LOOPE: Learnable Optimal Patch Order for Positional Encoders in Vision Transformers

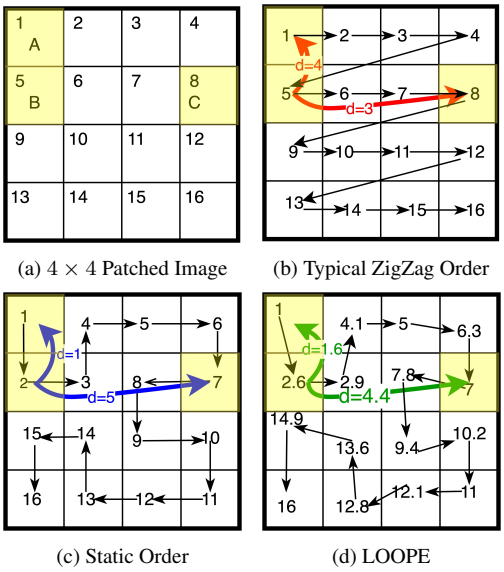
Anonymous CVPR submission

Abstract

Transformers are inherently permutation invariant, making positional encoding (PE) fundamental to their success. While prior work has focused on designing absolute or relative positional encoders, nearly all approaches implicitly assume that the 1D ordering of image patches is fixed. Yet for higher-dimensional data such as images, mapping a 2D grid into a 1D sequence is itself a geometric operation—one that can distort locality, disrupt neighborhood structure, and ultimately limit how effectively PEs capture spatial relations. Surprisingly, the impact of the order of patches on positional representation has received little attention. We revisit this overlooked design dimension and show that patch order is not merely a preprocessing choice but a learnable degree of freedom that fundamentally shapes positional embedding quality. We propose LOOPE, a lightweight framework that learns an image-dependent ordering of patches. Starting from a locality-preserving space-filling curve and applying small context-aware refinements, LOOPE yields an interpretable and content-adaptive coordinate system on which standard sinusoidal encodings operate robustly. Across multiple ViT and ViT-based architectures, LOOPE improves performance and stabilizes PEs under challenging settings such as non-optimal frequency schedules. To complement standard benchmarks, we introduce a simple diagnostic task revealing that effective PEs can preserve far more positional information than previously reported. Our findings highlight learnable patch ordering as a powerful and largely untapped tool for improving positional encoding in vision Transformers.

1. Introduction

Transformers have become the backbone of modern computer vision tasks—from image classification to image segmentation—yet they inherently lack explicit spatial ordering. Positional encoding (PE) is therefore essential for capturing spatial relationships. While numerous PE variants exist, effectively modeling 2D spatial structure remains challenging.



(a) 4×4 Patched Image (b) Typical ZigZag Order

(c) Static Order (d) LOOPE

Figure 1. Fig. 1(b) shows that under Generic ZigZag Order, the distances between patches are $d(B, A) = 4$ and $d(B, C) = 3$, even though B is spatially and contextually closer to A. Static Order (Fig. 1(c)) improves this with $d(B, A) = 1$ and $d(B, C) = 5$, but it remains static. LOOPE (Fig. 1(d)) dynamically reorders based on context, yielding $d(B, A) = 1.6$ and $d(B, C) = 4.4$.

Traditional PEs—such as sinusoidal encodings [1], learned absolute encodings [1], and relative positional encodings (RPEs) [25]—either impose rigid spatial biases or introduce substantial parameter overhead. Conditional encoders like CPE [5] incorporate local information but discard global coordinates, while frequency-based approaches [22, 31, 37] and Learnable Fourier Features (LFF) emphasize spectral design. However, a common assumption underlies all these methods: the 1D ordering of image patches is fixed (e.g., raster or zigzag) and independent of image content.

Most positional encoders can be represented by Eq. 1,

where PE construction depends on a frequency set, a sequence of patch indices, and an optional phase:

$$\mathbf{PE} = Policy\left(\begin{bmatrix} \omega_1 \\ \dots \\ \omega_j \\ \dots \\ \omega_L \end{bmatrix}, \underline{[\mathbf{x_1}, ., \mathbf{x_i}, ., \mathbf{x_N}]}, \frac{\pi}{2}\delta \right) \quad (1)$$

Yet none of these approaches examine **how the patch sequence itself should be constructed**. This is notable because 2D→1D flattening determines which patches become neighbors in the sequence, directly influencing locality preservation, similarity patterns, and how frequencies interact with spatial structure.

We observe that periodic PEs depend on three coupled factors: **context**, **order**, and **frequencies**. Contextual encoders [15] and frequency-based methods, including LFF, explore portions of this space, but the joint effect of context and patch order is almost entirely unexplored. Additionally, although RPEs often perform well on downstream tasks, they encode positional information inside attention weights and do not preserve a reusable absolute coordinate system. As shown later in the Three-Cell Experiment (Sec. 4.4), when labels depend purely on geometry, well-structured absolute PEs can outperform strong RPEs, highlighting the continuing importance of absolute encodings. In this work, we address this gap by treating patch ordering as a learnable, context-aware variable and by studying its interaction with sinusoidal PEs. To summarize,

1. Our primary contribution is the introduction of **LOOPE**, a lightweight patch-ordering framework that combines a topology-preserving static Gilbert curve X_G with a context-aware refinement X_C, yielding an image-dependent coordinate system that minimizes structural distortion and enhances robustness to frequency choices.To s
2. To prove the locality preserving nature of LOOPE, we additionally introduce lightweight structural probes (PESI) for analyzing monotonicity and symmetry in learned positional representations.
3. Finally, we propose a simple diagnostic task (Three-Cell Experiment) revealing that carefully structured absolute PEs can retain substantially richer spatial information than task-oriented RPEs.

Extensive experiments across ViT and ViT-based architectures demonstrate that LOOPE not only improves performance but also provides a clearer geometric understanding of how positional information is encoded within transformer architectures.

2. Related Works

Positional encoding (PE) is crucial for ViTs as self-attention lacks spatial awareness. Absolute positional encodings (APEs) [1], first introduced in NLP, were adapted for vision tasks where images are tokenized into patches. Models like ViT [10] and DeiT [32] use fixed sinusoidal or learnable encoders. APEs provide positional awareness but suffer from fixed-size constraints, limiting adaptability to varying resolutions in models such as CrossViT [2] and Swin [19], and fail to capture relative spatial relationships.

Relative positional encodings (RPEs) [26, 36] encode pairwise relationships, generalizing across sequence lengths for NLP [8] and vision tasks, e.g., axial attention [34] and 2D-aware encodings [23]. However, they discard absolute position information, limiting fine-grained spatial representation, particularly in object localization [27].

Hybrid PE strategies balance absolute and relative cues. CP-RPE [36] encodes horizontal and vertical distances separately but struggles with optimal mapping. CPE [5] generates encodings from local context, improving translation invariance but is hyperparameter-sensitive and lacks structured positional order. RoPE [12, 29] preserves relative positioning via rotational invariance but misses 2D spatial hierarchies, whereas AS2DRoPE [6] introduces scalable 2D priors. Learned Fourier Features (LFF) [14] further refine spatial representation by learning a spectral basis over fixed coordinates; however, LFF assumes a predetermined patch ordering and does not provide a mechanism for constructing or adapting the coordinate system itself.

ViTs are widely used for segmentation. SETR [38], Segmenter [28], and DPT [24] use ViT backbones with learned APEs. Mask2Former [4] uses Swin or ResNet [11] backbones, inheriting RPE-based design.

3. Methodology

3.1. Motivation

3.1.1. A Geometric Perspective on the Problem

Let an image be partitioned into an $h \times w$ grid of patches, giving $N = hw$ tokens. A Vision Transformer must process these tokens as a 1D sequence, which requires a bijection

$$\pi : \{1, \dots, h\} \times \{1, \dots, w\} \longrightarrow \{1, \dots, N\} \quad (2)$$

mapping every 2D patch coordinate to a unique sequence index. The quality of this mapping depends on how well it preserves spatial relationships: ideally, patches that are close in the grid remain close in the sequence. In geometric terms, the goal is to minimize distortion of Euclidean distances under the mapping π, i.e.

$$\|(i, j) - (i', j')\|_2 \approx |\pi(i, j) - \pi(i', j')| \quad (3)$$

so that neighborhood structure is not destroyed when flattening the image.

This is precisely the principle behind space-filling curves such as Hilbert [13] or Peano [21] traversals, which are

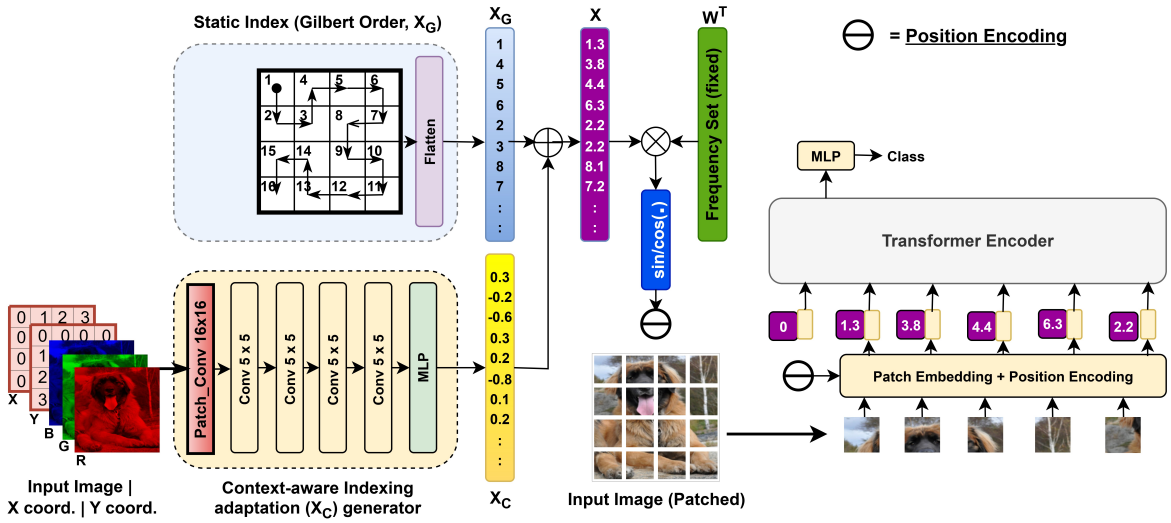


Figure 2. The complete architecture of LOOPE. For architectural details refer to *Supp. A.2*

designed to unfold two-dimensional domains into one-dimensional paths with strong locality preservation.

For ViTs, this ordering problem is not incidental but fundamental. Since self-attention is permutation-invariant, positional encodings are the only source of spatial information. If the initial 2D→1D mapping π introduces large distortions, the positional encoder must spend capacity correcting for these artifacts rather than representing meaningful structure—and even sophisticated APEs or RPEs cannot fully recover geometric relations once the coordinate system is distorted. Frequency-based methods such as LFF learn powerful positional functions but still assume a fixed ordering; our perspective instead optimizes the coordinate system itself.

In this light, patch ordering is not a cosmetic design choice but a structural constraint that determines how much geometric information a positional encoding can ultimately preserve.

3.1.2. Limitations of Existing Dynamic SFC Approaches

Prior attempts to learn adaptive space-filling curves (SFCs) [3, 35] highlight the importance of patch ordering, but they face two key obstacles. Most rely on discrete procedures such as Minimum Spanning Trees (MSTs) to build Hamiltonian traversals, which cannot be optimized end-to-end because they lack differentiable loss functions. GNN-based variants [35] estimate edge weights before applying MST heuristics, but this introduces substantial computational overhead. Moreover, these methods optimize curves independently of image content and therefore do not yield sample-specific orderings. While suitable for small graphs, such pipelines are far too heavy for positional encodings in ViTs, which must remain lightweight and efficient. In short, existing contextual SFC generators are either non-differentiable or computationally impractical—leaving open the need for a method that is both trainable and efficient in the transformer setting.

3.2. Proposed Method

To address these limitations, we propose **LOOPE**, a learnable patch-ordering framework that unifies two complementary components: a static space-filling order $\mathbf{X_G} \in \mathbb{Z}_+^{1\times N}$ and a context-aware refinement $\mathbf{X_C} \in [-1,1]^{1\times N}$, where N is the number of patches. **The static part ensures stability and locality preservation, while the dynamic part adapts to image content**, yielding an indexing scheme that is both efficient and trainable. In the following, we first describe the static order and then detail the context-aware refinement.

Static Patch Index (Gilbert Order, X_G): The Hilbert curve maps a $2^n \times 2^n$ grid, $n \in \mathbb{Z}_+$, into a 1D sequence while preserving locality, but it is restricted to square grids of power-of-two size. To generalize, we adopt the **Gilbert order** [40], which recursively partitions arbitrary rectangular grids while maintaining strong spatial coherence (see Fig. 1c). This improves over typical zigzag order by reducing locality distortion, but its effectiveness still depends on the chosen frequency set, as analyzed in Table 6.

Context-Aware Index Adaptation (X_C): The static Gilbert order X_G offers stability and locality preservation, but it remains tied to the hand-crafted choice of frequency set (Table 6). To address this, we introduce a learnable *context-aware bias* X_C that adapts the patch order to the input image itself. As shown in Fig. 2, the generator G takes as input the concatenated tensor $[\text{Image} = I_0 \in \mathbb{R}^{3\times H\times W}, \text{coordinates} = x, y \in \mathbb{R}^{1\times H\times W}] \in \mathbb{R}^{5\times H\times W}$

and outputs continuous index offsets. These offsets are added to X_G, yielding a refined ordering like in Fig. 3, 1d that is both content-adaptive and differentiable.

This refinement introduces two decisive properties. First, it **mitigates frequency sensitivity**: unlike methods tied to a fixed frequency design, X_C adaptively corrects the encoding, keeping it stable across different frequency sets (Table 6). Second, it enables **fractional indexing**, where patch indices are no longer restricted to integers—allowing patches to be pulled closer, pushed apart, or locally reordered while preserving differentiability. This flexibility gives X_C the freedom to offset poor frequency choices and to adjust neighborhood geometry in a lightweight manner (see $\partial \mathrm{PE}/\partial W$ analysis in *Supp.A.3*). Thus, while X_C introduces spatial awareness, its central role is to extend X_G into a geometry-preserving, *frequency-robust* framework—far more than a simple spatial inductive bias.

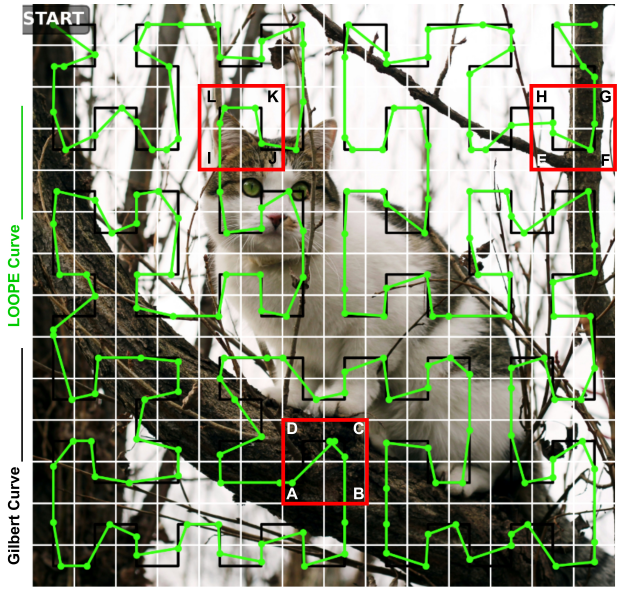


Figure 3. Here is an example of the complete generated curve. i) $\mathbf{d_{IL}} \downarrow$ to capture ear. ii) $\mathbf{d_{EH}} \downarrow$ (sky to sky) and $\mathbf{d_{EF}} \uparrow$ (sky to trunk). iii) $\mathbf{d_{CD}} \downarrow$ to swap C(Branch), D(Branch+Twig). Seemingly, A(Branch) is more similar to C than D, making $\underline{ADCB} \rightarrow \underline{ACDB}$. Contextually similar patches are drawn closer. Here, $\downarrow$= distance decreased, $\uparrow$= increased.

Final Formulation: Combining the static Gilbert order X_G with the context-aware refinement X_C, LOOPE defines the final positional encoding as

$$\mathbf{PE}(X) \;=\; \sin\Big(W(X_G + X_C) + \tfrac{\pi}{2}\,\delta\Big), \qquad (4)$$

where $W = [\omega_1, \ldots, \omega_L]^T$ denotes the chosen frequency set (a design choice) and $\delta \in \mathbb{Z}_2^{L\times N}$ is a fixed phase matrix with entries $\delta_{i,j} = (i+j) \bmod 2$.

To understand why X_C improves robustness to frequency choices, consider the j-th sinusoidal channel of Eq. 4: $\mathrm{PE}_j = \sin(\omega_j(X_G + X_C) + \frac{\pi}{2}\delta_j)$. A simple sensitivity analysis gives

$$\frac{\partial \mathrm{PE}_j}{\partial \omega_j} = \Big(X_G + X_C + \omega_j \tfrac{\partial X_C}{\partial \omega_j}\Big) \cos\big(\omega_j(X_G + X_C) + \tfrac{\pi}{2}\delta_j\big). \qquad (237)$$

Here X_C does not take ω_j as an explicit input; its dependence on ω_j is implicit through training, since different frequency sets lead to different learned refinements. This post-training sensitivity analysis (see Supplementary Sec. A.3 for the full derivation) highlights two terms: a static geometric component $(X_G + X_C)$ and an adaptive correction $\omega_j\, \partial X_C/\partial \omega_j$. The latter enables LOOPE to reshape its effective positional geometry under different frequency sets, explaining its empirical frequency robustness (Table 6).

Equation 4 summarizes the essence of LOOPE: a **stable static order** enriched by a **learnable, content-adaptive bias**, producing a lightweight yet flexible encoding that preserves geometry while reducing design sensitivity.

3.3. Positional Embeddings Structural Integrity (PESI) Metrics

PESI is designed as a set of structural probes that assess how well a positional embedding maintains radial and directional monotonicity, and how much radial symmetry it sacrifices to achieve higher precision.

Given a positional embedding tensor $P \in \mathbb{R}^{h\times w\times D}$, define the *cosine similarity matrix* centered at (x, y):

$$E_{(x,y)}(i,j) = \frac{P_{(x,y)} \cdot P_{(i,j)}}{\|P_{(x,y)}\|\|P_{(i,j)}\|}. \qquad (5)$$

Undirected monotonicity. The *radial average similarity function* is:

$$\mu_{(x,y)}(r) = \frac{1}{|B_r|} \sum_{(i,j)\in B_r} E_{(x,y)}(i,j) \qquad (6)$$

where B_r is the set of positions at radius r. We compute *Spearman's rank correlation*:

$$\rho_{(x,y)} = 1 - \frac{6\sum_r d_r^2}{|R|(|R|^2-1)} \qquad (7)$$

where d_r is the rank difference between $level(r)$ and $\mu_{(x,y)}(r)$, and $|R|$ is the total number of radial levels. The undirected monotonicity score is then

$$M_u = \frac{1}{hw} \sum_{x=0}^{h-1} \sum_{y=0}^{w-1} \big(1 - \rho_{(x,y)}\big), \qquad (8)$$

where a higher M_U indicates stronger undirected monotonicity across the grid. Ideally, it should approach 2.

Directed monotonicity. We quantize 2π into $N = \frac{2\pi}{\delta}$ directional buckets (with quantization angle δ). For each cell (i, j) relative to (x, y), compute:

$$\theta_{(x,y)}(i,j) = \operatorname{atan2}(j - y,\, i - x) \tag{9}$$

where $\operatorname{atan2}(y, x)$ returns the angle in $(-\pi, \pi]$. Assign the cell to bucket

$$k = \left\lfloor \frac{\theta_{(x,y)}(i,j)}{\delta} \right\rfloor \mod N. \tag{10}$$

Within each bucket k, order the cells by radial distance and compute Spearman's rank correlation:

$$\rho^{k}_{(x,y)} = 1 - \frac{6 \sum_r d_r^2}{|R_k|(|R_k|^2 - 1)}, \tag{11}$$

where d_r is the rank difference between the radius r and the corresponding similarity values in bucket k, and $|R_k|$ is the number of elements in bucket k. The mean correlation per cell is then

$$\bar{\rho}_{(x,y)} = \frac{1}{N} \sum_{k=0}^{N-1} \rho^{k}_{(x,y)} \tag{12}$$

and the global directed monotonicity measure is defined as

$$M_D = \frac{1}{hw} \sum_{x=0}^{h-1} \sum_{y=0}^{w-1} \big(1 - \bar{\rho}_{(x,y)}\big). \tag{13}$$

With varying N and δ, we can investigate how precisely the encoder maintains monotonicity along different directions. As $N \to 1$, M_D reduces to the undirected measure M_U. A higher M_D indicates stronger directional monotonicity; ideally it approaches 2.

Undirected asymmetry. For each center (x, y), let B_r be the set of cells at radius r from (x, y). We define the standard deviation of the cosine similarity values as

$$\sigma_{(x,y)}(r) = \sqrt{\frac{1}{|B_r|} \sum_{(i,j) \in B_r} \big(E_{(x,y)}(i,j) - \mu_{(x,y)}(r)\big)^2}. \tag{14}$$

The coefficient of variation at radius r is then

$$\mathrm{CV}_{(x,y)}(r) = \frac{\sigma_{(x,y)}(r)}{\mu_{(x,y)}(r)}. \tag{15}$$

Averaging over all radial distances $r \in R$ yields the undirected asymmetry measure at (x, y):

$$A'_{SU}(x,y) = \frac{1}{|R|} \sum_{r \in R} \mathrm{CV}_{(x,y)}(r), \tag{16}$$

and the global undirected asymmetry is defined as

$$A_{SU} = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} A'_{SU}(x,y). \tag{17}$$

For complete symmetry, $|A_{SU}| \to 0$. In practice, there is no single ideal value of undirected asymmetry, since many encoders increase A_{SU} to achieve sharper directional monotonicity M_D; if $A_{SU} = 0$, there is no directional information in the embedding. Detailed algorithms are provided in *Supp. A.5.1, A.5.2, A.5.3*.

4. Experiments

4.1. Experimental Setup

All experiments were conducted on a single NVIDIA RTX 5090 GPU (32 GB VRAM). Unless specified otherwise, we follow the standard training protocol used in prior ViT literature to ensure fair comparison with existing positional encoders.

We train all models for 150 epochs using the Adam optimizer with a cosine learning-rate schedule (max LR = 1×10^{-3}, min LR = 2.5×10^{-5}) and Cross-Entropy loss. Batch sizes are 96 for Oxford-IIIT [20] and 64 for CIFAR-100 [17], ImageNet-1k [9], and our Three-Cell probing dataset. For the resolution-scaling experiment, Oxford-IIIT images are trained at 384×384 with batch size 32. CrossViT models use 240×240 inputs with mixed patch sizes (12×12 and 16×16), following the original implementation.

To ensure a strictly controlled comparison across positional encodings, all models—including baseline APE/RPE variants—use the same ImageNet-1K pretrained backbones (DeiT-Base unless otherwise stated) while PEs are trained from scratch. This removes differences due to random initialization and isolates the contribution of the positional encoding itself, which is the primary quantity under investigation.

We adopt an 80-10-10 train–validation–test split for all datasets and apply standard augmentations (horizontal flips, rotations, brightness jitter, and elastic deformations) consistently across all PE variants.

ViT-Base contains **85.8M** parameters. LOOPE increases this to only **85.9M** (a **0.12%** overhead), confirming that the proposed content-aware patch-ordering mechanism adds negligible computational cost while providing significant gains in positional fidelity and downstream accuracy.

4.2. Experiments on Image Classification

4.2.1. Comparison against 1-D Positional Encoders

From Table 1, LOOPE achieves the highest accuracy across all models on Oxford-IIIT, most notably with ViT, indicating enhanced fine-grained feature learning. CIFAR-100 presents greater inter-class variability, still our PE outperforms all other PEs, especially with DeiT-Small. Finally, for ImageNet-1k, LOOPE again surpasses all PEs across all models. These results demonstrate that it effectively bal-

Dataset	Model	No PE	Learnable	Sinusoid	Static (X_G)	LOOPE ($X_G + X_c$)
Oxford-IIIT	ViT-Base [10]	83.6%	84.6%	85.3%	84.2%	**88.1%**
	DeiT-Base [32]	88.9%	89.4%	86.3%	89.0%	**89.8%**
	DeiT-Small [32]	83.8%	83.8%	83.7%	80.6%	**84.5%**
	CaiT [33]	87.4%	89.0%	90.0%	89.6%	**90.5%**
	Cross-ViT [2]	88.3%	90.9%	88.0%	89.3%	**91.0%**
CIFAR-100	ViT-Base [10]	79.8%	83.0%	85.2%	87.6%	**88.3%**
	DeiT-Base [32]	82.1%	86.3%	86.6%	86.9%	**87.1%**
	DeiT-Small [32]	68.6%	81.6%	71.9%	77.7%	**82.0%**
	CaiT [33]	77.3%	82.5%	82.3%	82.5%	**83.1%**
	Cross-ViT [2]	80.5%	84.6%	86.3%	85.3%	**86.8%**
ImageNet-1k	ViT-Base [10]	82.7%	83.6%	84.4%	84.5%	**84.7%**
	DeiT-Base [32]	81.7%	82.1%	82.0%	83.1%	**83.5%**
	DeiT-Small [32]	80.5%	82.2%	82.8%	82.8%	**83.0%**
	CaiT [33]	80.3%	81.6%	81.6%	82.8%	**83.1%**
	Cross-ViT [2]	82.4%	83.5%	83.1%	84.5%	**84.8%**

Table 1. Accuracy comparison of different ViT models with various PEs across Oxford-IIIT, CIFAR-100 and ImageNet-1k datasets.

ances structured spatial encoding with learnable adaptability, making it a robust solution.

4.2.2. More comparisons with Positional Encoders

PE	ImageNet-1k	Oxford-IIIT	CIFAR-100
CPE [5]	82.7%	83.9%	79.1%
RPE [36]	80.3%	80.5%	79.2%
LFF(Fourier)[18]	83.4%	**90.5%**	**89.1%**
2D Sinusoid [39]	83.2%	80.1%	86.3%
AS2DRoPE [6]	82.9%	88.7%	86.4%
Static (X_G) [13]	83.1%	89.0%	86.9%
LOOPE ($X_G + X_c$)	**83.5%**	89.8%	87.1%

Table 2. Accuracy comparison of LOOPE against other advanced PEs on DeiT-Base across Oxford-IIIT, CIFAR-100 and ImageNet-1k datasets

Table. 2 compares LOOPE against advanced PEs. Fourier PE achieves the highest accuracy for Oxford IIT and CIFAR-100, due to its rich frequency encoding properties. But, LOOPE outperforms Fourier in ImageNet-1k. These results validate the superiority of Fourier PE while demonstrating that LOOPE remains a strong alternative for vision tasks.

4.3. Experiments on Semantic Segmentation

To test the impact of LOOPE on downstream tasks, We used the Cityscapes [7] dataset for its strong sensitivity to positional information. SETR [38] and Segmenter [28] were chosen as base models for their state-of-the-art performance and exclusive use of ViT backbones, making them PE-sensitive. We excluded other SOTA models like Mask2Former [4] due to extra attention modules and custom loss, and SAM-2 [16] for its prompt encoder with cross-attention. All experiments were conducted with a batch size of 64 and trained for 100 epochs. We used a learning rate of 0.0001 and an 80-20 train-validation split.

Model	PE	Acc	mIoU	mDice	Boundary F1	MCC
SETR [38]	No PE	80.84%	79.56%	84.53%	3.48%	88.75%
	1D Sinusoid [25]	80.96%	78.33%	82.51%	3.55%	**91.99%**
	Learnable [25]	81.35%	74.41%	79.26%	3.88%	88.87%
	CPE [5]	80.67%	78.42%	82.71%	3.31%	87.26%
	LFF(Fourier) [18]	82.69%	80.41%	85.20%	**7.83%**	85.64%
	2D Sinusoid [39]	80.74%	78.26%	82.46%	3.04%	91.88%
	LOOPE	**83.63%**	**81.37%**	**86.17%**	7.44%	89.56%
Segmenter [28]	No PE	79.55%	79.13%	84.48%	6.23%	83.34%
	1D Sinusoid [25]	81.82%	80.59%	85.79%	18.37%	84.74%
	Learnable [25]	81.48%	80.46%	85.73%	15.33%	84.68%
	CPE [5]	79.19%	78.71%	84.00%	8.80%	82.89%
	LFF(Fourier) [18]	83.20%	81.41%	86.46%	26.60%	85.51%
	2D Sinusoid [39]	80.93%	79.90%	85.13%	9.58%	84.08%
	LOOPE	**84.05%**	**82.26%**	**87.31%**	**32.21%**	**86.37%**

Table 3. Performance comparison of PEs with SETR and Segmenter models on the Cityscapes dataset.

From Table 3, we observe that LOOPE outperforms all PEs by learning patch order, with LFF (Fourier) closely behind — surpassing LOOPE only once in Boundary F1 for SETR — and 1D Sinusoid outperforming LOOPE in MCC.

4.4. Three-Cell Experiment: Positional Probing Task

Standard vision benchmarks often show only modest gains from positional encodings (PEs), because natural images exhibit strong correlations between neighboring patches and transformers can infer rough spatial structure from content alone. This makes it difficult to isolate how much positional information different encoders truly preserve, and in particular to compare absolute PEs (APEs) with task-oriented relative PEs (RPEs). To obtain a more controlled view, we design a synthetic probing task where labels depend purely on geometry and cannot be solved from appearance.

Setup. Each 224×224 image is divided into a 14×14 grid of 16×16 patches. Three patches are randomly selected and colored red, green, and blue at coordinates $(x_r, y_r), (x_g, y_g), (x_b, y_b)$, while all other patches contain color black. The model receives only these images and must answer geometric questions about the positions of the three colored cells. We consider four query types (Fig. 4) and cast them jointly as a 6-way classification problem. Full formulas and the sampling procedure are provided in *Supp. A.2*.

Geometric queries. The four query types are designed to probe complementary aspects of positional structure: (i) *distance comparison* between d_{RG} and d_{RB}, testing undirected monotonicity; (ii) *orientation* (clockwise vs. counterclockwise) of the RGB triangle, testing signed relational structure; (iii) *shadow area comparison* under RG vs. RB,

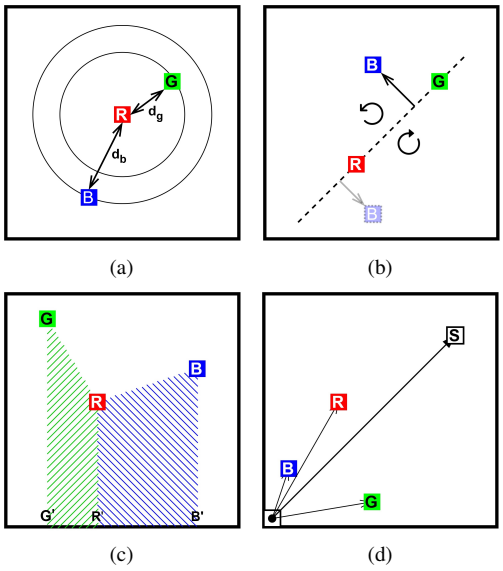


Figure 4. Three-Cell probing task. Given three colored patches (R,G,B), the model must: (a) decide whether d_{RG} or d_{RB} is larger; (b) predict the orientation of triangle RGB (clockwise vs. counter-clockwise); (c) compare "shadow areas" under lines RG and RB; (d) decide whether the sum vector $OR + OG + OB$ lies outside the grid. Labels depend purely on the coordinates of R, G, and B.

which requires combining offsets via addition; and (iv) *vector sum*, predicting whether $\mathbf{p}_r + \mathbf{p}_g + \mathbf{p}_b$ leaves the 14×14 grid, explicitly testing additive composition of coordinates. While the first two can in principle be derived from relative differences, the latter two are fundamentally difficult to reconstruct from purely RPE-based formulations.

	Relative Spatial Reasoning		Absolute Spatial Reasoning		
Models	Distance	Orientation	Area	Vector Sum	Average
ResNet-50[11]	90.8%	85.6%	89.7%	92.9%	89.8%
Inception-V3[30]	92.8%	96.1%	93.7%	95.3%	94.5%
No PE	61.9%	53.1%	60.1%	62.6%	59.4%
Learnable [1]	85.6%	89.3%	84.0%	92.2%	87.8%
1D Sinusoid [1]	90.9%	94.9%	91.3%	**95.3%**	93.1%
CPE [5]	72.8%	63.2%	72.6%	75.0%	70.9%
RPE [36]	73.7%	62.6%	71.0%	73.9%	70.3%
LFF(Fourier) [18]	93.1%	**96.7%**	92.4%	93.9%	94.0%
2D Sinusoid [18]	83.6%	60.0%	72.6%	92.3%	77.1%
Static (X_G) [13]	88.8%	95.0%	89.9%	93.8%	91.9%
LOOPE ($X_G + X_c$)	**93.4%**	95.8%	**93.3%**	94.6%	**94.3%**

Table 4. Accuracy comparison of baseline CNN models and PEs with DeiT-Base model on the Three-Cell Experiment dataset.

Findings. Table 4 shows three consistent trends. First, RPEs clearly outperform the no-PE baseline but remain close to chance on the area and vector-sum tasks, confirming that they struggle to encode additive coordinate structure. Second, APEs (sinusoidal, LFF) achieve much higher accuracy across all queries, indicating that absolute coordinates remain crucial even in modern ViTs. Third, LOOPE yields the highest geometric fidelity, suggesting that learning a content-adaptive patch ordering further improves how absolute PEs expose positional information to the transformer.

Overall, this probing task supports our central claim: rather than relying solely on task-oriented RPEs, developing stronger APEs—and in particular learnable patch orderings such as LOOPE—remains a promising direction for positional encoding research.

4.5. Positional Embedding Structural Integrity (PESI) Metrics

PE	Undirected Monotonicity $M_U \uparrow$	Directed Monotonicity $M_D \uparrow$	Undirected Asymmetry A_{SU}
Learnable [1]	1.7493	1.2003	-0.7272
1D Sinusoid [1]	1.9567	1.4905	0.1243
LFF(Fourier) [18]	1.9623	**1.5230**	0.2683
Static (X_G) [13]	1.9670	1.2897	0.0945
LOOPE ($X_G + X_c$)	**1.9674**	1.2900	**0.0939**

Table 5. Comparison of PEs in terms of Undirected Monotonicity, Directed Monotonicity, and Undirected Asymmetry.

Table. 5 presents the positional fidelity indices for various APEs. For calculating directed monotonicity, the number of buckets, N is set to 60 testing. So, the $\delta = 6^o$ The results indicate that LOOPE achieves the highest values in both **Undirected Monotonicity** and **Undirected Asymmetry**, demonstrating its robustness. Conversely, Learnable APE performs the worst across all three metrics, indicating that its embeddings are not highly monotone. A notable observation is the asymmetry value of Learnable, which is -0.72. This negative value arises because the average cosine similarity across all cells is predominantly negative, leading to an overall asymmetry value below zero. Meanwhile, Fourier exhibits strong **Directed Monotonicity** with stable results in the undirected setting. However, it compromises radial symmetry, meaning that values on a single radius show greater instability compared to other periodic APEs. In contrast, LOOPE demonstrates the most stable radial symmetry, reinforcing its reliability in positional encoding.

4.6. Ablation Studies

4.6.1. Robustness across varying Frequency Set

The results in Table 6 demonstrate that **LOOPE** consistently outperforms other PEs, even when using a non-optimal frequency set. Notably, there is no theoretically proven optimal frequency configuration. For instance, a simple geometric sequence achieves better accuracy than the original frequency set across all PEs. **This analysis clearly highlights the robustness of**

PE / Freq.	$\omega(i)$	1D Sinusoid	Static (X_G)	LOOPE ($X_G + X_c$)
Original	0.978^i	85.3%	84.6%	**88.1%**
Arithmetic	$1 - \frac{i(1-\lambda)}{L-1}$	73.6%	76.5%	**86.6%**
Geometric	$r^i,\ r = 0.9$	89.8%	87.3%	**89.9%**
Random	Uniform $(\lambda, 1)$	83.5%	87.5%	**88.0%**
Sensitivity to Freq. ↓		0.027760	0.022516	**0.005244**

Table 6. Robustness of LOOPE across different frequency sequences ($\lambda = 0.0001$, $L = 768$, $i \in 1, \ldots, L$) with ViT-Base as backbone on Oxford-IIIT Dataset. For Sensitivity computation, refer to *Supp.A.6.*

LOOPE—particularly in the case of the arithmetic sequence, where accuracy drops to 73.6%, yet LOOPE alone helps maintain it at 86.6%. This experiment indicates that **LOOPE** effectively mitigates frequency selection bias.

4.6.2. Impact of LOOPE on Variable Resolution

Resolution	Models	1D Sinusoid	LOOPE (Ours)
224×224	ViT-Base [10]	85.3%	88.1% (+2.8%)
	ViT-Small [10]	81.6%	83.8% (+2.2%)
	DeiT-Base [32]	86.3%	89.8% (**+3.5%**)
384×384	ViT-Base [10]	89.1%	92.2% (+3.1%)
	ViT-Small [10]	83.0%	86.1% (+3.1%)
	DeiT-Base [32]	88.5%	92.4% (**+3.9%**)

Table 7. Accuracy Comparison of different ViT models with Sinusoid and LOOPE PEs for different Image Resolutions on Oxford-IIIT. For both cases, patch shape is 16×16.

Table. 7 presents the performance of different ViT models with Sinusoid and LOOPE PEs across two different resolutions. Our method shows higher improvement in accuracy with bigger resolution.

4.6.3. Visualization of Positional Encodings

Figure: 5 clearly shows that LOOPE is generating more robust cosine similarity map for all positions. Due to traditional zigzag order in sinusoidal APE, the boundary and corner cells have inconsistent similarity pattern, as the order propagates from right to all way back to left boundary. For 1D learnable APE, it loosely generates map with a lots of anomaly in near to far distance. Also, other than, central cells, edge cells are having non-monotone similarity map. More visualization can be found in (*Supp. B*)

4.6.4. Trends in PESI metrics

Figure 6, shows the interesting facts about directed monotonicity, with this tool one can investigate how precisely the positional embeddings can maintain monotonicity. Clearly, increasing precision all positional encoders struggles to provide a monotone trend in cosine similarity. As $N \rightarrow 1, M_D \rightarrow M_U$ which is exactly what we expected. At

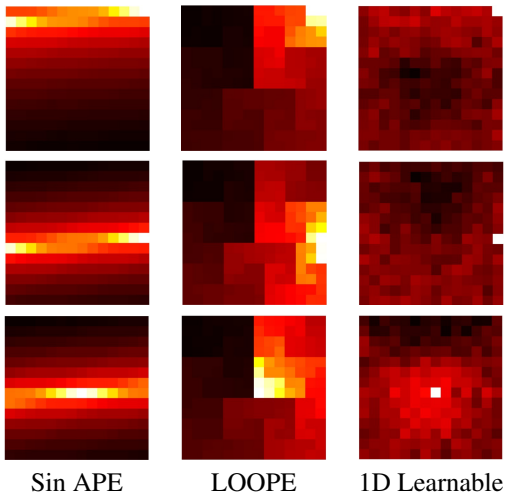


Figure 5. Cosine Similarity Maps for Three APEs: Top-Right Corner, Right-Boundary, and Middle Cell (Top to Bottom)

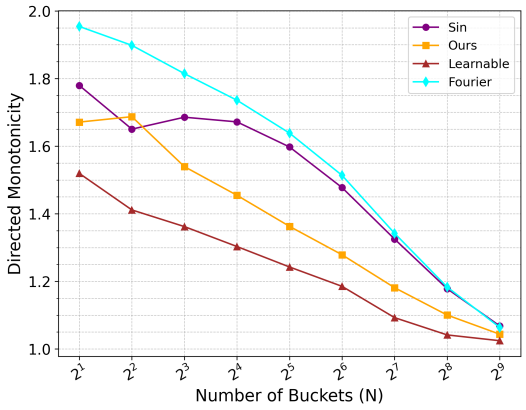


Figure 6. Trend of directed Monotonicity, M_D with increasing angle precision, $\delta = 2\pi/N$

$N = 4, \delta = \pi/2$, we can see that, LOOPE outperforms zigzag and learnable as it highly depends on hilbert order which propagates in square pattern.

5. Conclusion

We introduced **LOOPE**, a learnable framework for patch ordering that unifies geometric structure and context-aware adaptation, establishing reordering as a key design dimension with preserved geometry and frequency robustness. In addition, we proposed the **Three-Cell experiment** and **PESI metrics** as principled tools for evaluating PEs beyond downstream accuracy. This perspective opens new directions for rethinking positional information in transformers.

References

[1] Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017. 1, 2, 7

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2, 6

[3] Wanli Chen, Xufeng Yao, Xinyun Zhang, and Bei Yu. Efficient deep space filling curve. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17525–17534, 2023. 3

[4] Bowen Cheng, Alexander Schwing, and Alexander Kirillov. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 2, 6

[5] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 6, 7

[6] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified LLaMA backbone for vision tasks. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 6

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[10] 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. 2, 6, 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 7

[12] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer, 2024. 2

[13] David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. *Mathematische Annalen*, 38:459–460, 1891. 2, 6, 7

[14] Yun-Ning Hung, Ting-Yun Chang, Hsin-Ying Lee, Min Sun, and Sylvain Paris. Learnable fourier features for multi-dimensional spatial positional encoding, 2023. 2

[15] G Ke, D He, and TY Liu. Rethinking positional encoding in language pre-training. arxiv. *arXiv preprint arXiv:2006.15595*, 2021. 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5

[18] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021. 6, 7

[19] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2

[20] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V Jawahar. Cats and dogs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[21] Giuseppe Peano. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160, 1890. 2

[22] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 1

[23] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 2

[24] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 2

[25] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 1, 6

[26] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018. 2

[27] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021. 2

[28] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021. 2, 6

[29] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 7

[31] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 1

[32] 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. 2, 6, 8

[33] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 6

[34] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020. 2

[35] Hanyu Wang, Kamal Gupta, Larry Davis, and Abhinav Shrivastava. Neural space-filling curves, 2022. 3

[36] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. *CoRR*, abs/2107.14222, 2021. 2, 6, 7

[37] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13569–13578, 2021. 1

[38] Sixiao Zheng, Jiacheng Li, Yuandong Tian, Hengyu Cai, Ting Liu, Zhenguo Li, and Nanning Zheng. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2020. 2, 6

[39] Shengcao Zhou, Ziwei Liu, Xiaohang Zhan, Chen Change Loy, and Ping Luo. When vision transformers outperform resnets without pretraining or strong data augmentations, 2022. 6

[40] Jakub Červený. gilbert: Space-filling curve for rectangular domains of arbitrary size. https://github.com/jakubcerveny/gilbert. 3