Deep Model Reuse: Paving the Way for Efficient and Generalizable AI Systems

Xingyi Yang

The Hong Kong Polytechnic University
xingyi.yang@polyu.edu.hk

Usually, in deep learning, we develop a separate model for each task, which requires substantial computational resources and data. Although versatile foundation models exist, they are still expensive to train and deploy. To address these challenges, we propose **deep model reuse**. Instead of building from scratch, we leverage a library of pre-trained models by understanding their behavior, composing their expertise, and improving their reliability. It reduces costs while increasing flexibility. We show three key strategies for model reuse and discuss its applications in generative AI.

Reusing Strategies

Understanding the Behavior of Neural Networks.
My research begins with understanding the inherent knowledge and behavior of pre-trained models. For example, we use LLMs to understand the decision path within neural network (Yang and Wang 2024b). Additionally, we investigate the features of diffusion models by exploring its connection with discriminative models (Yang and Wang 2023). This understanding allows for later reuse of these models in new contexts, maximizing their potential.

Transforming and Composing Models.
Given the understanding, we hope to adapt models to new scenario while enhancing their performance. To do this, we design three key methods: *weight mapping* (Yang and Wang 2025b,a) to adjust the weight and load into another model; *knowledge distillation* to transfer knowledge from a "teacher" to a "student" model (Yang, Ye, and Wang 2022); and *dissecting and reassembling* (Yang et al. 2022) to break down models into components and then combine them for new tasks. These strategies enhance flexibility and efficiency, enabling reuse of existing networks with fewer resources.

Reliability and Behavior Editing.
Beyond adaptability, we also want to mitigate biases within AI models. In this direction, we study vulnerabilities in foundation models (e.g., Segment Anything) (Lu, Yang, and Wang 2024). Furthermore, we edited the behavior of generative models, enhancing their performance under complex object interactions (Yang and Wang 2024a; Wu, Yang, and Wang 2024). My goal is to detect errors and fix them, which ensures models are reliable and robust in diverse environments.

Applications in Generative Models

Generative models drive AI content creation. My research applies reuse principles to enhance their efficiency and compositional understanding.

Case 1: Efficient Diffusion Model.
My work explores how model reuse techniques, including knowledge distillation (Yang et al. 2023) and feature reuse (Yang, Liu, and Wang 2025), could accelerate diffusion models. For example, we found that compressed diffusion models often suffer from spectrum bias. To address this, we developed an efficient architecture and a knowledge distillation pipeline. It improve efficiency while maintaining visual quality.

Case 2: Compositional Understanding and Generation.
The next goal is to ensure the generative models to produce outputs with complex, compositional structures. This means the output should accurately reflect multiple objects and their spatial relationships in the input. To address this, we collect relevant data (Wu, Yang, and Wang 2024) and regulate the information (Yang and Wang 2024a) during the generation process. As a result, the generated outputs are both structurally complex and contextually accurate.

Case 3: Reusing 2D/3D model to perform 3D/4D tasks.
Generating 3D/4D content typically requires large computation and data. To overcome this, we repurpose pre-trained 2D/3D models for 3D/4D tasks. For example, we employed 2D segmentation model to sperate 3D objects (Shen, Yang, and Wang 2024), or used depth and optical flow from video to recover its 4D dynamics (Wang et al. 2025).

References

Lu, J.; Yang, X.; and Wang, X. 2024. Unsegment Anything by Simulating Deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24294–24304.

Shen, Q.; Yang, X.; and Wang, X. 2024. FlashSplat: 2D to 3D Gaussian Splatting Segmentation Solved Optimally. *European Conference of Computer Vision*.

Wang, S.; Yang, X.; Shen, Q.; Jiang, Z.; and Wang, X. 2025. GFlow: Recovering 4D World from Monocular Video. *The AAAI Conference on Artificial Intelligence*.

Wu, Y.; Yang, X.; and Wang, X. 2024. Relation Rectification in Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7685–7694.

Yang, X.; Liu, S.; and Wang, X. 2025. Hash3D: Training-free Acceleration for 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21481–21491.

Yang, X.; and Wang, X. 2023. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18938–18949.

Yang, X.; and Wang, X. 2024a. Compositional Video Generation as Flow Equalization. arXiv:2407.06182.

Yang, X.; and Wang, X. 2024b. Language Model as Visual Explainer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yang, X.; and Wang, X. 2025a. Kolmogorov-Arnold Transformer. In *The Thirteenth International Conference on Learning Representations*.

Yang, X.; and Wang, X. 2025b. Neural Metamorphosis. In *European Conference on Computer Vision*, 1–19. Springer.

Yang, X.; Ye, J.; and Wang, X. 2022. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, 73–91. Springer.

Yang, X.; Zhou, D.; Feng, J.; and Wang, X. 2023. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 22552–22562.

Yang, X.; Zhou, D.; Liu, S.; Ye, J.; and Wang, X. 2022. Deep model reassembly. *Advances in neural information processing systems*, 35: 25739–25753.