MobileSAM2: Lightweight Segment Anything for Spatial Intelligence

Kai Jiang[1], Jiaxing Huang[1,⋆], Jingyi Zhang[2], Weiying Xie[3], Yunsong Li[3], Yufei Wang[4], Aoran Xiao[2], and Dacheng Tao[2]

[1] Hong Kong Polytechnic University, HK, China
[2] Nanyang Technological University, Singapore
[3] Xidian University, China
[4] SparcAI Inc., USA
xdjiangkai@foxmail.com, jiaxing.huang@polyu.edu.hk

Abstract. The recent large video foundation model, SAM2, enables segment anything in both images and videos, serving as a powerful base model for various applications. However, many of such use cases require to operate on resource-constrained devices like mobile phones and laptops. In this work, we aim to make SAM2 more mobile-friendly by distilling the heavyweight SAM2 into a lightweight model, facilitating segment anything in both images and videos on mobile devices. To this end, we propose Hypergraphical Knowledge Distill (HyperKD), which introduces the idea of hypergraph into knowledge distillation, aiming to effectively model and transfer SAM2's generalizable and comprehensive knowledge. HyperKD consists of Temporal HyperKD and Granularity HyperKD that construct hypergraphs to explicitly model and extract the generalizable temporal knowledge and the comprehensive multi-granularity knowledge from SAM2 respectively, which are then distilled into the lightweight student model by aligning it with the constructed hypergraphs. Besides, we present MobileSAM2, a new family of lightweight SAM2 that balances efficiency and effectiveness via searching the best model architectures with HyperKD during model size reduction. Extensive experiments validate MobileSAM2 across multiple benchmarks and show promising generalization performance on embodied AI tasks.

Keywords: Segment anything model · Knowledge distillation · Lightweight model · Spatial intelligence · Embodied AI

1 Introduction

Large vision foundation models [26, 35, 50, 51] trained with large-scale visual data have achieved significant success across various areas of computer vision. A prominent example is Segment Anything Model (SAM) [35], the first foundation model for image segmentation, which learns over 11 million images with 1.1 billion mask annotations and thus enables segmenting anything. Recently, SAM2 [51]

⋆ Corresponding author.

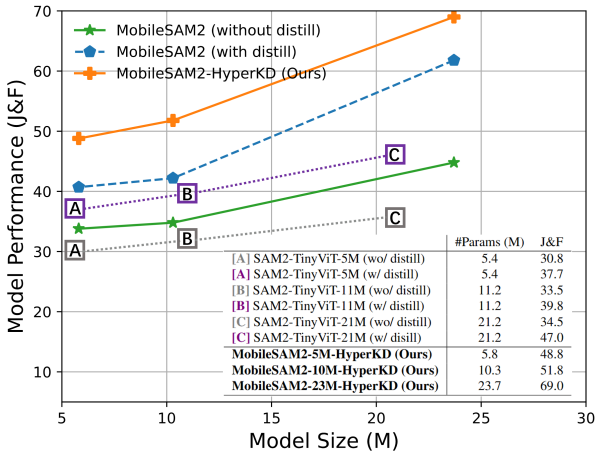

Fig. 1: Video Segmentation Comparison[5]**.** With the proposed HyperKD and the searched model architectures, our MobileSAM2 works effectively with limited parameters.

has extended SAM to video segmentation while keeping the image segmentation ability intact. SAM2 learns rich temporal and segmentation knowledge from an unprecedented dataset with 113.8K videos and 40.9M mask annotations (e.g., SA-V manual and internal data [51]) and outperforms traditional methods by substantial margins, showing great flexibility in segmenting anything in both still images and video streams and demonstrating strong generalization and zero-shot

Given its impressive capabilities, SAM2 has been employed as a powerful foundation model for a wide range of computer vision applications, such as video editing [5, 42], traffic surveillance [73, 77], autonomous vehicles [12, 47], robotics [11, 27], etc. On the other hand, many of these use cases require to operate on resource-constrained devices [31, 36, 55] like mobile phones, laptops, drones, robots, etc. In this work, we aim to make SAM2 more mobile-friendly by distilling it into a lightweight model, facilitating segmenting anything on mobile devices for both images and videos.

We begin by examining the key factors behind SAM2's success. Compared to previous video segmentation methods, SAM2, for the first time, scales up video training data to an unprecedented level, both in the number of videos and the granularity of mask annotations, learning from which two critical types of knowledge including (1) Temporal Knowledge: SAM2 captures rich temporal correlations across video frames by learning from over 100K videos (∼50 times of the previous largest dataset), ultimately acquiring generalizable temporal knowledge that enables robust segmentation along video frames; and (2) Multi-Granularity Knowledge: SAM2 captures diverse granularity correlations across multiple granularity levels of visual concepts from over 40M multi-level mask annotations (e.g, objects as well as parts and subparts), finally learning comprehensive

[5] SAM2-TinyViT denotes SAM2 using TinyViT [62] as backbones. Please refer to Table 2 for experiment details.

multi-granularity knowledge that enables comprehensive and coarse-to-fine understanding and segmentation of various visual concepts in videos. Inspired by these insights, we believe that one effective method of distilling SAM2's knowledge is to transfer such two types of knowledge comprehensively, such that the lightweight student model can inherit SAM2's robust and comprehensive video segmentation capabilities.

To this end, we propose Hypergraphical Knowledge Distill (HyperKD), which introduces the idea of hypergraph into knowledge distillation, aiming to effectively model and transfer SAM2's generalizable and comprehensive knowledge into a lightweight model. Since these two types of knowledge are implicitly encoded in SAM2's learnt parameters and thus difficult to directly access, HyperKD constructs hypergraphs to explicitly model and extract the generalizable temporal knowledge and the comprehensive multi-granularity knowledge from SAM2, which are then distilled into the lightweight student model by aligning it with the constructed hypergraphs.

We instantiate HyperKD with two examples, named Temporal HyperKD and Granularity HyperKD. Temporal HyperKD first considers patches/objects within a single video frame as nodes and the distances between their embeddings as edges, and then constructs a hypergraph by linking multiple related nodes across frames via hyperedges that are formed by fusing edges from each individual frame. Granularity HyperKD first treats segmentation entities within a single segmentation granularity level as nodes and the distances between their embeddings as edges, and then builds a hypergraph by linking multiple related nodes across segmentation granularity levels (e.g., objects, parts and subparts) via hyperedges that are formed by fusing edges from each individual segmentation granularity level.

Besides, based on our proposed HyperKD, we introduce MobileSAM2, a new family of lightweight segment anything models, which strikes a favorable trade-off between model size and performance. Specifically, we adopt a progressive model contraction strategy [53] to iteratively scale down a large model, generating a series of lightweight segment anything models by searching and identifying model architectures that best retain SAM2's temporal and multi-granularity knowledge (i.e., those can minimize HyperKD losses) during the model contraction process.

Our MobileSAM2 offers three desirable features: (1) Robust Temporal Segmentation: Temporal HyperKD models and transfers SAM2's temporal knowledge, enabling MobileSAM2 to capture rich correlations across video frames and achieve robust segmentation along video frames; (2) Comprehensive Multi-Granularity Segmentation: Granularity HyperKD enables MobileSAM2 to capture diverse correlations across multiple segmentation granularity levels, allowing for comprehensive and coarse-to-fine understanding and segmentation; (3) Efficiency&Effectiveness: By searching for the best model architectures via HyperKD, the resulting MobileSAM2 strikes an effective balance between model size and performance.

The main contributions of this work are threefold. First, we introduce the idea of hypergraph for knowledge distillation to explicitly model and transfer

SAM2's generalizable and comprehensive knowledge into a lightweight model. To our knowledge, this is the first work that explores hypergraph for video segmentation knowledge distillation. Second, we design HyperKD, including Temporal HyperKD and Granularity HyperKD that distill temporal and multi-granularity knowledge effectively. Third, we present MobileSAM2, a new family of lightweight segment anything models that effectively balances model size and performance. Fourth, extensive experiments demonstrate MobileSAM2's superior performance across multiple benchmarks.

2 Related Works

2.1 Segment Anything Model

Segmentation Anything Models (SAMs) have shown impressive zero-shot generalization and interactive segmentation across diverse datasets. Models like SAM [35], SEEM [81], and Semantic-SAM [38] advance segmentation by using a promptable architecture that processes spatial (points, boxes, masks) and/or semantic (text) prompts to generate segmentation masks. Fast-SAM [79], MobileSAM [74], and TinySAM [54] enhance inference speed, while HQ-SAM [32] improves segmentation quality. SAM2 [51] extends SAM [35] to video segmentation with a memory-based transformer, enabling robust and comprehensive segmentation across video frames by storing object information and handling past interactions. However, SAM2's large parameters and high computational demands make deployment on resource-constrained devices challenging. Compressing and accelerating SAM2 is a largely under-explored problem. We focus on distilling SAM2 into a lightweight model, enabling efficient segmentation on mobile devices for both images and videos.

2.2 Knowledge Distillation

Knowledge distillation (KD) [20] aims to train compact and efficient student models under the guidance of large teacher models. By learning from the teacher's soft labels, the student can outperform models trained only on hard labels. KD methods are broadly categorized into logit-based, feature-based, and relation-based approaches. Logit-based KD [6, 23, 30, 46, 76, 78] minimizes a distance metric, such as KL divergence, between the predicted probabilities of the student and teacher, typically derived from the softmax of logit outputs. Feature-based KD [7, 8, 21, 22, 40, 41, 43, 52, 56, 57, 70] uses intermediate feature maps or refined information as knowledge. Relation-based KD [39, 45, 48, 49, 58, 67] aligns instance correlations between the student and teacher networks. In this work, we aim to distill SAM2's generalizable temporal and multi-granularity knowledge learnt from over 100K videos and 40M annotations by capturing and encapsulating coarse-to-fine semantic relationships across frames, achieving robust and comprehensive video segmentation in a lightweight MobileSAM2.

2.3 Hypergraph

Hypergraph [1, 14–16, 18, 19, 19, 28, 34, 37, 61] is an advanced data structure that captures complex, high-order associations by allowing hyperedges to connect multiple nodes simultaneously. This capability makes hypergraphs particularly effective for modeling intricate relationships that traditional pairwise graph structures cannot adequately represent. Hypergraphs have been widely applied across domains such as social network analysis [66, 71], drug-target interaction modeling [29,59], and brain network analysis [63,82], where multi-way interactions are crucial for understanding underlying structures. In this work, we design a hypergraphical knowledge distillation method that constructs temporal and granularity hypergraphs to distill SAM2's generalizable knowledge learnt from a vast dataset. This approach effectively uncovers coarse-to-fine, high-order semantic relationships of various visual concepts across video frames, resulting in a lightweight MobileSAM2 with robust and comprehensive video segmentation capabilities.

3 Methodology

This section presents MobileSAM2, a new family of tiny and efficient segment anything models with effective and efficient distillation on video segmentation. We first introduce preliminaries in Section 3.1. Then we introduce the Hypergraphical Knowledge Distillation (HyperKD) framework for small segment anything model training in Section 3.2. After that, we design a new tiny model family that balances efficiency and effectiveness by progressively scaling down a large seed model in Section 3.4.

3.1 Preliminaries

Segment Anything Model 2 (SAM2) [51] is a unified model for video and image segmentation. SAM2 is designed to segment anything by utilizing a "Promptable Segmentation" scheme, where the model generates a segment of spatio-temporal mask (i.e., a 'masklet') based on given prompts on any frames of the video, such as points, boxes, or masks. SAM2 can segment the object of interest in single image and across video frames when appropriate prompts are provided. Additionally, SAM2 support "interactive segmentation", which allows users to iteratively refine segments and scale up training data through model-in-the-loop annotation. This process results in a more powerful SAM2 that can handle diverse segmentation tasks with improved flexibility and accuracy.

SAM2 typically works with two stages. It first encodes an input frame and a set of prompts into feature embeddings and then predicts the expected segmentation mask conditioned on these features and the memory context from previously observed frames. Specifically, SAM2 consists of an image encoder $\mathbf{Encoder}^I$, a memory module **MemModule**, a prompt encoder $\mathbf{Encoder}^P$, and a mask decoder **Decoder**. Given an input video $x^I \in \mathbb{R}^{T\times H\times W\times 3}$, where T is the number

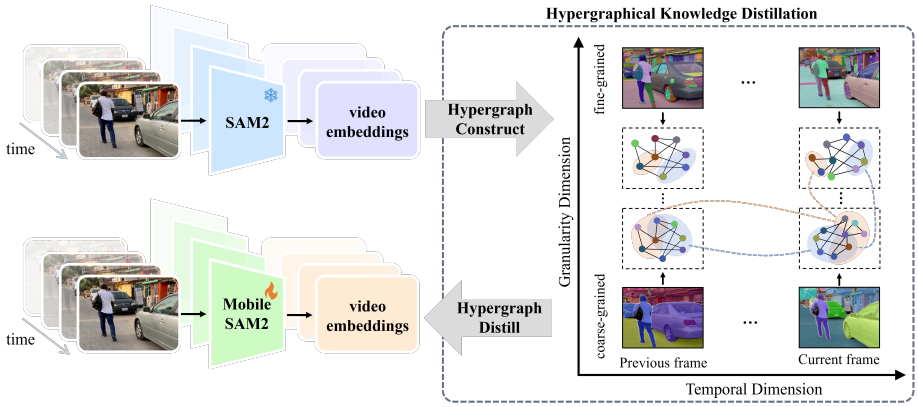

Fig. 2: Overview of HyperKD. HyperKD constructs hypergraphs to explicitly model and extract the generalizable temporal knowledge and the comprehensive multi-granularity knowledge from SAM2, which are then distilled into lightweight MobileSAM2 by aligning it with the constructed hypergraphs. Specifically, Temporal HyperKD considers objects within a single frame as nodes and constructs hypergraphs by linking multiple related nodes across frames via hyperedges, as shown along the temporal dimension. Granularity HyperKD treats segmentation entities within a single granularity level as nodes and builds hypergraphs by linking multiple relevant nodes across granularity levels (e.g., objects, parts and subparts) via hyperedges, as shown along the granularity dimension. In this way, the trained MobileSAM2 captures rich and diverse hypergraphical correlations across multiple frames and granularity levels, ultimately achieving robust and comprehensive video understanding and segmentation.

of frames, and a prompt x^P on the first frame, SAM2 first encodes x^I and x^P into embeddings as follows:

$$z^I = \textbf{Encoder}^I(x^I),\ z^P = \textbf{Encoder}^P(x^P). \tag{1}$$

The memory module then conditions the i-th frame embedding $z^I(i)$ on the past N frames embeddings $\{z^I(j)\}_{j=i-N}^{i}$ and predictions $\{m(j)\}_{j=i-N}^{i}$ as follows:

$$z_m^I(i) = \textbf{MemModule}\left(z^I(i) \mid \{z^I(j)\}_{j=i-N}^{i-1}, \{m(j)\}_{j=i-N}^{i-1}\right), \tag{2}$$

where $\{z^I(i)\}_{i=t-N}^{t-1}$ and $\{m(i)\}_{i=t-N}^{t-1}$ interact with $z^I(t)$, leading to a conditioned frame embedding z_m^I.

The mask decoder **Decoder** then decodes the conditioned frame embedding $z_m^I(t)$ conditioned on the prompt embedding z^P:

$$m(i), c(i) = \textbf{Decoder}\left(z_m^I(i) \mid z^P\right), \tag{3}$$

where z^P interacts with $z_m^I(i)$, producing binary segmentation predictions $m \in \mathbb{R}^{T\times M\times H\times W}$ that include M valid masks, each corresponding to different levels of segmentation granularity, along with the corresponding confidence scores $c \in \mathbb{R}^{T\times M}$.

Naïve Solution with FitNet [52]. In this paper, we adopt SAM2 [60] as the pretrained teacher model, using Hiera-L [53] as image encoder. We employ

the FitNet approach [52] for knowledge distillation, where a teacher model guides a student model by aligning the student's intermediate feature representations from the image encoder with those from the teacher. Given an unlabeled video x^I, the teacher model first generates feature embeddings through its image encoder, which serve as alignment targets for the corresponding layers in the student model's image encoder. This alignment of intermediate representations enhances knowledge transfer in our distillation setup. The unsupervised training of the student model on unlabeled data is formulated as:

$$\text{Loss} = \|z_t^I - z_s^I\|_1, \tag{4}$$

which aligns the intermediate feature embedding z_t^I and z_s^I from the teacher's image encoder $\mathbf{Encoder}_t^I$ and the student's image encoder $\mathbf{Encoder}_s^I$, respectively.

3.2 Hypergraphical Knowledge Distillation

We observe that SAM2's success is driven by its captured two key types of knowledge from large-scale, multi-granular video data: temporal knowledge for robust segmentation across frames and multi-granularity knowledge for detailed understanding of visual concepts at multiple levels. Motivated by this, we propose Hypergraphical Knowledge Distillation (HyperKD), which uses hypergraphs in knowledge distillation to effectively capture and encapsulate above two types of knowledge into a lightweight model. Since SAM2's temporal and multi-granularity knowledge are implicitly encoded in its parameters, HyperKD explicitly models them through hypergraphs, which the student model learns to emulate. We instantiate HyperKD with two methods: Temporal HyperKD that links related nodes across frames to capture temporal consistency, and Granularity HyperKD that connects nodes across segmentation levels to convey multi-level understanding.

Temporal HyperKD Temporal HyperKD works in a two-step manner. The first step is temporal hypergraph extraction with node and hyperedge construction that aims to uncover the implicitly encoded generalizable temporal knowledge in SAM2. The second step is temporal hypergraph encapsulation, which encapsulates the extracted temporal hypergraph into the student model, enabling the student model to generate more temporally consistent segmentation.

Temporal Hypergraph Extraction. With the generalizable temporal knowledge contained in video embedding z_t^I from teacher model, we formulate the proposed Patch-wise Temporal Hypergraph and Instance-wise Temporal Hypergraph as $\mathcal{G}_t^{Pa} = (\mathcal{V}_t^{Pa}, \mathcal{E}_t^{Pa})$ and $\mathcal{G}_t^{Ins} = (\mathcal{V}_t^{Ins}, \mathcal{E}_t^{Ins})$, which are capable of capturing temporal semantic relationships and associations across frames. For $\mathcal{G}_t^{Pa}$, we deconstruct the video embedding z_t^I as grid-based patches to constitute vertex set $\mathcal{V}_t^{Pa}$. For $\mathcal{G}_t^{Ins}$, we constitute vertex set $\mathcal{V}_t^{Ins}$ by pooling the video embedding z_t^I according to high confidence segmentation entries of each frame in teacher model prediction m_t.

Without loss of generality, to establish hyperedges that model relationships among vertices in $\mathcal{V}$, we apply an ϵ-ball distance threshold around each vertex. An

ϵ-ball forms a hyperedge that includes all vertices within a certain distance ϵ from a central vertex. The set of hyperedges $\mathcal{E}_t^{Pa}$ or $\mathcal{E}_t^{Ins}$ are therefore constructed as follows:

$$\mathcal{E} = \{\text{ball}(v, \epsilon) \mid v \in \mathcal{V}\}, \tag{5}$$

where $\text{ball}(v, \epsilon) = \{u \mid \text{dist}(x_u^I - x_v^I) < \epsilon,\ u \in \mathcal{V}\}$ represents the neighboring vertex set of a given vertex v, $\mathcal{V}$ is $\mathcal{V}_t^{Pa}$ or $\mathcal{V}_t^{Ins}$, and $\text{dist}(x_u^I - x_v^I)$ is the distance function used to determine proximity. In practical, the hypergraph $\mathcal{G}$ is represented by its incidence matrix H, where $H(u, v) = \text{dist}(x_u^I - x_v^I) = 1 - \langle x_u^I, x_v^I \rangle / (\|x_u^I\| \|x_v^I\|)$ if vertex u is part of the hyperedge centered at vertex v, and $H(u, v) = 0$ otherwise.

Temporal Hypergraph Encapsulation encapsulates both patch-level and instance-level generalizable temporal knowledge captured by the extracted hypergraph $\mathcal{G}_t^{Pa}$ and $\mathcal{G}_t^{Ins}$ into the student model, thereby preserving SAM2's generalizable temporal knowledge within the student model. Specifically, similar to the construction of $\mathcal{G}_t^{Pa}$ and $\mathcal{G}_t^{Ins}$, we construct $\mathcal{G}_s^{Pa}$ and $\mathcal{G}_s^{Ins}$ for video embeddings z_s^I from student model. Then, we encapsulate the extracted temporal hypergraph $\mathcal{G}_t^{Pa}$ and $\mathcal{G}_t^{Ins}$ into student model by minimizing:

$$\mathcal{L}_{\text{THKD}} = \|H_t^{Pa} - H_s^{Pa}\|_{\text{F}} + \|H_t^{Ins} - H_s^{Ins}\|_{\text{F}}, \tag{6}$$

where H_t^{Pa}, H_t^{Ins}, H_s^{Pa}, and H_s^{Ins} refer to the incidence matrices of $\mathcal{G}_t^{Pa}$, $\mathcal{G}_t^{Ins}$, $\mathcal{G}_s^{Pa}$, and $\mathcal{G}_s^{Ins}$, respectively.

In this way, Temporal HyperKD ensures that both patch-level and instance-level temporal hypergraphs capture neighborhood relationships within the feature space, effectively distilling SAM2's generalizable temporal knowledge into the student model, enabling the student model to maintain temporal consistency and achieve robust segmentation across video frames.

Granularity HyperKD As Temporal HyperKD distills only the generalizable temporal knowledge, we further design Granularity HyperKD to extract a granularity hypergraph and encapsulate it into the student model to improve segmentation precision across different levels of detail. Specifically, the granularity hypergraph captures hierarchical relationships between multi-granularity level segmentation entities, complementing the temporal hypergraph by providing orthogonal and multi-granularity knowledge.

Granularity Hypergraph Extraction. Given the i-th frame embedding $z_t^I(i) \in \mathbb{R}^{D \times h \times w}$ and its predicated binary segmentation masks $m(i) \in \mathbb{R}^{M \times H \times W}$ with M levels of segmentation granularity from teacher model, we instantiate granularity hypergraph as $\mathcal{G}_t^{Gran} = (\mathcal{V}_t^{Gran}, \mathcal{E}_t^{Gran})$, in which each node $v \in \mathcal{V}_t^{Gran}$ is initialized by pooling the frame embedding according to predicated binary segmentation masks as follows:

$$\mathcal{V}_t^{Gran} = \left\{ \textbf{MaskedAveragePooling}\left(z_t^I(i,j), m(i,j)\right) \mid j = 1, \ldots, M \right\}, \tag{7}$$

and the set of hyperedges $\mathcal{E}_t^{Gran}$ can be constructed with Eq. 5.

Granularity Hypergraph Encapsulation encapsulates the orthogonal and multi-granularity knowledge in the extracted granularity hypergraph into

the student model. This process complements the temporal hypergraph and enhances segmentation precision by enabling the model to segment objects at various levels of detail, such as objects, parts, and subparts. Specifically, similar to the construction of $\mathcal{G}_t^{Gran}$, we construct $\mathcal{G}_s^{Gran}$ for the video embeddings z_s^I from the student model, using the predictions $m(i)$ from the teacher model. We then encapsulate the extracted granularity hypergraph into the student model, following a similar approach as in Temporal Hypergraph Encapsulation, by minimizing the following loss function:

$$\mathcal{L}_{\text{GHKD}} = \|H_t^{Gran} - H_s^{Gran}\|_{\text{F}}, \tag{8}$$

where H_t^{Gran} and H_s^{Gran} are the incidence matrices of $\mathcal{G}_t^{Gran}$ and $\mathcal{G}_s^{Gran}$, respectively.

3.3 Overall Objective

In summary, the overall training loss of the proposed HyperKD can be formulated as:

$$\text{Loss} = \|z_t^I - z_s^I\|_1 + \alpha \cdot \mathcal{L}_{THKD} + \beta \cdot \mathcal{L}_{GHKD}, \tag{9}$$

where α and β are weighting coefficients.

3.4 Model Architectures

We introduce MobileSAM2, a lightweight version of SAM2 that strikes a balance between model size and performance. While SAM2's memory module, prompt encoder, and mask decoder are lightweight (under 12M parameters), its image encoder, based on Hiera-L [53], has over 212M parameters, making it too heavy for resource-constrained devices. Thus, the key instantiate MobileSAM2 is reducing the size of the image encoder while preserving SAM2's robust and comprehensive video segmentation capabilities.

Motivated by this observation, we present a new family of Hierarchical Vision Transformer [53] for SAM2 by scaling down a large model seed with a progressive model contraction approach [13]. Specifically, we begin with a manually designed Hiera backbone that serves as image encoder, which maintaining the core principles of the original Hiera design. Then, we establish a parameterized search space of contraction factors that includes critical architectural components such as embedding dimensions, layer counts, attention head configurations, and expansion ratios. The progressive model contraction process iteratively identifies promising architectures, ultimately resulting in a smaller Hiera variant that retains competitive performance while minimizing computational complexity and memory usage, making it suitable for efficient deployment. Hiera begins with a patch embedding layer that processes input images into tokens, followed by several stages where each stage progressively increases the channel dimensions while reducing spatial resolution through downsampling. The architecture emphasizes efficient parameter usage by employing lightweight MultiScaleBlock layers in early stages, transitioning to windowed self-attention in later stages. We consider the following contraction factors to form a model:

Table 1: Architectures of our searched lightweight MobileSAM2[6].

Contraction Factors	Γ_{Emb}	Γ_{Blk}	Γ_{WinSiz}	Γ_{HExp}	Params. (M)
MobileSAM2-5M	[48, 96, 192, 384]	[1, 2, 5, 2]	[8, 4, 14, 7]	2	5.84
MobileSAM2-10M	[64, 128, 256, 512]	[1, 2, 5, 2]	[8, 4, 14, 7]	2	10.37
MobileSAM2-23M	[96, 192, 384, 768]	[1, 3, 9, 1]	[8, 8, 14, 7]	2	23.74

- Γ_{Emb}: Embedding dimension of each stage. Decreasing it results in a thinner network with fewer heads in multi-head self-attention.
- Γ_{Blk}: The number of blocks in each stage. The depth of the model is decreased by reducing these values.
- Γ_{WinSiz}: Window size in the each stage. Smaller window size lead to fewer operations and model parameters during self-attention.
- Γ_{HExp}: Expansion ratio of attention heads in multi-head attention at each stage. Reducing the dimensionality of each attention head directly decreases the computation burden of self-attention, leading to lower computation cost.

We scale down the above factors by the progressive model contraction approach and search and identify a new family of lightweight SAM2, named MobileSAM2, as shown in Table 1.

4 Experiments

Table 2 show the benchmarking of our methods with state-of-the-art knowledge distillation methods, including FitNets [52], CIRKD [65], and FAKD [72], over 5 widely used video segmentation datasets including 2 general video object segmentation (VOS) dataset (i.e., MOSE [10], DAVIS 2017 [60]), 1 long-term VOS dataset (LVOS [24]), and 2 segment anything in videos dataset (SA-V) [51] (i.e., dataset SA-V-val [51] and SA-V-test [51]).

4.1 Implementation Details

We use our designed MobileSAM2 as the lightweight student model and SAM2-L [60] (with Hiera-L [53] as the image encoder) as the pretrained teacher model. Since SAM2's memory module (i.e., memory attention, memory encoder, and memory bank), prompt encoder and mask decoder are relatively lightweight (under 12M parameters), we optimize MobileSAM2 by training its image encoder while keeping its remaining modules (i.e., memory module, prompt encoder and mask decoder copied from the pretrained SAM2-L teacher model) frozen. We use only $\sim 10\%$ SA-V data [51] (i.e., 11K Videos) for efficient distillation training. Following SAM2 [51], we employ the AdamW [44] optimizer with an initial learning rate of 5×10^{-4}. All models are trained for 5 epochs on an A100 (80GB) GPU with a batch size of 5, using 8-frame sequences per sample. The training time is 65 hours. For each iteration, we use a 4×4 grid of spatial prompts for mask prediction by the teacher model. We set weighting coefficients $\alpha = 1$ and $\beta = 1$ to balance temporal consistency and hierarchical knowledge transfer.

[6] Please refer to appendix for more details of MobileSAM2 architecture.

Table 2: VOS comparison. With the proposed HyperKD and the searched model architectures, our MobileSAM2 works effectively with limited parameters on video segmentation. Note SAM2 Large [51] serves as the teacher model for all knowledge distill methods. For fair comparsion, all methods process video sequences.

Model	Distillation Method	Params. (M)	J&F				
			MOSE val	DAVIS 2017 val	LVOS val	SA-V val	SA-V test
100% Training Data with 256 A100 GPUs: 113.8K Videos (SA-V manual and internal data [51])							
SAM2 Base+ [51]	-	68.7	72.8	88.8	75.8	72.2	74.7
SAM2 Large [51]	-	212.1	74.6	89.2	81.7	74.5	76.0
∼ 10% Training Data with 1 A100 GPU: 11K Videos from SA-V manual data [51])							
SAM2-TinyViT-5M	No Distill	5.4	26.0	30.4	30.8	28.4	27.7
	FitNet [52]	5.4	33.8	38.7	37.7	34.3	35.1
MobileSAM2-5M	No Distill	5.8	29.7	36.6	33.8	32.7	33.0
	FitNet [52]	5.8	36.6	41.5	39.7	38.2	39.3
	CIRKD [65]	5.8	36.8	43.3	41.2	38.3	38.6
	CAT-KD [21]	5.8	35.0	41.9	40.3	38.3	39.8
	FAKD [72]	5.8	37.7	42.5	41.0	39.6	39.9
	HyperKD (Ours)	5.8	44.6	51.0	48.8	49.2	49.4
SAM2-TinyViT-11M	No Distill	11.2	27.9	29.6	32.5	28.9	30.8
	FitNet [52]	11.2	36.4	40.0	39.8	37.8	39.0
MobileSAM2-10M	No Distill	10.4	30.9	35.7	34.4	33.7	34.2
	FitNet [52]	10.4	38.9	44.3	42.2	41.6	41.4
	CIRKD [65]	10.4	40.8	45.8	43.7	42.4	43.2
	CAT-KD [21]	10.4	39.9	44.3	42.9	41.3	41.8
	FAKD [72]	10.4	41.3	46.7	44.0	42.4	43.3
	HyperKD (Ours)	10.4	48.7	54.2	51.8	49.9	50.0
SAM2-TinyViT-21M	No Distill	21.2	30.3	35.2	34.5	32.0	34.7
	FitNet [52]	21.2	44.9	49.0	47.0	45.8	45.6
MobileSAM2-23M	No Distill	23.7	39.5	44.7	44.8	42.4	44.3
	FitNet [52]	23.7	57.4	61.1	60.8	59.4	59.9
	CIRKD [65]	23.7	60.4	62.1	61.5	60.7	60.5
	CAT-KD [21]	23.7	60.2	62.1	61.1	59.7	59.9
	FAKD [72]	23.7	60.6	65.0	63.4	62.0	62.7
	HyperKD (Ours)	23.7	65.8	72.1	69.0	66.4	67.8

4.2 Results

In line with SAM2 [51], our proposed MobileSAM2 is designed for general, interactive, promptable video segmentation tasks, while we also address the semi-supervised video object segmentation (VOS) setting, where the prompt is a ground-truth mask on the first frame, as this is a common protocol in the field. We compare MobileSAM2 with existing state-of-the-art knowledge distillation methods as well as SAM2 equipped with the state-of-the-art lightweight backbone (Tiny-ViT [62]) in Table 2, reporting accuracy based on standard protocols. We evaluate three versions of MobileSAM2 with varying model sizes (5M, 10M, and 23M), each offering different size-accuracy trade-offs.

Results on general VOS dataset. Table 2 presents the video object segmentation results on two general VOS datasets: MOSE and DAVIS 2017. MobileSAM2 achieves substantial performance improvements over the baseline across these datasets. Specifically, MobileSAM2 outperforms state-of-the-art methods by a large margin in J&F score, highlighting the effectiveness of our proposed HyperKD in exploring efficient lightweight model architectures as well as

Table 3: Ablation studies of HyperKD with Temporal HyperKD and Granular HyperKD. The experiments are conducted on video object segmentation (VOS) over LVOS val.

	Temporal HyperKD		Granular HyperKD	J&F
	Patch-wise	Instance-wise		
MobileSAM2-23M (Baseline/no distill)	-	-	-	44.8
	✓			62.9
		✓		64.4
	✓	✓		64.4
			✓	63.1
MobileSAM2-23M	✓	✓	✓	69.0

distilling knowledge from the pretrained SAM2 model. The superior performance of MobileSAM2 is largely attributed to HyperKD's ability to capture generalizable temporal knowledge and comprehensive multi-granularity knowledge, enabling it to focus on cross-temporal and multi-granularity spatial cues of objects. This not only benefits the architecture search for lightweight MobileSAM2 but also enhances the distillation process from SAM2 to lightweight MobileSAM2 . All methods show performance gains. However, MobileSAM2 achieves the most significant improvements, demonstrating its capability to align closely with the SAM2's representation by effectively distilling both temporal and multi-granularity knowledge from SAM2.

Results on long-term VOS dataset. Table 2 also reports the experiments on long term VOS dataset, LVOS. MobileSAM2 surpasses all competing methods by a significant margin, demonstrating HyperKD's effectiveness and robustness in capturing SAM2's generalizable temporal and multi-granularity knowledge for long-term video segmentation. The substantial performance gains achieved by MobileSAM2 on LVOS highlight the efficiency of its lightweight architecture, explored by HyperKD's guidance, as well as the effectiveness of distilling knowledge from SAM2 by HyperKD.

Results on segment anything in videos dataset. We evaluate the effectiveness of our MobileSAM2 on segment anything in videos dataset, i.e., SA-V val and SA-V test, which measure performance for open-world segments of "any" object class [51]. Table 2 reports the VOS result, showcasing significant improvements over the baseline and outperforming state-of-the-arts thereby highlighting the superiority of MobileSAM2.

4.3 Ablation Study

In Table 3, we conduct ablation studies to assess the individual contributions of our proposed MobileSAM2 on the LVOS dataset. The baseline model, without distillation from SAM2, does not perform well due to its limited ability to capture high-order temporal and multi-granularity semantic relationships of visual concepts. By contrast, including Patch-wise Temporal HyperKD significantly improves the baseline, indicating that Patch-wise Temporal HyperKD effectively captures patch-wise temporal dependencies across frames. Additionally, applying

Table 4: Parameter analysis on hypergraph construction threshold ϵ in HyperKD on LVOS.

Model Architecture	MobileSAM2-23M				
ϵ	0.5	0.75	1.0	1.25	1.5
J&F	65.2	68.4	69.0	68.1	62.0

Table 5: Comparison of our HyperKD with traditional two-vertices graph-based knowledge distillation over LVOS.

Model Architecture	MobileSAM2-23M			
Distillation Method	No Distill	Context Matters [64]	IntRA-KD [25]	HyperKD (Ours)
J&F	44.8	60.7	60.9	69.0

Instance-wise Temporal HyperKD brings further improvements, showing that it enhances instance-level temporal knowledge across frames. Moreover, introducing Granularity HyperKD further boosts performance, demonstrating that it provides complementary multi-granularity knowledge that effectively benifits robust and comprehensive video segmentation capabilities.

4.4 Discussion

Parameter Study. In the construction of hyper graph, the predefined ϵ-ball distance threshold is used to establish hyperedges that model relationships among vertices in hypergraph. We studied ϵ by changing it from 0.5 to 1.5 with a step of 0.25. Table 4 reports the experiments over LVOS dataset. It is also observed that the performance of MobileSAM2 is relatively stable across from 0.75 to 1.25, with only minor variations, while there is a performance decline at the thresholds of 0.5 and 1.5. A higher threshold creates a more connected hypergraph, increasing the risk of over-smoothing, while a lower threshold may result in an under-connected hypergraph that fails to capture high-order relationships. To balance these factors, our HyperKD uses a distance threshold of 1.0 for hypergraph construction, chosen empirically to maintain connectivity without excessive smoothing.

HyperKD v.s. Graph-based Knowledge Distillation. We compare our HyperKD with the prior graph-based knowledge distillation method, Context Matters [64] and IntRA-KD [25]. As shown in Table 5, HyperKD outperforms Context Matters and IntRA-KD, primarily because the graphs of Context Matters and IntRA-KD cannot effectively capture high-order semantic relationships among visual concepts, which our hypergraph-based approach successfully achieves.

Distance Metrics for Constructing Hypergraph. We explore different feature distance metrics for hypergraph construction, conducting experiments with the following metrics: 1) Cosine Similarity [9], 2) Euclidean Distance [9], and 3) Manhattan Distance [9]. The results in Tab. 7 show that HyperKD performs effectively and consistently across all metrics. Among them, cosine similarity yields the best results, largely because it aligns with the attention mechanism,

Table 6: Efficiency comparison on mobile GPU.

Model	Params. (M)	Inference FPS	LVOS
SAM2 Base+	68.7	Out of Memory	-
SAM2 Large	212.1	Out of Memory	-
SAM2-TinyViT-5M	5.4	13.1	37.7
MobileSAM2-5M	5.8	13.3	48.8
SAM2-TinyViT-11M	11.2	11.3	39.8
MobileSAM2-10M	10.4	10.8	51.8
SAM2-TinyViT-21M	21.2	8.0	47.0
MobileSAM2-23M	23.7	8.2	69.0

Table 7: Analysis on hypergraph construction in HyperKD with different distance metrics on LVOS.

Model Architecture	MobileSAM2-23M		
Distance Metrics	Cosine Similarity (Ours)	Euclidean Distance	Manhattan Distances
J&F	69.0	67.2	65.7

making it a natural choice for the distillation of generalizable temporal knowledge and comprehensive multi-granularity knowledge in SAM2.

Efficiency Analysis. We evaluate model efficiency in inference speed for all models on a single RTX 4060 (8GB) mobile GPU using PyTorch 2.3.1 and CUDA 12.1 with automatic mixed precision (bfloat16). We compiled the image encoder with torch.compile for all models. FPS measurements for video object segmentation are taken with a batch size of 1, following the common protocol. Table 6 reports the results on LVOS dataset, demonstrating that MobileSAM2 achieves a significant reduction in computational overhead while maintaining competitive performance compared to the original SAM2, making it more mobile-friendly for resource-constrained devices.

Applicability of HyperKD to Other Foundation Models. We conduct experiments on TrackAnything (TAM) [68] as shown in Table 8. It demonstrates HyperKD's applicability beyond SAM2: applied to the structurally independent TrackAnything (TAM) [68] under the same limited-data setting ($\sim$10%), HyperKD improves TAM-TinyViT-21M from 58.9 to 63.1 J&F and MobileTAM-23M from 60.8 to 66.5 J&F on DAVIS-2017. More critically, it disentangles the distillation objective from the architecture design: since TAM shares no architectural lineage with SAM2 or MobileSAM2, the consistent gains confirm that the performance improvements stem from the HyperKD objective itself, not from co-adaptation with a specific backbone. Together, these findings establish HyperKD as a generalizable distillation framework, and validate that the gains on MobileSAM2 reflect genuine knowledge transfer rather than artifacts of the searched architecture.

Application of MobileSAM2 on Embodied AI. With the advent of reasoning models and agents [17, 69, 75], to verify the application of MobileSAM2 on embodied AI systems, we integrate MobileSAM2 into CaP-Agent0 [17] as the core perception model and evaluate manipulation performance on LIBERO-PRO [80].

Table 8: Applicability of HyperKD for TrackAnything (TAM) [68]. We report J&F scores on DAVIS-2017.

Model	Distillation Method	Params. (M)	DAVIS-2017 J&F
	100% *Training Data*		
TrackAnything (TAM) [68]	-	636	73.1
	$\sim$ 10% *Training Data*		
TAM-TinyViT-21M	No Distill	21.2	58.9
	HyperKD (Ours)	21.2	**63.1**
MobileTAM-23M	No Distill	23.7	60.8
	HyperKD (Ours)	23.7	**66.5**

Table 9: Application of MobileSAM2 on Embodied AI.

Method	Libero-Object		Libero-Goal		Libero-Spatial		Avg.
	Pos (Avg.)	Task (Avg.)	Pos (Avg.)	Task (Avg.)	Pos (Avg.)	Task (Avg.)	
OpenVLA [33]	0	0	0	0	0	0	0
π0 [3]	0	0	0	0	0	0	0
π0.5 [2]	17	1	38	0	20	1	12.8
SAM3 [4]-CaP-Agent0 [17]	27	31	29	16	13	23	23.2
SAM2-TinyViT-21M - CaP-Agent0	21	27	23	13	11	21	19.3
MobileSAM2-23M - CaP-Agent0	24	30	28	14	12	23	21.8

As shown in Table 9, the full SAM3-powered CaP-Agent0 achieves 23.2% average success rate, outperforming training-based VLA baselines like $\pi_{0.5}$ [2] (12.8%). At a comparable lightweight parameter scale, SAM2-TinyViT-21M built on the state-of-the-art Tiny-ViT [62] backbone drops to 19.3%. In comparison, our MobileSAM2-23M reaches 21.8% average success rate, retaining competitive performance and outperforming SAM2-TinyViT-21M on spatial reasoning tasks and embodied AI tasks. These results demonstrate that MobileSAM2 is an efficient perceptual backbone for resource-constrained robot platforms, enabling practical on-board deployment with minimal performance sacrifice.

5 Conclusion

This paper presents a new family of tiny and efficient video foundation models, MobileSAM2, distilled from large SAM2, using the proposed HyperKD. HyperKD consists of Temporal HyperKD and Granularity HyperKD that construct hypergraphs to explicitly model and extract the generalizable temporal knowledge and the comprehensive multi-granularity knowledge from SAM2 respectively, facilitating the knowledge distillation from SAM2 to MobileSAM2 as well as the model architecture searching of MobileSAM2. Extensive experiments on multiple video segmentation datasets demonstrate that MobileSAM2 consistently outperforms state-of-the-art techniques by clear margins.

Acknowledgment

[Re Prof Tao] This project is supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001. This work is also supported by PolyU Internal Fund.

References

1. Bai, S., Zhang, F., Torr, P.H.: Hypergraph convolution and hypergraph attention. Pattern Recognition **110**, 107637 (2021)
2. Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M.R., Finn, C., Fusai, N., Galliker, M.Y., et al.: $pi_{0.5}$: a vision-language-action model with open-world generalization. In: 9th Annual Conference on Robot Learning (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), `https://arxiv.org/abs/2511.16719`
5. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
6. Chen, D., Mei, J.P., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3430–3437 (2020)
7. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11933–11942 (2022)
8. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5008–5017 (2021)
9. Deza, E., Deza, M.M., Deza, M.M., Deza, E.: Encyclopedia of distances. Springer (2009)
10. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20224–20234 (2023)
11. Dupont, P.E., Nelson, B.J., Goldfarb, M., Hannaford, B., Menciassi, A., O'Malley, M.K., Simaan, N., Valdastri, P., Yang, G.Z.: A decade retrospective of medical robotics research from 2010 to 2020. Science robotics **6**(60), eabi8017 (2021)
12. Faisal, A., Kamruzzaman, M., Yigitcanlar, T., Currie, G.: Understanding autonomous vehicles. Journal of transport and land use **12**(1), 45–72 (2019)
13. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–213 (2020)
14. Feng, Y., Huang, J., Du, S., Ying, S., Yong, J.H., Li, Y., Ding, G., Ji, R., Gao, Y.: Hyper-yolo: When visual object detection meets hypergraph computation. arXiv preprint arXiv:2408.04804 (2024)
15. Feng, Y., Liu, S., Han, X., Du, S., Wu, Z., Hu, H., Gao, Y.: Hypergraph foundation model. arXiv preprint arXiv:2503.01203 (2025)
16. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3558–3565 (2019)

17. Fu, M., Yu, J., El-Refai, K., Kou, E., Xue, H., Huang, H., Xiao, W., Wang, G., Li, F.F., Shi, G., et al.: Cap-x: A framework for benchmarking and improving coding agents for robot manipulation. arXiv preprint arXiv:2603.22435 (2026)
18. Gao, Y., Feng, Y., Ji, S., Ji, R.: Hgnn+: General hypergraph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3181–3199 (2022)
19. Gao, Y., Zhang, Z., Lin, H., Zhao, X., Du, S., Zou, C.: Hypergraph learning: Methods and practices. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2548–2566 (2020)
20. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
21. Guo, Z., Yan, H., Li, H., Lin, X.: Class attention transfer based knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11868–11877 (2023)
22. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1921–1930 (2019)
23. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
24. Hong, L., Chen, W., Liu, Z., Zhang, W., Guo, P., Chen, Z., Zhang, W.: Lvos: A benchmark for long-term video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13480–13492 (2023)
25. Hou, Y., Ma, Z., Liu, C., Hui, T.W., Loy, C.C.: Inter-region affinity distillation for road marking segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12486–12495 (2020)
26. Huang, J., Zhang, J., Jiang, K., Lu, S.: Open-vocabulary object detection via language hierarchy. arXiv preprint arXiv:2410.20371 (2024)
27. Javaid, M., Haleem, A., Singh, R.P., Suman, R.: Substantial capabilities of robotics in enhancing industry 4.0 implementation. Cognitive Robotics **1**, 58–75 (2021)
28. Jiang, J., Wei, Y., Feng, Y., Cao, J., Gao, Y.: Dynamic hypergraph neural networks. In: IJCAI. pp. 2635–2641 (2019)
29. Jin, S., Hong, Y., Zeng, L., Jiang, Y., Lin, Y., Wei, L., Yu, Z., Zeng, X., Liu, X.: A general hypergraph learning algorithm for drug multi-task predictions in micro-to-macro biomedical networks. PLOS Computational Biology **19**(11), e1011597 (2023)
30. Jin, Y., Wang, J., Lin, D.: Multi-level logit distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24276–24285 (2023)
31. Kamath, V., Renuka, A.: Deep learning based object detection for resource constrained devices: Systematic review, future trends and challenges ahead. Neurocomputing **531**, 34–60 (2023)
32. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36** (2024)
33. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
34. Kim, S., Lee, S.Y., Gao, Y., Antelmi, A., Polato, M., Shin, K.: A survey on hypergraph neural networks: an in-depth and step-by-step guide. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6534–6544 (2024)

35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
36. Kornaros, G.: Hardware-assisted machine learning in resource-constrained iot environments for security: review and future prospective. IEEE Access **10**, 58603–58622 (2022)
37. Lee, G., Bu, F., Eliassi-Rad, T., Shin, K.: A survey on hypergraph mining: Patterns, tools, and generators. arXiv preprint arXiv:2401.08878 (2024)
38. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
39. Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., Liang, D.: Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 1306–1313 (2022)
40. Li, Z., Ye, J., Song, M., Huang, Y., Pan, Z.: Online knowledge distillation for efficient pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11740–11750 (2021)
41. Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G.: Knowledge distillation via the target-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10915–10924 (2022)
42. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
43. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2604–2613 (2019)
44. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
45. Mei, K., Delbracio, M., Talebi, H., Tu, Z., Patel, V.M., Milanfar, P.: Conditional diffusion distillation (2023)
46. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 5191–5198 (2020)
47. Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G.P., Cho, W.: A review on autonomous vehicles: Progress, methods and challenges. Electronics **11**(14), 2162 (2022)
48. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
49. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5007–5016 (2019)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
51. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and

videos. arXiv preprint arXiv:2408.00714 (2024), https://arxiv.org/abs/2408.00714
52. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
53. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: International Conference on Machine Learning. pp. 29441–29454. PMLR (2023)
54. Shu, H., Li, W., Tang, Y., Zhang, Y., Chen, Y., Li, H., Wang, Y., Chen, X.: Tinysam: Pushing the envelope for efficient segment anything model. arXiv preprint arXiv:2312.13789 (2023)
55. Shuvo, M.M.H., Islam, S.K., Cheng, J., Morshed, B.I.: Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. Proceedings of the IEEE **111**(1), 42–91 (2022)
56. Tian, Y., Zhang, C., Guo, Z., Zhang, X., Chawla, N.: Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency. In: The Eleventh International Conference on Learning Representations (2022)
57. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
58. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
59. Viñas, R., Joshi, C.K., Georgiev, D., Lin, P., Dumitrascu, B., Gamazon, E.R., Liò, P.: Hypergraph factorization for multi-tissue gene expression imputation. Nature machine intelligence **5**(7), 739–753 (2023)
60. Voigtlaender, P., Luiten, J., Torr, P.H., Leibe, B.: Siam r-cnn: Visual tracking by re-detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6578–6588 (2020)
61. Wang, Y., Kleinberg, J.: From graphs to hypergraphs: Hypergraph projection and its remediation. arXiv preprint arXiv:2401.08519 (2024)
62. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: European conference on computer vision. pp. 68–85. Springer (2022)
63. Xiao, L., Wang, J., Kassani, P.H., Zhang, Y., Bai, Y., Stephen, J.M., Wilson, T.W., Calhoun, V.D., Wang, Y.P.: Multi-hypergraph learning-based brain functional connectivity analysis in fmri data. IEEE transactions on medical imaging **39**(5), 1746–1758 (2019)
64. Yang, A., Lin, S., Yeh, C.H., Shu, M., Yang, Y., Chang, X.: Context matters: Distilling knowledge graph for enhanced object detection. IEEE Transactions on Multimedia (2023)
65. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12319–12328 (2022)
66. Yang, D., Qu, B., Yang, J., Cudré-Mauroux, P.: Lbsn2vec++: Heterogeneous hypergraph embedding for location-based social networks. IEEE Transactions on Knowledge and Data Engineering **34**(4), 1843–1855 (2020)
67. Yang, J., Martinez, B., Bulat, A., Tzimiropoulos, G., et al.: Knowledge distillation via softmax regression representation learning. International Conference on Learning Representations (ICLR) (2021)
68. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos (2023)

69. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. Advances in Neural Information Processing Systems **38**, 29918–29952 (2026)
70. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4133–4141 (2017)
71. Young, J.G., Petri, G., Peixoto, T.P.: Hypergraph reconstruction from network data. Communications Physics **4**(1), 135 (2021)
72. Yuan, J., Phan, M.H., Liu, L., Liu, Y.: Fakd: Feature augmented knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 595–605 (2024)
73. Zhang, B., Zhang, J.: A traffic surveillance system for obtaining comprehensive information of the passing vehicles based on instance segmentation. IEEE Transactions on Intelligent Transportation Systems **22**(11), 7040–7055 (2020)
74. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
75. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1859–1869 (October 2025)
76. Zhang, R., Shen, J., Liu, T., Liu, J., Bendersky, M., Najork, M., Zhang, C.: Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. arXiv preprint arXiv:2305.05010 (2023)
77. Zhang, X., Feng, Y., Angeloudis, P., Demiris, Y.: Monocular visual traffic surveillance: A review. IEEE Transactions on Intelligent Transportation Systems **23**(9), 14148–14165 (2022)
78. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4320–4328 (2018)
79. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
80. Zhou, X., Xu, Y., Tie, G., Chen, Y., Zhang, G., Chu, D., Zhou, P., Sun, L.: Liberopro: Towards robust and fair evaluation of vision-language-action models beyond memorization. [arXiv preprint arXiv:2510.03827] (2025)
81. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)
82. Zu, C., Gao, Y., Munsell, B., Kim, M., Peng, Z., Zhu, Y., Gao, W., Zhang, D., Shen, D., Wu, G.: Identifying high order brain connectome biomarkers via learning on hypergraph. In: Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings 7. pp. 1–9. Springer (2016)