Chapter 5

Transformer-Based Polarized Dust Emission Super-Resolution

Abstract

This study introduces a novel approach to generating high-resolution, non-Gaussian foreground models for cosmic microwave background (CMB) polarization studies. We develop a transformer-based model to increase the resolution of dust polarized emission images by a factor of 4, utilizing limited data and applying the same model across different resolutions. Our method fuses information from various sources, including Planck dust optical depth at 353 GHz (τ_{353}), and H I-based Stokes Q and U templates (Q^{HI} and U^{HI}), to predict small-scale dust structure. We quantify the relative importance of each input dataset, finding that τ_{353}, Q^{HI}, and U^{HI} contribute almost equally to the prediction of small-scale features. The model's attention map analysis

supports the assumption of scale-independence in dust polarization, consistent with the power-law approximation of dust polarization power spectra across angular scales. While our predictions serve as realistic non-Gaussian extrapolations useful as simulations, we emphasize that they may not represent actual small-scale dust polarized emission structure. This work contributes to the development of more accurate foreground models, potentially improving component separation, lensing reconstruction, and the detection of primordial *B* modes in future CMB polarization studies.

5.1 Paper Status and External Contributions

I wrote all the code, performed all the analysis, wrote all the text, and produced all the figures for this chapter. However, this work was performed under supervision from and the text has received extensive editorial input from my advisor Susan Clark.

5.2 Introduction

The cosmic microwave background (CMB) serves as a crucial probe into the early universe, offering insights into its initial conditions and subsequent evolution (Hu & Dodelson, 2002). While observations of CMB temperature anisotropies have yielded significant cosmological constraints, current research efforts are increasingly focused on measuring CMB polarization, particularly the *B*-mode component at large angular scales (Kamionkowski

& Kovetz, 2016). B-mode polarization is of particular interest due to its potential to detect primordial gravitational waves, a key prediction of inflationary theories (Guth, 1981; Linde, 1982). Numerous experiments, including ground-based (e.g., Liu et al., 2022; Ade et al., 2019; Abazajian et al., 2016; ACT Collaboration et al., 2024; SPT Collaboration et al., 2023; POLARBEAR Collaboration et al., 2022; BICEP/Keck Collaboration et al., 2021), balloon-borne (e.g., SPIDER Collaboration et al., 2022), and satellite-based (e.g., Collaboration et al., 2023) instruments, are actively pursuing inflationary B-mode detection.

However, the pursuit of primordial B-modes faces significant challenges. Gravitational lensing of CMB photons by large-scale structures converts some E modes to B modes that dominate over the primordial signal at small angular scales (Planck Collaboration et al., 2020h). While this lensing signal carries valuable cosmological information, it necessitates precise reconstruction and removal techniques to access the underlying primordial B-modes (Hu & Okamoto, 2002).

A more formidable obstacle is the presence of Galactic foregrounds, primarily thermal dust and synchrotron emission, which dominate the polarized signal at high ($\gtrsim$ 70 GHz) and low ($\lesssim$ 70 GHz) frequencies, respectively (Dunkley et al., 2009; Planck Collaboration et al., 2016e). These foregrounds exceed the expected CMB B-mode signal across all frequencies and sky positions, necessitating sophisticated component separation techniques for their removal (Delabrouille et al., 2009; Stompor et al., 2009).

Critically, Galactic foregrounds exhibit significant non-Gaussianity, particularly at large scales (Ade et al., 2019). This non-Gaussianity is expected to persist at smaller scales due to the complex distribution of the interstellar medium (ISM) structure and turbulent interstellar magnetic fields. The presence of non-Gaussian foregrounds introduces mode coupling in angular power spectra and potentially biases both primordial B-mode detection and lensing reconstruction (Beck et al., 2020).

Current foreground models are limited by the lack of high-resolution, large-area observations. Existing templates, primarily derived from extrapolating Planck and Wilkinson Microwave Anisotropy Probe (WMAP) data to other frequency bands through assumptions about their spectral energy distributions (SEDs), characterize foregrounds down to approximately 1° resolution (Planck Collaboration et al., 2020f; Bennett et al., 2013). Simulation packages like the Python Sky Model (PySM) extrapolate these templates to smaller scales using power-law fits and Gaussian realizations (Thorne et al., 2017; Zonca et al., 2021). However, this approach fails to capture the expected non-Gaussianity at small angular scales.

Alternative approaches to foreground modeling include data-driven methods (Clark & Hensley, 2019; BICEP/Keck Collaboration et al., 2023e; Halal et al., 2024a), phenomenological models (Hervías-Caimapo & Huffenberger, 2022), and magnetohydrodynamic (MHD) simulations (Kim et al., 2019). While each offers unique insights, they face limitations in either reproducing the observed morphology or achieving high resolution efficiently.

Recent advancements in machine learning techniques offer promising avenues for generating high-resolution, non-Gaussian foreground simulations. Generative Adversarial Networks (GANs) have been employed to inject small-scale features into low-resolution dust observations while preserving statistical properties (Krachmalnicoff & Puglisi, 2021; Yao et al., 2024). However, these works use high-resolution dust total intensity rather than polarized emission maps as the ground truth of the output. They, therefore, assume that the statistical properties of the small-scale thermal dust emission in polarization are the same as in total intensity.

In this work, we pursue a different approach to produce realistic non-Gaussian high-resolution maps of the polarized dust emission at small scales. To accomplish this, we introduce several techniques. We utilize ancillary high-resolution datasets described in Section 5.3 as additional inputs to the model. We perform different smoothings and projections on multi-resolution maps as described in Section 5.4. In Section 5.5, we describe how our model fuses high-angular-resolution information from ancillary datasets with the low-angular-resolution information of the polarized dust emission maps to generate high-angular-resolution maps of the polarized dust emission. We describe our results in Section 5.6 and conclude in Section 5.7.

5.3 Data

In this section, we describe the different datasets used in the inputs and outputs of our model and explain the motivation behind using each of them.

5.3.1 Planck Data Products

For the polarized dust emission, we use the multi-resolution R3.00 Planck data at 353 GHz, processed with the Generalized Needlet Internal Linear Combination (GNILC; Remazeilles et al., 2011) method to eliminate the Cosmic Infrared Background (CIB) radiation from the Galactic dust emission (Planck Collaboration et al., 2016c). In line with the fiducial offset corrections used by the Planck collaboration, we add a Galactic offset correction of 63 $\mu\mathrm{K}_{\mathrm{CMB}}$ then correct for the CIB monopole by subtracting 452 $\mu\mathrm{K}_{\mathrm{CMB}}$ from the GNILC total intensity map (Planck Collaboration et al., 2020a). These maps are provided in the COSMO convention, and we use them in that convention. These data are used to produce the low-angular-resolution input and high-angular-resolution output Stokes Q and U images as explained in Section 5.4.

We also use the accompanying masks indicating the angular resolution of each sky region to determine the ideal methodology for dividing the sky into patches as explained in Section 5.4. The mask is shown in Figure 5.1 with the Galactic plane masked out.

We use the R1.20 Planck dust optical depth at 353 GHz (τ_{353}; Planck Collaboration et al., 2014b) as an additional input to our model. Planck Collaboration et al. (2014b) fit

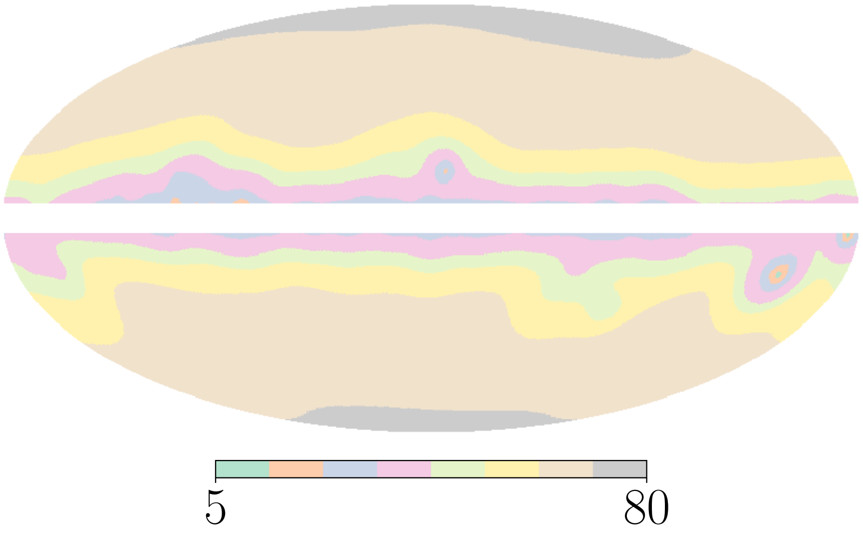

Figure 5.1: Mollweide map projection of the variable angular resolutions of the Planck GNILC polarization data. The colors on the colorbar correspond to 5′, 7′, 10′, 15′, 20′, 30′, 60′, and 80′. The region with Galactic latitudes $|b| < 5°$ is masked out.

a modified blackbody spectrum to Planck data at different frequencies to obtain a map of the optical depth. This is useful as a high-angular-resolution (FWHM=5′) tracer of the dust spatial distribution at 353 GHz.

5.3.2 H I-based Dust Polarization Templates

Galactic neutral hydrogen (H I) emission serves as an invaluable proxy for studying dust polarization in the interstellar medium (ISM). The strong correlation between H I and dust distributions in the diffuse ISM has been well-established (Boulanger et al., 1996b; Lenz et al., 2017). Both components exhibit filamentary structures that demonstrate significant alignment with the plane-of-sky magnetic field orientation (Clark et al., 2014, 2015).

The spectroscopic nature of H I observations, utilizing the 21 cm line, provides three-dimensional information (longitude, latitude, and radial velocity) about the ISM structure (Clark, 2018). This characteristic offers a distinct advantage over traditional dust emission measurements. Additionally, H I data are free from contamination by the cosmic infrared background (Chiang & Ménard, 2019) and are independent of broadband thermal dust emission observations, eliminating concerns about correlated instrumental systematics. Exploiting these properties, Clark & Hensley (2019) developed a model for dust polarization based solely on H I intensity measurements. Their approach, utilizing the Rolling Hough Transform algorithm (Clark et al., 2014, 2020), demonstrated significant correlations with Planck 353 GHz data, particularly at large angular scales. Specifically, they observed correlations of approximately 60% and 50% for E modes and B modes, respectively, at multipole $\ell = 50$ over high Galactic latitudes, with the correlation diminishing towards zero at $\ell = 1000$. This methodology has proven valuable for characterizing dust properties, including spectral indices, through cross-correlations with millimeter-wave polarization data (BICEP/Keck Collaboration et al., 2023e).

Given H I's advantages as a high-resolution tracer of the polarized dust emission, we utilize H I-based Stokes Q and U templates (Q^{HI} and U^{HI}) as additional inputs to the model. Although there exist higher angular resolution H I surveys than the H I 4π Survey (HI4PI; HI4PI Collaboration et al., 2016), such as the Galactic Arecibo L-Band Feed Array H I Survey (GALFA-H I; Peek et al., 2018) with an angular resolution of $4\overset{\prime}{.}1$, we utilize templates constructed with HI4PI data with an angular resolution of $16\farcm2$ in this analysis. This is because we are limited by the highest angular resolution polarized dust emission maps we use, which are at $15'$, and the HI4PI survey covers the full sky.

HI4PI combines data from the Effelsberg-Bonn H I Survey (EBHIS; Winkel et al., 2016) and the Parkes Galactic All-Sky Survey (GASS; McClure-Griffiths et al., 2009). It achieves a spectral resolution of 1.49 km s^{-1} and a normalized brightness temperature noise of $\sim$ 53 mK for a 1 km s^{-1} velocity channel.

For the H I-based polarization template we use the publicly available maps described in Halal et al. (2024a) constructed utilizing the version of the Hessian-based filament-finding algorithm described in Cukierman et al. (2023). These templates are divided into velocity bins with respect to the local standard of rest, inferred from the Doppler-shifted frequency along the line of sight. We use the version that integrates the templates along the range -13 km $\mathrm{s}^{-1} < v < 16$ km s^{-1}, which is found by Halal et al. (2024a) to result in the highest correlation with the Planck polarization data at 353 GHz. We convert these maps to the COSMO convention to match the convention of the Planck GNILC maps by flipping the sign of the Stokes U template.

5.4 Pre-processing

We divide the sky into overlapping patches and project them onto square images to perform the training. For the ground truth, we do not use simulations or a factor derived from the dust total intensity as in Krachmalnicoff & Puglisi (2021); Yao et al. (2024) in this analysis. Instead, we use the polarized dust emission data itself. We use the highest-resolution Planck GNILC data available as the high-resolution ground truth and an artificially smoothed version of this data as the input.

We train a model to increase the angular resolution of the input images by a factor of 4, no matter what the input angular resolution is. The motivation behind this approach is that the highest angular resolution ground truth data available has variable resolution across the sky as shown in Figure 5.1. Since deep learning models require a large sample size to train well, we train the same model on the different variable-resolution regions of the sky. To accomplish this, we vary the size of the projected patch based on its angular resolution, with larger patches for lower-angular-resolution regions and vice versa. Note that this approach assumes scale invariance. To weaken this assumption, we add an input to the network that is dependent on the resolution used for a given patch. This allows the network to learn a scale-dependent term and is explained further in Section 5.5.

5.4.1 Masks

We use the angular resolution map shown in Figure 5.1 to inform our projection and training strategies. Note from Figure 5.1 that the $5'$, $7'$, and $10'$ regions of the sky are not large enough to create enough training patches, so we limit the training to output 353 GHz Stokes Q and U images at $15'$, $20'$, $25'$, and $30'$, given the same images at $60'$, $80'$, $100'$, and $120'$, respectively, i.e., a 4× increase in resolution. We mask out Galactic latitudes $|b| < 5^\circ$ as shown in Figure 5.1. This is to remove regions associated with the inner Galactic plane. We exclude the $60'$ and $80'$ regions from the training samples. Therefore, the resulting sky area from which we project onto square patches covers about 40% of the sky.

For each angular resolution $\theta \in \{15', 20', 25', 30'\}$, we project patches of size $20\,\theta \times 20\,\theta$ to 80×80 pixel images. The longitudinal and latitudinal distances between neighboring patch centers are $8\ \theta$, i.e., twice the low angular resolution associated with the patches. We consider projections within 3 overlapping sky masks. The first is a combination of all the masks with an angular resolution equal to or higher than $15'$ and is used for predicting images with angular resolution $15'$, given images with angular resolution $60'$. The second is a combination of all the masks with an angular resolution higher than or equal to $20'$, which includes the first mask, and is used for predicting images with angular resolutions $20'$ and $25'$, given images with angular resolutions $80'$ and $100'$, respectively. The third and last is a combination of all the masks with an angular resolution higher or equal to $30'$, which includes the first and second masks. This is used for predicting images with angular resolution $30'$, given images with angular resolution $120'$.

5.4.2 Patch Projections

Within each of the 3 masks described in the previous subsection, we smooth the data to a constant resolution equal to the lowest angular resolution within that mask. For instance,

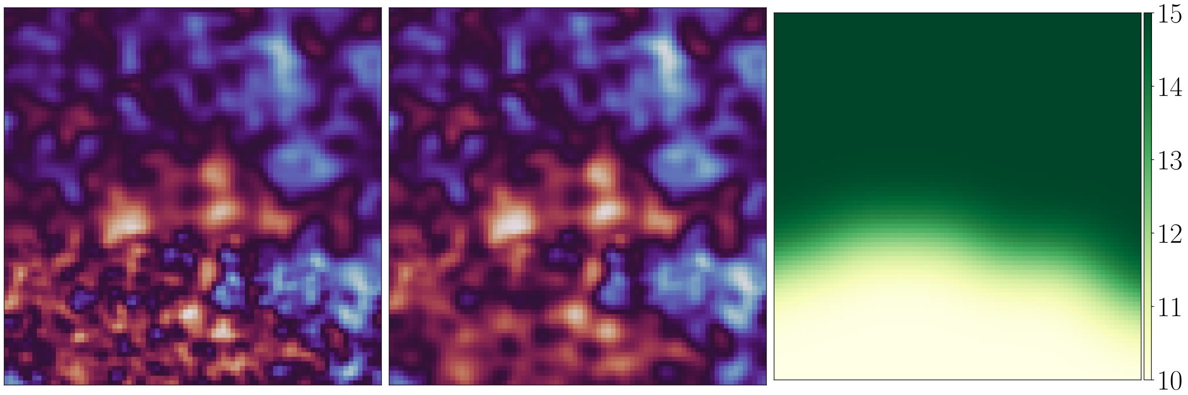

Figure 5.2: Example 80×80 pixel projected patch centered at $(l,\ b)\ =\ (210°,\ -26°)$ with a pixel width of $3.\!'75$. The projected patch spans both the $10'$ and $15'$ angular resolution regions of the variable resolution Planck GNILC map. Left: a projection of the original variable resolution GNILC Stokes Q map. Right: the Planck GNILC variable resolution mask, smoothed by a 1° Gaussian smoothing kernel. Middle: a weighted average of the map in the left panel and the same map smoothed with a Gaussian smoothing kernel from $10'$ to $15'$ according to the weights in the right panel, where $10'$ corresponds to the smoothed map and $15'$ corresponds to the original map.

we need to use the data in the regions corresponding to the $5'$, $7'$, and $10'$ resolutions as part of the mask used for predicting $15'$ resolution images. We start by smoothing the angular resolution map shown in Figure 5.1 using a Gaussian kernel with FWHM = 1° as shown in an example patch on the right in Figure 5.2. We do this to avoid sharp transitions when combining maps as follows.

We smooth the entire multi-resolution map with a Gaussian kernel that corresponds to the smoothing scale that would degrade a map resolution from $\mathrm{FWHM}_{\mathrm{current}}$ to $\mathrm{FWHM}_{\mathrm{desired}}$, i.e., a kernel with

$$\mathrm{FWHM} = \sqrt{\mathrm{FWHM}^2_{\mathrm{desired}} - \mathrm{FWHM}^2_{\mathrm{current}}}, \tag{5.1}$$

where $\mathrm{FWHM}_{\mathrm{desired}}$ is $7'$ and $\mathrm{FWHM}_{\mathrm{current}}$ is $5'$ in this case. This is useful in the $5'$ region, where the data are correctly smoothed to $7'$. We blend the original and smoothed maps in the regions corresponding to the $5'$ and $7'$ resolutions, using the smoothed resolution map to determine the weights for the weighted average. In the regions of the smoothed resolution map where the pixel values are $7'$ or higher, we use the original map. In the regions where the pixel values are $5'$, we use the map smoothed from $5'$ to $7'$. For the other pixel values between $5'$ and $7'$ in the smoothed resolution map, i.e., at the transition between the two angular resolution regions, we replace the pixels with a weighted average between the original and the smoothed data, where the weights are determined by the relative distances of the pixel values in the smoothed resolution map from $5'$ and $7'$. We iteratively repeat this process, combining the $7'$ region, which now includes the original $5'$ region, with the $10'$ region, and so on. An example patch where the $10'$ and $15'$ angular resolution regions are combined is shown in Figure 5.2. Through this process, we can project patches that overlap with regions of different angular resolution after having combined them into a common resolution.

When performing the patch projections for different projection sizes and centers, we ignore patches that contain regions outside the 3 aforementioned masks, i.e., if parts of the patch are at Galactic latitudes $|b| < 5°$ or in angular resolution regions outside the defined mask. For the projection schema, we perform a bilinear interpolation with a zenithal equal area (ZEA) projection. This projection preserves the area of the sky region being projected

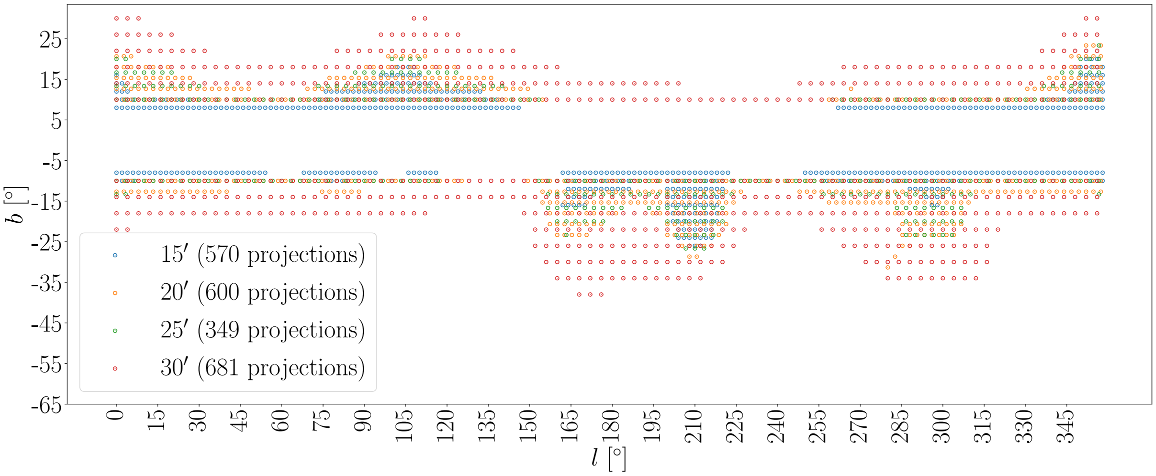

Figure 5.3: Scatter plot of the projection centers of patches used for different resolutions. The legend has the number of patches for each resolution. These are not the total number of patches used for the training since augmentations including flipping and rotations are also used.

while minimizing distortions near the center of the projection. Figure 5.3 shows the resulting projection centers for the patches corresponding to the different angular resolutions used in the training. We repeat this projection procedure for patches rotated by 45°. This is a form of augmentation used to increase the size of the dataset. Unlike 90° rotations, which can be performed after the projections, the 45° rotations include different parts of the sky. Therefore, not every 45° rotated counterpart of every patch is included in the training set if the rotated patch contains a part of the sky that is outside the mask. Figure 5.4 shows the resulting projection centers for the 45° rotated patches corresponding to the different angular resolutions used in the training.

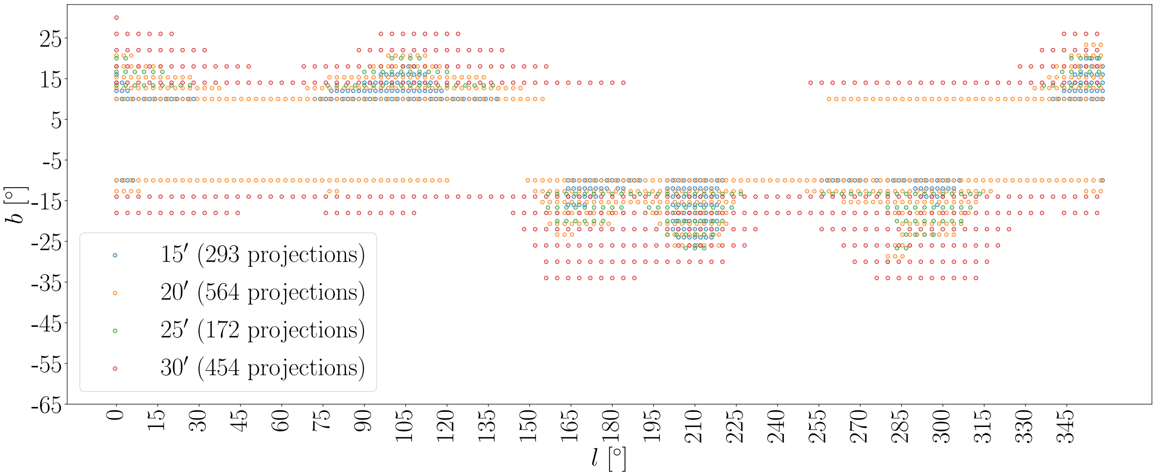

Figure 5.4: Scatter plot of the projection centers of patches rotated by 45° used for different resolutions. The legend has the number of patches for each resolution. These are not the total number of patches used for the training since augmentations including flipping and rotations are also used.

5.4.3 Training, Validation, and Testing Data

For each of the patches in Figures 5.3 and 5.4, we perform 7 augmentations, resulting in 8 images. These include 90°, 180°, and 270° rotations and a mirror flip for each rotation including the original image. Therefore, we have 3,683 patches with 8 augmentations for each, resulting in 29,464 training samples. Normally, not all augmentations are applied to all images in computer vision tasks. Instead, images are randomly augmented by the CPU before being passed to the GPU for training (Buslaev et al., 2020). This avoids overfitting the model on the training data. However, in this case, we do use all augmentations for all images because the number of images available is small otherwise. We test for overfitting by splitting the dataset into 80% (23,576 samples) for training, 10% (2,944 samples) for validation, and 10% (2,944 samples) for testing. We ensure all augmentations of the same patch are in the same data split to avoid data leakage between the different splits.

Each training sample consists of 7 images and a scale-dependent term that indicates to the model the desired output angular resolution. The 5 images used as inputs in addition to the angular resolution are the smoothed low-resolution Planck GNILC Stokes Q and U 353 GHz images (Q_{353}^{LR} and U_{353}^{LR}), the H I-based Stokes Q and U templates (Q^{HI} and U^{HI}), and the optical depth at 353 GHz (τ_{353}). The output consists of the high-resolution Planck GNILC Stokes Q and U 353 GHz images (Q_{353}^{HR} and U_{353}^{HR}). Therefore, we have 206,248 images in total. We use float-32 as the data type. This corresponds to 80 pixels × 80 pixels × 2 bytes = 2.6 GB for all the images. Note that the input and output images have the same pixelization, so no downsampling and upsampling is required.

The reasoning behind including the τ_{353} map as an input is to provide the network with a high-resolution tracer of the spatial distribution of the dust at 353 GHz. Similarly, Q^{HI} and U^{HI} provide the network with a high-resolution polarization template that is highly correlated with the polarized dust emission.

For different input and output angular resolutions for the Planck GNILC 353 GHz Stokes images, we also use differently smoothed τ_{353}, Q^{HI}, and U^{HI} images. For the desired output resolution θ, we smooth the τ_{353}, Q^{HI}, and U^{HI} images according to Equation 5.1, where $\mathrm{FWHM}_{\mathrm{current}}$ is $5'$, $16.'2$, and $16.'2$, respectively, and $\mathrm{FWHM}_{\mathrm{desired}}$ is $\sqrt{\theta^2 - 15^2 + 5^2}$, $\sqrt{\theta^2 - 15^2 + 16.2^2}$ and $\sqrt{\theta^2 - 15^2 + 16.2^2}$, respectively. That is, when the desired output resolution is $15'$, we use the τ_{353}, Q^{HI}, and U^{HI} images at their native resolutions. When the desired output resolution θ is a value larger than $15'$, the τ_{353}, Q^{HI}, and U^{HI} images are smoothed proportional to the distance between θ^2 and $15'^2$. The smoothing is done on the full-sky maps before projecting the maps into patches to avoid any edge effects.

5.4.4 Normalization

We normalize the images to a common scale to accelerate the optimization and convergence of the model. It is common in computer vision tasks to normalize images to the range [0, 1] (Krizhevsky et al., 2012). However, since Stokes Q and U can be positive or negative, we scale, Q^{HI}, and U^{HI} to the range [-1, 1], where 0 in the original maps remains 0 after the normalization. Since τ_{353} is positive, we scale those data to the range [0, 1]. This scaling is done on the sky map level rather than on the patch level, i.e., the ratio of the pixel values across different patches remains constant after scaling. We only consider pixels within the sky mask from which we perform the patch projections. Also, we scale the Stokes Q and U maps of the same dataset together, i.e., we consider the pixel values of both the Stokes Q and U maps together when scaling. This preserves the ratio of Q to U and thus the polarization angle. For the Planck GNILC data, we determine the normalization based on the unsmoothed maps, Q_{353}^{HR} and U_{353}^{HR}, and use the same normalization for the smoothed maps, Q_{353}^{LR} and U_{353}^{LR}. Therefore, we only need 3 values to perform the normalization for our datasets. These are the values that will map to 1 in the normalizations, while 0 remains 0 for all the maps. One is obtained from the distribution of the τ_{353} pixel values, one from the distribution of the Q_{353}^{HR} and U_{353}^{HR} pixel values, and one from the distribution of the Q^{HI} and U^{HI} pixel values. These 3 distributions contain outlier pixel values. We empirically find that the 99.9995th percentiles of the absolute values of these distributions separates the bulk of the distributions from these outlier pixel values in the tails of the distributions. This percentile corresponds to 10,020 outliers in each distribution. We set all pixel values with magnitudes higher than this percentile to the value at that percentile in the positive and negative directions before scaling. These thresholds are 1.0×10^{-3} K, 3.2×10^{-3} K, and 4.8×10^{7} K km s^{-1} for the τ_{353} data, the Q_{353}^{HR} and U_{353}^{HR} data, and the Q^{HI} and U^{HI} data, respectively.

5.5 Model

5.5.1 Super-Resolution Techniques

Image super-resolution, a fundamental task in computer vision, encompasses two primary approaches: single image super-resolution (SISR) and multi-image super-resolution (MISR). SISR techniques aim to enhance the resolution of a single low-quality image, while MISR methods leverage multiple low-resolution images of the same scene to reconstruct a high-resolution output. MISR is particularly prevalent in remote sensing applications, where satellites capture multiple temporal views of the same geographical area (e.g., An et al., 2022).

Deep-learning-based MISR techniques have significantly advanced the field by introducing sophisticated mechanisms for fusing information from multiple low-resolution inputs. These fusion approaches vary in their architectural designs and information integration strategies (e.g., Deudon et al., 2020; Bordone Molini et al., 2020; Dorr, 2020). For instance, Arefin et al. (2020) employed a Gated Recurrent Unit (GRU), a type of recurrent neural network, as the fusion module.

More recent work has explored the use of transformers and the attention mechanism as fusion modules. The attention mechanism, originally introduced in natural language processing, allows a model to focus on different parts of the input when producing each part of the output. In the context of image processing, attention enables the model to selectively emphasize or suppress different spatial regions or feature channels. Mathematically, the attention mechanism can be expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q\,K^T}{\sqrt{d_k}}\right) V \tag{5.2}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors (Vaswani et al., 2017). The query, key, and value matrices are learnable linear weights applied to the input sequence.

Several studies have applied attention mechanisms to MISR tasks. Salvetti et al. (2020)

introduced a 3D convolutional feature attention mechanism with embedded residual connections, enabling the network to focus on extracting high-frequency information in both the temporal and spatial dimensions. Valsesia & Magli (2021) utilized self-attention mechanisms and permutation invariance for temporal images, allowing their model to process input frames in any order. An et al. (2022) were the first to apply a transformer, which is based on self-attention, to the remote sensing image MISR task. Li et al. (2023) extracted image features from different scales, and used attention in channel and spatial dimensions and across images, allowing their model to capture multi-scale information and inter-image relationships.

5.5.2 Architecture

Rather than training two separate networks, one for the Stokes Q and one for the Stokes U image, we train the same model on both. This is because information in the Stokes Q and U images are not physically independent. We utilize convolution-based encoders and decoders, that sandwich a transformer-based fusion module, which also takes in an embedding vector corresponding to the desired output resolution as an input. The overall architecture of the model we use in this work is shown in Figure 5.5. All of the convolutions used in this model have a kernel size of 3 × 3 pixels, a stride of 1 pixel, and a zero-padding of 1 pixel. The result of this convolution has the same dimensions as the original image. Therefore, the output images of this model have 80 × 80 pixels, which is the same as the

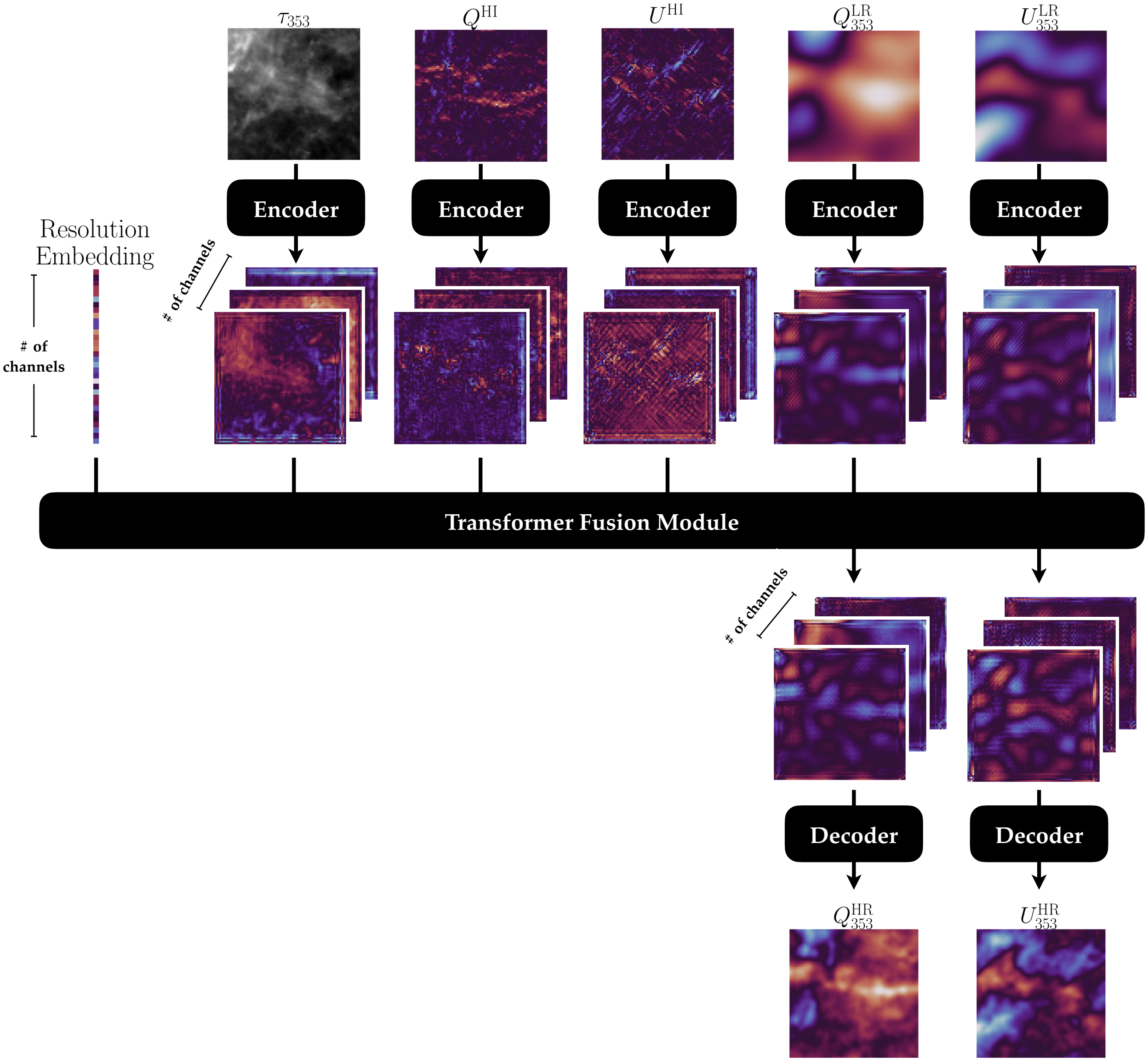

Figure 5.5: The full model architecture. The input images are each processed separately by an encoder, whose architecture is shown in Figure 5.6. The processed images along with an angular resolution embedding vector are then processed by a transformer fusion module, which consists of transformer layers (Figure 5.7). The transformer module's outputs corresponding to the Planck GNILC Stokes Q and U images are then processed through decoders, whose architecture is shown in Figure 5.6.

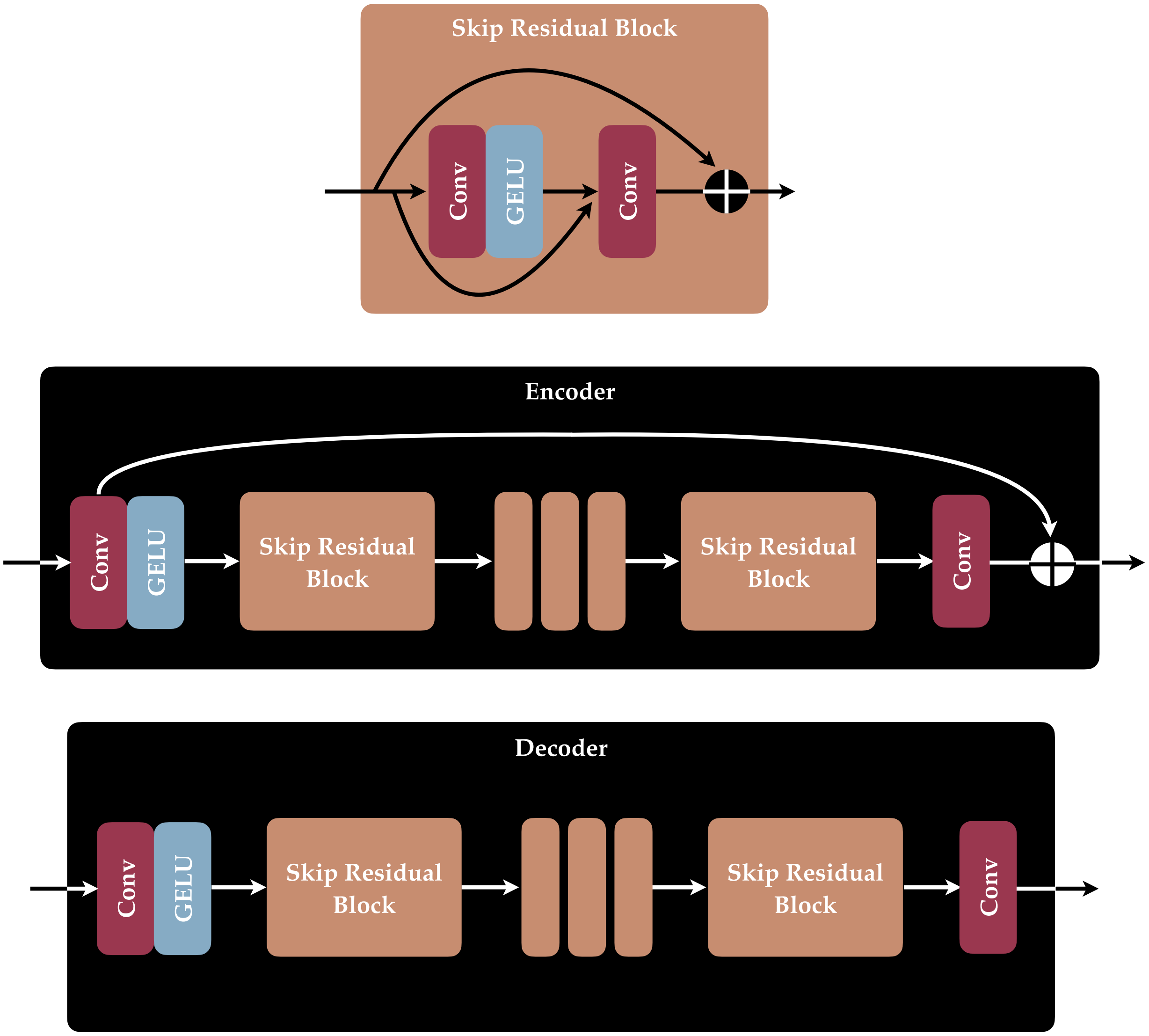

Figure 5.6: The architectures of the encoder (middle), of the decoder (bottom), and of their main component, the Skip Residual Block (top). The Skip Residual Block is made up of two convolutional layers with a non-linearity and a skip connection between them and a residual connection over the entire block. The encoder and decoder are made up of a convolutional layer, followed by a non-linearity, a number of Skip Residual Blocks, and a final convolutional layer. The encoder additionally has a residual connection from the output of the first convolutional layer to the end.

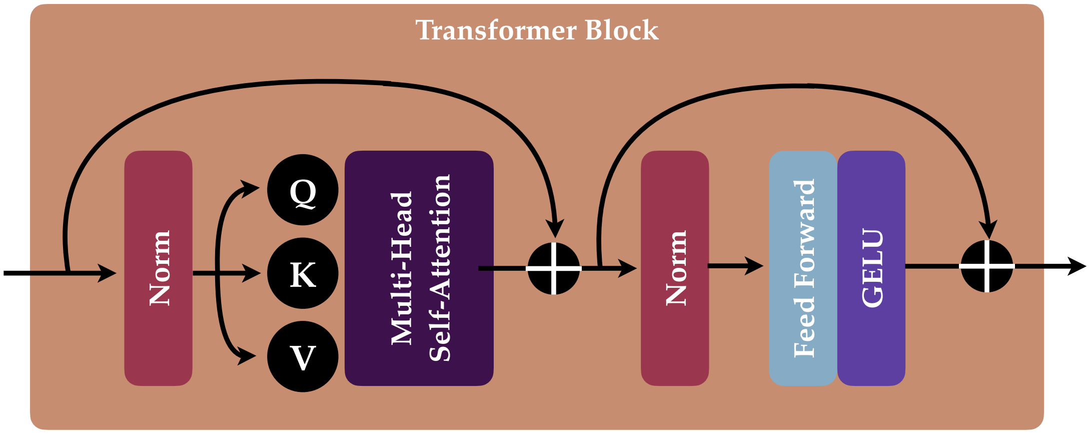

Figure 5.7: The transformer architecture used in this paper. It is made up of multi-head self-attention, where Q, K, and V are the query, key, and value tensors and a feed-forward layer followed by a non-linearity. Both components are preceded by a layer normalization and enveloped by a residual connection.

input images. No downsampling or upsampling is performed in this model. While each convolutional layer only has a receptive field of 3 × 3 pixels, stacking multiple convolutional layers in a sequence allows later convolutional layers to have a wider receptive field of the original images. We use the Gaussian Error Linear Unit (GELU; Hendrycks & Gimpel, 2016) as the non-linearity throughout this model. It applies a smooth approximation to the identity function, which enhances model performance by enabling better gradient flow during training.

The main building block in the encoder and decoder architectures is the Skip Residual Block. The architecture of this block is shown in Figure 5.6. It consists of convolutional layers, a non-linearity, and skip and residual connections. Skip and residual connections have several useful properties, including accelerating the convergence of the training process and avoiding the common vanishing and exploding gradient problems (He et al., 2015). Skip connections concatenate the output of an earlier layer with the output of the current layer before feeding them into the following layer. Residual connections add the output of an earlier layer with the output of the current layer before feeding them into the following layer. The first convolutional layer has the same number of input and output channels, $\mathrm{N}_{\mathrm{channels}}$. Due to the concatenation performed before the second convolutional layer, the input of the second convolutional layer is double the size of the input of the first convolutional layer, i.e., 2 $\mathrm{N}_{\mathrm{channels}}$, but its output has $\mathrm{N}_{\mathrm{channels}}$. Therefore, the residual connection is performed element-wise between two sets of $\mathrm{N}_{\mathrm{channels}}$. The non-linearity only follows the first convolutional layer.

The same patch of the sky for each dataset is processed by the same encoder architecture, resulting in a set of encoded images equal in number to the number of convolutional channels in the last convolutional layer of the encoder, $\mathrm{N}_{\mathrm{channels}}$. The encoder architecture is shown in Figure 5.6. The images are first passed through a convolutional layer with $\mathrm{N}_{\mathrm{channels}}$ channels. The output of this layer is both passed through subsequent layers of the encoder and added to the output of the encoder through a residual connection. The results of the first convolutional layer are passed through a non-linearity, followed by a number of Skip Residual Blocks, $\mathrm{N}_{\mathrm{encoder\ blocks}}$. The output is then passed through a convolution layer with an equal number of input and output channels, $\mathrm{N}_{\mathrm{channels}}$.

The desired output resolution is embedded by an embedding layer into a vector with size N_{channels}, equal to the number of channels in the last convolutional layer of the encoder. The embedding layer learns a separate embedding vector for each of the 4 angular resolutions used in the training: $15'$, $20'$, $25'$, and $30'$.

A transformer module is used to fuse information from the different datasets. It accepts a sequence of 6 input vectors of size N_{channels}. The first vector in this sequence is always the embedding of the angular resolution. The five vectors after that correspond to the five input datasets. Each training sample is split into 80×80 training samples, one for each pixel. For each pixel, a vector of size N_{channels} is constructed from the pixels at the same location across the channel dimension in the output of the encoder. Therefore, the five vectors in the transformer sequence after the resolution embedding vector correspond to a pixel in the same location across the five input datasets. In practice, we treat the 80×80 pixels as different samples in a batch, i.e., the batch size is scaled by a factor of 80×80 and passed through the transformer in a single pass rather than stacking the results of 80×80 passes. The output of the transformer also has a sequence length of 6 vectors. We stack a number of transformer layers $N_{\text{transformer layers}}$ back-to-back. For the last transformer layer, we discard the first 4 outputs and only pass the output vectors corresponding to the Planck GNILC 353 GHz Stokes Q and U images through decoders.

The architecture of each transformer layer we use is shown in Figure 5.7. Its main components are the multi-head self-attention, the feed-forward layer, and residual connections. The feed-forward layer is followed by a non-linearity. The number of self-attention

Hyper-parameter	Value
N_{channels}	32
$N_{\text{encoder blocks}}$	4
$N_{\text{decoder blocks}}$	5
$N_{\text{transformer layers}}$	1
N_{heads}	1
Batch size	1
Learning rate	0.0001
Loss function	L1
Optimizer	AdamW (β_1=0.9, β_2=0.999)
N_{epochs}	209

Table 5.1: List of hyper-parameters used to train the model.

heads N_{heads} is a hyper-parameter. We apply layer normalization (Lei Ba et al., 2016) before each layer to control the scale of the gradients.

The decoder architecture is shown in Figure 5.6. It is very similar to the encoder architecture but with three differences. First, the number of Skip Residual Blocks $N_{\text{decoder blocks}}$ is another hyper-parameter that does not need to be equal to $N_{\text{encoder blocks}}$. Second the output number of channels of the last convolutional layer is 1, the target image. Because this is not equal to N_{channels}, the third difference is that we do not use a residual connection here.

5.6 Results

We experiment with different hyper-parameters and loss functions while attempting to overfit the model on 100 training samples. Once we find a set of hyper-parameters for which the model is able to overfit, we use those hyper-parameters for training the model on the full training dataset. The set of hyper-parameters we use, including the loss function, are summarized in Table 5.1. With these hyper-parameters, the model has 928,194 tunable parameters. We find that increasing the batch size degrades the model's performance so we use a batch size of 1. We also see no improvement in the model's loss when increasing $\mathrm{N}_{\text{transformer layers}}$ and $\mathrm{N}_{\text{heads}}$ from 1 to 2 each for the same number of epochs. Therefore, we use 1 for both since that decreases the number of tunable parameters and enables interpretability of the attention weights.

While training this model, we save its weights whenever the validation loss reaches a new minimum. We also log the output of the same 4 validation patches (1 for each angular resolution) once every 50 training steps. We stopped the training when those model outputs visually look similar enough to the ground truth with no visible artifacts. The last saved model checkpoint is at epoch 209 after the training is stopped. We note that the validation loss was still decreasing when we stopped the training. The model can, therefore, be trained for longer for more optimal results. Using the last saved model weights, we show the model's outputs for 4 randomly selected test patches, 1 for each angular resolution, in Figure 5.8. Note that there is no visible distinction in the model's performance on the different angular resolutions.

The attention weights, i.e., the softmax $\left(\frac{Q\,K^T}{\sqrt{d_k}}\right)$ part of Equation 5.2, is useful for interpreting how much attention each input element receives from each output element. The softmax ensures that these weights sum to 1 across all positions, effectively creating a probability distribution. Because the query and key matrices are dependent on the input encoded images, the attention weights vary pixel-by-pixel and patch-by-patch. Therefore, to visualize the attention map, we average the attention weights over all the pixels and over 100 different patches. We find no measurable difference in the attention weights between the different angular resolution patches. Therefore, for the 100 different patches we average over for visualizing the attention map in Figure 5.9, we select 25 randomly from each angular resolution.

Figure 5.9 shows how important each of the 6 inputs to the transformer block are to each of the 2 outputs that are passed through decoders. The importance is quantified as weights that sum to 1. Note that the encoder outputs corresponding to Q_{353}^{LR} and U_{353}^{LR} are the least important for predicting U_{353}^{HR} and Q_{353}^{HR}, respectively. We also note that the angular resolution embedding is also unimportant. If the resolution embedding adequately captures the scale dependence in the data, then this result supports the assumption of scale independence. Conversely, if the resolution embedding is not sufficient, the model would be imposing scale independence and the result is evidence that the imposed scale independence is working as intended. The most important encoder outputs for predicting Q_{353}^{HR} and U_{353}^{HR} are those corresponding to Q_{353}^{LR} and U_{353}^{LR}, respectively. This is expected since the output images are just higher-resolution versions of the input images. It is interesting that both Q^{HI} and U^{HI} are important for predicting both Q_{353}^{HR} and U_{353}^{HR}, and τ_{353} is also as important. This result highlights the importance of these datasets in predicting dust structure at small scales. It is still the case, however, that Q^{HI} is slightly more informative than U^{HI} for predicting Q_{353}^{HR} and vice versa, as expected.

5.7 Conclusions, Limitations, and Future Work

The development of accurate high-resolution foreground models is crucial for the future of CMB polarization studies. Such models will enable more robust component separation, improved lensing reconstruction, and ultimately, a clearer path to detecting primordial B modes. In this work, we contribute to this effort by exploring novel techniques for generating realistic, non-Gaussian foreground simulations at small angular scales. We demonstrated how the attention mechanism can be used to fuse information from images of different sources to increase the resolution of polarized dust emission images by a factor of 4. We also demonstrated how to achieve this with limited data, using the same model for different resolutions. We quantify the importance of each input dataset in predicting the small-scale dust structure and find the τ_{353}, Q^{HI}, and U^{HI} to all be almost as important as the low-resolution version of the map.

Given that no ground truth is available for the actual small-scale structure of the dust polarized emission, our predictions are only to be considered as realistic non-Gaussian extrapolations from large-scale structure that follow the correct statistics. We emphasize that these predictions are to be used as simulations rather than actual predictions of the small-scale structure of the dust polarized emission. One of the weak assumptions made in our modeling approach is that the mapping between two scales which are a multiplicative factor apart is similar to the mapping between two different scales with the same multiplicative factor between them. We weaken this assumption by adding a scale-dependent term as an input to the model. However, by studying the attention map of our model, we find that this term is given little attention. This either implies that the assumption of scale independence is justified or that encoding the scale dependence in this term is insufficient for capturing the scale dependence in the data. This is in line with the fact that the power spectra of the observed dust polarization can, at first order, be approximated as a power law as a function of angular scales (Córdova Rosado et al., 2024).

The next step in this analysis is to run the model on patches of the entire sky to create a higher-resolution version of the Planck GNILC map. Given that the full sky Planck GNILC map includes regions with angular resolutions other than the 4 resolutions we train our model on and the fact that the attention weights corresponding to the resolution embedding are very low, we could retrain the model without this embedding.

Several options are possible for extending the analysis described in this chapter. The main one is to turn this model into a generator by adding randomness to the input images, adding a critic model, and performing adversarial training using techniques like the Wasserstein generative adversarial network with gradient penalty (WGAN-GP). This would allow the model to generate different possible small-scale predictions for the same patch of sky, which would be useful when different realizations are needed. Moreover, we have not performed extensive hyper-parameter tuning of the model architecture. We only experimented with different configurations until we found one that produces satisfactory results. Therefore, extensive hyper-parameter tuning, including experimenting with different loss functions and image quality metrics, can be helpful for producing the optimal predictions. Another possible analysis extension would be to experiment with changes to the input datasets. For instance, the Stokes Q and U images can be combined into $Q + i\,U$ images with complex-valued neural networks. Another example is that the integrated H i-based polarization templates can be replaced with different H i-based polarization templates divided across the velocity dimension. Additionally, the τ_{353} data could be replaced with data from the WISE experiment, which measures emission from polycyclic aromatic hydrocarbon (PAH) molecules at 12 microns (Meisner & Finkbeiner, 2014). These molecules are intermixed with the dust in the interstellar medium (Córdova Rosado et al., 2024). The benefit of this data is its angular resolution at $5''$. The possibilities listed here may not improve the model's predictions but are worth experimenting with in case they do.

5.8 Acknowledgements

This work was supported by the National Science Foundation under grant No. 2106607. This work utilizes data from Planck, an ESA science mission funded by ESA Member States, NASA, and Canada.

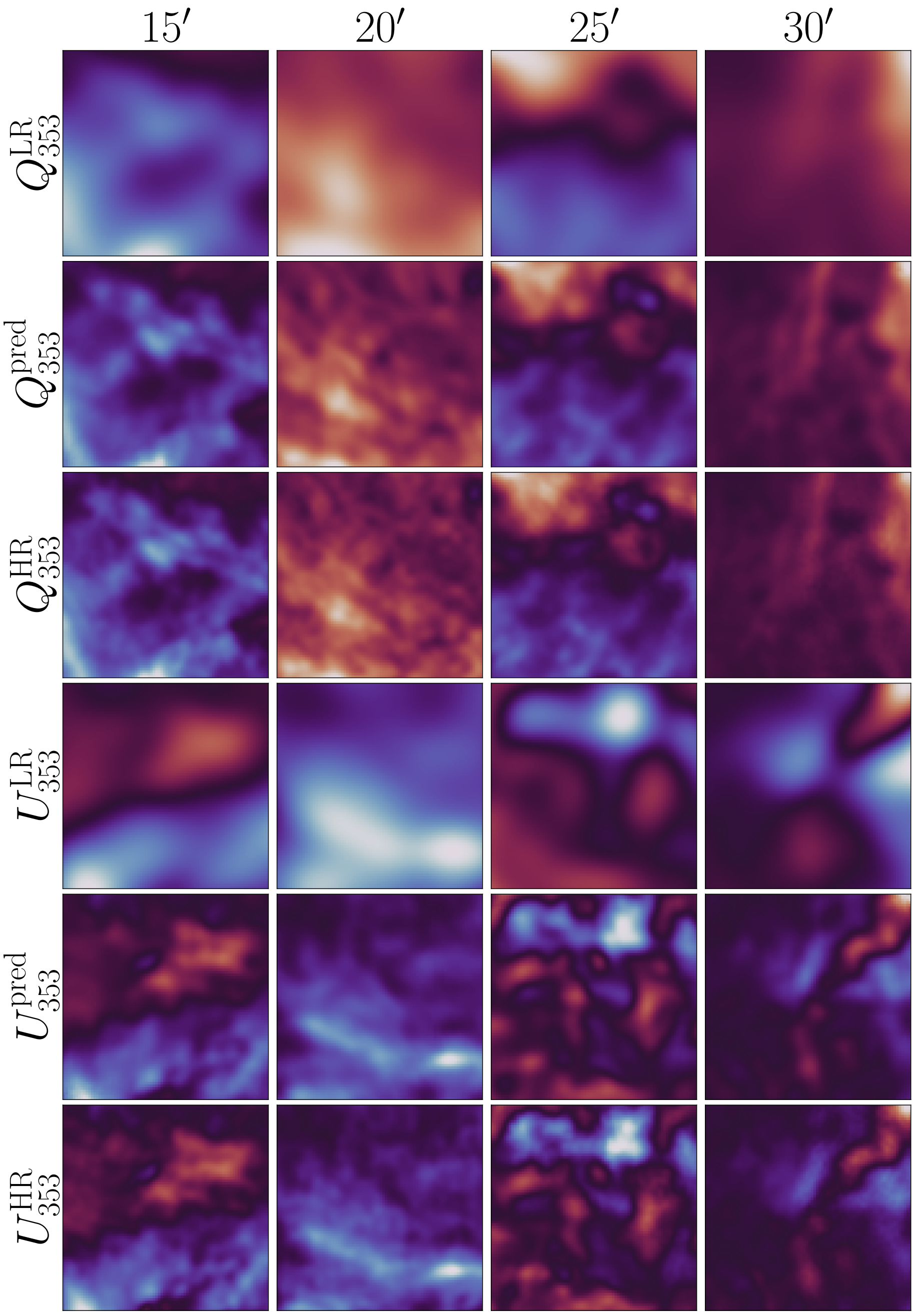

Figure 5.8: Example low-resolution input and high-resolution predictions and target 353 GHz Stokes Q (top 3 rows) and U (bottom 3 rows) patches of sky. The same patch of sky is shown across each column with its corresponding high angular resolution denoted at the top. The colorbars are centered at zero (darkest) and brighter red (blue) corresponds to higher positive (negative) values.

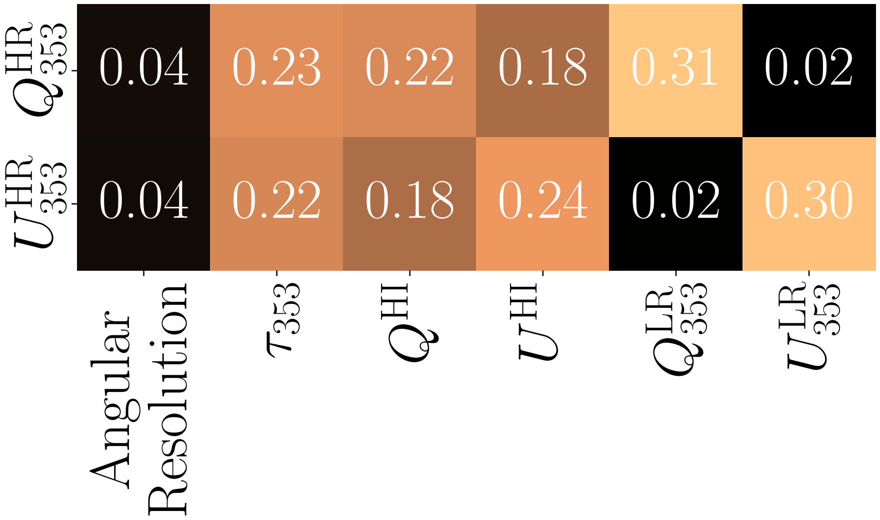

Figure 5.9: Attention map averaged over 80 × 80 pixels and over 100 test sample patches, 25 for each angular resolution. The values in each row sum up to 1 because of the softmax operation from Equation 5.2.