RECOVERING CLOUD MICROSTRUCTURES WITH CASCADED DIFFUSION INVERSION

Hanan Gani[1,3] **Guy Pulik**[2] **Daniel Rosenfeld**[2] **Duncan Watson-Parris**[3] **Salman Khan**[1]
Correspondence to: {hanan.ghani, salman.khan}@mbzuai.ac.ae
[1]Mohamed Bin Zayed University of Artificial Intelligence [2]Hebrew University of Jerusalem
[3]University of California, San Diego

ABSTRACT

High-resolution satellite imagery is critical for observing fine-scale cloud structures that inform weather modification strategies like cloud seeding for rain-enhancement. However, the spatial resolution of current geostationary and polar-orbiting satellites is often insufficient for capturing small cloud features. Current super-resolution methodologies are suited for natural images and, therefore, struggle to generalize to satellite-captured spectral images of cloud cover. To address this, we propose a two-stage diffusion-based super-resolution framework to enhance the resolution of multi-spectral cloud microstructures by a factor of 4×. Specifically, we use inverse diffusion to recover the high resolution properties from low resolution. Stage 1 utilizes real-world paired data to learn robust degradation handling and inter-sensor alignment, while Stage 2 employs a self-supervised internal downgrading of high resolution data to refine structural learning and texture synthesis. Our approach outperforms the state-of-the-art transformer and diffusion-based baselines in both reconstruction accuracy and visual quality. We demonstrate that the two-stage method better captures fine cloud microstructures (e.g. convective turrets and cloud gaps) that are crucial for effective cloud seeding decisions. Ablation studies confirm the complementary benefits of the two stages: Stage 1 excels in coarse structural fidelity, while Stage 2 contributes enhanced detail and realism. These results highlight a practical path toward improving cloud microphysics analysis and as a step towards utilizing AI for climate and sustainability. Our code and models are available at: https://github.com/hananshafi/superresolution-cloud-microphysics

1 INTRODUCTION

High-resolution meteorological satellite cloud imagery is essential for diagnosing localized weather phenomena and for operations such as weather modification (e.g., cloud seeding), which relies on identifying suitable microphysical targets such as supercooled liquid regions (World Meteorological Organization, 2025; NOAA Atlantic Oceanographic and Meteorological Laboratory). However, spatial and temporal resolution trade off sharply across current observing systems. Modern geostationary (GEO) imagers provide rapid refresh but remain spatially coarse: for example, GOES-R ABI offers ∼0.5–2 km sampling depending on band (NOAA/NASA GOES-R Series), while Meteosat Second Generation (MSG) provides full-disk imagery every 15 minutes from geostationary orbit (EUMETSAT), and its SEVIRI imager samples at ∼3 km at the sub-satellite point (World Meteorological Organization). Moreover, the effective GEO pixel footprint grows away from the sub-satellite point (CEDA, 2017), blurring sub-kilometer structures that are often diagnostically important.

Low-Earth-orbit (LEO) instruments such as VIIRS resolve substantially finer detail (375 m and 750 m native resolutions at nadir) but provide limited temporal sampling at a fixed location, since global coverage is achieved via orbital revisits rather than continuous staring (NASA JPL (PO.DAAC)). For cloud-focused applications, this motivates computational super-resolution (SR) that aims to recover LEO-like spatial detail while retaining GEO-like update frequency.

This setting is particularly challenging because the SEVIRI $\rightarrow$ VIIRS (and MSG $\rightarrow$ MTG) mapping is cross-sensor and time-shifted, with imperfect alignment and non-stationary degradations. Deterministic SR methods optimized for pixel fidelity, including transformer-based models (Liang et al., 2021a), tend to oversmooth high-variance cloud textures and degrade under cross-sensor distribution shift. Diffusion-based (Song et al., 2021; Ho et al., 2020b) SR can synthesize sharper detail, but may introduce artifacts or inconsistent structure when applied outside the natural-image priors they inherit (Ho et al., 2020a; Wang et al., 2024a;b).

We address these limitations with a two-stage curriculum for diffusion inversion SR (Yue et al., 2025) that explicitly decouples cross-sensor robustness from texture recovery (See Fig. 1). Stage 1 trains on real paired SEVIRI $\rightarrow$ VIIRS / MSG $\rightarrow$ MTG data to learn a physically consistent mapping under temporal and geometric mismatch. Stage-2 uses prior from stage-1 and trains on HR-only imagery with synthetic degradations to learn cloud-specific fine-scale structure in a perfectly aligned setting. The resulting cascade improves both distortion and structure preservation, producing sharper yet more consistent reconstructions for multi-spectral cloud imagery.

2 Related Work

Image Super-Resolution. Deterministic SR methods—from bicubic interpolation and early CNN baselines (Dong et al., 2016; Kim et al., 2016) to modern transformer architectures (Liang et al., 2021b; Conde et al., 2022)—typically minimize pixel-wise distortion, which often yields oversmoothed textures under large scaling factors and imperfect supervision. Diffusion-based (Ho et al., 2020c; Song et al., 2020) SR (Wang et al., 2023; 2021)) can synthesize sharper detail by leveraging strong generative priors, but their stochastic generation can introduce artifacts or structurally inconsistent textures when transferred to new domains. We build on diffusion inversion (Xiao et al., 2020), which frames SR through an invertible latent variable formulation. This perspective explicitly models lost high-frequency information and enables a curriculum that separates cross-sensor robustness from detail synthesis, improving structural fidelity without relying on unconstrained hallucination.
Multi-Spectral Cloud Microphysics. Multi-spectral satellite observations are widely used to study cloud microphysics (e.g., optical thickness and effective radius). Geostationary sensors such as SEVIRI on MSG provide high temporal sampling for tracking cloud evolution but at coarse spatial resolution (Schmetz et al., 2002). In contrast, polar-orbiting instruments such as VIIRS provide finer spatial detail but with limited revisit frequency (Cao et al., 2013). Bridging this spatio-temporal gap is important for cloud-focused applications, motivating cross-sensor super-resolution. Our work targets this setting and additionally considers next-generation Meteosat Third Generation satellite system (MTG) (Holmlund et al., 2021).

3 Methodology

We formulate cloud super-resolution as an inverse problem: given a low-resolution (LR) observation $\mathbf{x} \in \mathbb{R}^{h\times w\times C}$ (SEVIRI/MSG), we recover a high-resolution (HR) microphysical field $\mathbf{y} \in \mathbb{R}^{H\times W\times C}$ (VIIRS/MTG), with $H \gg h$. To solve this, we adopt the framework of Xiao et al. (2020), which leverages the rich priors of pre-trained diffusion model (e.g., SD-Turbo) via a diffusion inversion mechanism. Unlike standard diffusion methods that start from random Gaussian noise, our approach learns to predict a deterministic starting point conditioned on the LR input, ensuring fidelity to the observed satellite data. We propose *Cascaded Invertible Framework* that decouples this problem into two sub-tasks: (1) Cross-sensor domain alignment, & (2) High-frequency structure restoration.
Stage 1: Cross-Sensor Alignment (Learning the Prior). The primary challenge in satellite SR is the domain gap caused by sensor differences and spatiotemporal misalignment between geostationary and polar imagery. We train the first network, $\mathcal{M}_1$, on paired real-world data $\mathcal{D}_{\text{real}} = \{(\mathbf{x}_{\text{LR}}, \mathbf{y}_{\text{HR}})\}$. In the forward training pass, we enforce that the downscaled representation of HR matches the LR input, effectively teaching the model to register the HR physics to the LR geometry. Given an LR–HR pair, we first reconstruct the HR image $\hat{\mathbf{y}}_{\text{HR}\leftarrow t}$ by initializing the diffusion reverse process at a selected starting timestep t using the LR-conditioned noise predictor, then running a short reverse sampling chain: $\mathcal{L}_{\text{S1}} = \mathcal{L}_2(\hat{\mathbf{y}}_{\text{HR}\leftarrow t}, \mathbf{y})$.
Stage 2: Structural Refinement. While Stage 1 aligns the data, the inherent noise in real-world pairings prevents it from learning sharp high-frequency textures. Stage 2, $\mathcal{M}_2$, focuses purely on recovering convective structures. We construct a synthetic dataset $\mathcal{D}_{\text{syn}}$ by applying a degradation

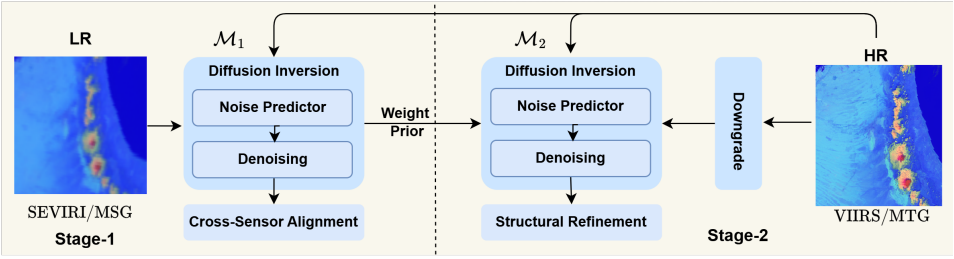

Figure 1: **Two-stage InvSR training.** Stage 1 trains InvSR on paired SEVIRI→VIIRS/MTG→MSG data to learn cross-sensor alignment. Stage 2 trains the same InvSR formulation on HR-only data with synthetic degradations to emphasize high-frequency cloud structures.

Table 1: Quantitative SR results on SEVIRI → VIIRS and MSG → MTG.

	SEVIRI→VIIRS			**MSG→MTG**		
Model	PSNR↑	Grad.(ideal=1.0)	Percep↓	PSNR↑	Grad.(ideal=1.0)	Percep↓
SwinIR	18.37	0.19	0.45	15.91	0.53	0.75
StableSR	20.52	1.72	0.44	18.71	2.94	0.42
SinSR	20.69	0.46	0.37	24.0	0.96	0.30
Ours	**21.25**	**1.06**	**0.28**	**24.0**	**1.03**	**0.29**

kernel k to HR images such that, $\mathbf{x}_{\text{syn}} = \mathbf{y}_{\text{HR}} \downarrow_k$. $\mathcal{M}_2$ is trained to map $\mathbf{y}_{\text{HR}}$ to $(\mathbf{x}_{\text{syn}}, \mathbf{z})$. Because the input and target are perfectly aligned, the network focuses entirely on modeling the distribution of fine details in $\mathbf{z}$. We follow the same optimization procedure as in stage-1. Further, in both stages 1 and 2, we include an additional perceptual similarity loss.

Inference. At test time, given a real LR input $\mathbf{x}_{\text{LR}}$, we run the standard InvSR inversion-and-sampling procedure using the learned noise predictor (typically with 1-step sampling for efficiency). Optionally, we apply a cascaded pass $\mathbf{x}_{\text{LR}} \xrightarrow{\mathcal{M}_1} \hat{\mathbf{y}}_{\text{base}} \xrightarrow{\mathcal{M}_2} \hat{\mathbf{y}}_{\text{final}}$ when additional cross-sensor correction is beneficial.

4 Experimentation

We base our approach on Xiao et al. (2020). The data for training and validation multi-spectral cloud cover is sourced from UAE's Rain Enhancement Program under National Center for Metrology (UAEREP). Our data includes 9k train and 1k validation low and high resolution pairs corresponding to SEVIRI and VIIRS respectively. We further scale our approach to channel-wise data obained across MSG and MTG satellites, for which we obtained around 2250 training pairs and 500 validation pairs. The training for both stages takes around 5 days on two Nvidia A100 40GB GPUs. We compare against an exsting transformer-based approach (Liang et al., 2021b) and two recent diffusion-based approaches (Wang et al., 2023; 2021). For empirical evaluation, we use PSNR, Gradient Preservation and Perceptual distance ratio to adjudge the quality of super-resolution. Refer to Appendix for additional details on data specifications and channel alignment.

Quantitative Results. Table 1 summarizes performance on SEVIRI→VIIRS pairs and channel-wise MSG→MTG pairs using three complementary metrics: PSNR (distortion), gradient preservation ratio (structure; ideal = 1), and perceptual distance (lower is better). On SEVIRI→VIIRS, our method achieves the best overall balance, improving PSNR to 21.25 dB while producing the lowest perceptual distance (0.28). In contrast, SwinIR underperforms on cross-sensor data (18.37 dB) and severely under-recovers high-frequency structure (grad. ratio 0.19). StableSR yields a high gradient ratio (1.72), indicating over-sharpening relative to the target gradient statistics with high perceptual distance (0.44) while taking 100 denoisign steps. SinSR improves over SwinIR in both PSNR and perceptual distance, but remains structurally under-sharp (gradient ratio 0.46) compared to our near-ideal 1.06. On MSG→MTG, our method matches the best PSNR (24.0 dB) while achieving the lowest perceptual distance (0.29) and the closest-to-ideal gradient ratio (1.03). StableSR again exhibits a strongly inflated gradient ratio (2.94), consistent with oversharpening/artifacts, whereas SwinIR lags substantially in both PSNR (15.91 dB) and perceptual distance (0.75). Overall, the

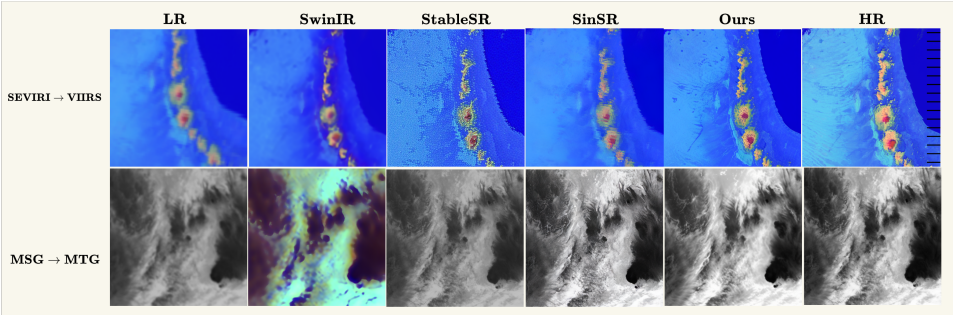

Figure 2: **Qualitative Comparison.** Most baselines produce blurry and over-sharpened outputs, however, our approach ticks the correct balance between sharpness and smoothing, constantly adhering to the ground-truth HR structure (best viewed in zoom).

results support our proposed two-stage diffusion-inversion training which improves fidelity (PSNR) and preserves physically meaningful structure (grad. ratio$\sim$1) without increasing perceptual error.

Qualitative Analysis. Fig. 2 shows visual comparisons for SEVIRI$\rightarrow$VIIRS (top) and MSG$\rightarrow$MTG (bottom). On SEVIRI$\rightarrow$VIIRS, SwinIR largely preserves the low-frequency appearance but oversmooths the narrow cloud band and convective cores, failing to recover the fine filaments visible in HR. StableSR produces noticeably sharper outputs, but the sharpening comes with widespread speckle / grid-like artifacts and noisy textures (inflated gradient ratio), which do not correspond to the HR structures. SinSR recovers some mid-frequency details but still misses thin streaks and tends to smooth background texture. In contrast, our method reconstructs coherent high-frequency structure (crisper boundaries and more faithful cloud-cell shapes) while avoiding spurious texture, yielding the closest visual match to HR. On MSG$\rightarrow$MTG, SwinIR exhibits severe appearance/contrast artifacts under the cross-sensor setting, while StableSR and SinSR sharpen cloud regions but can introduce granularity or wash out fine wisps. Our approach preserves the mesoscale cloud streaks and local gradients more consistently, producing fine-scale patterns that align with the HR target without obvious hallucinations.

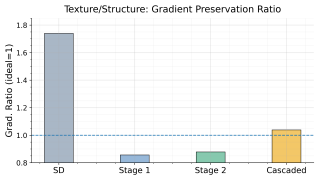

Figure 3: Effectiveness of each stage.

Discussion. Most SR baselines implicitly assume (a) a known, stationary degradation (often bicubic), (b) tightly aligned LR/HR supervision, and (c) a single-sensor imaging model. In SEVIRI $\rightarrow$ VIIRS and MSG $\rightarrow$ MTG, these assumptions are violated: the mapping is cross-sensor, time-shifted, and affected by different PSFs, noise characteristics, and spectral responses. As a result, distortion-driven models tend to blur (regression-to-the-mean under misalignment), while perceptual/diffusion baselines can generate sharp but unreliable textures. Fig. 3 helps disentangle the role of each training stage using the gradient preservation ratio (ideal = 1.0). The single-stage variants remain under-sharp: Stage-1 improves robustness but retains smoothing (ratio 0.86), and Stage-2 slightly improves structure (ratio 0.88) but is still conservative. The cascaded two-stage training moves the ratio close to the physically desirable regime (1.04), indicating substantially better recovery of fine-scale cloud boundaries and filaments. In contrast, the diffusion baseline (SD) overshoots the target gradients (ratio 1.74), consistent with over-sharpening artifacts rather than faithful structure. Overall, the plot supports our main claim: Stage-1 provides cross-sensor robustness, Stage-2 injects aligned structural supervision, and their combination is necessary to achieve near-ideal performance.

5 Conclusion

We presented a two-stage diffusion-inversion super-resolution framework for multi-spectral cloud imagery in challenging cross-sensor settings, including SEVIRI$\rightarrow$VIIRS and MSG$\rightarrow$MTG. The key idea is to decouple *robust cross-sensor mapping* from *fine-detail recovery*: Stage 1 leverages real paired observations to handle temporal and geometric mismatch, while Stage 2 uses HR-only self-supervision with synthetic degradations to recover high-frequency cloud structure under aligned training. Across both benchmarks, our approach achieves a stronger trade-off between distortion, structural fidelity (gradient preservation), and perceptual distance than transformer- and diffusion-based baselines, producing sharper yet more consistent reconstructions.

REFERENCES

Changyong Cao, Frank J De Luccia, Xiaoxiong Xiong, Robert Wolfe, and Fuzhong Weng. Early on-orbit performance of the visible infrared imaging radiometer suite onboard the suomi national polar-orbiting partnership (s-npp) satellite. *IEEE Transactions on Geoscience and Remote Sensing*, 52(2):1142–1156, 2013.

CEDA. Seviri (spinning enhanced visible and infrared imager) fire radiative power (frp) data from the meteosat second generation (msg) satellite. https://www.data.gov.uk/dataset/dfffd1c3-101b-44f1-bda9-0c43c6a281c9/, 2017. Accessed: 2026-02-06.

Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pp. 669–687. Springer, 2022.

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, 2016. doi: 10.1109/TPAMI.2015.2439281.

EUMETSAT. Meteosat second generation. https://www.eumetsat.int/meteosat-second-generation. Accessed: 2026-02-06.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020c.

Kenneth Holmlund, Jochen Grandell, Johannes Schmetz, Rolf Stuhlmann, Bojan Bojkov, Rose Munro, et al. Meteosat third generation (mtg): Continuation and innovation of geostationary earth observation from europe. *Bulletin of the American Meteorological Society*, 102(5):E990–E1015, 2021.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1833–1844, October 2021a.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021b.

NASA JPL (PO.DAAC). Suomi national polar-orbiting partnership (suomi npp) — po.daac / jpl / nasa. https://podaac.jpl.nasa.gov/S-NPP. Accessed: 2026-02-06.

NOAA Atlantic Oceanographic and Meteorological Laboratory. Cloud seeding. https://www.aoml.noaa.gov/hrd/hrd_sub/cseed.html. Accessed: 2026-02-06.

NOAA/NASA GOES-R Series. Abi — goes-r series (advanced baseline imager). https://www.goes-r.gov/spacesegment/abi.html. Accessed: 2026-02-06.

Johannes Schmetz, Paolo Pili, Stephen Tjemkes, Dieter Just, Jochen Kerkmann, S Rohn, and A Roth. The meteosat second generation (msg) seviri instrument. *Bulletin of the American Meteorological Society*, 83(7):977–992, 2002.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Jianyi Wang, Zongsheng Dong, Liang Xie, Ying Shan, and Anup Basu. Exploiting diffusion prior for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2657–2666, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132:5929–5949, 2024a. doi: 10.1007/s11263-024-02168-7.

Longguang Wang, Yingqian Wang, Zaiping Dong, Qingyu Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4801–4810, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25796–25805, June 2024b.

World Meteorological Organization. Wmo oscar — details for instrument seviri. https://space.oscar.wmo.int/instruments/view/seviri. Accessed: 2026-02-06.

World Meteorological Organization. Wmo statement on weather modification. https://wmo.int/content/wmo-statement-weather-modification, 2025. Accessed: 2026-02-06.

Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaonan Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *European Conference on Computer Vision (ECCV)*, pp. 126–144. Springer, 2020.

Zongsheng Yue, Xinyuan Chen, Shangchen Zhou, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17746–17755, June 2025.

A Appendix

A.1 Band Selection and Pair Construction

A.1.1 Channels used and cross-sensor correspondence

A central challenge in cross-sensor super-resolution is that "matching channels" are only approximate: sensors differ in their spectral response functions (SRFs), viewing geometry, calibration, point-spread functions (PSFs), and noise characteristics. Consequently, even when two bands are nominally similar (e.g., "0.6μm reflectance"), the measured radiance/reflectance/brightness temperature will not be identical. We therefore choose channels that (i) measure the *same physical quantity* (reflectance or brightness temperature), and (ii) are the *closest available spectral counterparts* across sensors. Any residual mismatch is treated as part of the domain gap handled by Stage 1 real-pair training.

SEVIRI→VIIRS (composite). For SEVIRI→VIIRS we represent each scene using a 3-channel "day microphysics" style composite consisting of two reflective bands and one thermal band. This choice provides a compact but informative representation of cloud morphology and thermodynamic structure: the reflective channels emphasize cloud optical properties and fine boundaries, while the thermal channel captures cloud-top temperature patterns and mesoscale organization.

- **VIIRS.** We use I-band SDRs: *I_01* (reflectance), *I_03* (reflectance), and *I_05* (brightness temperature).
- **SEVIRI.** We use the closest SEVIRI counterparts: *VIS006* (0.6μm reflectance), *IR_016* (1.6μm reflectance), and *IR_108* (10.8μm brightness temperature).

In our preprocessed SEVIRI tensor (25 layers), longitude/latitude are stored at indices $[0, 1]$, and the three composite channels used for learning correspond to indices $[6, 8, 22]$ (i.e., `-3` for the thermal band), consistent with the VIS006/IR_016/IR_108 triplet. We apply fixed clipping and normalization to stabilize training and enable consistent dynamic ranges across sensors: reflective channels are scaled to $[0, 1]$ and $[0, 0.8]$, and the thermal channel is clipped to $[203, 323]$ K prior to normalization.

Quantity	SEVIRI (MSG)	VIIRS
Visible reflectance	VIS006 (0.6μm)	I_01 (0.64μm)
NIR/SWIR reflectance	IR_016 (1.6μm)	I_03 (1.61μm)
Thermal IR (BT)	IR_108 (10.8μm)	I_05 (11.45μm)

Table 2: Channel correspondence used to form the SEVIRI→VIIRS 3-channel composites. Channels are matched by physical quantity and nearest spectral counterpart. Differences in SRF/PSF remain and constitute part of the cross-sensor gap addressed in Stage 1.

MSG→MTG (channel-wise pairs). For MSG→MTG we create *channel-wise* paired samples to evaluate and train across a broader set of spectral bands. We load MTG/FCI L1C using Satpy (`reader=fci_l1c_nc`) and MSG/SEVIRI L1B using Satpy (`reader=seviri_l1b_native`), then match each MTG band to its closest MSG counterpart by spectral region and physical quantity (reflectance vs brightness temperature). We restrict to the overlapping subset of channels for which a clear correspondence exists.

Reflectance channels are scaled to $[0, 1]$ (including percent-to-fraction conversion when needed). Brightness temperature channels are normalized using a fixed physical range of $[180, 330]$ K before saving patches (16-bit PNG), ensuring consistent scaling across channels.

A.1.2 Grid alignment for paired samples

SEVIRI–VIIRS alignment (geolocation-based regridding). To create aligned SEVIRI–VIIRS pairs we use per-pixel geolocation grids. VIIRS longitude/latitude are read from the VIIRS geolocation product, while SEVIRI longitude/latitude are read from the SEVIRI grid (stored as the first two layers in the SEVIRI tensor). We first identify a SEVIRI crop whose latitude/longitude extent

Quantity	MTG/FCI	MSG/SEVIRI
Visible reflectance	vis_06	VIS006
Visible reflectance	vis_08	VIS008
NIR reflectance	nir_16	IR_016
Water vapor (BT)	wv_63	WV_062
Water vapor (BT)	wv_73	WV_073
Thermal IR (BT)	ir_38	IR_039
Thermal IR (BT)	ir_87	IR_087
Thermal IR (BT)	ir_97	IR_097
Thermal IR (BT)	ir_105	IR_108
Thermal IR (BT)	ir_123	IR_120
Thermal IR (BT)	ir_133	IR_134

Table 3: Channel-wise correspondence used for MSG→MTG. Channels are matched by physical quantity and nearest spectral region; exact SRFs differ across sensors.

lies within the SEVIRI scene bounds (via crop-wise min/max checks) to guarantee overlap. We then reproject VIIRS onto the SEVIRI grid using nearest-neighbor interpolation in $(\mathrm{lon}, \mathrm{lat})$ space (via `scipy.interpolate.griddata`), producing co-registered pairs on a shared grid for training and evaluation. Remaining pixel-level discrepancies can persist due to time offset, differing view angles, and residual navigation errors; these are handled implicitly by Stage-1 training on real paired data.

MSG–MTG alignment (Satpy area resampling). For MSG–MTG we rely on Satpy's projection-aware resampling. We load MTG/FCI and MSG/SEVIRI scenes and resample one scene to the other's native `area` (default: `MSG_TO_MTG`) using Satpy's `Scene.resample` with a configurable resampler (bilinear in our generation pipeline). Optionally, we crop both scenes to a lat/lon bounding box (`crop(ll_bbox=...)`). After resampling, arrays share the same spatial grid and shape, enabling direct extraction of aligned channel-wise LR/HR patches.

A.2 Metric Definitions

We evaluate super-resolution quality using three complementary metrics. **Peak Signal-to-Noise Ratio (PSNR)** measures pixel-level reconstruction fidelity between the super-resolved output $\hat{y}$ and the high-resolution reference y:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}(\hat{y},\, y)}\right), \tag{1}$$

where MAX is the maximum pixel intensity and MSE is the mean squared error. Higher values indicate closer pixel-level agreement. **Gradient Preservation Ratio (Grad.)** measures how faithfully the model recovers physically meaningful fine-scale cloud structure, defined as the ratio of mean gradient magnitudes in the super-resolved output versus the HR reference:

$$\mathrm{Grad.} = \frac{\mathbb{E}[\|\nabla \hat{y}\|]}{\mathbb{E}[\|\nabla y\|]}, \tag{2}$$

where ∇ denotes image gradients computed via a Sobel operator. A ratio of 1.0 is ideal: values below 1 indicate over-smoothing, while values above 1 indicate over-sharpening or spurious texture. **Perceptual Distance Ratio (Percep.)** quantifies perceptual similarity of the super-resolved output relative to a bicubic baseline using deep feature distances. Let $d(A, B) = \|\phi(A) - \phi(B)\|_2$ denote the LPIPS distance between images A and B, where $\phi(\cdot)$ extracts VGG feature embeddings. The ratio is defined as:

$$\mathrm{Percep.} = \frac{d(\hat{y},\, y)}{d(\hat{y}_{\mathrm{bic}},\, y)}, \tag{3}$$

where $\hat{y}_{\mathrm{bic}}$ is the bicubic upsampled output. Values below 1 indicate better perceptual fidelity than the bicubic baseline; lower is better.

A.3 Additional Qualitative Results

Figures 4 and 5 present further visual comparisons across representative cloud scenes for both SEVIRI→VIIRS and MSG→MTG settings, respectively. Across both configurations, our method consistently recovers coherent high-frequency cloud structures including convective turrets, thin filaments, and sharp cloud-gap boundaries. Our approach produces outputs that most closely match the ground-truth HR structure in both spatial organisation and local gradient fidelity, with less hallucinating spurious textures.

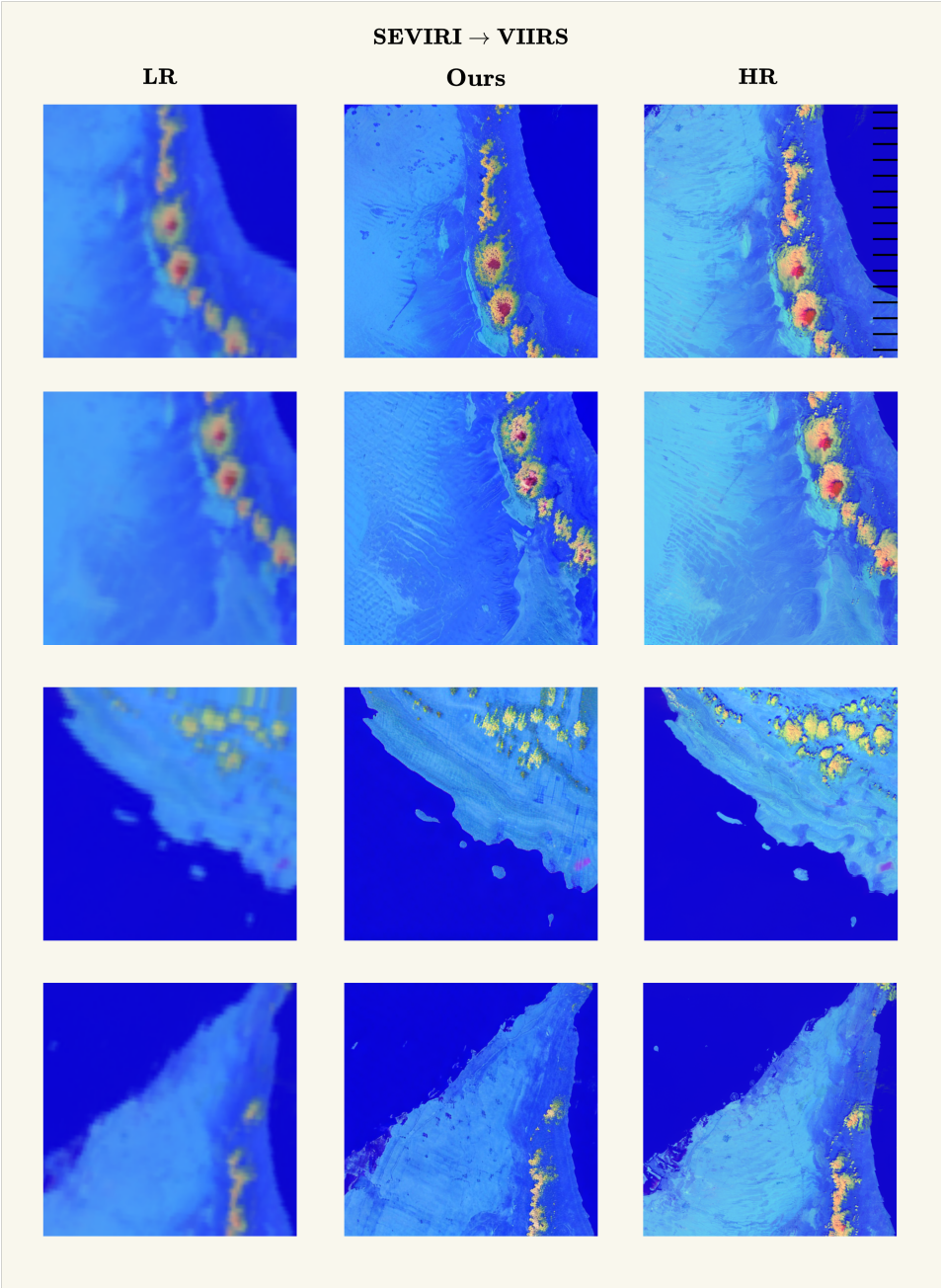

Figure 4: **Qualitative Comparison SEVIRI→VIIRS** (best viewed in zoom).

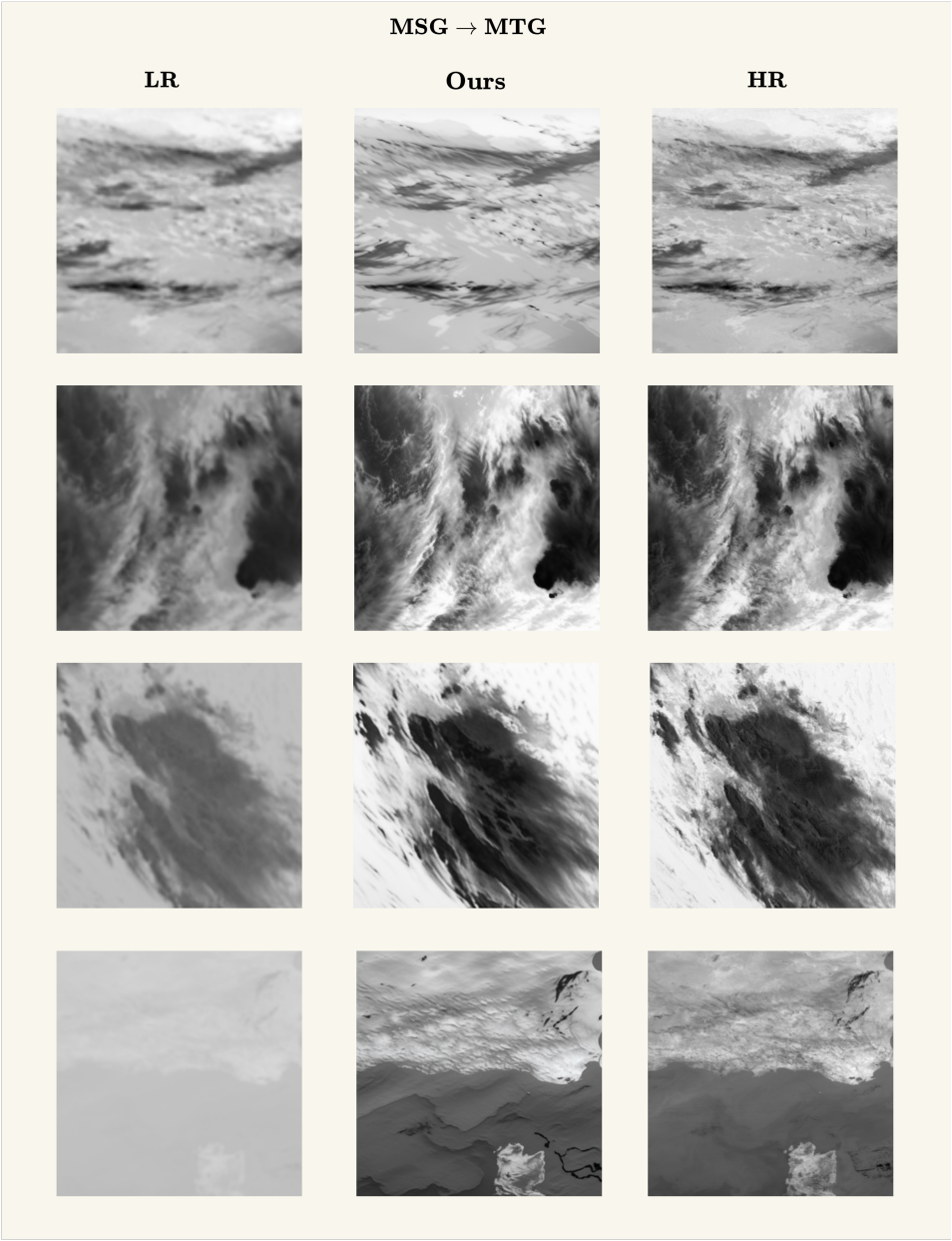

Figure 5: **Qualitative Comparison MSG→MTG** (best viewed in zoom).

A.4 Reproducibility Statement

Our code is publicly available at `https://github.com/hananshafi/superresolution-cloud-microphysics`. While the UAEREP dataset is subject to institutional data-sharing agreements and cannot be released in full, we will release our trained model checkpoints and evaluation scripts upon publication to enable usage on publicly available data samples.