + Ref, no embeddings	
Layer	Dimension
Input Noised HR	(480,480,3)
Input Ref	(480,480,3)
Input LR	(32,80,80,12)
Input Timestep	(1,)
Reshaped and Upsampled LR	(480,480,384)
Concatenated Encoder Input	(480,480,390)
:	:
Convolution Block	(480,480,3)

(b) Config B: Only S2 images, no embeddings	
Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	(32,80,80,12)
Input Ref	NA
Input Timestep	(1,)
Upsampled LR	(120,120,384)
Concatenated Encoder Input	(120,120,4224)
:	:
Convolution Block	(120,120,3840)

(a) Config C: Only LR Embeddings	
Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	(32,32,64)
Input Ref	NA
Input Timestep	(1,)
Upsampled LR	(120,120,64)
Concatenated Encoder Input	(120,120,3904)
:	:
Convolution Block	(120,120,3840)

(b) Config D: Only Ref Embeddings	
Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	NA
Input Ref	(120,120,3840)
Input Timestep	(1,)
Upsampled LR	NA
Concatenated Encoder Input	(120,120,7680)
:	:
Convolution Block	(120,120,3840)

(a) Config E: Ref Embeddings with S2 Images	
Layer	Dimension
Input Noised HR	(120,120,3840)
Input Ref	(120,120,3840)
Input LR	(32,80,80,12)
Input Timestep	(1,)
Reshaped and Upsampled LR	(120,120,384)
Concatenated Encoder Input	(120,120,8064)
:	:
Convolution Block	(120,120,3840)