# Fast Dynamic IR-Drop Prediction with Dual-path Spatial-Temporal Attention

Bangqi Fu, Lixin Liu, Qijing Wang, Yutao Wang, Martin D.F. Wong, Evangeline F.Y. Young
Department of Computer Science and Engineering, The Chinese University of Hong Kong
{bqfu21,lxliu,qjwang21,ytwang1,mdfwong,fyyoung}@cse.cuhk.edu.hk

***Abstract*—The analysis of IR-drop stands as a fundamental step in optimizing the power distribution network (PDN), and subsequently influences the design performance. However, traditional IR-drop analysis using commercial tools proves to be exceedingly time-consuming. Fast and accurate IR-drop analysis is desperately in demand to achieve high performance on timing and power. Recently, machine learning approaches have garnered attention owing to their remarkable speed and extensibility in IC designs. However, prior works for dynamic IR-drop prediction presented limited performance since they did not exploit the time-varying activities. In this paper, we proposed a dual-path model with spatial-temporal transformers to extract the static spatial features and dynamic time-variant activities for dynamic IR drop prediction. Experimental results on the large-scale advanced dataset CircuitNet show that our model significantly outperforms the state-of-the-art works.**

***Index Terms*—IR drop, Power, Machine Learning, Physical Synthesis**

## I. Introduction

IR drop analysis is a crucial step in the integrated circuits (ICs) design in order to improve the design performance like power and timing. This analysis specifically focuses on the voltage drop occurring within the power delivery networks (PDNs) when current flows through the power grids. Mitigation of IR drop is essential to ensure reliable power distribution and avoid performance degradation in ICs.

As illustrated in Fig. 1, the wires on PDNs are modeled as segments of resistors. When current flows from the power source to standard cells, voltage drops along these resistor-modeled wires, leading to a reduced working voltage at the standard cell.

IR drop must be restricted under constraints to achieve timing closure and correct functionality of a circuit. The circuit design process consists of stages from initial placement to final signoff, during which IR drop is continually estimated and optimized to prevent performance degradation.

However, the IR drop analysis algorithms become increasingly complex and time-consuming as the design scale grows, since they solve a large number of linear equations. Thus, accurate and fast IR drop analysis is essential for IC design. The work [1] attempts to accelerate the IR drop analysis by employing a fast algorithm that separates the traversal of power grids. While it shows a significant speedup compared to traditional algorithms, the runtime cost remains exceedingly expensive.

Machine learning approaches have demonstrated strong potential to accelerate the IR drop analysis with reliable accuracy throughout design cycles. Many works have been studied in recent years for IR drop prediction.

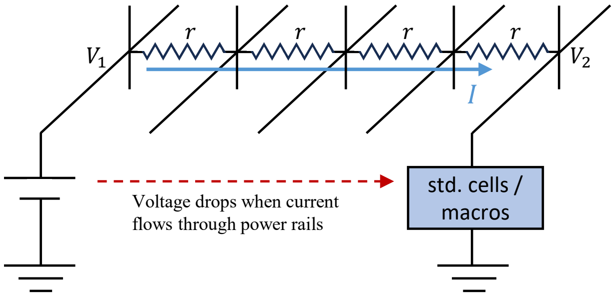


Fig. 1: An illustration of IR drop in a circuit.

The work [2] proposed a framework using the Encoder-Decoder structure model for design IR drop prediction. [3] presented a decomposition of power maps to obtain the time-varying values of power grids for dynamic IR drop prediction. [4] introduced a 3D convolution network to extract the temporal features for dynamic IR drop prediction. [5] implemented a Recurrent model by connecting a series of U-Net models in series for dynamic IR drop prediction. [6] proposed a spatial-attention-gated U-Net to better capture the contextual information in the decoder. [7] proposed a framework with circuit PDN structure information. [8] proposed a Graph Neural-Network (GNN) enhanced model with PDN structure information.

Previous works primarily focus on spatial feature representation but lack consideration for the temporal information in dynamic IR prediction tasks. The [3] takes the maximum of time frames, the [4] conducts a local 3D convolution on neighboring time frames, whereas the [5] merges the hidden information from the previous time frames sequentially. The aforementioned works only aggregate power map features through local neighbor time-frames. However, dynamic IR drop prediction requires capturing power activities from a global temporal perspective

To tackle the limits of the prior works, we present our dual-path spatial-temporal model which extracts time-varying power features. Our main contribution can be summarized as follows:

- We propose a dual-path model incorporating spatial and temporal transformer blocks, leveraging self-attention across the dynamic power feature maps.
- We employ a multi-scale hierarchical encoder and 2D/3D shifting window to effectively extract the long-range in-

formation across the feature maps.

- We adopt a multi-level decoder with fused feature maps to recover the IR drop hotspot prediction.
- We evaluated our model on large-scale advanced datasets and demonstrated better quality compared to the state-of-the-art works.

## II. Preliminaries

### A. Overview of Dynamic IR Drop Prediction

Dynamic IR drop takes simulation patterns as input and estimates the power demand of standard cells caused by switching activities and current fluctuations [9]. The IR drop analysis plays a critical role in VLSI design to mitigate the impact of IR drop on the design stability and reliability. However, the IR drop analysis requires substantial computation and simulation and thus is very time-consuming. This results in heavy runtime costs in the design cycles. Thus, an accurate and fast IR drop analysis is in great demand in the early stages of the design to reduce the turnaround time of a circuit.

Recently, Machine Learning (ML) based methods for IR drop prediction have been widely studied. The ML methods leverage power-related features as inputs and predict the IR drop hotspot. In dynamic IR drop prediction, the power-related features include cell internal power, switching power, leakage power, and toggle rates [3]. The clock period is decomposed into even timing windows and each type of power map is reported in a timing window. The overall power reports and the timing window decomposed power reports provide abundant information in both spatial and temporal domains.

In conclusion, the design layout is split into uniform grid tiles, and the cell power information is accumulated into the corresponding tiles. The features power maps can be categorized into:

- $power_i = p_i$
- $power_s = p_s$
- $power_{all} = p_i + p_s + p_l$
- $power_{sca} = (p_i + p_s) \times r_{tog} + p_l$
- $power_t[0, ..., T-1]$

where the $p_i, p_s, p_l, r_{tog}$ are the overall cell internal, switching, leakage power, and toggles rate. The $power_i$ and $power_s$ are power maps derived from internal and switching power, the $power_{all}$ is the overall power, and the $power_{sca}$ is the toggle rate scaled power. The $power_t$ is the toggle rate scaled power map at a certain timing window.

Current methods for dynamic IR drop prediction can be categorized into 3 types:

- Treat dynamic power maps as spatial feature channels and perform 2D Convolution. [3]
- Expand the feature dimension and treat dynamic power maps as temporal channels to perform 3D Convolution. [4], [8]
- Conduct sequential models with power maps in neighboring timing windows. [5]

However, prior works have not fully exploited the dynamic activities across the timing windows and thus have limited performance in dynamic IR drop prediction

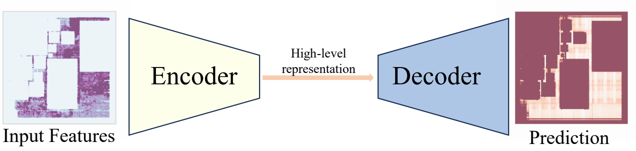


Fig. 2: The structures of encoder-decoder model for the image prediction task.

### B. Problem Formulation

Given the image-based input feature maps $x_i \in \mathbb{R}^{H\times W\times C}$ with $C$ channels of $H \times W$ size feature map, we aim to perform an image-to-image mapping $f : \mathbb{R}^{H\times W\times C} \Rightarrow \mathbb{R}^{H\times W\times 1}$ that minimize the Mean Absolute Error (MAE) between the predicted IR drop maps and ground-truth IR drop maps:

$$\mathcal{L}_1(\mathbf{y_i}, f(\mathbf{x_i})) = ||f(\mathbf{x_i}) - \mathbf{y_i}||_1 \tag{1}$$

where $\mathbf{x}_i, \mathbf{y}_i, f$ denote the $i$-th input feature map, ground-truth IR drop map, and the model.

### C. Encoder-decoder Model

The encoder-decoder structure is widely utilized in ML tasks. It is acknowledged for its effectiveness in image-to-image tasks, preserving high-dimensional representation and reconstructing images with the original information through a sequence of decoding layers.

The encoder captures high-dimensional spatial information and reduces the image size via downsampling convolution blocks and max pooling layers.

Conversely, in the decoder, the compressed information is upsampled through convolution blocks to recover the original input size.

To mitigate the loss of spatial information during the decoding stage, models such as U-Net [10] employ a skip-connection by concatenating the encoded feature maps with their corresponding decoded feature maps to preserve critical spatial details.

## III. Methods

Our proposed model addresses the challenge of dynamic IR drop prediction by leveraging a dual-path Vision Transformer (ViT) based architecture [11] which extracts multi-scale spatial-temporal information so that the influence of timing-variant switching activity can be better captured. A shifting window is adopted to perceive the features in a global sense. The dual-path features are fused and decoded to predict the IR drop hotspot.

In the following sections, we will present the ViT-based multi-scale feature extraction, the shifted window-based transformer, the overall architecture of the dual-path spatial-temporal model, and the hybrid feature decoder.

### A. Multi-scale Feature Encoder

Features with different scales involve abundant information from global, low-frequency components to local, high-frequency contexts. To effectively capture the multi-scale information, we adopt the Swin Transformer architecture as the encoder [12], [13].

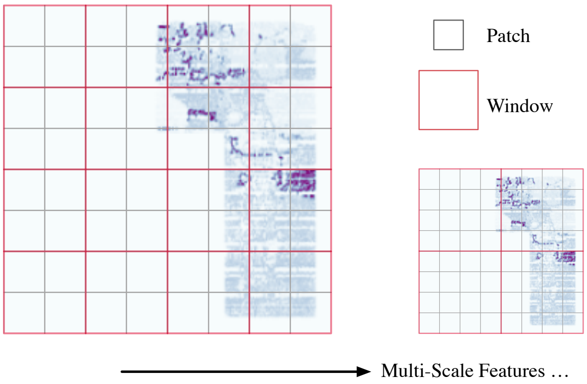


Fig. 3: An illustration of multi-scale transformer. Input feature map is split into patches and a local window performs self-attention on the patches.

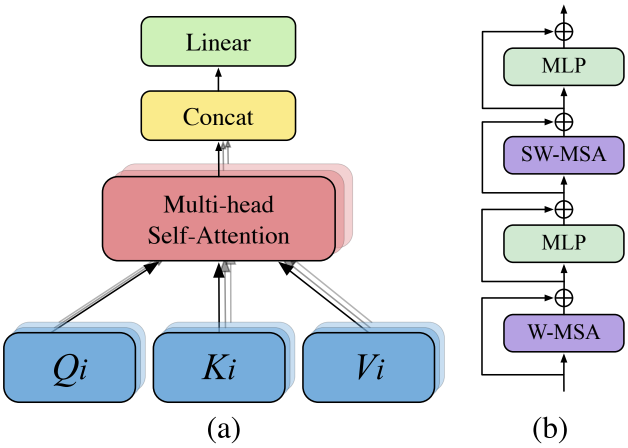


Fig. 4: (a) An illustration of Multi-head Self Attention block. (b) The structure of a transformer block.

The encoder first splits the input feature maps into small patches as a sequence of tokens [11], as illustrated in Fig. 3. These tokens are projected into high-dimension space as hidden embeddings. The feature maps are then partitioned into local windows to perform Multi-Head Self-Attention (MSA) on a group of patches in the window. The self-attention is conducted on the entire sequence of tokens, allowing the model to capture the global dependencies between input and output across the entire map, which traditional CNNs cannot achieve. The Multi-Head Self-Attention consists of 3 linear transform layers and an attention layer. The 3 linear transforms correspond to 3 input sequences: Query(Q), Key(K) and Value(V), which are projected into multiple heads as illustrated in Fig. 4(a). These heads will conduct self-attention in parallel. A self-attention layer can be formulated as:

$$\begin{aligned} \text{Attention}(Q_i, K_i, V_i) &= \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \\ \text{MultiHead}(Q, K, V) &= \text{Concat}(head_1, ..., head_h) \\ head_i &= \text{Attention}(Q_i, K_i, V_i) \end{aligned} \tag{2}$$

where $d_k$ is the normalization factor and $i$ is the head number.

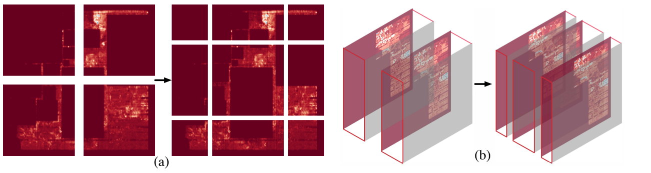


Fig. 5: An illustration of (a) 2D Shifted Window. (b) 3D Shifted Window on the temporal direction.

A hierarchical architecture enables feature extraction in various scales across different stages of the encoder. The patches will gradually merge with neighboring patches to scale down the feature map size and increase the number of feature channels. By applying the hierarchical structure, a window is able to capture multi-scale feature representation, with abundant local and global image information.

To further enhance global representation, the Swin Transformer employs a Shifted Window-based MSA (SW-MSA) scheme, facilitating connections across windows. At each stage of the transformer, the feature map is partitioned into non-overlapping windows. Each window contains a group of patches to perform self-attention and the windows will shift between consecutive layers as illustrated in Fig. 5(a).

A transformer block consists of 2 consecutive MSA layers with shifted windows as illustrated in Fig. 4(b). After each stage of the transformer block, a Multi-layer Linear Perceptron (MLP) serves to refine and transform the generated contextualized embeddings in MSA layers.

### B. 3D Shifted Window based MSA

The dynamic IR drop prediction takes power maps of different time frames $power_t[0, ..., T-1]$ as input feature maps. To capture the dynamic activities across time, 3D convolution [4] has been applied to aggregate the information of neighboring time frames. The 3D convolution makes a local region aware of its neighbor's activities so that the output can be predicted based on the timing variance. However, the 3D convolution only aggregates the local neighboring features of power maps and lacks a global view. To better extract the spatial-temporal information, we apply a 3D Shifted Window based MSA [14].

The 3D SW-MSA is similar to 2D SW-MSA, except that it performs patch embedding and local windowing in 3D space. The input temporal feature map of $T \times H \times W \times 1$ is lifted $T \times H \times W \times E$ first and split into $2 \times 4 \times 4$ patches, where $E$ is the number of hidden embeddings. Each 3D patch is regarded as a token in the patch sequence and MSA is performed on the whole sequence based on a 3D local window of $4 \times 8 \times 8$. After each MSA layer, the window is shifted by half of its size in the 3 directions so that the whole feature space can have interaction.

### C. Dual Path Spatial-Temporal Model

The overall architecture of our proposed model is shown in Fig. 6. The model aim to capture both the spatial power features and also the internal temporal activities of cell powers. Given

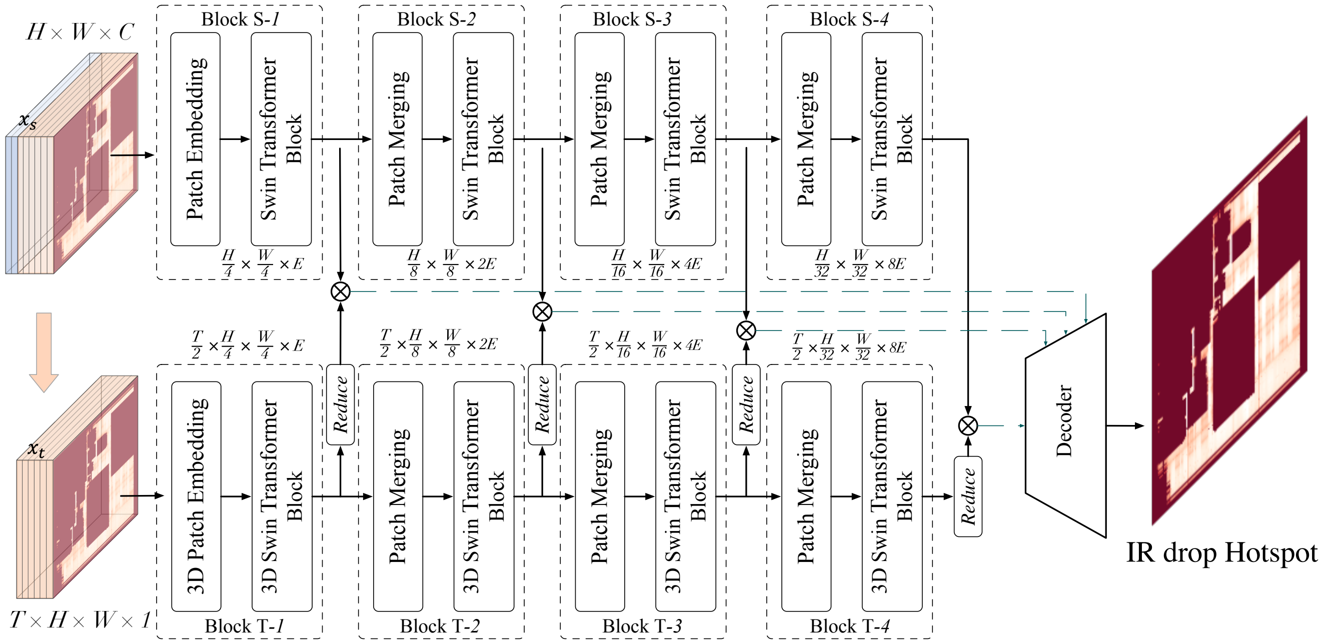

$H \times W \times C$
$x_s$
$T \times H \times W \times 1$
$x_t$
Block S-1
Patch Embedding
Swin Transformer Block
$\frac{H}{4} \times \frac{W}{4} \times E$
Block S-2
Patch Merging
Swin Transformer Block
$\frac{H}{8} \times \frac{W}{8} \times 2E$
Block S-3
Patch Merging
Swin Transformer Block
$\frac{H}{16} \times \frac{W}{16} \times 4E$
Block S-4
Patch Merging
Swin Transformer Block
$\frac{H}{32} \times \frac{W}{32} \times 8E$
$\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times E$
$\frac{T}{2} \times \frac{H}{8} \times \frac{W}{8} \times 2E$
$\frac{T}{2} \times \frac{H}{16} \times \frac{W}{16} \times 4E$
$\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8E$
Reduce
Reduce
Reduce
Reduce
3D Patch Embedding
3D Swin Transformer Block
Block T-1
Patch Merging
3D Swin Transformer Block
Block T-2
Patch Merging
3D Swin Transformer Block
Block T-3
Patch Merging
3D Swin Transformer Block
Block T-4
Decoder
IR drop Hotspot

Fig. 6: The overall architecture of our proposed model.

the input features maps $x \in \mathbb{R}^{H \times W \times C}$, where $x$ is the the concatenated power maps of $power_i$, $power_s$, $power_{all}$, $power_{sca}$ and $power_t[0, ..., T-1]$. We reconstruct and transpose the feature maps to obtain the 3D temporal maps $x_t \in \mathbb{R}^{T \times H \times W \times 1}$, which is derived from the $power_t[0, ..., T-1]$, where $T$ denotes the total number of time frames of power analysis. And we define the spatial feature $x_s = x$, which are the overall spatial feature maps.

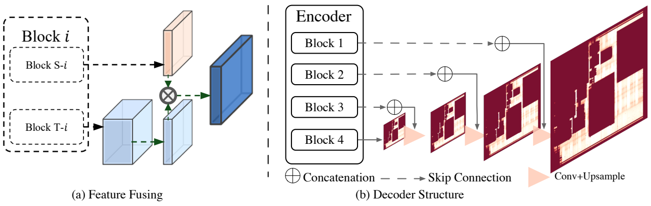
The spatial features $x_s$ and temporal features $x_t$ are fed into two paths to perform encoding. We first conduct the patch embedding for both spatial and temporal feature maps to partition the input features into $4 \times 4$ and $2 \times 4 \times 4$ patches correspondingly and lift the embedding channels into $E$.

The patches are flattened into embedded tokens and fed into the encoders, where self-attention is applied to extract semantic information. This is achieved using both 2D and 3D Swin Transformer blocks, which process the spatial and temporal dimensions respectively

After each Swin Transformer block, the patching merging aggregates local patches linearly so that the feature map resolutions are downsampled, allowing multi-scale feature representation. Note that since the time frame number $T$ is not a large number, we do not merge the temporal dimension after the 3D Swin Transformer blocks. $T$ is 20 in our framework.

In the spatial path, we perform MSA with 2D windows to aggregate 2D local features, whereas in the temporal path, the 3D windows bridge between different time frames so that the dynamic power activities can be aware.

We conduct 4 transformer blocks in both paths to obtain hierarchical multi-scale information, which is then recovered to the prediction IR-drop map in the decoder.



(a) Feature Fusing

(b) Decoder Structure

Fig. 7: The U-Net decoder with fusing channels.

### D. Hybrid Channel Decoder

To recover the original image size, we adopt the U-Net as the decoder. The U-Net progressively upsamples the high-dimension feature maps with the skip connection from different stages of encoders. To combine the multi-scale feature maps of the dual paths, we perform a feature fusing on the spatial path feature map $z_s$ and temporal path feature map $z_t$ as illustrated in Fig. 7(a):

$$\begin{aligned} \hat{z}^l &= z_s^l \cdot \text{Reduce}(z_t^l) \\ \text{Reduce}(z_t^l) &= \text{Conv3D}(z_t^l, \frac{T}{2}, 1) \end{aligned} \tag{3}$$

where $l$ is the block number and $z_s, z_t$ are the internal feature maps in spatial and temporal paths. The $\text{Conv3D}(\cdot, \frac{T}{2}, 1)$ is applied on the feature maps of size $\frac{T}{2} \times W^l \times H^l \times E^l$ to perform a linear transform and reduce the feature dimension to $1 \times W^l \times H^l \times E^l$ so that it can match the feature map resolution as the spatial path. The fused feature maps are upsampled and recovered in the decoder as illustrated in Fig. 7(b).

## IV. Experiments

We develop our framework based on PyTorch with CUDA, and the experiments are conducted on a Linux machine with

TABLE I: Model parameters

| Parameter | | Value |
|---|---|---|
| Model | Transformer depth | [2,2,2,2] |
| | #Attention heads | [3,6,12,24] |
| | Patch size | $4 \times 4, 2 \times 4 \times 4$ |
| | Window size | $8 \times 8, 4 \times 8 \times 8$ |
| | #Embeddings | 96 |
| Training | Epoch | 50 |
| | Optimizer | AdamW |
| | Learning rate | 1.00E-03 |
| | Weight decay | 0.01 |

TABLE II: CircuitNet-N28 dataset statistics

| Set | #Samples | Design | #Cells | #Nets | Cell Area |
|---|---|---|---|---|---|
| Train | 7078 | RISCY-a<br>RISCY-FPU-a<br>RISCY-b<br>RISCY-FPU-b | 45717<br>65793<br>31311<br>51126 | 47759<br>68351<br>33970<br>54327 | 65739<br>75985<br>69779<br>80030 |
| Test | 3164 | zero-riscy-a<br>zero-riscy-b | 34299<br>20946 | 33970<br>22692 | 58631<br>62648 |

a 2.90GHz Intel Xeon CPU and an Nvidia RTX 3090 GPU. Our model settings are listed in Table I, where the transformer depth denotes the number of transformer layers in each block, number of attention heads denotes the number of parallel self-attention in the transformer. The 2D/3D patch and window size correspond to spatial and temporal channels.

We test our performance on the dynamic IR drop prediction task of CircuitNet [9]. CircuitNet is an open-source dataset for VLSI CAD applications. It provides over 10K samples of circuits with different technology nodes. The circuits are synthesized from 6 different RTL designs with variant synthesis configurations like number of macros, clock frequency, utilization PDN settings, etc. The large variation of design features makes CircuitNet a favored dataset to test the model performance.

We select CircuitNet-N28 with 28nm planar technology as our dataset, which splits 6 RTL designs into 4 training designs with 7078 samples and 2 testing designs with 3164 samples. The dataset statistics are listed in Table II.

We compare our model with the state-of-the-art open-source dynamic IR drop prediction method MAVIREC [4] and the multiscale spatial attention-gated model MAUnet [6].

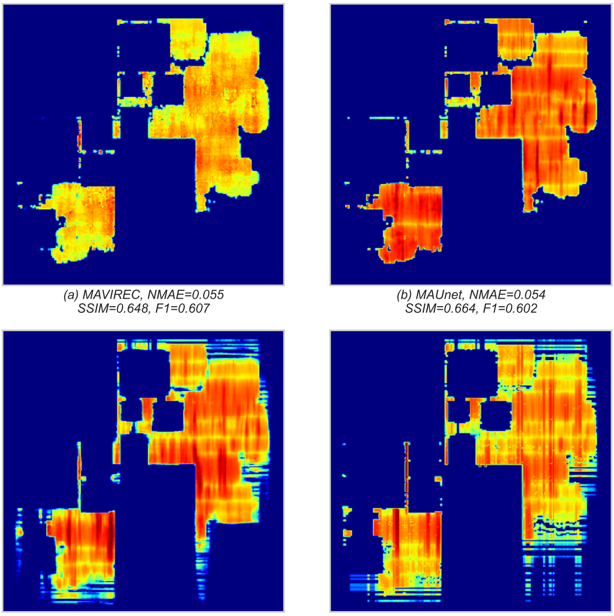
We evaluate the performance of models with Normalized Mean Absolute Error (NMAE), Normalized Root Mean Square Error (NRMS), $R^2$ correlation, F1 score, and Structural Similarity (SSIM).

The threshold of F1 score hotspot is defined as the top 10% of the IR drop tiles. IR drop maps are classified into binary values given the threshold and the F1 score is computed as:

$$\begin{aligned} F1 =& \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Precision} =& \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN} \end{aligned} \tag{4}$$

where the $TP$ is true positive, $FP$ is false positive, and $FN$ is false negative value. The NRMS is defined as:

$$NRMS(x,y) = \frac{\|x-y\|2}{(y_{\text{max}} - y_{\text{min}})\sqrt{N}} \tag{5}$$


(a) MAVIREC, NMAE=0.055 SSIM=0.648, F1=0.607

(b) MAUnet, NMAE=0.054 SSIM=0.664, F1=0.602

(c) Ours, NMAE=0.045 SSIM=0.716, F1=0.719

(d) Ground-truth

Fig. 8: The IR drop hotspots of (a)MAVIREC; (b)MAUnet; (c)Ours; (d)Ground-truth.

where $y_{\text{max}}, y_{\text{min}}$ are the data range of $y$ and $N$ is number of pixels. And the SSIM is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

where $\mu_x, \mu_x$ are the mean values of $x, y$, $\sigma_x^2, \sigma_y^2$ are their variance, $\sigma_{xy}$ is the correlation between $x, y$, and $C_1, C_2$ are constant parameters.

The NMAE, NRMS, and $R^2$ provide a straightforward view of the accuracy of the IR drop values, whereas the F1 score and SSIM give a global view of the structural accuracy of IR drop hotspots. Note that lower NMAE and NRMS scores are better, while higher F1, SSIM, and $R^2$ scores are better.

### A. *Dynamic IR drop prediction*

The results of dynamic IR drop prediction compared with other SOTA models are shown in Table III. Performance is examined on the 2 sets of test designs (zero-riscy-a , zero-riscy-a ) with 3164 samples in total.

Our model displayed better performance in all metrics. The NRMS and NMAE are improved by over 30%, highlighting a better accuracy of our prediction. We also demonstrate a better F1 score and SSIM score, which means our model presents a better global sense across the whole circuit. This is achieved because our model extracts the global feature representation in both spatial and temporal spaces so that the dynamic power activities can be well captured.

We demonstrate the IR drop hotspots of the 3 models in Fig. 8. Our model presents a better prediction quality in a global view. Besides, the model also better captures the local detailed features which the other models do not.

TABLE III: Experimental results on CircuitNet for dynamic IR drop prediction. Best results are highlighted in **brown** .

| Design | MAVIREC [4] | | | | | MAUnet [6] | | | | | Ours | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NRMS↓ | NMAE↓ | F1↑ | SSIM↑ | R2↑ | NRMS↓ | NMAE↓ | F1↑ | SSIM↑ | R2↑ | NRMS↓ | NMAE↓ | F1↑ | SSIM↑ | R2↑ |
| zero-riscy-a | 0.101 | 0.032 | 0.642 | 0.765 | 0.919 | 0.100 | 0.030 | 0.722 | 0.805 | 0.921 | 0.076 | 0.022 | 0.776 | 0.851 | 0.955 |
| zero-riscy-b | 0.126 | 0.038 | 0.710 | 0.762 | 0.826 | 0.129 | 0.038 | 0.735 | 0.773 | 0.817 | 0.103 | 0.030 | 0.786 | 0.818 | 0.882 |
| Total | 0.111 | 0.035 | 0.663 | 0.761 | 0.882 | 0.111 | 0.033 | 0.718 | 0.790 | 0.881 | **0.086** | **0.025** | **0.780** | **0.839** | **0.929** |
| Ratio | 1.30 | 1.40 | 0.85 | 0.91 | 0.95 | 1.30 | 1.34 | 0.92 | 0.94 | 0.95 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

↓ means the smaller the better, ↑ means the larger the better.

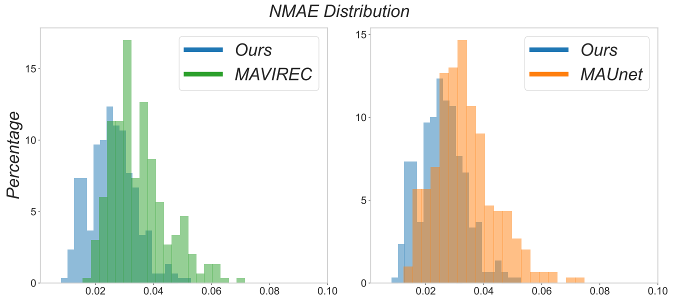


Fig. 9: The normalized error distribution of 3 models.

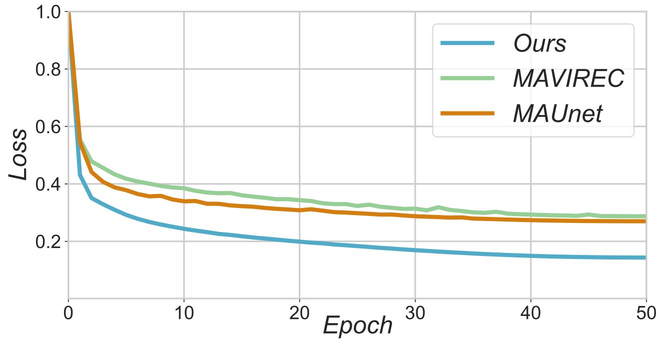


Fig. 10: The training loss of our model compared with other methods.

The normalized error distribution in Fig. 9 indicated that our prediction has a greater correlation with the accurate value, with the $R^2$ reaching up to 93%.

### B. Training and Runtime Analysis

We show the training loss of our model compared with the other 2 models in Fig. 10. Our model reveals faster and better convergence than the other 2 models. The inference runtime comparison is also demonstrated in Fig. 11. Our model has a much shorter inference time compared with MAVIREC because we only perform 3D convolution in the patch embedding layer so that the runtime of our model is linear to the 2D models. The runtime overhead compared with MAUnet is small and tolerable in IC design cycles, with commercial tools running over hours.

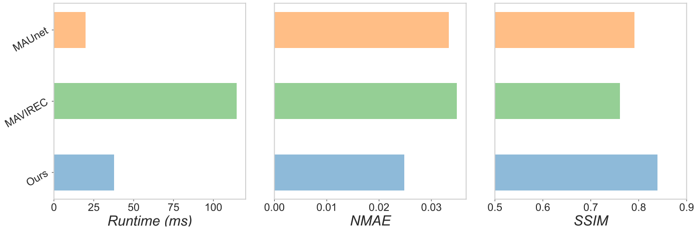


Fig. 11: The inference runtime and quality comparison with other methods.

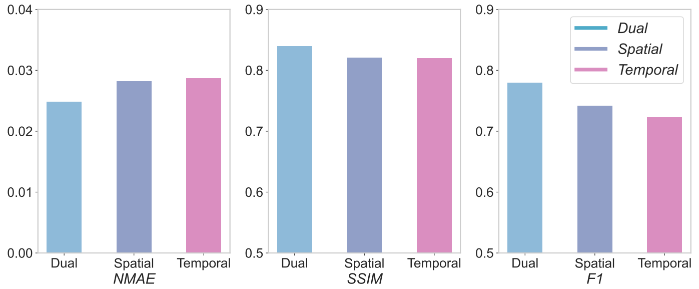


Fig. 12: Quality metrics on our dual path model and single path models.

### C. Ablation Studies

To measure the contribution of our proposed model, we trained our model with a single spatial and temporal path separately and tested the performance. The results on NMAE, SSIM and F1 score shown in Fig. 12 indicate a degradation in quality compared with the dual-path model, which suggests the effectiveness of our proposed dual-path model framework.

## V. Conclusion

In this paper, we propose a dual-path spatial-temporal model. We present strong quality and runtime compared to the SOTA works in IR drop prediction tasks. Results on large-scale advanced datasets exhibit better performance and reliability of our model. The future work would be on optimizing the design IR drop based on the prediction results.

## References

[1] Y. Zhong and M. Wong, "Fast algorithms for ir drop analysis in large power grid," in *ICCAD-2005. IEEE/ACM International Conference on Computer-Aided Design, 2005.*, 2005.

[2] V. A. Chhabria, V. Ahuja, A. Prabhu, N. Patil, P. Jain, and S. S. Sapatnekar, "Thermal and ir drop analysis using convolutional encoder-decoder networks," in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021, pp. 690–696.

[3] Z. Xie, H. Ren, B. Khailany, Y. Sheng, S. Santosh, J. Hu, and Y. Chen, "Powernet: Transferable dynamic ir drop estimation via maximum convolutional neural network," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE Press, 2020.

[4] V. A. Chhabria, Y. Zhang, H. Ren, B. Keller, B. Khailany, and S. S. Sapatnekar, "Mavirec: Ml-aided vectored ir-drop estimation and classification," in *2021 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2021, pp. 1825–1828.

[5] Y. Kwon and Y. Shin, "Fast prediction of dynamic ir-drop using recurrent u-net architecture," in *2022 ACM/IEEE 4th Workshop on Machine Learning for CAD (MLCAD)*, 2022.

[6] M. Wang, Y. Cheng, Y. Lin, K. Peng, Y. Shunchuan, Z. Jin, and W. Xing, "Maunet: Multiscale attention u-net for effective ir drop prediction," in *2024 61st ACM/IEEE Design Automation Conference (DAC)*. ACM/IEEE, 2024, pp. 1–6.

[7] Y. Meng, R. Lyu, Z. Bi, C. Yan, F. Yang, W. Hu, D. Zhou, and X. Zeng, "Circuits physics constrained predictor of static ir drop with limited data," in *2024 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2024, pp. 1–2.

[8] Y. Zhao, Z. Chai, X. Jiang, Y. Lin, R. Wang, and R. Huang, "Pdnnet: Pdn-aware gnn-cnn heterogeneous network for dynamic ir drop prediction," 2024. [Online]. Available: https://arxiv.org/abs/2403.18569

[9] Z. Chai, Y. Zhao, W. Liu, Y. Lin, R. Wang, and R. Huang, "Circuitnet: An open-source dataset for machine learning in vlsi cad applications with improved domain-specific evaluation metric and learning strategies," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 12, pp. 5034–5047, 2023.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[13] S. Zheng, L. Zou, P. Xu, S. Liu, B. Yu, and M. Wong, "Lay-net: Grafting netlist knowledge on layout-based congestion prediction," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–9.

[14] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3192–3201.