韓 曦（ハン・シー）

39 Montclair Dr, Selden, New York 11784, United States
(+1) 631-710-8313 | xihan1@cs.stonybrook.edu | https://axihixa.github.io/

学　歴

ニューヨーク州立大学ストーニーブルック校・コンピュータサイエンス学科 **2019年8月 ～ 現在**
(Stony Brook University, Department of Computer Science, Stony Brook, New York, United States)
哲学博士（コンピュータサイエンス）| 進行中、2026年春修了予定 | GPA: 3.9/4.0

清華大学・コンピュータサイエンス学科 **2015年8月 ～ 2019年7月**
(Tsinghua University, Department of Computer Science and Technology, Beijing, China)
工学学士（コンピュータサイエンス）| GPA: 3.25/4.0

出版物

- **Xi Han**, Fei Hou and Hong Qin, "UGrid: An Efficient-And-Rigorous Neural Multigrid Solver for Linear PDEs", In *Proceedings of the 41st International Conference on Machine Learning*, pp. 17354 – 17373, July 2024.
- Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, **Xi Han**, Dingcheng Yang, Hao-Zhi Huang and Shi-Min Hu, "Pose2Seg: Detection Free Human Instance Segmentation", In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 889 – 898, June 2019.

職　歴

ニューヨーク州立大学ストーニーブルック校・コンピュータグラフィックス研究室 **2019年8月 ～ 現在**
研究助手（リサーチアシスタント）・ティーチングアシスタント

- コンピュータグラフィックス（物理的モデリング）の研究。研究方向： PDEをベースの差分可能なベクトルグラフィックス・データ駆動型ニューラルPDEソルバー。
- コンピュータビジョン研究室と協力し、AIモデルのトレーニング/推論効率を最適化。CUDAカーネル融合・キャッシュ効率の良いAIオペレーター（例: モンテカルロ積分器・2D Mambaスキャンナー）の開発・調整・パフォーマンス最適化。
- ティーチングアシスタントとして、OpenGL/C++/Pythonなどの講義。

テキサス大学ダラス校・コンピュータグラフィックス・アニメーション研究室 **2018年9月 ～ 2018年11月**
研究助手（リサーチアシスタント）

- サムスン研究所との共同研究で3D顔再構築プロジェクトに参加。人顔モデルデータセットを構築。
- Linuxワークステーションを設定し、ニューラルネットワークモデルをデプロイ。

清華大学・グラフィックス・幾何・計算研究室 **2017年1月 ～ 2019年7月**
研究助手（リサーチアシスタント）

- Apple CoreMLフレームワークを使用し、iOSプラットフォームでMobileNetモジュールをデプロイ。人間セグメンテーション用アプリを開発。
- アプリで使用するモデルを最適化し、精度向上およびFPS 10倍の高速化を実現。

スキル

- **数値解析・高性能計算・コンピューターグラフィックス・機械学習・Linuxシステム：**
  - **コンピューターグラフィックス**：物理的モデリング・ベクターグラフィックス
  - **数値解析**：ニューラルPDEソルバー、CUDAオペレーターのカスタマイズ。
  - **高性能計算とAI**： AIモデルのトレーニング・推論効率の最適化。関与技術：PyTorch C++/CUDAエクステンション、CUDAオペレーターの融合・キャッシュとメモリーの効率最適化・プロファイリング。
  - **プログラミング言語**：C/C++（OOP・STL・メタプログラミング・並列処理）・CUDA（PTXを含む）・Python。
  - **ツール**：PyTorch Profiler・CUDA-GDB・Nsight Compute・NVIDIA Compute Sanitizer。
  - **フレームワーク**：PyTorch・OpenGL・Qt。
  - **その他**：Bash・CMake・Assembly・MATLAB・Java・Objective C/C++・Swift。

- **言語**
  - **中国語**（母語）。
  - **英語**（業務での使用可。TOEFL: 106/120・GRE: 324/340）。
  - **日本語**（基本業務に対応可。JLPT N1: 173/180, N2: 169/180）。

プロジェクト例

UGrid：収束保証ありの高効率的なニューラルマルチグリッドPDEソルバー

- TL;DR: UGridは、機械学習をベースに、収束保証ありの効率的なニューラルマルチグリッドPDEソルバー。
- ユーネット（U-Net）とマルチグリッド法を融合し、伝統的なソルバーに比べて、同じ精度（1e-5の残差）で速度を20倍ブースト。不規則な境界形状やトポロジーへの汎化能力を持ち、再学習を必要とせずにスケーラビリティも備えています。
- 関与技術：収束性に関する数値解析、カスタマイズAIオペレーター（PythonおよびCUDAベース）。
- オープンソース：https://github.com/AXIHIXA/UGrid

2DMamba：ハードウェア効率ありの並列2D Mambaスキャンナー

- TL;DR: 2DMambaスキャンナーは1D Mambaを2Dに拡張し、モデリング能力を高める同時に、速度やメモリー効率も最適化へ。
- 1D Mambaスキャン操作を2Dに拡張しつつ、トレーニングおよび推論効率を維持します。単純的な実装に比べて、スループットは10倍に達し、GPUメモリ消費はわずか10%になります。
- 関与技術: ワープシャッフルと並列スキャン、2Dタイル処理とキャッシング、HBMアクセスの最適化、CUDAカーネルおよびAIモデルのプロファイリング、PyTorch CUDAエクステンションのカプセル化。
- オープンソース：https://github.com/AtlasAnalyticsLab/2DMamba （CUDAエクステンションの部分）

GPUアルゴリズム・CUDAオペレーターの実装・最適化実験

- Parallel reduction（loop unrolling and warp shuffle primitives)。
- Histogram and Copy-If（原子オペレーションを使用）。
- Parallel scan（WarpScan・Rakingの二つの方法）。
- Fused Biased-Mask-Scale-Add（fp32とfp16）。
- SGEMM and GEMV（Loop unrolling・SMEM padding・warp tiling・double buffer optimizations・cuBLASの90%の効率を達成）。
- Dropout（cuRAND APIを使用）。
- Fused SoftMax/LayerNorm/RMSNorm・Im2Col・Matrix transpose・その他。
- オープンソース：https://github.com/AXIHIXA/CudaDemo