SELF-SUPERVISED PRETRAINING FOR POWER GENERATION VIA DYNAMIC GUIDED MASKING

Ananya Shriram, Advika Srivastava[†], Ishita Choudhary[†],
Ananya Das[‡], Sarvesh Kumar[‡]
Manipal Institute of Technology, Manipal, India
{ananya4.mitmpl2024, advika.mitmpl2024, ishita2.mitmpl2024,
ananya6.mitmpl2024, sarvesh.mitmpl2023}@learner.manipal.edu

ABSTRACT

The escalating reliance on fossil fuel power plants remains a critical driver of global greenhouse gas emissions, necessitating precise and scalable monitoring systems for climate change mitigation. Traditional power generation estimation methods rely on bottom-up self reporting methods, which are time-consuming and subjective. These limitations have motivated growing interest in remote sensing based approaches, where satellite observations provide an objective, scalable, and globally consistent alternative. We present a self-supervised learning (SSL) approach tailored for multi-spectral satellite imagery. Unlike traditional methods, our methodology utilizes a dynamic and guided masking pretext task that forces the model to internalize latent features from high-priority spectral bands. Experimental results demonstrate that the proposed framework beats existing benchmarks in regression and segmentation by 7.7% and 14.1% respectively, achieving superior accuracy in power generation estimation without requiring auxiliary segmentation or classification labels.

1 INTRODUCTION

Rising CO_2 and other greenhouse gas emissions are the primary drivers of extreme climate events such as intense hurricanes, glacier loss, and flooding (Solomon et al., 2009). Fossil fuel based power generation remains the largest contributor to these emissions (Janssens-Maenhout et al., 2019), making accurate and continuous monitoring of individual power plants essential for climate mitigation and treaty enforcement.

Early power estimation relied on bottom up inventories and self reporting, which were time consuming and lacked transparency (Nassar et al., 2017). More recent deep learning approaches using satellite imagery, particularly multi task frameworks (Hanna et al., 2023), have proven effective for independent CO_2 estimation. However, despite improved performance, these methods often suffer from task interference and negative gradient transfer (Liu et al., 2019).

In this work, we adopt a self-supervised framework with a dynamic, sharpness-driven masking strategy that prioritizes high-fidelity spectral features, achieving superior performance across diverse plant configurations.

2 RELATED WORK

Early satellite-based power generation estimates relied on proxies such as cooling-tower plume characteristics, but varying cooling technologies led to inconsistent signals and uneven data (Couture et al., 2024). Convolutional neural networks better capture activity-related patterns missed by such methods while transformers handle long range dependencies efficiently (Khirwar & Narang, 2024), while multitask learning further improves power estimation by exploiting shared operational correlations (Hanna et al., 2023).

[†]Second authors
[‡]Third authors

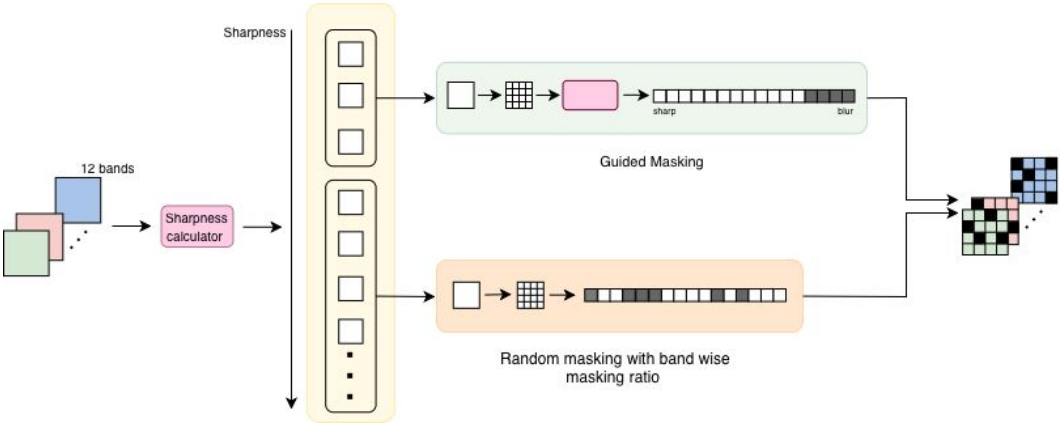

Figure 1: Sharpness-aware dynamic and guided masking strategy.

Self-supervised learning, particularly Masked Autoencoders (MAEs), mitigates limited labeled data by learning robust representations from unlabeled satellite imagery (Deb & Das, 2025). Extensions such as SatMAE (Cong et al., 2023), SSMAE (Lin et al., 2023), SpectralMAE (Zhu et al., 2023), Scale-MAE (Reed et al., 2023), and PRESTO (Tseng et al., 2023)adapt MAEs to multi-spectral, geospatial, and multi-sensor satellite data, but rely largely on random masking, which ignores varying information density across spectral bands. Recent work explores informed masking using Histogram of Oriented Gradients(HOG) features (Wang et al., 2024) or attention-based dynamic strategies (Chatterjee et al.). Accounting for spectral heterogeneity and spatial quality is crucial, as degradation like blur and loss of high-frequency detail reduces perception reliability (Pagaduan et al., 2020). Building on this, we introduce a sharpness driven metric to guide band selection and masking based on spatial informativeness.

3 Dataset

Our approach is evaluated on the dataset presented in Hanna et al. (2023) which was built on the dataset from Hanna et al. (2021) The dataset includes 2204 satellite observations of 153 different fossil fuel power plants utilizing 4 different fuel types: hard coal(41%), gas(29%), lignite(29%) and peat(<=1%). Each sample comprises of thirteen spectral bands with ground truth annotations for binary plume segmentation, power generation rate, fuel type, and associated meteorological variables including temperature, relative humidity, and wind speed. Additional information on data quality and resolution have been described in the appendix.

4 Methodology

We utilize a self-supervised, reconstruction-based pretext task with a dynamic guided masking strategy to learn representations from multi-spectral imagery. The pretext network follows a U-Net style encoder decoder architecture composed of symmetric double convolution blocks. We add skip connections to transfer intermediate encoder feature maps to corresponding decoder stages, preserving spatial detail and stabilizing training (Ronneberger et al., 2015).

4.1 Sharpness-Aware Dynamic and Guided Masking

The pretext task introduces structured masking across spectral bands to encourage the encoder to exploit spatial detail and inter-band relationships when learning representations from multi-spectral imagery. Reconstruction is supervised using a Huber loss (Gokcesu & Gokcesu, 2021) applied between the reconstructed output and the original multi-spectral image. To estimate the spatial informativeness of each spectral band, a channel-level sharpness score is computed using the equally weighted linear combination of two complementary measures: the variance of the Laplacian and the 2D Tenengrad gradient energy (Pagaduan et al., 2020). The Laplacian responds to localized intensity changes and the Tenengrad measure reflects the presence of consistent gradient structures(Bansal

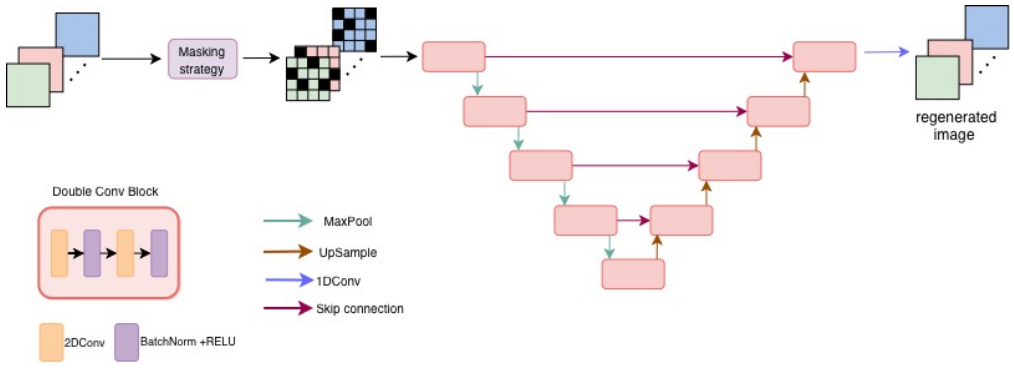

Figure 2: The masked multi-spectral input reconstructed using a UNet style encoder-decoder with skip connections.

et al., 2016). The Laplacian-based measure is obtained by applying the Laplacian operator to each band and computing the variance of the resulting response, while the Tenengrad measure evaluates first-order spatial variation by aggregating squared Sobel gradient responses along the horizontal and vertical directions (Gao et al., 2018). These measures are defined in Eq. (1) and Eq. (2), respectively.

$$S_{\text{Lap}}(b) = \text{Var}\left(\nabla^2 I_b\right) \tag{1}$$

$$S_{\text{Ten}}(b) = \sum_{x,y} \left(G_x^2(x,y) + G_y^2(x,y)\right) \tag{2}$$

where b indexes the spectral band, I_b denotes the input image corresponding to band b, ∇^2 is the Laplacian operator, G_x and G_y are the horizontal and vertical Sobel gradients at spatial location (x, y), and the summation is performed over all spatial locations.

Our dual stage masking strategy consists of a dynamic masking stage followed by a dynamic guided stage. In the first stage, dynamic masking assigns band-specific masking ratios using a normalized band-wise sharpness score. Bands exhibiting stronger spatial structure are masked less, while bands with weaker spatial content are masked more, ensuring that reliable spatial cues remain available for reconstruction. The second stage introduces guided masking only where it is most necessary. The three sharpest spectral bands are selected and partitioned into non-overlapping patches, and patch-level sharpness is computed to identify spatially weak regions within otherwise informative bands. Masking is then concentrated on the least informative patches, with the number of masked patches governed by the band-specific ratios from the first stage. All remaining bands continue to follow dynamic masking while safeguards ensure that no spatial location is masked more than once.

4.2 Downstream

To evaluate the proposed pretraining strategy, the encoder is reused as a shared backbone for three downstream tasks, regression, classification, and segmentation with separate task-specific heads trained independently, enabling consistent reuse of learned representations while allowing task-wise optimization.

The classification head aggregates multi-level features from the three deepest encoder layers at different spatial resolutions using global pooling and concatenation to improve scale-invariant robustness (Wijaya et al., 2022). The downstream task is trained with categorical cross-entropy loss, while the lightweight head employs batch normalization (Ioffe & Szegedy, 2015) and dropout for stable training and reduced overfitting.

For plume segmentation, the pretext decoder is reused, and the final 1×1 convolution is repurposed to generate dense binary plume masks rather than input reconstructions. Training uses pixel-wise cross-entropy loss, with performance evaluated using IoU and Dice coefficient to measure region overlap and boundary quality, respectively.

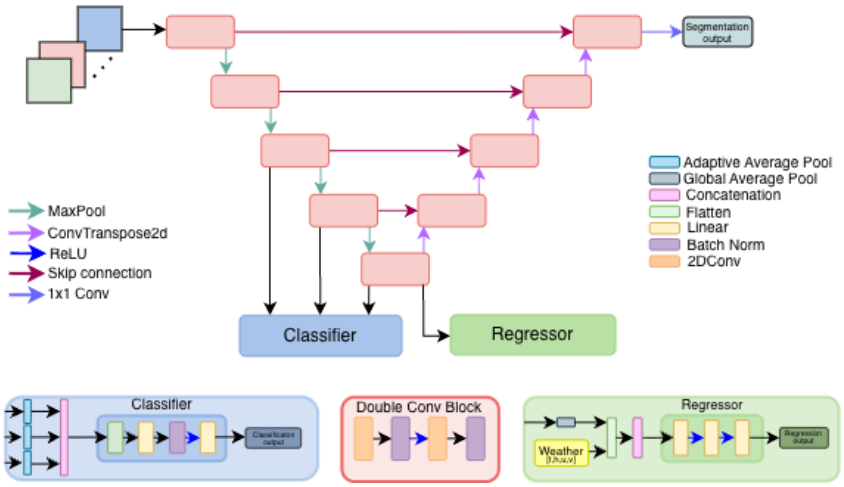

Figure 3: Downstream architecture with shared encoder and task specific heads.

The regression task leverages the pretrained representations from the last encoder layer which are globally averaged pooled and concatenated with a 4-dimensional normalized weather vector, subsequently passed through a lightweight MLP head. The encoder weights are frozen for the first 8 epochs to stabilize training and prevent catastrophic forgetting of representations (Yosinski et al., 2014). The network is then jointly optimized with differential learning rates, a smaller value (1e-5) for encoder to preserve pretrained representations and a larger value (3e-4) for the regression head to enable rapid task specific adaptation (Chen et al., 2020) (Howard & Ruder, 2018), with gradient clipping(Ramaswamy, 2023) applied for stability. Huber loss is used to provide robustness to outliers (Gokcesu & Gokcesu, 2021). Implementation and evaluation details can be viewed in the appendix.

5 Results

We evaluate the pretrained representations on plume segmentation, fuel classification, and power generation, as summarized in Table 1.The proposed dynamic guided approach achieves the strongest overall performance across tasks, establishing a new benchmark on the dataset. In particular, guided dynamic masking yields the largest improvements over the existing benchmark from Hanna et al. (2023), increasing R^2 by 7.7% and IoU by 14.1%. Fuel classification performance remains broadly comparable across methods. While MAE-based pretraining improves segmentation and regression performance relative to the existing benchmark, the introduction of dynamic masking leads to further gains. The guided dynamic strategy achieves the highest classification accuracy of 0.96, while also attaining the best IoU and R^2 values and reducing prediction error.

Table 1: Downstream performance comparison across masking strategies. * refers to the masking ratio range.

Task	Metric	Existing Benchmark (Hanna et al., 2023)	Without Pretext	MAE 75%	Dynamic (25%–75%)*	DynGuided (25%–75%)*
Seg	IoU ↑	0.64	0.69	0.69	0.71	**0.73**
	Dice ↑	–	0.75	0.75	0.76	**0.79**
Reg	R^2 ↑	0.78	0.80	0.83	0.79	**0.84**
	MAE ↓	157	176	151	170	**147**
Cls	F1 ↑	–	0.95	0.95	**0.96**	0.95
	Acc ↑	0.94	0.95	0.95	**0.96**	0.95

6 Conclusion

Our work presents a method for power estimation, using sharpness-aware dynamic and guided masking to extract hidden features from multi-spectral data without relying on auxiliary segmentation

and classification labels, or computationally intensive physics-based modules. Strategically masked spectral bands enable our model to surpass prior benchmarks in power estimation, plume segmentation, and fuel type classification. This approach offers a solution for global emission surveillance. Future work will explore regression performance through targeted extensions to the current model.

REFERENCES

Raghav Bansal, Gaurav Raj, and Tanupriya Choudhury. Blur image detection using laplacian operator and open-cv. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, pp. 63–67. IEEE, 2016.

Abhiroop Chatterjee, Susmita Ghosh, and Ashish Ghosh. Adaptive attention-guided masking in vision transformers for self-supervised hyperspectral feature learning.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, 2023. URL https://arxiv.org/abs/2207.08051.

Heather D Couture, Madison Alvara, Jeremy Freeman, Aaron Davitt, Hannes Koenig, Ali Rouzbeh Kargar, Joseph O'Connor, Isabella Söldner-Rembold, André Ferreira, Jeyavinoth Jeyaratnam, et al. Estimating carbon dioxide emissions from power plant water vapor plumes using satellite imagery and machine learning. *Remote Sensing*, 16(7):1290, 2024.

Dibyabha Deb and Kamal Das. Improving power plant co2 emission estimation with deep learning and satellite/simulated data. *arXiv preprint arXiv:2502.02083*, 2025.

Shuqin Gao, Min Han, and Xu Cheng. The fast iris image clarity evaluation based on tenengrad and roi selection. In *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, volume 10615, pp. 1391–1396. SPIE, 2018.

Kaan Gokcesu and Hakan Gokcesu. Generalized huber loss for robust learning and its efficient minimization for a robust statistics. *arXiv preprint arXiv:2108.12627*, 2021.

Joëlle Hanna, Michael Mommert, Linus Mathias Scheibenreif, and Damian Borth. Multitask learning for estimating power plant greenhouse gas emissions from satellite imagery. Tackling Climate Change with Machine Learning workshop at NeurIPS., 2021.

Joëlle Hanna, Damian Borth, and Michael Mommert. Physics-guided multitask learning for estimating power generation and co 2 emissions from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. URL https://arxiv.org/abs/1801.06146.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

Greet Janssens-Maenhout, Monica Crippa, Diego Guizzardi, Marilena Muntean, Edwin Schaaf, Frank Dentener, Peter Bergamaschi, Valerio Pagliari, Jos GJ Olivier, Jeroen AHW Peters, et al. Edgar v4. 3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012. *Earth System Science Data*, 11(3):959–1002, 2019.

Madhav Khirwar and Ankur Narang. Geovit: versatile vision transformer architecture for geospatial image analysis. In *2024 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS)*, pp. 1–3. IEEE, 2024.

Junyan Lin, Feng Gao, Xiaocheng Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial-spectral masked auto-encoder for multi-source remote sensing image classification, 2023. URL https://arxiv.org/abs/2311.04442.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9977–9978, 2019.

Ray Nassar, Timothy G Hill, Chris A McLinden, Debra Wunch, Dylan BA Jones, and David Crisp. Quantifying co2 emissions from individual power plants from space. *Geophysical Research Letters*, 44(19):10–045, 2017.

Roxanne A Pagaduan, R Aragon, and Ruji P Medina. iblurdetect: Image blur detection techniques assessment and evaluation study. In *International Conference on Culture Heritage, Education, Sustainable Tourism, and Innovation Technologies (CESIT2020)*, pp. 286–291, 2020.

Arunselvan Ramaswamy. Gradient clipping in deep learning: A dynamical systems perspective. In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pp. 107–114. INSTICC, SciTePress, 2023. ISBN 978-989-758-626-2. doi: 10.5220/0011678000003411.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Susan Solomon, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingstein. Irreversible climate change due to carbon dioxide emissions. *Proceedings of the national academy of sciences*, 106 (6):1704–1709, 2009.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

Kevin Tirta Wijaya, Dong-Hee Paek, and Seung-Hyun Kong. Advanced feature learning on point clouds using multi-resolution features and learnable pooling, 2022. URL https://arxiv.org/abs/2205.09962.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014. URL https://arxiv.org/abs/1411.1792.

Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7), 2023. ISSN 1424-8220. doi: 10.3390/s23073728. URL https://www.mdpi.com/1424-8220/23/7/3728.

A APPENDIX

A.1 DATASET DETAILS

Each sample in the dataset contains thirteen spectral bands in visible, near-infrared and short-wave infrared regions with a revisit frequency of approximately 5 days. These bands have a pixel resolution of up to 10m. These images have then been interpolated to the same spatial resolution(10m) for each spectral band and cropped to 120 X 120 pixel regions focusing on the geographic locations of the power plants.

A.2 PRETEXT METHODOLOGY

The final channel-level sharpness score for spectral band b is obtained by combining the Laplacian-based and Tenengrad-based measures using an equal-weighted linear combination:

$$S(b) = 0.5\, S_{\mathrm{Lap}}(b) + 0.5\, S_{\mathrm{Ten}}(b). \tag{3}$$

This formulation integrates second-order and first-order spatial gradient information into a single scalar score per band.

Masking ratio is calculated using the following formula:

$$r_b = r_{\min} + (r_{\max} - r_{\min})\,(1 - S(b))\,, \tag{4}$$

where r_b denotes the masking ratio assigned to spectral band b, $S(b) \in [0, 1]$ is the normalized sharpness score of band b, and $r_{\min}$ and $r_{\max}$ denote the minimum and maximum allowable masking ratios, respectively.

A.3 DOWNSTREAM LOSS FORMULATION

We optimize the downstream tasks using task-specific loss functions for regression, segmentation, and classification. For the regression task, we employ the Huber loss,

$$\mathcal{L}_{\mathrm{reg}}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \le \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise,} \end{cases} \tag{5}$$

where y denotes the ground-truth regression target, $\hat{y}$ is the model prediction, $\delta > 0$ is a threshold parameter controlling the transition between quadratic and linear regimes. This combines the sensitivity of the ℓ_2 loss for small errors with the robustness of the ℓ_1 loss for larger deviations.

For the segmentation task, we use a pixel-wise cross-entropy loss averaged over spatial locations and samples:

$$\mathcal{L}_{\mathrm{seg}} = -\frac{1}{BHW} \sum_{b=1}^{B} \sum_{p=1}^{HW} \log \left(\frac{\exp\left(z_{b,p,y_{b,p}}\right)}{\sum_{c=1}^{C} \exp\left(z_{b,p,c}\right)} \right) \tag{6}$$

where B denotes the batch size, H and W are the spatial dimensions, C is the number of classes, $z_{b,p,c}$ represents the logit for class c at pixel p of batch element b, and $y_{b,p} \in \{0, \ldots, C-1\}$ is the ground-truth class label.

For the classification task, we adopt the standard categorical cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log\left(\hat{y}_{i,c}\right). \tag{7}$$

where N denotes the number of samples, C is the number of classes, $y_{i,c} \in \{0, 1\}$ is the one-hot encoded ground-truth label indicating whether sample i belongs to class c, and $\hat{y}_{i,c} \in (0, 1)$ is the predicted probability for class c, obtained via a softmax operation.

A.4 PRETEXT EVALUATION

Pretraining uses a masked reconstruction objective evaluated with SSIM, FSIM, and LPIPS to quantify structural and perceptual recovery under different masking strategies.

Table 2: Pretext reconstruction quality.

Metric	MAE 75%	DynGuided (25%–75%)	Dynamic (25%–75%)
SSIM ↑	0.48	0.80	0.81
FSIM ↑	0.70	0.83	0.85
LPIPS ↓	0.20	0.15	0.12

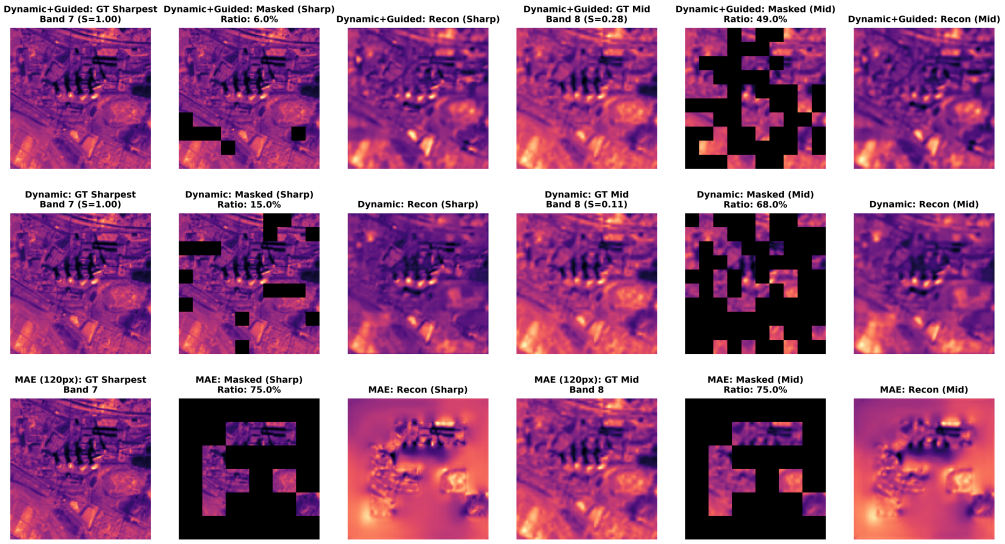

Figure 4: Comparative reconstruction results across masking strategies and pretext tasks.

A.5 IMPLEMENTATION DETAILS

All the pretext, classification and segmentation tasks are trained for 50 epochs while regression is trained for 80 epochs. Our experiments use a learning rate of 1e-4 employing AdamW optimizer with a weight decay of 1e-2 for regression and Adam for all other tasks. A batch size of 64 and a fixed random seed value is 42 to ensure reproducibility across all experiments on an NVIDIA P100 GPU. All images inputs are of resolution 120x120.