How to Understand Masked Autoencoders

Shuhao Cao
Washington University in St. Louis
s.cao@wustl.edu

Peng Xu[*]
University of Oxford
peng.xu@eng.ox.ac.uk

David A. Clifton
University of Oxford
david.clifton@eng.ox.ac.uk

Abstract

"Masked Autoencoders (MAE) Are Scalable Vision Learners" revolutionizes the self-supervised learning method in that it not only achieves the state-of-the-art for image pre-training, but is also a milestone that bridges the gap between visual and linguistic masked autoencoding (BERT-style) pre-trainings. However, to our knowledge, to date there are no theoretical perspectives to explain the powerful expressivity of MAE. In this paper, we, for the first time, propose a unified theoretical framework that provides a mathematical understanding for MAE. Specifically, we explain the patch-based attention approaches of MAE using an integral kernel under a non-overlapping domain decomposition setting. To help the research community to further comprehend the main reasons of the great success of MAE, based on our framework, we pose five questions and answer them with mathematical rigor using insights from operator theory.

1 Introduction

"Masked Autoencoders (MAE) Are Scalable Vision Learners" [31] (illustrated in Figure 1) recently introduces a ground-breaking self-supervised paradigm for image pretraining. This seminal method makes great contributions at least in the following respects:

(1) MAE achieves the state-of-the-art on self-supervised pretraining on ImageNet-1K dataset [21], outperforming the strong competitors (*e.g.*, BEiT [5]) by a clear margin in a simpler and faster manner. Particularly, a *vanilla* Vision Transformer (ViT) [23] Huge backbone based MAE achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. This inspiring phenomenon motivates the researchers to re-consider the ViT variants in self-supervised contexts. Moreover, MAE achieves good transfer performance in downstream tasks beating the supervised pretraining and shows promising scaling behavior.

(2) MAE is a generative learning method and beats the contrastive learning competitors (*e.g.*, MoCo v3 [18]) that have dominated vision self-pretraining in recent years.

(3) MAE is a mile-stone that bridged the gap between the visual and linguistic masked autoencoding (BERT-style [22]) pretrainings. Actually, the previous work prior to MAE fails to apply the masked autoencoding pretrainings on the visual domain. Thus, MAE paves a path that *"Self-supervised learning in vision may now be embarking on a similar trajectory as in NLP"* [31].

The great success of MAE is interpreted by its authors as *"We hypothesize that this behavior occurs by way of a rich hidden representation inside the MAE"* [31]. However, to the best of our knowledge, to
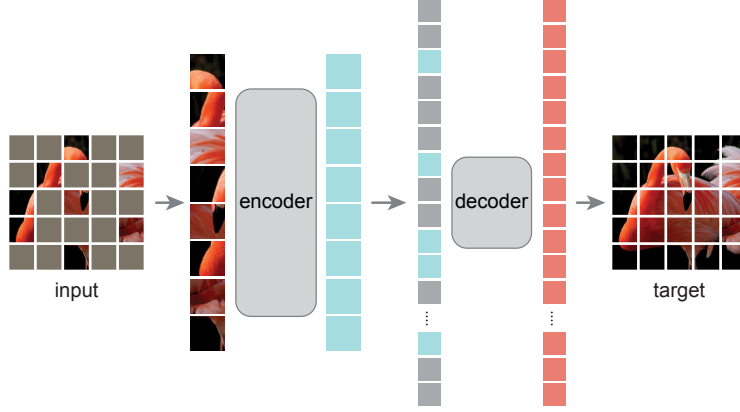
---

[*]Corresponding Author

Technical Report.

Figure 1: Architecture of the Masked Autoencoder (MAE) in [31], with Vision Transformer (ViT) backbone. (Used under CC BY-NC 4.0 license from `https://github.com/facebookresearch/mae`.) Best viewed in color.

date, there are no theoretical viewpoints to explain the powerful expressivity of MAE. Unfortunately, the theoretical analysis for the expressivity of ViT based models is so challenging that it is still under-studied.

Our Contributions In this paper, we, for the first time, propose a unified theoretical framework that provides a mathematical understanding for MAE. Particularly, we not only rethink MAE by regarding each image's embedding as a learned basis function in certain Hilbert spaces instead of a 2D pixel grid, but also explain the patch-based attention approaches of MAE from the operator theoretic perspective of an integral kernel under a non-overlapping domain decomposition setting.

To help the researchers to further grasp the main reasons for the great success of MAE, based on our proposed unified theoretical framework, we contribute five questions, and for the first time answer them partially by insights from rigorous mathematical theory:

Q1: *How is the representation space of MAE formed, optimized, and propagated through layers?*
A: We illustrate that the attention mechanism in MAE is equivalent to a learnable integral kernel transform, and its representation power is dynamically updated by the Barron space with the positional embeddings that work as the coordinates in a high dimensional feature space. See Section 4.1.

Q2: *Why and how does patchifying contribute to MAE?*
A: We prove that the random patch selecting of MAE preserves the information of the original image, while reduces the computing costs, under common assumptions on the low-rank nature of images. This paves a theoretical foundation for the patch-based neural networks/models including but not limited to MAE or ViT. See Section 3.

Q3: *Why are the internal representations in the lower and higher layers of MAE not significantly different?*
A: We have provided a new theoretical justification for the first time by proving the stability of the internal presentations. The great success of MAE benefits from a main reason that the scaled dot-product attention in the ViT-backbone provides stable representations during the cross-layer propagation. Furthermore, we view the skip-connection for the attention mechanism from a completely new perspective: representing an approximated solution explicitly to a Tikhonov-regularized Fredholm integral equation. See Section 5.

Q4: *Is decoder unimportant for MAE?*
A: No. We argue that the decoder is vital to helping encoder to build better representations, even if decoder is discarded after pretraining. Due to the bigger patch dimension in the MAE decoder, it allows the representation space in the encoder enriched much more often by functions from the Barron space to learn a better basis. See Section 6.

Q5: *Does MAE reconstruct each masked patch merely inferred from its adjacent neighbor patches?*
A: No. We prove that the latent representations of the masked patches are interpolated globally based on an inter-patch topology that is learned by the attention mechanism. See Section 6.

2

Overall, our proposed unified theoretical framework provides a mathematical understanding for MAE and can be used to understand the intrinsic traits of the extensive patch and self-attention based models, not limited to MAE or ViT.

The rest of this paper is organized as follows: Section 2 summarizes related work. Based on our proposed theoretical framework, some mathematical understandings for the encoder and decoder of MAE are presented in Section 4 through Section 6. We draw some conclusions in Section 7.

2   Related Work

Vision Transformer (ViT) [23] is a strong image-oriented network based on a standard Transformer [44] encoder. It has an image-specific input pipeline where the input image needs to be split into fixed-size patches. After going through the linearly embed layer and adding with position embeddings, all the patch-wise sequences will be encoded by a standard Transformer encoder. ViT and its variants have been widely used in various computer vision tasks (*e.g.*, recognition [43], detection [7], segmentation [36]), and meanwhile work well for both supervised [43] and self-supervised [18, 15] visual learning. Recently, some pioneering works provide further understanding for ViT, *e.g.*, its internal representation robustness [38], the continuous behavior of its latent representation propagation [39]. However, the theoretical analysis for the expressivity of ViT based models is so challenging that it is still under-studied.

Masked Autoencoder (MAE) [31] is essentially a denoising autoencoder [45], which has a straight-forward motivation that randomly mask patches of the input image and reconstruct the missing pixels. MAE works based on two key designs: (i) asymmetric encoder-decoder architecture where the encoder takes in only the visible patches and the lightweight decoder reconstructs the target image, (ii) high masking ratio (*e.g.*, 75%) for the input image yields a nontrivial and meaningful self-supervisory task. Its great success is attributed to *"a rich hidden representation inside the MAE"* [31]. However, to the best of our knowledge, to date, there are no theoretical viewpoints to explain the powerful expressivity of MAE.

Mathematical Theory Related to Attention Self-attention [44] is essentially processing each input as a fully-connected graph [51]. Therefore, as aforementioned, we start from a more general perspective of topological spaces [13] to rethink MAE, by regarding each image as a graph connecting patches instead of 2D pixel grid. Meanwhile, we study the patch-based attention approaches of MAE through the operator theory for the Fredholm integral equations [1], to formulate the dot-product attention matrix as an integral kernel [29], *i.e.*, a learned Green's function to reconstruct solutions to the partial differential equations *e.g.*. [26]. The skip-connection can be then explained through two perspectives, first as the term corresponding to the Tikhonov regularization [27, 48], or Neural Ordinary Differential Equations (ODE) [17]. On the other hand, the patchification is connected with the Domain Decomposition Methods (DDM) [42, 52].

3   Patch is All We Need?

Patchifying has become a standard practice in Transformer-based CV models since ViT [23], see also [30, 37]. In this section, we try to answer the question "Why and how does patchifying contribute to MAE?" from a perspective of the domain decomposition method to solve an integral equation.

We first consider the full grid $\Omega \simeq I \otimes I$ with $I = [0, 1/N, \ldots, (N-1)/N]$, then the patchification of Ω is essentially a non-overlapping domain decomposition method (DDM) [16]. DDM is commonly used in solving integral/partial differential equations bearing similar forms with (10), *e.g.*, see [42], as well as image denoising problems [52].

DDM practices "divide and conquer", which is a general methodology suitable for lots of science and engineering disciplines: the original domain of interest is decomposed into much smaller subdomains. Each subproblem associated with subdomains can be more efficiently solved by the same algorithm, especially when the algorithm has a super-linear dependence on the grid size. The attention mechanism computational complexity and storage requirement have a quadratic dependence on n, and this translates a quartic dependence on the coarse grid size p (how many patches along one axis) as $n = p^2$. Without patchification, we can see the operation in (6) is of $\mathcal{O}(N^4)$ that renders the algorithm unattainable.

In this setup, Ω is first partitioned to equal-sized non-overlapping subdomains: $\Omega = \cup_{i=1}^{n} \Omega_i$, $\text{int}(\Omega_i) \cap \text{int}(\Omega_j) = \emptyset$ as well as $\text{int}(\Omega_i) \simeq \text{int}(\Omega_j)$ if $i \neq j$. For the simplicity of the presentation, we consider a single-channel image and define the space of bounded variation (BV) on Ω as $BV(\Omega)$, and for any unpatchified image $\mathbf{u} \in BV(\Omega) \subset \mathbb{R}^{N \times N}$

$$|\mathbf{u}|_{BV(\Omega)} := \sum_{x_i \in \Omega} \left\{ \sum_{x_j \in N_{x_i}} w_{ij}(\mathbf{u}(x_i) - \mathbf{u}(x_j))^2 \right\}^{\frac{1}{2}}, \tag{1}$$

where N_{x_i} denotes the grid points that are connected with x_i through an undirected edge $x_i x_j$, and w_{ij} is some positive weights measuring the interaction strength. When being viewed in the continuum such that $\mathbf{u}$ is a discrete sampling of the function $u : \mathbb{R}^2 \to \mathbb{R}^+$, this norm approximates $|\nabla u|_{L^1}$ and measures the smoothness of an image, and is widely used in graph/image recovery e.g., see [8].

After the patchification we have $\mathbf{u} = \sum_{i=1}^{n^2} E_i \mathbf{u}_i$, where $E_i : \Omega_i \to \Omega$ is an extension operator that maps the patch pixel values on Ω_i to Ω by zero-padding. It is straightforward to see that the representation consist of the concatenated patch matrices, and is in $\Pi_{i=1}^{n^2} BV(\Omega_i)$. Here we denote $\mathcal{Y} := \oplus_{i=1}^{n^2} E_i \, BV(\Omega_i)$. By (1), the original BV space is only a subspace of the decomposed product space, i.e., $BV(\Omega) \subset \mathcal{Y}$, while the reverse is not true. The reason is that by patchification, the underlying function spaces have completely lost the inter-patch topological relations: e.g., how the inter-patch pixel intensities change? There is also another analytical set of questions to answer: how to choose the size of the patch? If bigger patches are used, it then requires a bigger embedding dimension, thus harder to formulate the bases for the representation space. On the other hand, if smaller patches are used, because of the aforementioned quartic dependence on the grid size, the inter-patch topology is much harder and more expensive to learn.

In the following proposition, we shall show that, if an image has a low rank structure, then there exists a semi-randomly chosen set of patches that can represent the original image. This selection has a very high mask ratio, that is, the representation using a small fraction e.g., p patches out of all the patches. Meanwhile, we assume that there exists an embedding that is able to represent the original image, in the sense that the difference between the reconstruction and the original image is small measured under $|\cdot|_{BV(\Omega)}$.

Following the standard practices [23, 43], we consider the ViT-base case in MAE where the dimension of the feature embedding matches the original dimension of the pixel counts in the any given patch: $\mathbf{u} \in \mathbb{R}^{N \times N} \simeq \mathbb{R}^{n^2 \times d}$, i.e., for each patch's embedding in the MAE encoder, there exists a continuous embedding that maps $\mathbb{R}^{n^2 \times d} \hookrightarrow \mathbb{R}^{N \times N}$, from the patch embeddings to images.

Proposition 3.1 (Existence of a near optimal patch embedding). *For any $\mathbf{u} \in BV(\Omega) \subset \mathbb{R}^{N \times N}$, and an equal-sized $n \times n$ patchification of $\mathbf{u}$, let the patch grid size $N_c := N/n$, assuming there exists a rank r approximation to $\mathbf{u}$ with $r < N_c$ and error ϵ, and a unique patch is randomly chosen from each row such that the final selection's columns set is a permutation of $\{1, \cdots, n^2\}$, then there exists an embedding $\mathbf{y} \in \mathbb{R}^{p \times d}$ such that*

$$\|\mathbf{u} - R(\mathbf{y})\|_{BV(\Omega)} < C(r, N_c, p)\epsilon, \tag{2}$$

where R is a rank preserving reconstruction operator $R : \mathbb{R}^{p \times d} \to \mathbb{R}^{N \times N}$.

4 Attention in MAE: a Kernel Perspective

In this section, we reexamine the attention block present in both encoder and decoder layers in MAE from multiple perspectives. The first and foremost understanding comes from the operator theory for the integral equations. By formulating the scaled dot-product attention as a nonlinear integral transform, learning the latent representation is equivalent to learning a set of basis in a Hilbert space. Moreover, upon adopting the kernel interpretation of the attention mechanism, a single patch is characterized by not only its own learned embedding, but also how it interacts with all other patches (reproducing property of a kernel).

4.1 Self-attention: A Nonlinear Integral Kernel Transform

For MAE [31], each input image patch is projected as a 1D token embedding. Following the practice in [31], we also omit the shared class token added to the embedding that can be removed from the
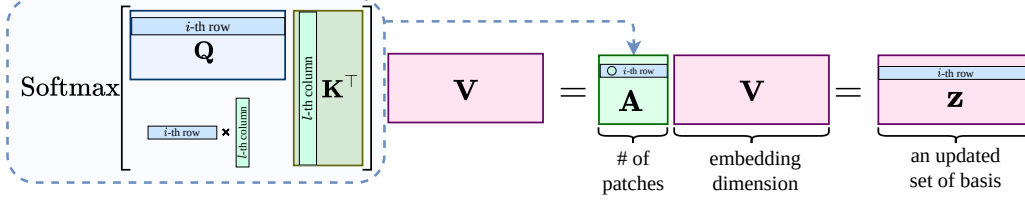
Figure 2: The i-th patch embedding from the scaled dot-product attention is a convex combination of the attention weights, which encodes the interactions among all patch embeddings. Best viewed in color.

pretraining of MAE. Given an attention block in an encoder or a decoder layer of MAE, the input and output are defined respectively as $\mathbf{y}_{\text{in}} := \mathbf{y}, \mathbf{y}_{\text{out}} \in \mathbb{R}^{p \times d}$. When an image has not been masked, it has $p = n^2$ patches, where n is a positive integer. The embedding dimension is d. In $\mathbf{y}$, a positional embedding $\mathcal{X} := \{x_i\}_{i=1}^p \subset \mathbb{R}^{1 \times d}$, such that x_i is associated with the i-th patch, is added. Each x_i can be viewed as a coordinate in high dimension. We note that the patch ordering map $i \mapsto x_i$ is injective, since the first component of x_i is the polar coordinate in 2D, and as such the relative position of each patch can be recovered from x_i. For simplicity, the analysis done in this section applies to a single head.

Here we briefly review the self-attention from [44]. The query $\mathbf{Q}$, key $\mathbf{K}$, value $\mathbf{V}$ are generated by three learnable projection matrices $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d}$: $\mathbf{Q} = \mathbf{y}\mathbf{W}^Q, \mathbf{K} = \mathbf{y}\mathbf{W}^K, \mathbf{V} = \mathbf{y}\mathbf{W}^V$. The scaled dot-product attention is to obtain $\mathbf{z} \in \mathbb{R}^{p \times d}$:

$$\mathbf{z} = \text{Attn}_s(\mathbf{y}) := \text{Softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d}\right)\mathbf{V}. \tag{3}$$

Then the softmax attention $\text{Attn}(\cdot)$ with a global receptive field works as the following nonlinear mapping:

$$\text{Attn} : \mathbb{R}^{p \times d} \to \mathbb{R}^{p \times d},$$
$$\mathbf{y} \mapsto \text{LN}\left(\mathbf{y} + \mathbf{z} + \text{FFN}\left(\text{LN}(\mathbf{y} + \mathbf{z})\right)\right), \tag{4}$$

where $\text{LN}(\cdot)$ denotes the Layer Normalization (LN) [2] that essentially is a learnable column scaling with a shift, and $\text{FFN}(\cdot)$ is a standard two-layer feedforward neural network (FFN) applied to the embedding of each patch.

Upon closer inspection of the scaled dot-product attention (3), for $\mathbf{z}$, the j-th element of its i-th row $\mathbf{z}_i$ is obtained by

$$(\mathbf{z}_i)_j = \mathbf{A}_{i\bullet} \cdot \mathbf{v}^j, \tag{5}$$

and the i-th row $\mathbf{A}_{i\bullet}$ of the attention matrix $\mathbf{A}$ is

$$\mathbf{A}_{i\bullet} = \frac{e^{(\mathbf{Q}\mathbf{K}^\top / \sqrt{d})_{i\bullet}}}{\sum_{j=1}^p e^{(\mathbf{Q}\mathbf{K}^\top / \sqrt{d})_{ij}}} =: \text{Softmax}(\mathbf{q}_i \mathbf{K}^\top / \sqrt{d}), \tag{6}$$

i.e., using the diagram in Figure 2, we have

$$\mathbf{z}_i = \mathbf{A}_{i\bullet} \begin{pmatrix} \text{—} & \mathbf{v}_1 & \text{—} \\ \text{—} & \vdots & \text{—} \\ \text{—} & \mathbf{v}_n & \text{—} \end{pmatrix} = \sum_{j=1}^p \mathbf{A}_{ij} \mathbf{v}_j. \tag{7}$$

Due to the softmax, $\mathbf{A}_{i\bullet}$ contains the coefficients for the convex combination of the vector representations $\{\mathbf{v}_j\}_{j=1}^p$. This basis set $\{\mathbf{v}_j\}_{j=1}^p$ form the $\mathbf{V}$'s row space, and it further forms each row of the output $\mathbf{z}$ by multiplying with $\mathbf{A}$. As a result, the scaled dot-product attention has the following basis expansion form:

$$\mathbf{z}_i := \sum_{j=1}^p A(\mathbf{q}_i, \mathbf{k}_j) \mathbf{v}_i, \tag{8}$$

where $\mathbf{u}_i$ denotes the i-th row of certain latent representation $\mathbf{u} \in \{\mathbf{z}, \mathbf{Q}, \mathbf{K}, \mathbf{V}\}$. The key term in (8), $A(\cdot, \cdot)$, denotes the attention kernel, which maps each patch's embedding represented by the rows of

$\mathbf{Q}, \mathbf{K}$ to a measure of how they interact. From (8), we shall see that the representation space for an encoder layer in MAE is spanned by the row space of $\mathbf{V}$, and is being nonlinearly updated layer-wise. This means, the embedding for each patch serves as a basis to form the representation space for the current attention block, whereas the row-wise attention weights are the Barycentric coordinates for a simplex in $\mathbb{R}^{1 \times d}$.

Here we further assume that there exist a set of feature maps for query, key, and value *e.g.*, see [20]. For $\mathbf{z}$, the feature map $z \in \mathcal{V}$ that maps $\mathbb{R}^{1 \times d} \to BV(\Omega_i)$, *i.e.*, $\mathbf{z}_i = z(x_i)$, we can then define an asymmetric kernel function $\tilde{\kappa}(\cdot, \cdot) : \mathbb{R}^{1 \times d} \times \mathbb{R}^{1 \times d} \to \mathbb{R}$,

$$
\begin{aligned}
A(\mathbf{q}_i, \mathbf{k}_j) &= \alpha^{-1}(x_i) \langle \mathbf{q}_i, \mathbf{k}_j \rangle \\
&= \alpha^{-1}(x_i) \langle q(x_i), k(x_j) \rangle \\
&=: \alpha^{-1}(x_i) \, \tilde{\kappa}(x_i, x_j),
\end{aligned}
\tag{9}
$$

where $\tilde{\alpha}(x_i) := \sum_{l=1}^{d} (e^{\mathbf{q}_i \mathbf{K}^\top / \sqrt{d}})_l$. Now the discrete kernel $A(\cdot, \cdot)$ with vectorial input is rewritten to an integral kernel $\tilde{\kappa}(\cdot, \cdot)$ whose inputs are positions. As a result, using the formulation above, we can express the scaled dot-product attention approximately as a nonlinear integral transform:

$$
\begin{aligned}
z(x) &= \alpha^{-1}(x) \sum_{x' \in \mathcal{X}} \big(q(x) \cdot k(x') \big) v(x') \, \delta_{x'} \\
&\approx \alpha^{-1}(x) \int_\omega \tilde{\kappa}(x, x') v(x') \, d\mu(x'),
\end{aligned}
\tag{10}
$$

where δ_x is the Dirac measure associated with the position at x. For the second equation, with a slight abuse of the order presentation, we assume that there exists $\tilde{\kappa}(\cdot, \cdot)$ and a Borel measure $\mu(\cdot)$ such that $\tilde{\kappa}(\cdot, \cdot)$ has a reproducing property under the integral by $d\mu(\cdot)$, which shall be elaborated in the next paragraph. Here in this integral, the image domain ω is approximated by a patchified grid $\mathcal{X} \simeq (0, \frac{1}{n} \cdots, \frac{n-1}{n}) \otimes (0, \frac{1}{n} \cdots, \frac{n-1}{n})$. Returning to the perspective of "embedding $\approx$ basis" for the representation space, the backpropagation from the output $z(\cdot)$ to update weights to obtain a new $v(\cdot)$ during pretraining can be interpreted as an iterative procedure to solve the Fredholm integral equation of the first kind in (10).

How to Obtain a Translation Invariant Kernel. We note that the formulation above in (10) is closely related to Reproducing Kernel Hilbert Space (RKHS) [9]. First, the dot-product in (6) is re-defined to be

$$
\mathbf{A}_{i\bullet} := \mathrm{Softmax} \Big(-\gamma (\mathbf{q}_i - \mathbf{k}_i)(\mathbf{Q} - \mathbf{K})^\top \Big).
\tag{11}
$$

Then, a pre-layer normalization scheme in [4, 19, 46, 49] or pre-inner product normalization in [33, 12] and can be a cheap procedure to re-center the kernel. Rewritting the original dot-product for $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{1 \times d}$ as follows

$$
\mathbf{q}\mathbf{k}^\top = \frac{1}{2}\mathbf{q}\mathbf{q}^\top + \frac{1}{2}\mathbf{k}\mathbf{k}^\top - \frac{1}{2}(\mathbf{q} - \mathbf{k})(\mathbf{q} - \mathbf{k})^\top.
\tag{12}
$$

Thus, a layer normalization for each row in $\mathbf{Q}$ and $\mathbf{K}$ before $\mathbf{Q}\mathbf{K}^\top$ results the two moment terms above, $\mathbf{q}\mathbf{q}^\top$ and $\mathbf{k}\mathbf{k}^\top$, in (12) being relatively small versus the third term. Consequently, using merely the third term in (12) to form (11) suffices to provided a good approximation to the scaled dot-product attention. As an important consequence, the kernel becomes symmetric, and translation-invariant, which is arguably one of the nicest traits of CNN kernels. As a result, the normalized attention kernel $\tilde{\kappa}(x, x')$ based on (11) is

$$
(\mathbf{A})_{ij} = \gamma \langle \mathbf{q}_i - \mathbf{k}_i, \mathbf{q}_j - \mathbf{k}_j \rangle =: \kappa(x_i, x_j).
\tag{13}
$$

Moreover, $\kappa(x, x')$ can be written as $K(x - x')$ for $K : \mathbb{R}^{1 \times d} \to \mathbb{R}$ being the 1-dimensional Gaussian radial basis function (RBF) kernel, $K(\theta) = e^{-\gamma \|\theta\|^2}$. By the positivity and the symmetrcity, $\kappa(\cdot, \cdot)$ becomes a reproducing kernel. We can define the following nonlinear integral operator $\mathcal{K}$:

$$
\mathcal{K}(v)(x) := \int_\omega \kappa(x, x') v(x') \, d\mu(x').
\tag{14}
$$

In the traditional settings such that $\kappa(x, x')$ is not explicitly dependent on $v(\cdot)$, the vector version of the Mercer representation theorem for the integral kernel can be exploited [24, 14], and there exists

an optimal representation space to approximate the solution to this integral equation: the eigenspace the nonlinear integral operator $\mathcal{K}$. While for the scaled dot-product attention's integral transform interpretation, explicitly pinning down the eigenspace is impossible during the dynamic training procedure. Nevertheless, we can still obtain a representation expansion to show that the internal representations are propagated in a stable fashion (Theorem 5.1).

Relation to Other Methods. The integral transform formula derived in (10) resembles the nonlocal means method (NL-means) in tasks of traditional image signal processing such as denoising, *e.g.*, [10]. In NL-means, $\kappa(x, x')$ is a non-learnable integral kernel, which measures the similarity between pixels at x and x'. While in ViT, the layer-wise kernel $\kappa(x, x'; \theta)$ is learnable, and establishes the inter-patch topology (continuity, total variation, *etc*). One key common trait is that they are both normalized in the sense that for any $x \in \omega$

$$\alpha^{-1}(x) \int_\omega \kappa(x, x') \, d\mu(x') = 1, \tag{15}$$

which is facilitated by the row-wise softmax operation to enforce a unit row sum. We shall see in Section 5 that this plays a key role in obtaining a stable representation in ViT layers mathematically.

Compared with another popular approach, the Convolutional Neural Network (CNN)-based models, the MAE has several differences. In CNN, the translation-invariance is inherited automatically from the convolution operation, *i.e.*, the CNN kernel depends only on the difference of the positions $(x - x')$. Moreover, $K(\cdot)$ is acting on a pixel level and locally supported, thus having a small receptive field. For CNN to learn long-range dependencies, the deep stacking of convolutions is necessary but then renders the optimization algorithms difficult [40]. Moreover, repeating local operations around a small pixel patch is computationally inefficient in that the convolution filter sizes being small in CNN makes the "message passing" from distant positions back and forth difficult [47]. In MAE, the translation invariance is obtained through proper normalization. The learned kernel acts on the basis function representing each patch. Moreover, the kernel map is globally supported, which means it can learn effectively the interaction between even far away patches. As a result, this global nature makes it easier to learn a better representation, thus greatly reduces the number of layers needed to complete the same task.

5 Stable Representation Propagation in the Attention Block

In this section, we try to explain why the representations are continuously changing in a stable fashion in both the encoder and decoder of MAE, which is first observed in [39]. In the following theorem, we have proved a key result that the softmax normalization plays a vital role in stabilizing the propagation of the representations through the layers.

Theorem 5.1. *Assume that a trained attention block in the MAE/ViT encoder layer such that $\|\mathbf{W}^Q - \mathbf{W}^K\|_2$ are uniformly bounded above and below by two positive constants, and $\mathbf{W}^V = \mathbf{I}$, and the attention kernel being in the form of (13) such that it is translation invariant and symmetric, let $v^{(t)}(\cdot)$ and $v^{(t+1)}(\cdot)$ be the feature maps for the input and output of the scaled dot-product attention (8), and $v^{(t)} \in BV(\omega) \cap \mathcal{V}$, then we have*

$$\|v^{(t+1)} - v^{(t)}\|_\mathcal{V} \leq C(d) \frac{1}{n} \left(\|v^{(t)}\|_\mathcal{V} + |v^{(t)}|_{BV} \right). \tag{16}$$

Proof Sketch and an Interpretation for Theorem 5.1. The softmax operation in (6) makes the attention matrix to have a row sum being 1. This normalization further translates to the integral kernel's integration being 1 in (15). This enables us to estimate of how the internal representations are propagated in a stable evolution from $v^{(t)}$ to $v^{(t+1)}$. We can rewrite $v^{(t)}(x)$

$$v^{(t)}(x) = \alpha^{-1}(x) \int_\omega \kappa(x, x') v^{(t)}(x) \, d\mu(x'), \tag{17}$$
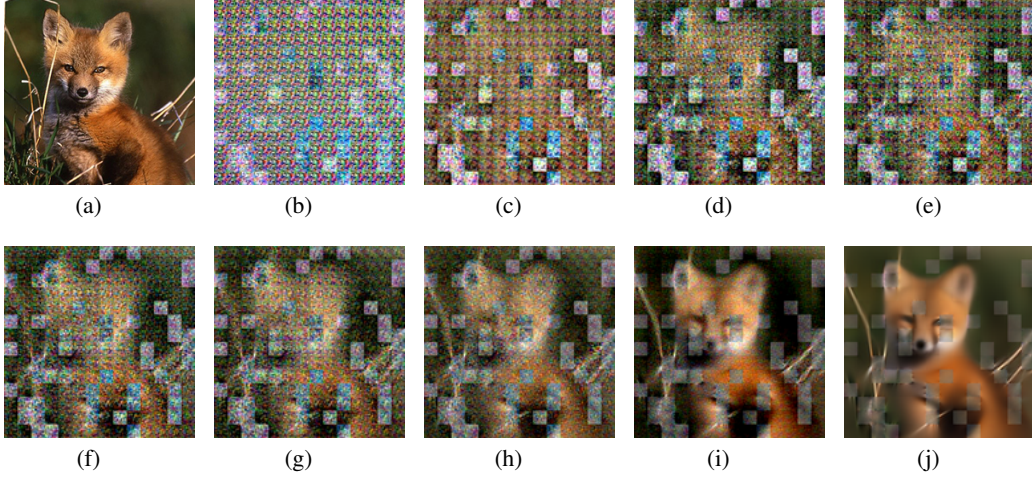
7

Figure 3: The internal representation propagation during evaluation for a random sample from ImageNet-1K validation set. (a) input of MAE encoder; (b) output of MAE encoder, concatenated with the learnable mask token embedding that is shared among all masked patches; (c)–(i) the latent representations, *i.e.*, the outputs of the decoder layers #1 to #7; (j) output of the last decoder layer (#8). (We use the official released model and pretrained weights at https://github.com/facebookresearch/mae.) See texts for details. Best viewed in color.

Thus, the propagation can lead the following modulus of continuity representation

$$
\begin{aligned}
&v^{(t+1)}(x) - v^{(t)}(x) \\
&= \alpha^{-1}(x) \int_\omega \kappa(x, x') v^{(t)}(x') \, d\mu(x') - v^{(t)}(x) \\
&= \alpha^{-1}(x) \int_\omega \kappa(x, x') \big(v^{(t)}(x') - v^{(t)}(x)\big) \, d\mu(x').
\end{aligned}
\tag{18}
$$

Using the argument in Section 4.1, $\kappa(x, x') = K(x - x')$ thus a simple substitution can be exploited to get the following representation:

$$
\begin{aligned}
&\alpha^{-1}(x) \int_\omega K(x - x') \big(v^{(t)}(x') - v^{(t)}(x)\big) \, d\mu(x') \\
&= \alpha^{-1}(x) \int_{\omega_\xi} K(\xi) \big(v^{(t)}(x + \xi) - v^{(t)}(x)\big) \, d\mu(\xi).
\end{aligned}
\tag{19}
$$

By Mercer's theorem, there exists an eigenfunction expansion for $\kappa(x, x') = \sum_{i=1}^\infty a_i \psi_i(x) \psi_i(x')$ that has a spectral decay. Then, the propagation bounds can be proved under the assumption that the current layer's learned feature map offers a "reasonable" approximation to the eigenspace of $\kappa(\cdot, \cdot)$.

Overall, this bound describes that the propagation of the internal representation is continuously changing based on the inter-patch interaction encoded in the attention kernel. In the kernel interpretation of the attention mechanism, there is one key difference with the conventional kernel method: the conventional kernel measures the inter-sample similarity in the whole dataset; while the attention kernel $\kappa(\cdot, \cdot)$ or $\tilde{\kappa}(\cdot, \cdot)$ is built for a single instance, and learns inter-patch topological relations to build a better representation space for the MAE. This learned representation for a specific data sample determines how amenable it is for the downstream tasks.

Additionally, we note that this result applies to the layer-wise representation propagation in the decoder layers as well. An enlightening illustration is demonstrated in Figure 3. There are multi-fold interesting aspects about this evolution diagram shown: (1) in the ViT-base MAE, the embedding dimension of its decoder is merely 512, which is less than the embedding dimension in the encoder ($d = 768$). Note that $768 = N_c \times N_c \times 3$ which enables that the latent representation (a row vector) in a single patch can be directly reshaped to an image patch. Thus, we need a patch-wise upsampling projection; (2) this projection, which maps vectors in $\mathbb{R}^{196 \times 512}$ to those in $\mathbb{R}^{196 \times 768}$

resides in the MAE decoder. It should be noted that this upsampling projection is only connected with the last decoder layer. Heuristically, this projection may only upsample the last decoder layer's output $\mathbf{z}_8 \in \mathbb{R}^{196 \times 512}$ to the desired embedding dimensions. Nevertheless, due to the stable representation propagation, the outputs from the previous decoder layers can also benefit from the last layer projection weights to obtain sensible reconstruction results, as illustrated in Figure 3.

Skip-Connection as a Tikhonov Regularizer After presenting Theorem 5.1, one might ask: given the integral transform representation is already stable, what is the role of the skip-connection? Diverting from the original interpretation of battling the diminishing gradient for the skip-connection [32], here we offer a new perspective and some heuristics inspired by functional analysis to articulate the reason why using the skip-connection would make the representation propagation more stable, which is also observed in [39].

Knowing that the integral kernel interpretation of attention (10) resembles the Fredholm integral equation of the first kind, we can interpret the skip-connection as a layer-wise Tikhonov regularization in the Fredholm integral equation of the second kind. Starting from (10)

$$z(x) = \alpha^{-1} \int_\omega \kappa(x, x')v(x')d\mu(x'). \tag{20}$$

This Fredholm intergral equation of the first kind is usually extremely ill-posed [28], in the sense that the solution, even if it exists, does not depend continuously on $v(\cdot)$.

If $\mathbf{W}^V \approx \beta^{-1}\mathbf{I}$, then for $x \in \mathcal{X}$, we have that

$$z(x) \approx \beta v(x) + \alpha^{-1} \int_\omega \kappa(x, x')v(x')d\mu(x'). \tag{21}$$

This is the Fredholm integral equation of the second kind witha variable coefficient, the extra $\beta v(\cdot)$ term not only contributes to the well-posedness of this equation, but renders the numerical scheme to approximate a better $v(\cdot)$ more stable. Tikhonov firstly introduced this method in [41], see also [35, Chapter 16].

Theorem 5.2 (Skip-connection can represent the minimizer to a Tikhonov-regularized integral equation functional)**.** *For $\mathcal{K}(\cdot)$ using $\kappa(\cdot, \cdot)$ as the integral kernel in (13), define the following functional*

$$\mathcal{L} := \frac{1}{2}\|\alpha^{-1}(\cdot)\mathcal{K}(v) - z\|_{\mathcal{V}*}^2 + \beta\langle \mathcal{K}(v), v\rangle, \tag{22}$$

where $\|u\|_{\mathcal{V}'} := \sup_{v \in \mathcal{V}} |\langle u, v\rangle|/\|v\|_{\mathcal{V}}$ denotes the dual norm. The Euler-Lagrange equation associated with $\min_v \mathcal{L}$ has the following form:

$$\beta u + \alpha^{-1}\mathcal{K}(u) = z \quad in \ (\mathcal{K}^*(\mathcal{V}))'. \tag{23}$$

Note that (23) bears exactly the same form with (21). Moreover, instead of the conventional L^2-type Tikhonov regularizer $\|v\|_{\mathcal{V}}^2$, here $\langle \mathcal{K}(v), v\rangle$ is an induced norm by the positive (semi)definite integral operator $\mathcal{K}(\cdot)$. The skip-connection term in (23) comes from the Tikhonov regularization term in (22), and without it, the Euler-Lagrange equation reverts to the Fredholm integral equation of the first kind.

6 MAE Decoder: Low-Rank Reconstruction Through Global Interpolation

During pretraining, the major function of the MAE decoder is to map the low-rank representation obtained by the MAE encoder to a reconstruction. The encoder embedding has a bigger dimension, yet is define on only a fraction of patches. The decoder reduces the embedding dimension, but is able to obtain the embedding for all $p \times p$ patches including the masked ones. Despite of the fact that the decoder in MAE is only used in pretraining, not downstream tasks, it plays a vital role in learning a "good" representation space for the MAE encoder.

Enrichment of the Representation Space through Positional Embedding Because the positional embedding $\mathbf{x} := \|_{i=1}^n x_i \in \mathbb{R}^{p \times d}$ is added to the latent representation $\mathbf{y}$ in (3), the nonlinear universal approximator (FFN) in each attention block shall also contribute to learning a better

representation. In every decoder layer, the basis functions in $\mathcal{V}$ are being constantly enriched by $\text{span}\{w_j \in \mathcal{X} : w_j(x_i) = (\text{FFN}(\mathbf{y} + \mathbf{x}))_{ij}, 1 \leq j \leq d\}$. If we treat the positional embeddings $\{x_i\}$ as coordinate again, $\text{FFN}(\mathbf{y} + \mathbf{x})$ is a subset of the famous Barron space [6, 3], which has a rich and powerful representation power that can approximate smooth function arbitrarily well, *e.g.*, see [34, 50]. As a result, the representation space is dynamically updated layer by layer to try to build a more expressive representation to better characterize the inter-patch topology. FFNs themselves have no linear structure, however, the basis functions produced this way act as a building block to update the linear space for the expansion in (8) and (10). In the MAE decoder, the number of basis ($p^2 = 196$) is much greater than that of the MAE encoder ($0.25 \cdot p^2 = 49$), thus heuristically speaking, this basis update mechanism is mostly working in the decoder.

Reconstruction is a Global Interpolation.   In MAE decoder, the mask token, a learnable vector shared by every masked patch, is concatenated to the unshuffled representation of the unmasked patches. To demonstrate that the reconstructed image is a global interpolation by using the unmasked patch embeddings, we first need to prove the following lemma, which states that the learned mask token embedding can be simply replaced by zero with an updated set of attention weights. For the simplicity, we denote the embedding dimension in the MAE decoder still as d.

Lemma 6.1. *Let* $\mathbf{m} \in \mathbb{R}^{1 \times d}$ *be the learned mask token embedding that is shared by all masked patches, and let* $m(\cdot)$ *be its feature map. Denote the set of masked and unmasked patches' index as* $\mathcal{M}$ *and* $\mathcal{N}$*, respectively, and we assume that* $|\mathcal{N}| \geq 2$*. For the input* $\mathbf{y} \in \mathbb{R}^{p \times d}$ *that already contains the masked patches, i.e., all* $i \in \mathcal{M}$*,* $\mathbf{y}_i = \mathbf{m}$*, , there exists a new set of affine linear maps to generate* $\{\widehat{\mathbf{Q}}, \widehat{\mathbf{K}}, \widehat{\mathbf{V}}\}$*, such that for any* $i = 1, \ldots, n$*,*

$$\left\| \sum_{j=1}^{n} A(\mathbf{q}_i, \mathbf{k}_j) \, \mathbf{v}_i - \sum_{j \in \mathcal{N}} A(\hat{\mathbf{q}}_i, \hat{\mathbf{k}}_j) \, \hat{\mathbf{v}}_i \right\| < Cn^{-1}. \tag{24}$$

With this lemma, we are already present our final result: for the first layer of the MAE decoder, the network interpolates the representation using global information from the embeddings learned by the MAE encoder, not just the ones from the nearby patches. Moreover, with Theorem 5.1, the MAE decoders continue to perform such a global interpolation in subsequent layers. For an empirical evidence, please refer to Figure 3.

Proposition 6.2 (Interpolation results for masked patches)**.** *For the embedding of every masked patch* $i \in \mathcal{M}$*, let* $v_i^{(t+1)}$ *be the output embedding of a decoder layer for this patch, and let* $v_j^{(t)}$ *be the input from the encoder for* $j = 1, \ldots, p$*, then* $v_i^{(t+1)}$ *is*

$$v_i^{(t+1)} = \sum_{j \in \mathcal{N}} a_j v_j^{(t)}, \tag{25}$$

for a set of weights based on unmasked patches $a_j(v_{i_1}, \ldots, v_{i_k})$*,* $\mathcal{N} := \{i_1, \ldots, i_k\}$*. Moreover, the reconstruction quality is bounded above by the global reconstruction error of the unmasked patches,*

$$\|Rv_i - \mathbf{u}_i\|_{BV(\Omega_i)} \leq C \sup_{j \in \mathcal{N}} \|Rv_j - \mathbf{u}\|_{BV(\Omega_j)} + Cn^{-1}\|\mathbf{u}\|. \tag{26}$$

7   Conclusion

To the best of our knowledge, to date, there are no theoretical viewpoints to explain the powerful expressivity of MAE. In this paper, we, for the first time, propose a unified theoretical framework that provides a mathematical understanding for MAE. Particularly, we explain the patch-based attention approaches of MAE from a perspective of an integral kernel under a non-overlapping domain decomposition setting. To help the researchers to further grasp the main reasons for the great success of MAE, our mathematical proof contributes the following major conclusions:

(1) The attention mechanism in MAE is a learnable integral kernel transform, and its representation power is dynamically updated by the Barron space with the positional embeddings that work as the coordinates in a high dimensional feature space.

(2) The random patch selecting of MAE preserves the information of the original image, while reduces the computing costs, under common assumptions on the low-rank nature of images. This

paves a theoretical foundation for the patch-based neural networks/models including but not limited to MAE or ViT.

(3) The great success of MAE benefits from a main reason that the scaled dot-product attention built-in ViT provides the stable representations during the cross-layer propagation.

(4) In MAE, the decoder is vital to helping the encoder to build better representations, while decoder is discarded after pretraining.

(5) The latent representations of the masked patches are interpolated globally based on an inter-patch topology that is learned by the attention mechanism.

Furthermore, our proposed theoretical framework can be used to understand the intrinsic traits of not only the ViT-based models but also even the extensive networks/models made by patch and self-attention.

References

[1] Kendall Atkinson and Weimin Han. Numerical solution of fredholm integral equations of the second kind. In *Theoretical Numerical Analysis*, pages 473–549. Springer, 2009.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[4] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*, 2018.

[5] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. 2021.

[6] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[7] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

[8] Peter Berger, Gabor Hannak, and Gerald Matz. Graph signal recovery via primal-dual algorithms for total variation minimization. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):842–855, 2017.

[9] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[10] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale modeling & simulation*, 4(2):490–530, 2005.

[11] I.U.D. Burago, J.D. Burago, and V.G. Maz'ya. *Potential Theory and Function Theory for Irregular Regions*. Seminars in mathematics. Consultants Bureau, 1969.

[12] Shuhao Cao. Choose a transformer: Fourier or galerkin. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[13] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[14] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[16] Tony F Chan and Tarek P Mathew. Domain decomposition algorithms. *Acta numerica*, 3:61–143, 1994.

[17] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[19] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[20] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] JC Ferreira and VA2501172 Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

[25] C. Gerhardt. *Analysis II*. Analysis. International Press, 2006.

[26] Craig R Gin, Daniel E Shea, Steven L Brunton, and J Nathan Kutz. Deepgreen: deep learning of green's functions for nonlinear boundary value problems. *Scientific reports*, 11(1):1–14, 2021.

[27] CW Groetsch. The theory of tikhonov regularization for fredholm equations. *104p, Boston Pitman Publication*, 1984.

[28] Jacques Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Courier Corporation, 2003.

[29] Paul Richard Halmos and Viakalathur Shankar Sunder. *Bounded integral operators on L 2 spaces*, volume 96. Springer Science & Business Media, 2012.

[30] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, Online, November 2020. Association for Computational Linguistics.

[34] Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, and Tuan Anh Nguyen. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN partial differential equations and applications*, 1(2):1–34, 2020.

[35] Rainer Kress. *Linear Integral Equations*. Springer New York, 2014.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[38] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *AAAI*, 2022.

[39] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] Andrei Nikolajevits Tihonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

[42] Andrea Toselli and Olof Widlund. *Domain decomposition methods-algorithms and theory*, volume 34. Springer Science & Business Media, 2004.

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[45] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[46] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[48] Jürgen Weese. A reliable and fast method for the solution of fredhol integral equations of the first kind based on tikhonov regularization. *Computer physics communications*, 69(1):99–111, 1992.

[49] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

[50] Jinchao Xu. Finite neuron method and convergence analysis. *Communications in Computational Physics*, 28(5):1707–1745, 2020.

[51] Peng Xu, Chaitanya K Joshi, and Xavier Bresson. Multigraph transformer for free-hand sketch recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[52] Xiaoqun Zhang, Martin Burger, Xavier Bresson, and Stanley Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences*, 3(3):253–276, 2010.

A   Proof of Theorem 5.1

Assumption A.1 (Assumption of Theorem 5.1). To prove Theorem 5.1, we assume the following conditions hold true,

(B_1) The underlying Hilbert space for the latent representation is $\mathcal{V} \subset \left(L^2(\omega, \mu)\right)^d$, i.e., there is no *a priori* inter-channel continuity, and the inter-channel topology is learned from data.

(B_2) For a BV function on ω, there exists a smooth extension to $\mathbb{R}^2$ such that the extension is stable in $|\cdot|_{BV(\Omega)}$ (see *e.g.*, [25], [11, Theorem 2]).

(B_3) For each $v \in \mathcal{V}$, there exists a smooth K such that $\kappa(x, x') = K(x - x')$, and $K(x - x') \le C_\gamma e^{-\gamma \|x - x'\|}$ with a uniform constant C_γ that only depends on γ.

(B_4) For each $v \in \mathcal{V}$, $\alpha(x) := \int_\omega \kappa(x, x') \, d\mu(x')$, and α is bounded below by a positive constant C_α uniformly.

Theorem 5.1 (Stability result, rigorous version). *Under the assumption of Assumption A.1, we have*

$$\|v^{(t+1)} - v^{(t)}\|_{\mathcal{V}} \le C(\gamma, d) \frac{1}{n} \left(\|v^{(t)}\|_{\mathcal{V}} + |v^{(t)}|_{BV}\right). \tag{27}$$

Proof. First we extend the kernel function from ω to the whole $\mathbb{R}^2$ by

$$\text{Ext}(\kappa(x, x')) = \begin{cases} \kappa(x, x') & \text{if } x \text{ and } x' \in \omega, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

With a slight abuse of notation we still denote the extension $\text{Ext}(\kappa(\cdot, \cdot))$ still as $\kappa(\cdot, \cdot)$, and the normalization (15) still holds:

$$\alpha^{-1}(x) \int_\omega \kappa(x, x') \, d\mu(x') = \alpha^{-1}(x) \int_{\mathbb{R}^2} \kappa(x, x') \, d\mu(x') = 1. \tag{29}$$

Without loss of generality, we also assume that the Thus, under assumption (B_1)–(B_2), we have the original integral on ω can be written as on the full plane

$$v^{(t+1)}(x) = \alpha^{-1}(x) \int_{\mathbb{R}^2} \kappa(x, x') v^{(t)}(x') \, d\mu(x'). \tag{30}$$

Similarly by (29),

$$v^{(t)}(x) = \alpha^{-1}(x) \int_{\mathbb{R}^2} \kappa(x, x') v^{(t)}(x) \, d\mu(x'). \tag{31}$$

Therefore,

$$
\begin{aligned}
&v^{(t+1)}(x) - v^{(t)}(x) \\
&= \alpha^{-1}(x) \int_{\mathbb{R}^2} \kappa(x, x')\big(v^{(t)}(x') - v^{(t)}(x)\big)\, d\mu(x').
\end{aligned}
\tag{32}
$$

Using $\kappa(x, x') = K(x - x')$, let $\xi = x' - x$:

$$
\begin{aligned}
&\alpha^{-1}(x) \int_{\mathbb{R}^2} K(x - x')\big(v^{(t)}(x') - v^{(t)}(x)\big)\, d\mu(x') \\
&= \alpha^{-1}(x) \int_{\mathbb{R}^2} K(\xi)\big(v^{(t)}(x + \xi) - v^{(t)}(x)\big)\, d\mu(\xi).
\end{aligned}
\tag{33}
$$

By $v^{(t)} \in BV(\mathbb{R}^2)$, and ω being compact, there exists $\delta > 0$ such that

$$
\sup_{x \in \omega} \sup_{\|\xi\| < \delta/n} \|v^{(t)}(x + \xi) - v^{(t)}(x)\| < \frac{1}{n}|v^t|_{BV(\mathbb{R}^2)}.
\tag{34}
$$

Thus, for any $x \in \omega$

$$
\begin{aligned}
&\|v^{(t+1)}(x) - v^{(t)}(x)\| \\
&\leq |\alpha^{-1}| \int_{\mathbb{R}^2} \|K(\xi)\big(v^{(t)}(x + \xi) - v^{(t)}(x)\big)\|\, d\mu(\xi) \\
&\leq |\alpha^{-1}| \sup_{\|\xi\| < \delta/n} \|v^{(t)}(x + \xi) - v^{(t)}(x)\| \int_{\omega} |K(\xi)|\, d\mu(\xi) \\
&\quad + |\alpha^{-1}| \int_{\|\xi\| \geq \delta/n} |K(\xi)|\, 2\|v^{(t)}\|_\infty d\mu(\xi) \\
&\leq 2 C_\alpha^{-1} C_\gamma \frac{1}{n} \left(|v^{(t)}|_\infty + |v^{(t)}|_{BV}\right).
\end{aligned}
\tag{35}
$$

Taking $\sup_{x \in \omega}$ yields the result. $\qquad\square$

B   Proof of Theorem 5.2

Proof. Define $\eta(\cdot) : \mathcal{V} \to \mathbb{R}$ such that $\eta_u(v) := \langle z, v \rangle - \langle \alpha^{-1}\mathcal{K}u, v \rangle$, where $u \in \mathcal{W} \subseteq \mathcal{V}$ is the solution to the integral equation in (23), and $\mathcal{W}$ is the solution subspace. Clearly, $\eta_u \in \mathcal{V}'$ defines a bounded functional on $\mathcal{V}$ for $(\mathcal{V}, \langle \cdot, \cdot \rangle)$. By Riesz representation theorem, there exists an isomorphism $G : \mathcal{V} \to \mathcal{V}'$ such that $\phi_u := G^{-1}(\eta) \in \mathcal{V}$ and

$$
\eta_u(v) = \langle \phi_u, v \rangle = \langle z, v \rangle - \langle \alpha^{-1}\mathcal{K}(u), v \rangle.
\tag{36}
$$

Then, define $\|\cdot\|_{\mathcal{V}}^2 := \langle \cdot, \cdot \rangle$, we have

$$
\min_{u \in \mathcal{V}} \|\eta(\cdot)\|_{\mathcal{V}'}^2 = \min_{u \in \mathcal{V}} \|\phi_u\|_{\mathcal{V}}^2 = \min_{u \in \mathcal{V}} \|G^{-1}(\langle z, \cdot \rangle - \langle \alpha^{-1}(\cdot)\mathcal{K}(u), \cdot \rangle)\|_{\mathcal{V}}^2.
$$

Thus, the functional $\mathcal{L}(v)$ can be written as:

$$
\mathcal{L}(v) = \frac{1}{2} \left\{ \|\alpha^{-1}(\cdot)\mathcal{K}(v) - z\|_{\mathcal{V}'}^2 + \beta\langle \mathcal{K}(v), v \rangle \right\}
\tag{37}
$$

Taking the Gateaux derivative $\lim_{\tau \to 0} d\mathcal{L}(u + \tau w)/d\tau$ in order to find the critical point(s) $u \in \mathcal{W}$, we have for any perturbation $w \in \mathcal{W}$ such that $u + \tau w \in \mathcal{W}$

$$
\begin{aligned}
0 = \lim_{\tau \to 0} \frac{d}{d\tau} \bigg\{ &\langle \alpha^{-1}(\cdot)G^{-1}(\langle z, \cdot \rangle - \alpha^{-1}(\cdot)\mathcal{K}(u + \tau v), \cdot), G^{-1}(\langle z, \cdot \rangle - \alpha^{-1}(\cdot)\mathcal{K}(u + \tau v), \cdot) \rangle, \\
&+ \beta\langle \mathcal{K}(u + \tau v), u + \tau v \rangle \bigg\}
\end{aligned}
$$

14

and applying G^{-1} on $\mathcal{K}(\cdot, u) \in \mathcal{V}'$, it reads for any $w \in \mathcal{W}$

$$
\begin{aligned}
&\langle \alpha^{-1}(\cdot) G^{-1}\big(\langle z, \cdot \rangle - \langle \mathcal{K}(u), \cdot \rangle \big), G^{-1}(\langle \mathcal{K}(w), \cdot \rangle) \rangle + \langle \mathcal{K}(u), v \rangle \\
&= \langle \alpha^{-1} \phi_u, G^{-1}(\langle \mathcal{K}(u), \cdot \rangle) \rangle + \langle \mathcal{K}(u), w \rangle \\
&= \langle \alpha^{-1} \mathcal{K}(w), \phi_u \rangle + \langle \mathcal{K}(u), w \rangle = 0.
\end{aligned}
\tag{38}
$$

As a result, combining (38), (36), and the self-adjointness of $\mathcal{K}$, we have the system in the continuum becomes

$$
\begin{cases}
\langle \phi_u, v \rangle + \langle \alpha^{-1} \mathcal{K}(u), v \rangle = \langle z, v \rangle, & \forall v \in \mathcal{V}, \\
\langle \alpha^{-1} \mathcal{K}(w), \phi_u \rangle + \beta \langle \mathcal{K}(u), w \rangle = 0, & \forall q \in \mathcal{Q}.
\end{cases}
\tag{39}
$$

The first equation implies that:

$$
\phi_u + \alpha^{-1} \mathcal{K}(u) = z \quad \text{in } \mathcal{V}'.
\tag{40}
$$

Hence, by $w \in \mathcal{W} \subseteq \mathcal{V}$, when plugging w to the second equation in (36) we have

$$
\langle z - \alpha^{-1} \mathcal{K}(u) + \beta u, \mathcal{K}(w) \rangle = 0 \quad \text{for } w \in \mathcal{W},
\tag{41}
$$

and the theorem follows.

$\square$