

# GANGR: GAN-Assisted Scalable and Efficient Global Routing Parallelization

Hadi Khodaei Jooshin

*Department of Electrical and Computer Engineering  
University of Illinois Chicago  
Chicago, USA  
hjoos@uic.edu*

Inna Partin-Vaisband

*Department of Electrical and Computer Engineering  
University of Illinois Chicago  
Chicago, USA  
vaisband@uic.edu*

**Abstract**—Global routing is a critical stage in electronic design automation (EDA) that enables early estimation and optimization of the routability of modern integrated circuits with respect to congestion, power dissipation, and design complexity. Batching is a primary concern in top-performing global routers, grouping nets into manageable sets to enable parallel processing and efficient resource usage. This process improves memory usage, scalable parallelization on modern hardware, and routing congestion by controlling net interactions within each batch. However, conventional batching methods typically depend on heuristics that are computationally expensive and can lead to suboptimal results (oversized batches with conflicting nets, excessive batch counts degrading parallelization, and longer batch generation times), ultimately limiting scalability and efficiency. To address these limitations, a novel batching algorithm enhanced with Wasserstein generative adversarial networks (WGANs) is introduced in this paper, enabling more effective parallelization by generating fewer higher-quality batches in less time. The proposed algorithm is tested on the latest ISPD’24 contest benchmarks, demonstrating up to 40% runtime reduction with only 0.002% degradation in routing quality as compared to state-of-the-art router.

**Index Terms**—Wasserstein GANs, batching, parallelization, global routing, machine learning EDA

## I. INTRODUCTION

Global routing algorithms target runtime efficiency, scalability (i.e., the ability to handle increasingly larger and more complex designs), and design quality (i.e., routability, congestion, power dissipation and wirelength). Existing approaches typically employ graph-based heuristics [5], [11], iterative maze routing [7], [11], and hierarchical decomposition [8], [18], yielding improved yet limited routing performance. Achieving scalable routing requires more than improved algorithms—effective parallelization with hardware accelerators (GPUs, TPUs, etc.) is essential.

To parallelize global routing execution, traditional batching algorithms group nets that do not share routing resources into parallelizable batches. However, forming such batches is computationally expensive. For example, in InstantGR framework [8], this process consumes up to 17.2% of the overall routing runtime. Efficient batching is therefore a primary

challenge in large, congested integrated circuits (ICs). Several studies have exploited GPU acceleration for maze routing, Steiner tree construction, and related subproblems [6], [7], [9], [10], however, there hasn’t been a GPU exploitation approach for batching nets using deep learning models. For example, maze routing and net batching is accelerated with a GPU-based algorithm, GAMER [6], achieving a  $2.7\times$  speedup compared to traditional CUGR [9]. GPU-accelerated FastGR router achieves a  $2.489\times$  speedup compared with CUGR [11]. A  $13\times$  speedup over multithreaded CUGR is reported in [7] as a result of parallelization techniques. A scalable data partitioning and memory management schemes in the most recent InstantGR [18] facilitate mapping of global routing tasks onto thousands of GPU threads, while balancing computational load across nets to avoid thread divergence. These improvements increase throughput compared to the CPU-based version, particularly on large three-dimensional (3D) benchmarks, positioning InstantGR as the top performing global router. A primary limitation with the existing GPU-based routers is their heavy reliance on time-consuming and inefficient heuristics used to achieve net-level parallelism. Thus, more efficient batching algorithms are required to reduce batch count, improve parallelism, and leverage GPU-like architectures more effectively. Some algorithms try to implement deep learning methods into the global routing or pathfinding problems; however, none of them take advantage of deep learning models to increase parallelization [12]–[17].

In this paper, a batching framework for **fast generation of fewer and larger routing batches** is proposed. The key idea is to enhance parallelism by grouping a larger set of non-overlapping nets into fewer batches more effectively. The primary contributions are as follows:

- **Enhanced parallelism:** a novel batching algorithm partitions nets into fewer larger batches, reducing inter-batch conflicts.
- **Faster batching:** WGAN [3] is used to learn complex net-interference patterns, enabling faster and more accurate batch formation compared to existing heuristics. To the best of our knowledge, this is the first application of learning-based methods to routing batching.
- **Runtime improvements:** achieves faster global routing

This work was supported in part by the National Science Foundation under Grant No. 2151854, titled End-to-End Global Routing with Reinforcement Learning in VLSI Systems.

(due to WGAN-based accelerated batch generation and increased parallelism with fewer batches) with similar wirelength as compared with InstantGR.

The rest of the paper is organized as follows. The proposed framework is described in Section II. The experimental results are presented in Section III. The paper is concluded in Section IV.

## II. PROPOSED FRAMEWORK

The proposed approach advances the parallelism principles introduced by InstantGR [8] in two key ways. First, the rule-based batching of nets is replaced with a WGAN-based mechanism trained to group non-overlapping nets more effectively, producing fewer and larger highly parallelizable batches. Second, CPU-only multithreaded batch processing is replaced with GPU acceleration, significantly increasing throughput. The WGAN-based batching algorithm is integrated into the first version of the InstantGR [8], yielding up to 40% faster runtime without additional overheads.

### A. Overview of Net Overlap Analysis

Net batching is a traditional approach for reducing global routing runtime [1], [2], [4], [6]–[9], [11], [18]. The main idea is to parallelize the routing of nets that neither overlap nor share routing resources. Owing to their simplicity, bounding boxes (i.e., the smallest rectangles enclosing the routing graphs of individual nets) are often used to detect potential net overlaps [6], [7], [9], [11]. This approach, however, has two main limitations: 1) it achieves a lower degree of parallelism, and 2) analyzing bounding box overlaps typically requires complex data structures (e.g., R-trees) that are difficult to optimize.

To enhance parallelization, InstantGR routes nets simultaneously if their vertical segments do not overlap and their horizontal segments do not overlap, since these segments use different routing resources. However, it treats nets as overlapping whenever their segments are vertically or horizontally aligned, even if the segments are placed on different metal layers, thereby ignoring layer distinctions during parallelization. To address this issue, metal layers are considered in this paper independently when analyzing overlaps, allowing nets to be routed in parallel, provided that no overlaps exist within the same metal layer. To identify overlaps, the algorithm compares horizontal (vertical) segments only with horizontal (vertical) segments of other nets on the same metal layer.

These three batching approaches (i.e., bounding box-based, metal layer agnostic, and the proposed 3D-aware) are illustrated with four nets in Figure 1. In this case, the four nets cannot be parallelized with the bounding box and InstantGR batching approaches because of the overlaps in the bounding boxes (see Figure 1(a)) or the individual segments, i.e., horizontal for the black/blue nets and vertical for the green/black and red/blue nets, (see Figure 1(b)). Alternatively, if the vertical and horizontal segments of the nets that are considered overlapped by InstantGR are placed in different layers, routing of these nets can be parallelized (see Figure 1(c)).

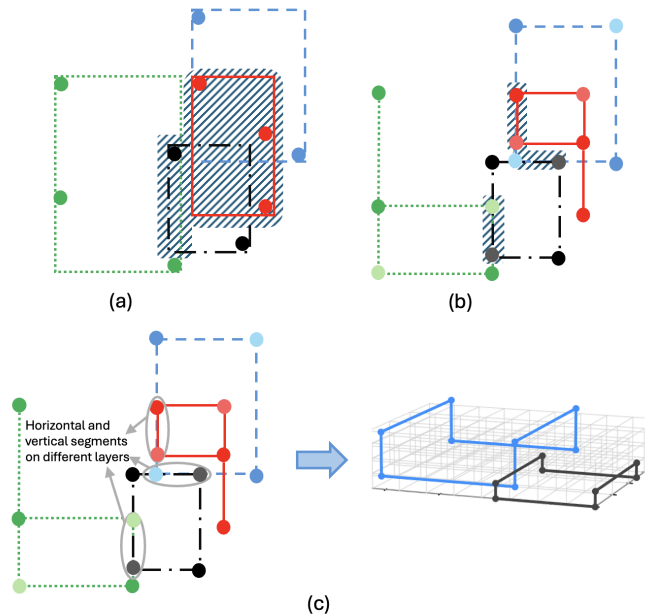


Fig. 1. Net overlap analysis with conflicts (excluded from parallel routing) shown as hatched regions, (a) the bounding box method flags large overlap, limiting parallelism (used by many algorithms, e.g., CUGR2 [9]), (b) the layer-agnostic analysis in InstantGR [8], which conservatively marks vertical and horizontal conflicts across all layers, underutilizing available parallelism, and (c) the proposed layer-aware method considers overlaps only within the same metal layer, avoiding false conflicts between nets on different layers, as illustrated by the black and blue nets in the 3D view.

As previously noted, net overlap analysis is a primary concern in batching, since it is used both to construct batches and to detect conflicts that require reallocating overlapping nets. Consequently, overlap analysis directly impacts batching quality while also being computationally expensive. The proposed algorithm efficiently detects 3D spatial overlaps among nets within each batch, leading to improved batch validity and reduced conflicts. Dynamic scheduling is used to distribute batch processing across multiple threads, where each thread maintains a local 3D visibility array<sup>1</sup> spanning the routing grid dimensions ( $x_G \cdot y_G \cdot z_G$ ). For each net-batch pair, a layer-aware conflict detection is performed. If no conflicts are detected, the positions of all horizontal and vertical segments along with the pins are recorded in the visibility map to ensure proper isolation of routing elements across different metal layers in 3D space.

### B. Wasserstein GAN-Based Layer-Aware Batching

To the best of our knowledge, the proposed WGAN batching algorithm is the first algorithm that leverages GPUs using a deep learning-based approach. It operates in three stages.

- 1) **WGAN-based initial batching** (see Algorithm 1): groups  $\sim 90\%$  of nets correctly (without overlaps).

<sup>1</sup>a Boolean data structure indexed over the routing grid dimensions, where each entry indicates whether a grid cell is free (visible) or occupied, enabling constant-time conflict and overlap detection.

- 2) **Adaptive evaluation of WGAN batches** (see Algorithm 2): identifies conflicts within the individual batches based on layer-aware overlap criteria.
- 3) **Greedy layer-aware nets reallocation** (see Algorithm 3): reallocates the remaining  $\sim 10\%$  overlapping nets into suitable WGAN-generated batches, creating a new batch only when no fit exists.

**B.1. Strategies:** The following strategies are utilized throughout the framework to optimize memory usage and facilitate synchronization.

*Data structures:* By default, batches are stored in a dense 3D array, offering  $\mathcal{O}(1)$  access time at the expense of increased memory usage, which is suitable for smaller ICs. When the total memory requirement (number of batches  $\cdot x_G \cdot y_G \cdot l_G$ ) exceeds a predefined threshold, the algorithm switches to a sparse hash-based representation. Although slightly slower, this approach is far more memory-efficient and prevents overflow in large-scale designs while preserving overall efficiency. Each hash set is pre-allocated for 1,000 elements that represent typical net occupancy patterns observed in industrial routing scenarios.

*Memory management:* Efficient data movement (i.e., C++ move semantics) and bulk operations are utilized to reduce memory allocation overhead.

*Thread Synchronization:* To maximize computational efficiency and avoid race conditions, parallel processing with carefully designed thread-local storage is employed for analyzing the individual batches for conflicts. Thread-local results are aggregated in shared data structures to avoid conflicting writes to shared memory.

**B.2. WGAN-Based Initial Batching (Algorithm 1):** The WGAN model is designed to efficiently batch nets in large-

scale routing tasks. To optimize memory usage and computational efficiency with multithreading, nets are processed in chunks, with chunk size adaptively determined based on the total number of nets. For larger ICs, smaller chunks are used to prevent GPU out-of-memory issues and maintain scalability. This strategy enables processing of millions of nets within hardware limits. For each chunk, feature extraction (pin coordinates) and inference are performed sequentially. Extracted features are normalized by the grid dimensions to ensure consistent input scaling. Nets with fewer features than required are padded or repeated. All features are aggregated in a chunk into a feature matrix, converted into a tensor, and transferred to the GPU. The WGAN is used to predict batch assignments for all nets in the chunk, after which results are moved back to the CPU to free GPU memory for subsequent chunks. To capture inter-chunk dependencies, the predicted assignments for each net are stored in a global assignment list.

Following this procedure, batches are generated with net indices assigned to each batch. The method enables scalable, memory-efficient batching by combining neural network inference, multithreading, and adaptive chunking.

**B.3. Adaptive Evaluation of WGAN Batches (Algorithm 2):** The WGAN-generated batches are evaluated for 3D-aware routing overlaps, with memory management strategies that adapt dynamically to problem size, ensuring scalability. To enable efficient spatial queries across multi-layer routing paths, 3D pin coordinates are mapped into unified spatial indices, based on the pin layer and its 2D position within the layer (see lines 15-16).

Dense array access or sparse hash lookup are prototyped based on the preferred data structure strategy (see Section B.1). The dense representation detects conflicts through direct array

---

#### Algorithm 1 WGAN-Based Initial Batching

---

**Require:** *nets* (net IDs), *model* (trained WGAN cached in GPU)

**Ensure:** *batches* (batched nets for parallel processing)

- 1: **Memory Management:** Compute adaptive chunk size:  $\text{MAX\_CHUNK\_SIZE} \leftarrow \min(10^5, \max(10^4, |nets|/10))$
  - 2: **for** each chunk of nets **do**
  - 3:   **Feature Extraction:**
  - 4:     Extract normalized coordinates from net pin locations
  - 5:     Create 16-dimensional feature vector per net
  - 6:     Pad feature vector as needed
  - 7:   **Tensor Creation:** Convert feature vectors to PyTorch tensor and transfer to GPU memory
  - 8:   **ML Inference:** Net-batch distribution probability  $P \leftarrow model$
  - 9:   **Batch Assignment:** For each net  $n$ ,  $n$ 's batch index is  $i = \arg \max(P)$ ,  $batches[i] \leftarrow batches[i] + n$
  - 10:   **Memory Cleanup:** Transfer results to CPU; clear GPU tensor
  - 11: **end for**
  - 12: **Batch Aggregation:** Collect batch IDs from all chunks into unified result vector
  - 13: **Batch Organization:** Group nets by predicted batch IDs with memory pre-allocated based on expected batch sizes to reduce dynamic allocation overhead
  - 14: **Return:** *batches* for subsequent parallel routing processing
- 

---

#### Algorithm 2 Adaptive Evaluation of WGAN Batches

---

**Require:** *batches* (batched nets), *grid\_dimensions* ( $x_G, y_G, l_G$ )

**Ensure:** *accepted\_batches* (conflict-free batches), *nets2reroute* (conflicting nets)

- 1: **for** each batch  $i$  in *batches* **in parallel do**
  - 2:   Initialize *conflicts* (conflict detection structure in dense or sparse representation; see Data structure strategy, Section B.1)
  - 3:   **for** each pin  $(x_p, y_p)$  of each net  $n$  in batch  $i$  **do**
  - 4:     **if**  $\text{CONFLICT\_DETECTED}((x_p, y_p))$  **then**
  - 5:        $nets2reroute \leftarrow nets2reroute + n$
  - 6:     **else**
  - 7:        $accepted\_batches \leftarrow accepted\_batches + n$
  - 8:       Mark  $conflicts[n.pins]$  as occupied
  - 9:     **end if**
  - 10:   **end for**
  - 11: **end for**
  - 12: Synchronize shared data structures across threads (see Thread synchronization strategy, Section B.1)
  - 13: **Return:** *accepted\_batches*, *nets2reroute*
  - 14: **Function**  $\text{CONFLICT\_DETECTED}(pin(x_p, y_p))$
  - 15:   Extract layer  $\ell \leftarrow \lfloor (x_p, y_p) / (x_G \cdot y_G) \rfloor$
  - 16:    $pos_{3D} \leftarrow \ell \cdot x_G \cdot y_G + x_p \cdot y_G + y_p$
  - 17:   Check conflicts in layer  $\ell$ ; return Boolean yes/no
-

indexing with bounds checking, offering constant-time access but at the cost of high memory usage. In contrast, the sparse representation relies on hash set membership queries, which reduce memory overhead but introduce hashing latency. This trade-off favors dense methods for performance in heavily populated grids, while sparse methods scale better in sparse routing scenarios.

Per-thread data structures are maintained independently and the overall results are aggregated across threads in shared data structures using thread synchronization mechanisms. To address variability in batch sizes, dynamic scheduling of parallel loops is employed to ensure load balancing across threads.

The adaptive approach ensures that the conflict detection overhead remains manageable even for large-scale industrial designs while maintaining the accuracy necessary for successful parallel routing execution, as demonstrated in the Results section.

---

### Algorithm 3 Greedy Layer-Aware Nets Reallocation

---

**Require:**  $nets2reroute$ ,  $MAX\_BATCH\_SIZE$ ,  $batches$  (batched nets),  $grid\_dimensions(x_G, y_G, l_G)$

**Ensure:**  $accepted\_batches$  (conflict-free batches with all nets assigned)

```

1: for each net  $n$  in  $nets2route$  do
2:   for each batch  $b$  in  $batches$  do
3:     if  $|b| + |batch\_assignments[b]| < MAX\_BATCH\_SIZE$ 
       and  $!CONFLICT\_DETECTED(n, b)$  then
4:        $assigned\_batch \leftarrow b$ 
5:     end if
6:   end for
7:   if  $not\_assigned\_batch$  then
8:      $unassigned\_nets \leftarrow unassigned\_nets + n$ 
9:   else
10:     $batch\_assignments[assigned\_batch] \leftarrow$ 
        $batch\_assignments[assigned\_batch] + n$ 
11:   end if
12: end for
13: for each batch  $i$  in  $batch\_assignments$  in parallel do
14:   Initialize  $conflicts$  (conflict detection structure in dense or
       sparse representation; see Data structure strategy, Section B.1)
15:   for each pin  $(x_p, y_p)$  of each net  $n$  in batch  $i$  do
16:     if  $CONFLICT\_DETECTED((x_p, y_p))$  then
17:        $nets2reroute \leftarrow nets2reroute + n$ 
18:     else
19:        $accepted\_batches \leftarrow accepted\_batches + n$ 
20:       Mark  $conflicts[n.pins]$  as occupied
21:       Mark  $conflicts[n.RSMT\_segments]$  as occupied
22:     end if
23:   end for
24: end for
25: Synchronize shared data structures across threads (see Thread
       synchronization strategy, Section B.1)
26: while  $unassigned\_nets \neq \emptyset$  do
27:   Exhaustively batch remaining nets into new batches until no
       more assignments possible
28: end while
29: Consolidate small batches ( $\leq 5$  nets for datasets  $< 10M$  nets,
        $\leq 10$  nets for datasets  $> 10M$  nets) and report statistics

```

---

**B.4. Greedy Layer-Aware Nets Reallocation (see Algorithm 3):** To resolve the spatial conflicts arising from the WGAN-based batching, conflicting nets are (i) identified and reassigned to conflict-free batches while preserving parallelization. The approach combines adaptive conflict resolution with efficient batch management to maximize routing throughput.

As a first step, a systematic assignment of conflicting nets into existing batches is attempted. Then, parallel processing with thread-local 3D arrays is employed to maintain isolation between concurrent evaluations. All pins, along with the horizontal and vertical segments of the rectilinear Steiner minimal tree (RSMT) are recorded in the local arrays to ensure complete coverage of potential routing paths.

For nets that cannot be assigned without conflict within existing batches, an iterative greedy procedure is applied to generate additional conflict-free batches. The process is repeated until all nets are either assigned or identified as unassignable due to inherent constraints. Results from parallel processing phases are combined through thread synchronization mechanisms, ensuring data consistency and sustaining computational efficiency.

In the final phase, small batches are merged to reduce overhead: batches with at most five nets are combined for datasets containing up to 10M nets, and batches with at most ten nets are combined for larger datasets.

### C. ML Training with WGAN-Based Gradient Penalty

An application-specific Wasserstein GAN with gradient penalty (WGAN-GP) architecture is designed for training the WGAN batching model. Training involves (i) data preprocessing, and (ii) adversarial learning under domain-specific constraints.

**C.1. Training Dataset Construction:** The training dataset is generated by integrating the proposed layer-aware overlap batching criteria into the InstantGR algorithm [8], and routing 177k nets from the NVLDA benchmark [5] with the adjusted InstantGR algorithm. The resulting output comprises pin coordinates, horizontal and vertical routing segments, and the batch number of each net.

**C.2. Data Preprocessing and Feature Engineering:** The training process begins with parsing routing batch files containing nets' pin coordinates, horizontal and vertical routing segments, half-perimeter wirelengths (HPWL), and batch assignments. To ensure statistical significance and maintain the number of batches consistent with the WGAN clustering output, only batches with more than 160 nets in the NVDLA benchmark are retained. The number of batches generated by the WGAN model (i.e., the WGAN batch number) is a key parameter influencing both batching quality and inference runtime. In this work, a fixed WGAN batch number is used across all evaluated benchmarks. The chosen value is experimentally determined to balance quality and runtime, as detailed in Section III. A feature graph is then constructed, where each

net is represented by a feature vector that incorporates pin coordinates. To reduce the WGAN input size, preprocessing time, and inference latency, only pin coordinates are used to construct the feature graph. Although this feature reduction lowers accuracy, it provides an acceptable tradeoff between runtime and accuracy. To improve the accuracy of the model at runtime cost, additional spatial characteristics can be included in the feature vector, such as bounding box dimensions, net centers, and vertical and horizontal segment perimeters. Additional preprocessing can also be added, including (i) feature detection through intersection-over-union (IoU) calculations, and (ii) segment overlap analysis to identify nets that cannot be assigned to the same batch due to resource conflicts. The size of the feature vector for each net is restricted to eight pins (i.e., 16 coordinates): nets with fewer than eight pins (i.e., about 92% in the NVDLA benchmark) are fully represented, while for nets with more than eight pins, only the first eight pins are included in the feature graph.

**C.3. Training Loss Function:** The proposed loss function balances three objectives. The segment overlap loss penalizes nets within the same batch whose routing segments overlap, encouraging spatial separation. The center penalty loss discourages grouping nets with overlapping segments by applying a penalty proportional to the squared Euclidean distance between their pin centers, promoting compact batch formation. The pin overlap loss penalizes nets that share pins but assigned to the same batch, reducing pin congestion. Each term is weighted and normalized by the number of nets, yielding a differentiable loss that jointly enforces spatial separation, batch compactness, and balanced pin distribution for large-scale routing tasks. The proposed loss function is:

$$\begin{aligned} \mathcal{L}_{\text{final}} = & \frac{w_{\text{seg}}}{N} \sum_{(i,j) \in \mathcal{N}_{\text{seg}}} \left( \sum_{b=1}^B (p_i^b p_j^b)^2 \right) \\ & + \frac{w_{\text{ctr}}}{N} \sum_{(i,j) \in \mathcal{N}_{\text{seg}}} \left( \sum_{b=1}^B (p_i^b p_j^b) \|c_i - c_j\|^2 \right) \\ & + \frac{w_{\text{pin}}}{N} \sum_{p \in \mathcal{P}} \sum_{(i,j) \in \mathcal{N}_p} \left( \sum_{b=1}^B (p_i^b p_j^b)^2 \right). \end{aligned} \quad (1)$$

- $N$ : Total number of nets in the dataset.
- $w_{\text{seg}}, w_{\text{ctr}}, w_{\text{pin}}$ : Weighting coefficients for segment overlap, center penalty, and pin overlap loss terms, respectively.
- $\mathcal{N}_{\text{seg}}$ : Set of pairs of nets with overlapping routing segments, as defined by the conflict graph.
- $\mathcal{P}$ : Set of all unique pin locations in the design.
- $\mathcal{N}_p$ : Set of pairs of nets that share the same pin  $p \in \mathcal{P}$  location.
- $B$ : Total number of batches
- $p_i^b$ : Assignment probability of net  $i$  to batch  $b$ , as predicted by the WGAN model.
- $c_i$ : Center coordinate (mean pin location) of net  $i$ .

**C.4. Adversarial Training with Spatial Constraints:** The WGAN framework consists of a generator and a critic. The generator is a 5-layer fully connected network with residual connections, LeakyReLU activations, and layer normalization, producing soft clustering probability distributions for input nets. The critic is an 8-layer network designed for spatial feature extraction, distinguishing generator outputs from reference clustering patterns. Training alternates between five

TABLE I  
BENCHMARK DETAILS [ISDP2024] [5].

IC	Benchmark	#Nets	#Pins	Gcell Grid
0	Ariane_sample	129K	420K	844 × 1144
1	MemPool-Tile_sample	136K	500K	475 × 644
2	NVDLA_sample	177K	630K	1240 × 1682
3	BlackParrot_sample	770K	2.9M	1532 × 2077
4	MemPool-Group_sample	3.3M	10.9M	1782 × 2417
5	MemPool-Cluster_sample	10.6M	40.2M	3511 × 4764
6	TeraPool-Cluster_sample	59.3M	213M	7891 × 10708
7	Ariane_rank	128K	435K	716 × 971
8	MemPool-Tile_rank	136K	483K	429 × 581
9	NVDLA_rank	174K	610K	908 × 1682
10	BlackParrot_rank	825K	2.9M	1532 × 2077
11	MemPool-Group_rank	3.2M	10.9M	1782 × 2417
12	MemPool-Cluster_rank	10.6M	40.2M	4113 × 5580
13	TeraPool-Cluster_rank	59.3M	213M	9245 × 12544

TABLE II  
SCORE AND RUNTIME PERFORMANCE OF GANGR COMPARED WITH INSTANTGR [8]

Bench.	InstantGR		GANGR	
	Score	Time (s)	Score	Time (s)
0	19715069	1.51	19714965	1.51
1	15124499	1.67	15126701	1.28
2	47979984	2.65	48038696	1.74
3	112468847	9.44	112463498	6.82
4	397600677	27.83	398550511	15.24
5	1623738805	90.30	1644464118	59.21
7	22544301	1.82	22539569	1.58
8	13772197	1.62	13780323	1.27
9	43044750	2.64	43140388	2.14
10	109840647	5.73	109855333	4.33
11	382576253	23.42	383171693	15.06
12	1780759982	97.65	1788086646	59.58
Avg. Ratio	<b>0.998</b>	<b>1.395</b>	<b>1.000</b>	<b>1.000</b>

critic updates and one generator update, following the WGAN-GP protocol. The critic loss maximizes the Wasserstein distance<sup>2</sup> [3] between real and generated assignments with a gradient penalty ( $\lambda_{gp} = 10.0$ ) to enforce the Lipschitz constraint<sup>3</sup>. The generator optimizes a composite objective combining adversarial loss with domain-specific constraints: segment overlap (routing conflicts), pin overlap (resource contention), center distance (spatial locality), and cluster balance, with tuned loss weights. Both networks are trained with Adam (learning rate = 0.0003,  $\beta_1 = 0.0$ ,  $\beta_2 = 0.9$ ). Early stopping based on overlap reduction metrics prevents overfitting and ensures convergence to physically realizable batching solutions.

### III. EXPERIMENTAL RESULTS

The proposed algorithms are integrated into the InstantGR framework [8]. Experiments are performed on a 64-bit Linux workstation with an AMD Ryzen 9 7950X 16-core processor, 124 GB of memory, and a single NVIDIA GeForce RTX 4090 using 8 CPU threads. The ISPD'24 global routing contest [5] benchmarks (see Table I) and evaluator [5] are used to evaluate the existing [8] and proposed frameworks. Performance is measured with the ISPD'24 contest metric—a weighted cost

<sup>2</sup>a metric used to measure the dissimilarity between the distribution of real data and the distribution of generated data

<sup>3</sup>a mathematical condition applied to the critic function

TABLE III  
WIRELENGTH, VIA COUNT, AND OVERFLOW PERFORMANCE OF GANGR COMPARED TO INSTANTGR [8].

Bench	Wirelength		Via Count		Overflow	
	InstantGR	GANGR	InstantGR	GANGR	InstantGR	GANGR
0	9361015	9361764	2794988	2798052	7560742	7555149
1	8350182	8352319	3259740	3262132	3516588	3512251
2	21289207	21279195	4302208	4301116	22390810	22458385
3	58157063	58158384	18932720	18935032	35385097	35370082
4	260156132	260428582	71296860	71485016	66205020	66636913
5	1091435102	1092394481	256292732	256701240	276218463	295368397
7	11972097	11973312	2827892	2828236	7745811	7738022
8	7553701	7555546	3212488	3217760	3008600	3007016
9	21582358	21578627	4411648	4408724	17053141	17153037
10	55660842	55665285	18993980	18999924	35189233	35190124
11	248207884	248426271	70609280	70845020	63822725	63900402
12	1189011781	1189873247	267495852	267769824	324389756	330443575
Avg. Ratio	<b>1.000</b>	<b>1.000</b>	<b>0.999</b>	<b>1.000</b>	<b>0.992</b>	<b>1.000</b>

TABLE IV  
SCORE AND RUNTIME PERFORMANCE OF GANGR AS A FUNCTION OF THE WGAN BATCH NUMBER.

Benchmark	$n = 20$		$n = 30$		$n = 40$		$n = 50$		$n = 60$	
	Score	Time (s)	Score	Time (s)	Score	Time (s)	Score	Time (s)	Score	Time (s)
0	<b>19712211</b>	<b>1.35</b>	19714965	1.51	19715790	1.69	19716959	1.85	19719862	2.03
1	<b>15119754</b>	<b>1.16</b>	15126701	1.28	15130345	1.41	15132497	1.52	15129001	1.62
2	48176669	2.08	48038696	<b>1.74</b>	48028344	1.89	48008122	2.02	<b>48007290</b>	2.17
3	112530612	<b>6.35</b>	112463498	6.82	<b>112461801</b>	7.12	112479244	7.42	112468346	7.83
4	400002097	15.45	398550511	15.24	398189929	<b>15.17</b>	398241192	15.23	<b>398141390</b>	15.52
5	1755161764	61.8	1644464118	59.21	1645831061	<b>58.35</b>	1631026125	58.73	<b>1627691371</b>	59.61
7	22542373	<b>1.44</b>	22539569	1.58	22538789	1.77	<b>22536818</b>	1.92	22541876	2.1
8	13764719	<b>1.18</b>	13780323	1.27	13773858	1.41	<b>13763651</b>	1.48	13767970	1.58
9	43270730	1.84	43140388	2.14	43107865	<b>1.77</b>	43113495	1.89	<b>43062996</b>	1.99
10	<b>109827103</b>	<b>4.2</b>	109855333	4.33	109852896	4.53	109857836	4.73	109856100	4.91
11	383748478	<b>14.88</b>	383171693	15.06	383049653	15.01	383069104	15.22	<b>383016832</b>	15.38
12	1949165122	61.54	1788086646	59.58	1789137454	<b>59.02</b>	1786770109	59.66	<b>1782743886</b>	59.11
Avg. Ratio	<b>1.000</b>	<b>1.000</b>	<b>0.987</b>	<b>1.035</b>	<b>0.987</b>	<b>1.071</b>	<b>0.986</b>	<b>1.124</b>	<b>0.986</b>	<b>1.182</b>

function combining total wirelength, via count, and overflow penalty [5].

The score and runtime performance of GANGR compared with InstantGR [8] is presented in Table II. For a fair comparison, InstantGR, the only publicly available version among the current SOTA frameworks, was re-evaluated on the previously described local workstation together with GANGR. As compared with InstantGR, the proposed GANGR framework achieves  $\sim 40\%$  improvement in runtime with only 0.002% reduction in routing score. GANGR achieves speedup through both efficient batch generation and parallelization. On benchmark #12, segment-based InstantGR produces 383 batches in 9.3 s, while GANGR generates only 68 batches (82% fewer) in 4.4 s (52% faster), enabling greater parallelization.

A detailed comparison of GANGR with InstantGR is presented in Table III. The average wirelength of GANGR matches that of InstantGR. The average via count is 0.001% higher than InstantGR, while the overflow is 0.008% higher. Given the runtime improvement achieved, these small degradations are considered negligible.

As noted earlier, in this work, a fixed WGAN batch number is used across all evaluated benchmarks. Experimental results for different WGAN batch number are summarized in Table IV. Increasing the number of batches reduced the average score while increasing runtime. However, the strong postprocessing stage mitigates the effect of fewer batches on quality, resulting in only a minor score reduction. Thus, a smaller WGAN batch number is preferred. To maintain consistency

across experiments, 30 batches ( $n = 30$ ) is selected as it offers the best balance between quality and runtime. While assigning benchmark-specific values could further optimize results, this option is not considered here for fair comparison.

#### IV. CONCLUSION

In this paper, GANGR is introduced, a deep learning-assisted batching framework that integrates Wasserstein generative adversarial networks (WGANs) with a newly developed batching approach in the global routing pipeline. Unlike heuristic-based batching methods, the proposed framework learns complex net-interference patterns and partitions nets into fewer batches, thereby improving parallelism and scalability. With layer-aware overlap detection and adaptive memory-efficient validation strategies, preprocessing and batch generation time are reduced while routing quality is maintained across large-scale industrial benchmarks. Experimental results on the ISPD'24 benchmarks show that GANGR achieves runtime improvements of up to 40% over state-of-the-art global routers, while maintaining competitive wirelength, via count, and congestion metrics.

#### REFERENCES

- [1] Gengjie Chen, Chak-Wa Pui, Haocheng Li, Jingsong Chen, Bentian Jiang, and Evangeline FY Young. Detailed routing by sparse grid graph and minimum-area-captured path search. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 754–760, 2019.

- [2] Gengjie Chen, Chak-Wa Pui, Haocheng Li, and Evangeline FY Young. Dr. cu: Detailed routing by sparse grid graph and minimum-area-captured path search. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(9):1902–1915, 2019.
- [3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.
- [4] Haocheng Li, Gengjie Chen, Bentian Jiang, Jingsong Chen, and Evangeline FY Young. Dr. cu 2.0: A scalable detailed routing framework with correct-by-construction design rule satisfaction. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–7. IEEE, 2019.
- [5] Rongjian Liang, Anthony Agnesina, Wen-Hao Liu, and Haoxing Ren. Gpu/ml-enhanced large scale global routing contest. In *Proceedings of the 2024 International Symposium on Physical Design*, pages 269–274, 2024.
- [6] Shiju Lin, Jinwei Liu, Evangeline FY Young, and Martin DF Wong. Gamer: Gpu-accelerated maze routing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(2):583–593, 2022.
- [7] Shiju Lin and Martin DF Wong. Superfast full-scale cpu-accelerated global routing. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, pages 1–8, 2022.
- [8] Shiju Lin, Liang Xiao, Jinwei Liu, and Evangeline FY Young. Instantgr: Scalable gpu parallelization for global routing. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pages 1–8, 2024.
- [9] Jinwei Liu, Chak-Wa Pui, Fangzhou Wang, and Evangeline FY Young. Cugr: Detailed-routability-driven 3d global routing with probabilistic resource model. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.
- [10] Jinwei Liu and Evangeline FY Young. Edge: Efficient dag-based global routing engine. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023.
- [11] Siting Liu, Yuan Pu, Peiyu Liao, Hongzhong Wu, Rui Zhang, Zhitang Chen, Wenlong Lv, Yibo Lin, and Bei Yu. Fastgr: Global routing on cpu-gpu with heterogeneous task graph scheduler. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(7):2317–2330, 2022.
- [12] Inna Partin-Vaisband. Shf: Small: End-to-end global routing with reinforcement learning in vlsi systems. *NSF Award Number 2151854. Directorate for Computer and Information Science and Engineering*, 21(2151854):51854, 2022.
- [13] Inna Partin-Vaisband. High-resolution ic net routing system, components and methods with deep neural networks, June 15 2023. US Patent App. 18/065,068.
- [14] Dmitry Utyamishv and Inna Partin-Vaisband. Progressive vae training on highly sparse and imbalanced data. *arXiv preprint arXiv:1912.08283*, 2019.
- [15] Dmitry Utyamishv and Inna Partin-Vaisband. Late breaking results: A neural network that routes ics. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–2. IEEE, 2020.
- [16] Dmitry Utyamishv and Inna Partin-Vaisband. Late breaking results: Parallelizing net routing with cgans. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 1372–1373. IEEE, 2021.
- [17] Dmitry Utyamishv and Inna Partin-Vaisband. Multiterminal pathfinding in practical vlsi systems with deep neural networks. *ACM Transactions on Design Automation of Electronic Systems*, 28(4):1–19, 2023.
- [18] Liang Xiao, Shiju Lin, Jinwei Liu, Qinkai Duan, Tsung-Yi Ho, and Evangeline FY Young. Instantgr: Scalable gpu parallelization for 3-d global routing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2025.