# GeoX-2: High-Throughput Driving Geometry through Geometry-Preserving Compression

**Anonymous authors**
Paper under double-blind review

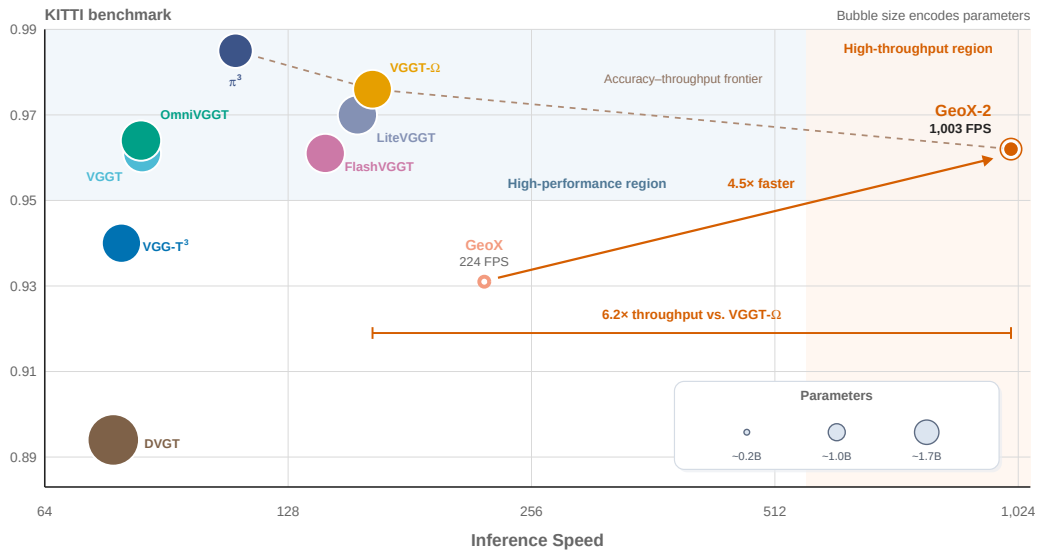


Figure 1: **Accuracy–throughput comparison on KITTI video depth.** GeoX-2 reaches 1003.1 FPS while approaching the depth accuracy of models with nearly $4\times$ more parameters, achieving a favorable balance between geometric performance and inference efficiency.

## Abstract

Feed-forward visual geometry models jointly recover camera pose, depth, and dense 3D structure, but their computational cost limits efficient scaling to synchronized driving views. We present **GeoX-2**, which compresses the complete inference graph for efficient driving geometry. Structurally, an Efficient Hybrid Encoder reduces the token count of each view, while a lightweight dense head supports efficient geometry inference. To compress the decoder, GeoX-2 treats each Frame–Global block pair as a removable unit and applies Geometry-Derived Structural Slimming. The Pair Importance Score combines LayerScale-based structural saliency with depth-intervention sensitivity to retain important pairs from the full decoder. Prior-free geometry distillation then transfers geometry from a prior-conditioned teacher to the selected compact RGB-only student, removing the conditional adapters and learned point head while recovering 3D points from predicted depth and camera poses. Extensive experiments across multiple driving benchmarks demonstrate strong geometry accuracy. GeoX-2 reaches 1003.1 FPS offline sequence-batched throughput on KITTI and provides leading throughput and peak-memory efficiency across multi-view inference settings with varying view counts.

## 1 Introduction

Feed-forward visual geometry models such as VGGT recover camera pose, depth, and dense 3D structure directly from images (Wang et al., 2024; 2025a; 2026b). This unified output space is par-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ticularly useful for autonomous driving, where a single model can organize observations from a synchronized multi-camera rig without being specialized for a particular downstream task. Driving-oriented extensions further accommodate wide camera baselines and limited overlap between adjacent views (Zuo et al., 2026; Lian et al., 2026). However, each view introduces a long token sequence that repeatedly traverses within-view and cross-view attention, making computation grow rapidly with the number of cameras.

VGGT-specific efficiency methods reduce token computation by removing, merging, or caching spatial features (Shen et al., 2025; Shu et al., 2026), but largely preserve decoder depth. Structured pruning offers a complementary direction, yet existing methods are designed mainly for large language models and rank layers by hidden-state change rather than final geometry (Men et al., 2025), leaving their suitability for VGGT-family geometry models unverified. Some methods also condition prediction on camera or depth priors (Peng et al., 2026; Lian et al., 2026). Their prior-processing paths and learned point heads add computation for geometry that can otherwise be recovered from depth and camera-pose predictions. Efficient driving geometry therefore requires compression across the complete inference graph.

Motivated by these challenges, we build GeoX-2 upon GeoX (Lian et al., 2026) to compress computation across the complete inference graph. Structurally, a hybrid encoder reduces the token count of each view, while a lightweight dense head supports efficient geometry inference. To compress the decoder, GeoX-2 treats each Frame–Global block pair as a removable unit and uses Geometry-Derived Structural Slimming, whose Pair Importance Score (PIS) combines structural saliency with depth-intervention sensitivity, to select important pairs from the full decoder. Prior-free geometry distillation then transfers geometry from the camera- and depth-conditioned teacher to the selected compact RGB-only student, removing the conditional adapters and learned point head while recovering 3D points from predicted depth and camera poses.

Figure 1 summarizes the geometry–efficiency trade-off achieved by this design. On KITTI video depth, GeoX-2 reaches 1003.1 FPS offline sequence-batched throughput with an Abs Rel of 0.071. Across five driving datasets, it achieves the best reconstruction results among the compared methods on four of the five datasets and 100.0 RRA on all four pose benchmarks. Controlled ablations further attribute the acceleration to compression of tokens, decoder pairs, and output paths, and validate the proposed pair selection and prior-free distillation. Our main contributions are:

- We introduce GeoX-2, a complete-graph compression framework that jointly reduces per-view tokens, decoder computation, and dense output paths for efficient driving geometry.
- We propose Geometry-Derived Structural Slimming, which combines LayerScale structural saliency with the geometric response to pair removal to select important Frame–Global pairs.
- We develop prior-free geometry distillation to transfer a conditioned teacher into a compact RGB-only student without conditional adapters or a learned point head, yielding 1003.1 FPS offline KITTI throughput together with competitive multi-dataset geometry results.

## 2 Related Work

**Feed-forward visual geometry.** Multi-view reconstruction has progressed from cost-volume stereo (Yao et al., 2018) to direct feed-forward geometry prediction. DUSt3R reconstructs dense 3D without calibrated cameras, MASt3R connects image matching with metric geometry, and VGGT jointly predicts camera poses, depth, and point maps (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a). Subsequent models extend this paradigm: $\pi^3$ introduces permutation equivariance, CUT3R maintains a persistent scene state, and VGGT-$\Omega$ combines register attention with lighter dense decoding (Wang et al., 2026b; 2025b; 2026a). Together, these models broaden feed-forward prediction across multi-view geometry settings.

**Geometric representations for driving.** Driving perception commonly transforms synchronized multi-camera images into bird's-eye-view features (Philion & Fidler, 2020; Huang et al., 2021; Li et al., 2024; Liang et al., 2022) or dense occupancy representations (Huang et al., 2023; Li et al., 2023; Wei et al., 2023; Tian et al., 2023). Complementing these task-oriented representations, DVGT and GeoX adapt feed-forward visual geometry to driving scenes (Zuo et al., 2026; Lian et al., 2026),

with GeoX jointly predicting depth, camera pose, and point-level structure from wide-baseline, weak-overlap views. Scaling VGGT-family geometry models to synchronized multi-view inputs requires per-view tokens to repeatedly traverse within-view and cross-view attention, causing inference cost to grow rapidly with the number of views. GeoX-2 preserves the unified geometric output space while reducing computation across the complete inference graph.

**Efficient geometry and structured slimming.** Existing methods reduce token computation through redundancy-aware processing or cached token merging (Shen et al., 2025; Shu et al., 2026). Hybrid backbones shorten spatial sequences before global modeling in general vision (Wu et al., 2021; Zhu et al., 2024; Mehta & Rastegari, 2021), but remain underexplored in feed-forward geometry. Structured pruning has meanwhile studied variable-depth training, hidden-state change, and output-distribution divergence (Fan et al., 2019; Men et al., 2025; Liu et al., 2025). These criteria do not directly target repeated Frame–Global pairs or final dense geometry. GeoX-2 instead compresses the complete inference graph by reducing per-view tokens, decoder computation, and dense output paths, while ranking removable pairs through per-view log-depth responses and sparsified LayerScale saliency.

**Distillation with geometric priors.** Knowledge distillation transfers predictive knowledge to compact students, while generalized distillation permits training-only priors that are unavailable at inference (Hinton et al., 2015; Lopez-Paz et al., 2015). In 3D perception, prior work distills sensor, BEV, or point representations (Chen et al., 2022; Wang et al., 2023; Zhou et al., 2023; Klingner et al., 2023; Zhang et al., 2023). OmniVGGT and GeoX incorporate camera and depth priors into geometry prediction, with GeoX further combining mask-guided depth-aware distillation and relative-pose relational supervision (Peng et al., 2026; Lian et al., 2026). GeoX-2 extends this setting with prior-free geometry distillation, using depth, camera-pose, point, and relational supervision from ground truth and teacher predictions. For efficient inference, the selected RGB-only student removes the camera/depth prior pathways and learned point head, then analytically unprojects its predicted depth and camera poses into 3D points.

## 3 METHOD

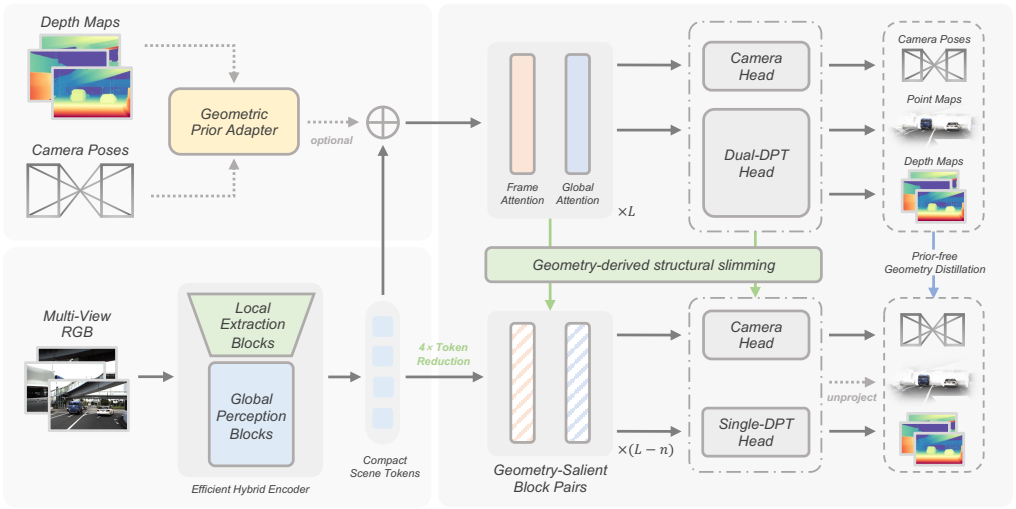


Figure 2: **Overview of GeoX-2.** The Efficient Hybrid Encoder converts multi-view RGB images into compact scene tokens. A camera- and depth-conditioned teacher is compressed by Geometry-Derived Structural Slimming, which retains salient Frame–Global block pairs. Prior-free geometry distillation then trains an RGB-only student initialized from these pairs. At inference, the student predicts depth and camera poses and recovers point maps by analytic unprojection, without prior adapters or a learned point head.

### 3.1 OVERVIEW

Figure 2 summarizes the GeoX-2 framework. Let $\mathcal{I} = \{I_i\}_{i=1}^{N}$ denote synchronized RGB views. A token-efficient encoder maps each view to a compact sequence processed by the $L$-pair teacher $F_T$, whose Frame–Global pairs alternate within-view and cross-view interaction. Camera and depth priors enter through conditional adapters, and the teacher predicts depth $D^T$, camera poses $C^T$, and point maps $P^T$. The RGB-only student $F_S$ retains $(L-n)$ selected pairs and predicts depth $D^S$ and camera poses $C^S$, from which it reconstructs the point map $P^S = \Pi(D^S, C^S)$. Its deployed graph contains neither prior-conditioned adapters nor a learned point head.

Training proceeds in three stages. First, the conditioned teacher learns coupled depth, camera, and point geometry. Second, LayerScale sparsity provides pair-level structural saliency, while depth interventions measure each pair's effect on final geometry and determine the retained subnetwork. Third, the retained weights initialize an RGB-only student trained by prior-free geometry distillation with ground-truth and teacher supervision for depth, camera, points, and inter-view relations.

### 3.2 TOKEN-EFFICIENT HYBRID ENCODER

Each Frame–Global pair repeatedly processes tokens from every view, making the encoder sequence length a direct multiplier of decoder cost. We therefore use an Efficient Hybrid Encoder (Mu et al., 2026) to extract hierarchical features and output a compact stride-32 grid. For $I_i \in \mathbb{R}^{H\times W\times 3}$,

$$Z_i = E_{\text{hybrid}}(I_i), \qquad Z_i \in \mathbb{R}^{(H/32)\times(W/32)\times C_4}. \tag{1}$$

As shown in Figure 3, the four stages contain $(3, 6, 16, 8)$ blocks with channel widths $(C_1, C_2, C_3, C_4) = (80, 160, 320, 768)$. Stages 1–2 use RepConv blocks for local extraction, followed by Adaptive Bidirectional Mamba blocks in Stage 3 and Transformers in Stage 4 for global modeling (Ding et al., 2021; Zhu et al., 2024). Stage-wise downsampling progressively reduces the spatial grid, yielding at most 256 tokens per view for a $512 \times 512$ input.

At the same input scale, GeoX's DINOv2-Small produces up to 1,024 tokens per view; the Efficient Hybrid Encoder therefore reduces every subsequent decoder sequence to one quarter. Before geometry training, the encoder is pretrained on approximately 15 billion image–text pairs with a CLIP-style contrastive objective and final-feature MSE distillation from DINOv2-Large and SigLIP-Large (Radford et al., 2021; Oquab et al., 2023; Zhai et al., 2023):

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{DINO}}\mathcal{L}_{\text{MSE}}^{\text{DINO}} + \lambda_{\text{SigLIP}}\mathcal{L}_{\text{MSE}}^{\text{SigLIP}}. \tag{2}$$

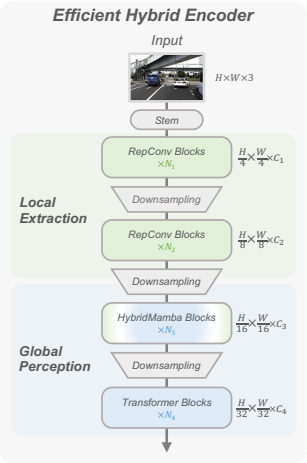


Figure 3: **Token-efficient hybrid encoder.** Local extraction and global modeling progressively reduce each $512 \times 512$ image to a compact $16 \times 16$ token grid.

### 3.3 GEOMETRY-DERIVED STRUCTURAL SLIMMING

Token reduction does not shorten the decoder. We therefore treat each complete Frame–Global pair, comprising the Frame and Global self-attention and feed-forward residual branches, as a removable unit. Pair-wise removal preserves regular execution and yields a static inference graph. Figure 4 summarizes how complementary structural and output-space evidence ranks these pairs.

**Structural saliency from LayerScale.** For pair $l$, denote the LayerScale vectors (Touvron et al., 2021) of Frame Self-Attention, Frame FFN, Global Self-Attention, and Global FFN by $\gamma_l^{\text{F-SA}}$, $\gamma_l^{\text{F-FFN}}$, $\gamma_l^{\text{G-SA}}$, and $\gamma_l^{\text{G-FFN}}$. We concatenate them as

$$\mathbf{g}_l = \text{Concat}\big(\gamma_l^{\text{F-SA}}, \gamma_l^{\text{F-FFN}}, \gamma_l^{\text{G-SA}}, \gamma_l^{\text{G-FFN}}\big). \tag{3}$$

Starting from the trained $L$-pair teacher, LayerScale sparsity training jointly updates the LayerScale parameters and the rest of the network using

$$\mathcal{L}_{\text{sparse}} = \mathcal{L}_{\text{task}} + \lambda_s(t)\sum_{l=0}^{L-1}\|\mathbf{g}_l\|_1, \tag{4}$$

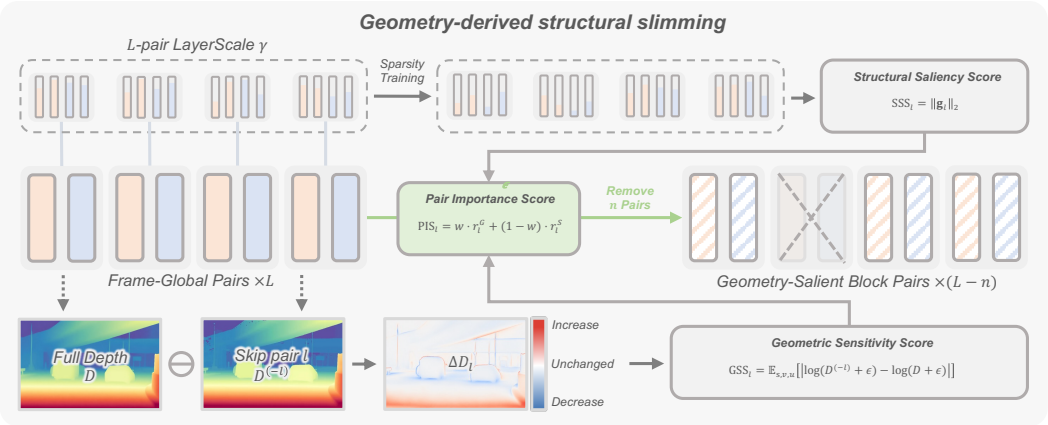


Figure 4: **Geometry-derived structural slimming.** LayerScale sparsity exposes pair-level structural saliency, while independent pair bypasses measure geometric sensitivity from the final-depth response. Their ranked fusion removes $n$ pairs and retains the remaining $(L-n)$ in their original order.

where $\lambda_s(t)$ is linearly warmed up during sparsity training. We define the Structural Saliency Score as

$$\mathrm{SSS}_l = \|\mathbf{g}_l\|_2. \tag{5}$$

A low SSS indicates weak aggregate residual scaling across the pair. Because this parameter-space evidence does not directly measure output geometry, we complement it with a depth intervention.

**Geometric sensitivity from depth intervention.** Let $D_{s,v}$ be the intact teacher's depth prediction for view $v$ of calibration sample $s$, and let $D_{s,v}^{(-l)}$ be the prediction after bypassing only pair $l$. Depth provides a dense final-geometry signal, while its log response measures relative change. Over valid ground-truth depth pixels $\Omega_{s,v}$, we average the absolute log-depth response within each view, then across views and samples:

$$\mathrm{GSS}_l = \frac{1}{|\mathcal{C}|}\sum_{s\in\mathcal{C}}\frac{1}{|\mathcal{V}_s|}\sum_{v\in\mathcal{V}_s}\frac{1}{|\Omega_{s,v}|}\sum_{u\in\Omega_{s,v}}\left|\log\left(D_{s,v}^{(-l)}(u)+\epsilon\right)-\log(D_{s,v}(u)+\epsilon)\right|. \tag{6}$$

where $u$ indexes a pixel location in the valid-depth set $\Omega_{s,v}$, and $\epsilon$ ensures numerical stability. The calibration set $\mathcal{C}$ comprises 5% of the training data, uniformly sampled across multiple view-count configurations. Low GSS indicates insensitivity to removing pair $l$. Scoring requires one intact pass and $L$ single-pair interventions per calibration batch, performed once offline with no deployment cost.

**Pair selection.** To remove the scale mismatch between GSS and SSS, we map them to normalized ascending ranks,

$$r_l^G = \frac{\mathrm{rank}_\uparrow(\mathrm{GSS}_l)-1}{L-1}, \qquad r_l^S = \frac{\mathrm{rank}_\uparrow(\mathrm{SSS}_l)-1}{L-1}, \tag{7}$$

and define the Pair Importance Score for $w \in [0,1]$ as

$$\mathrm{PIS}_l = w r_l^G + (1-w) r_l^S. \tag{8}$$

Lower PIS indicates a more removable pair. We remove the $n$ lowest-ranked pairs and copy the remaining $(L-n)$ pairs to the student in their original execution order. The instantiated architecture and selected indices are reported in the supplementary material.

### 3.4 Prior-Free Geometry Distillation

Pair removal alone degrades the coupled geometry outputs. We recover them by distilling the conditioned full teacher into the selected compact student. The teacher receives ground-truth intrinsics and extrinsics together with each dataset's available dense or sparse depth. Separate adapters add

camera conditioning to camera tokens and depth conditioning to patch tokens. The teacher follows the GeoX depth, point, and camera-pose objectives, with its Mask-Guided Depth-Aware Adapter used only during teacher training.

After pair selection, the retained $(L - n)$ pairs initialize the student, while the teacher is frozen. Their interfaces are

$$(D^T, C^T, P^T) = F_T(\mathcal{I}; C_{\text{prior}}, D_{\text{prior}}), \tag{9}$$

$$(D^S, C^S) = F_S(\mathcal{I}). \tag{10}$$

The student receives RGB images only and contains neither conditional adapters nor a learned point head. It shares VGGT-Ω's use of analytic point recovery with retained point supervision (Wang et al., 2026a):

$$P^S = \Pi(D^S, C^S), \tag{11}$$

where $\Pi$ denotes the unprojection operator.

In PFD, ground truth and predictions from the frozen conditioned teacher supervise depth and camera-pose outputs. The analytic point map $P^S$ is independently matched to ground-truth points $P^*$ and points $P^T$ from the teacher's learned head. We reuse GeoX's Relative-Pose Relational (RPR) loss without modification, applying MSE to off-diagonal inter-view cosine relations (Lian et al., 2026). The student objective is

$$\begin{aligned}\mathcal{L}_S = {} & \lambda_D^{\text{gt}}\mathcal{L}_D(D^S, D^*) + \lambda_D^{\text{kd}}\mathcal{L}_D(D^S, \text{sg}(D^T)) \\ & + \lambda_C^{\text{gt}}\mathcal{L}_C(C^S, C^*) + \lambda_C^{\text{kd}}\mathcal{L}_C(C^S, \text{sg}(C^T)) \\ & + \lambda_P^{\text{gt}}\mathcal{L}_P(P^S, P^*) + \lambda_P^{\text{kd}}\mathcal{L}_P(P^S, \text{sg}(P^T)) \\ & + \lambda_R\mathcal{L}_{\text{RPR}},\end{aligned} \tag{12}$$

where $\text{sg}$ stops gradients through the teacher, and superscripts $\text{gt}$ and $\text{kd}$ denote ground-truth and teacher supervision. Both $\mathcal{L}_D$ and $\mathcal{L}_P$ are confidence-aware and jointly optimize their predictions and associated confidence estimates; explicit confidence arguments are omitted for compactness. No teacher–student feature MSE is used. Camera and depth priors remain teacher-side, while teacher predictions supervise only the student's outputs and inter-view relations; neither enters the deployed graph.

### 3.5 Artifact-Aware Dense Prediction Head

VGGT-Ω replaces high-resolution decoder convolutions with an MLP and $4\times$ Pixel Shuffle (Wang et al., 2026a). GeoX-2 shares this lightweight decoding principle, using multi-scale fused features $F \in \mathbb{R}^{H/8\times W/8\times C}$ and projections for $8\times$ Pixel Shuffle (Shi et al., 2016), as shown in Figure 5. Inconsistent sub-pixel responses can appear as periodic checkerboard artifacts after rearrangement (Aitken et al., 2017).

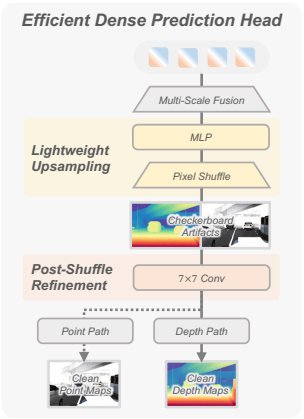


Figure 5: **Efficient dense prediction head.** GFRR refines full-resolution Pixel Shuffle outputs, while the deployed student removes the learned point head.

To suppress these artifacts at their native resolution, we apply Gaussian-Initialized Full-Resolution Refinement (GFRR) immediately after Pixel Shuffle and before the final activation. For depth,

$$\widehat{D} = \exp(\text{GFRR}_D(\text{PS}_8(Q_D(F)))), \tag{13}$$

Here $Q_D$ is the depth projection, and $\text{GFRR}_D$ is a learnable $7 \times 7$ convolution on the single-channel depth output, initialized by a Gaussian kernel with standard deviation 1.5. The depth-confidence branch uses an independent GFRR layer with the same design, while the teacher applies channel-wise GFRR to its three-channel point output. The deployed student omits the point and point-confidence projections, retains the depth and depth-confidence branches, and recovers points by analytic unprojection. Under matched training and evaluation protocols, the supplementary ablation shows that adding GFRR reduces Waymo depth Abs Rel from 0.143 to 0.110.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets.** Training uses Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), PandaSet (Xiao et al., 2021), Lyft (Houston et al., 2021), DDAD (Guizilini et al., 2020), Virtual KITTI (Gaidon et al., 2016), and OpenScene (Contributors, 2023). Training samples contain 2, 5, 6, 7, or 8 synchronized views, with images resized to a width of 512 while preserving their aspect ratios. We evaluate surround-view geometry on Waymo, DDAD, Lyft, PandaSet, and nuScenes, and video depth on KITTI (Geiger et al., 2012). Dataset sensor configurations are summarized in the supplementary material.

**Baselines.** We compare GeoX-2 with VGGT, $\pi^3$, FastVGGT, OmniVGGT, FlashVGGT, LiteVGGT, DVGT, VGG-T$^3$, and VGGT-$\Omega$ using each method's official checkpoint and inference implementation (Wang et al., 2025a; 2026b; Shen et al., 2025; Peng et al., 2026; Wang & Xu, 2026; Shu et al., 2026; Zuo et al., 2026; Elflein et al., 2026; Wang et al., 2026a). Reported baseline results are reproduced using these official model configurations and the evaluation protocols described here and in the supplementary material.

**Metrics.** For 3D reconstruction, we report Accuracy (Acc), Completeness (Comp), and Normal Consistency (NC) after Umeyama alignment (Umeyama, 1991). Camera-pose evaluation reports Relative Rotation Accuracy (RRA), Relative Translation Accuracy (RTA), and the area under the pose-accuracy curve over $0°$–$30°$. Depth evaluation uses Abs Rel and $\delta < 1.25$. Efficiency protocols and timing details are provided in the supplementary material.

## 4.2 MAIN RESULTS

**Surround-view reconstruction.** As shown in Table 1, GeoX-2 improves both Acc and Comp over GeoX on all five datasets and achieves the best results on eight of the ten metrics with 0.298B parameters. On Waymo, FlashVGGT gives the lowest Acc and VGGT-$\Omega$ the lowest Comp.

Table 1: **Surround-view 3D reconstruction across five driving datasets.**

| Method | Params | Waymo | | DDAD | | Lyft | | PandaSet | | nuScenes | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | (B) | Acc ↓ | Comp ↓ | Acc ↓ | Comp ↓ | Acc ↓ | Comp ↓ | Acc ↓ | Comp ↓ | Acc ↓ | Comp ↓ |
| VGGT | 1.10 | 0.389 | 1.565 | 1.413 | 4.010 | 1.660 | 3.789 | 1.402 | 2.082 | 1.191 | 2.417 |
| $\pi^3$ | 0.959 | 0.494 | 1.399 | 1.017 | 1.281 | 1.247 | 1.632 | 1.238 | 1.600 | 1.101 | 1.194 |
| FastVGGT | 1.10 | 0.409 | 1.647 | 1.376 | 5.554 | 1.652 | 4.845 | 1.476 | 2.952 | 1.154 | 3.532 |
| OmniVGGT | 1.21 | 0.553 | 2.017 | 1.600 | 3.303 | 1.838 | 4.686 | 1.847 | 3.451 | 1.336 | 3.260 |
| FlashVGGT | 1.16 | **0.378** | 1.162 | 0.984 | 1.982 | 1.592 | 2.796 | 1.396 | 1.994 | 0.908 | 2.254 |
| LiteVGGT | 1.10 | 0.400 | 1.420 | 1.446 | 3.656 | 1.723 | 4.214 | 1.533 | 2.257 | 1.151 | 1.983 |
| DVGT | 1.73 | 0.493 | 3.283 | 1.172 | 3.718 | 1.217 | 6.459 | 1.399 | 2.321 | 1.060 | 2.152 |
| VGGT-$\Omega$ | 1.14 | 0.494 | **0.819** | 1.457 | 1.105 | 1.074 | 0.586 | 1.030 | 1.566 | 1.139 | 0.814 |
| GeoX | 0.155 | 0.547 | 1.729 | 1.012 | 1.174 | 1.426 | 2.762 | 1.629 | 2.078 | 1.139 | 1.729 |
| GeoX-2 | 0.298 | 0.451 | 1.603 | **0.319** | **0.425** | **0.210** | **0.358** | **0.479** | **0.672** | **0.321** | **0.394** |

Table 2: **Surround-view relative camera-pose estimation across four driving datasets.**

| Method | PandaSet | | | DDAD | | | Lyft | | | nuScenes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RRA ↑ | RTA ↑ | AUC ↑ | RRA ↑ | RTA ↑ | AUC ↑ | RRA ↑ | RTA ↑ | AUC ↑ | RRA ↑ | RTA ↑ | AUC ↑ |
| VGGT | 98.66 | 30.66 | 13.35 | 64.49 | 27.50 | 5.928 | 94.41 | 41.25 | 15.32 | 76.50 | 44.00 | 13.19 |
| $\pi^3$ | **100.0** | 38.67 | 21.13 | **100.0** | 39.36 | 20.38 | **100.0** | 52.66 | 25.82 | 98.00 | 45.00 | 19.30 |
| FastVGGT | 89.25 | 35.58 | 12.07 | 57.12 | 27.71 | 3.113 | 87.16 | 39.00 | 13.19 | 68.50 | 38.50 | 9.627 |
| OmniVGGT | 86.00 | 34.41 | 11.20 | 55.08 | 20.35 | 3.957 | 77.25 | 33.53 | 7.334 | 48.00 | 32.66 | 5.027 |
| FlashVGGT | 95.00 | 34.08 | 12.99 | 87.59 | 33.54 | 12.12 | 95.08 | 46.42 | 19.54 | 72.33 | 35.67 | 10.54 |
| LiteVGGT | 98.83 | 33.00 | 15.00 | 94.17 | 38.31 | 14.07 | 99.58 | 39.25 | 18.43 | 97.33 | 40.16 | 16.58 |
| VGGT-$\Omega$ | **100.0** | 61.33 | 32.13 | **100.0** | 66.18 | 36.77 | **100.0** | 69.25 | 39.44 | **100.0** | 69.67 | 40.73 |
| GeoX | **100.0** | **100.0** | 91.24 | **100.0** | 99.92 | 94.94 | **100.0** | 98.75 | 92.14 | **100.0** | 98.33 | 92.17 |
| GeoX-2 | **100.0** | **100.0** | **97.73** | **100.0** | **100.0** | **96.72** | **100.0** | **99.08** | **96.05** | **100.0** | **98.66** | **94.92** |

**Camera-pose estimation.** Table 2 shows that the RGB-only GeoX-2 reaches 100.0 RRA on all four datasets and achieves the highest AUC throughout, improving over GeoX by 1.78–6.49 points. It also matches or exceeds GeoX in RTA. These results indicate that prior-free geometry distillation transfers camera-pose supervision from the conditioned teacher without requiring camera priors at inference.

**Depth estimation and offline throughput.** As shown in Table 3, GeoX-2 improves both depth metrics over GeoX on all four surround-view datasets and KITTI, while increasing KITTI throughput from 223.8 to 1003.1 FPS ($4.48\times$). It achieves the best Abs Rel and $\delta < 1.25$ on Lyft, PandaSet, and nuScenes, while VGGT-$\Omega$ leads on Waymo and $\pi^3$ on KITTI.

Table 3: **Depth estimation and offline throughput.** KITTI FPS measures offline sequence-batched model throughput.

| Method | Lyft | | PandaSet | | Waymo | | nuScenes | | KITTI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Abs Rel ↓ | $\delta < 1.25$ ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ | FPS ↑ |
| VGGT | 0.158 | 0.801 | 0.243 | 0.729 | 0.176 | 0.811 | 0.337 | 0.579 | 0.058 | 0.961 | 84.46 |
| $\pi^3$ | 0.352 | 0.468 | 0.252 | 0.651 | 0.187 | 0.747 | 0.230 | 0.634 | **0.039** | **0.985** | 110.20 |
| OmniVGGT | 0.270 | 0.622 | 0.241 | 0.658 | 0.281 | 0.623 | 0.269 | 0.618 | 0.053 | 0.964 | 84.18 |
| FlashVGGT | 0.312 | 0.531 | 0.295 | 0.614 | 0.134 | 0.817 | 0.204 | 0.735 | 0.068 | 0.961 | 142.29 |
| LiteVGGT | 0.368 | 0.555 | 0.319 | 0.739 | 0.134 | 0.818 | 0.195 | 0.711 | 0.057 | 0.970 | 155.93 |
| DVGT | 0.224 | 0.685 | 0.268 | 0.635 | 0.148 | 0.832 | 0.218 | 0.648 | 0.109 | 0.894 | 77.80 |
| VGG-T$^3$ | 0.352 | 0.535 | 0.274 | 0.741 | 0.119 | 0.874 | 0.226 | 0.654 | 0.084 | 0.940 | 79.62 |
| VGGT-$\Omega$ | 0.226 | 0.644 | 0.153 | 0.786 | **0.087** | **0.915** | 0.178 | 0.706 | 0.060 | 0.976 | 162.78 |
| GeoX | 0.136 | 0.795 | 0.104 | 0.876 | 0.114 | 0.856 | 0.115 | 0.873 | 0.084 | 0.931 | 223.80 |
| GeoX-2 | **0.043** | **0.971** | **0.053** | **0.972** | 0.110 | 0.886 | **0.048** | **0.970** | 0.071 | 0.962 | **1003.1** |

### 4.3 Ablation and Analysis

**Does combined structural and geometric evidence improve pair selection?** Table 4 compares compact students formed by selecting 6 pairs from the same 12-pair teacher and calibration subset using independent interventions. Score-based strategies outperform uniform and random selection. Block Influence (BI) ranks pairs by their input–output hidden-state change (Men et al., 2025), whereas Task-Loss Sensitivity (TLS) uses the task-loss increase after bypassing each pair and improves every metric over BI. GSS favors reconstruction and depth, whereas SSS favors camera pose. Their PIS fusion achieves the best Comp, NC, and camera-pose results, ties the best Acc, and remains close to GSS on depth, demonstrating the complementarity of structural and geometric evidence.

Table 4: **Pair-selection ablation on Waymo.** All configurations retain 6 Frame–Global pairs. Random reports mean $\pm$ standard deviation over five independently sampled pair subsets.

| Configuration | 3D Reconstruction | | | Camera Pose | | | Depth | |
|---|---|---|---|---|---|---|---|---|
| | Acc ↓ | Comp ↓ | NC ↑ | RRA ↑ | RTA ↑ | AUC ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ |
| Uniform spacing | 0.509 | 1.747 | 0.588 | 97.59 | 92.88 | 75.55 | 0.129 | 0.844 |
| Random | $0.489 \pm 0.032$ | $1.708 \pm 0.036$ | $0.573 \pm 0.032$ | $98.95 \pm 1.050$ | $94.99 \pm 2.785$ | $76.79 \pm 0.660$ | $0.129 \pm 0.011$ | $0.838 \pm 0.031$ |
| BI | 0.492 | 1.694 | 0.601 | 98.34 | 94.53 | 76.40 | 0.123 | 0.858 |
| TLS | 0.478 | 1.652 | 0.609 | 98.81 | 95.46 | 77.23 | 0.119 | 0.862 |
| SSS | 0.465 | 1.627 | 0.618 | 99.41 | 96.98 | 77.87 | 0.114 | 0.877 |
| GSS | **0.451** | 1.612 | 0.621 | 97.29 | 93.36 | 75.70 | **0.109** | **0.889** |
| PIS (GSS + SSS) | **0.451** | **1.603** | **0.631** | **100.0** | **98.33** | **78.70** | 0.110 | 0.886 |

Table 5: **Cumulative speed attribution on KITTI.** Hybrid denotes the Efficient Hybrid Encoder. Token counts follow the actual KITTI inputs: $518 \times 164$ with patch size 14 for DINOv2-S and $512 \times 160$ with output stride 32 for Hybrid. Each row adds one structural change to the preceding configuration; latency is $1000/\mathrm{FPS}$, and speedup is measured relative to GeoX.

| Configuration | Encoder | Tokens/view | Pairs | Cond. adapter | Point head | Time (ms) ↓ | FPS ↑ | Speedup ↑ |
|---|---|---|---|---|---|---|---|---|
| GeoX | DINOv2-S | 407 | 12 | ✓ | ✓ | 4.47 | 223.8 | $1.00\times$ |
| + token-efficient encoder | Hybrid | 80 | 12 | ✓ | ✓ | 1.29 | 774.3 | $3.46\times$ |
| + pair slimming | Hybrid | 80 | 6 | ✓ | ✓ | 1.12 | 893.3 | $3.99\times$ |
| GeoX-2 | Hybrid | 80 | 6 | – | – | **1.00** | **1003.1** | $\mathbf{4.48}\times$ |

**Where does the speedup come from?** Table 5 reports cumulative throughput after each structural change under the same KITTI protocol. At the respective KITTI input sizes, the efficient hybrid encoder reduces each view from 407 to 80 tokens and provides the largest gain, from 223.8 to 774.3 FPS. Slimming the decoder from 12 to 6 pairs raises throughput to 893.3 FPS, while removing the conditional adapters and learned point head further reaches 1003.1 FPS ($4.48\times$).

**Does Prior-Free Geometry Distillation recover geometry in the compact student?** Table 6 compares three variants of the same 6-pair RGB-only graph under the same optimizer, learning rate, training steps, data, augmentation, and ground-truth supervision. The scratch-initialized direct student loads the pretrained encoder but initializes the remaining modules from scratch. The teacher-initialized direct student inherits the same teacher components as PFD, while both direct students use only ground-truth supervision. Teacher initialization improves every metric over scratch initialization. PFD further improves every metric over the teacher-initialized direct student, reducing Acc from 0.493 to 0.451 and Abs Rel from 0.193 to 0.110 while raising RTA from 87.20 to 98.33. This controlled comparison shows that teacher-derived supervision contributes beyond the shared warm start. All deployed variants exclude camera/depth priors and the learned point head.

Table 6: **Prior-free geometry distillation on Waymo.** Prior input and point head indicate components retained at inference. Boldface compares the three RGB-only students; the conditioned teacher is shown only as a reference.

| Model | Pairs | Deployed Graph | | | 3D Reconstruction | | | Camera Pose | | | Depth | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Distill. | Prior input | Point head | Acc ↓ | Comp ↓ | NC ↑ | RRA ↑ | RTA ↑ | AUC ↑ | Abs Rel ↓ | $\delta < 1.25$ ↑ |
| 12-pair teacher | 12 | – | ✓ | ✓ | 0.382 | 1.600 | 0.637 | 100.0 | 98.33 | 78.74 | 0.093 | 0.916 |
| Direct student (scratch-init.) | 6 | – | – | – | 0.540 | 2.473 | 0.537 | 93.25 | 63.18 | 66.47 | 0.284 | 0.605 |
| Direct student (teacher-init.) | 6 | – | – | – | 0.493 | 1.821 | 0.586 | 98.71 | 87.20 | 75.33 | 0.193 | 0.764 |
| PFD student (GeoX-2) | 6 | ✓ | – | – | **0.451** | **1.603** | **0.631** | **100.0** | **98.33** | **78.70** | **0.110** | **0.886** |

### 4.4 Multi-View Inference Efficiency

We measure throughput and peak memory per complete synchronized multi-view timestep at each method's native input resolution. Table 7 shows that GeoX-2 sustains 18.44–19.54 timesteps/s with 1.277–1.383 GB peak memory and leads every evaluated view count in both metrics. Timing settings and input resolutions are detailed in the supplementary material.

Table 7: **Multi-view inference efficiency across varying numbers of synchronized views.** Metrics are reported per complete timestep at each method's native input resolution.

| Method | 5 Views | | | 6 Views | | | 7 Views | | | 8 Views | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Latency (ms) ↓ | FPS ↑ | Mem. (GB) ↓ | Latency (ms) ↓ | FPS ↑ | Mem. (GB) ↓ | Latency (ms) ↓ | FPS ↑ | Mem. (GB) ↓ | Latency (ms) ↓ | FPS ↑ | Mem. (GB) ↓ |
| $\pi^3$ | 148.27 | 6.74 | 3.747 | 165.40 | 6.05 | 3.779 | 178.29 | 5.61 | 3.823 | 197.50 | 5.06 | 3.871 |
| OmniVGGT | 193.29 | 5.17 | 6.282 | 218.84 | 4.57 | 6.612 | 243.18 | 4.11 | 6.956 | 270.67 | 3.69 | 7.324 |
| FlashVGGT | 171.14 | 5.84 | 5.367 | 194.27 | 5.15 | 5.571 | 209.76 | 4.77 | 5.790 | 230.23 | 4.34 | 6.013 |
| LiteVGGT | 227.66 | 4.39 | 3.126 | 226.91 | 4.41 | 3.299 | 233.93 | 4.27 | 3.474 | 247.85 | 4.03 | 3.663 |
| DVGT | 199.69 | 5.01 | 8.246 | 236.36 | 4.23 | 8.597 | 259.62 | 3.85 | 8.963 | 285.62 | 3.50 | 9.307 |
| VGGT-Ω | 103.09 | 9.70 | 4.795 | 129.60 | 7.72 | 4.896 | 146.40 | 6.83 | 5.013 | 164.81 | 6.07 | 5.129 |
| GeoX | 93.72 | 10.67 | 1.665 | 103.32 | 9.68 | 1.877 | 110.03 | 9.09 | 2.112 | 119.65 | 8.36 | 2.346 |
| GeoX-2 | **51.18** | **19.54** | **1.277** | **52.32** | **19.11** | **1.303** | **51.81** | **19.30** | **1.332** | **54.22** | **18.44** | **1.383** |

## 5 Conclusion

GeoX-2 compresses per-view tokens, Frame–Global decoder pairs, and dense output paths across the complete driving-geometry inference graph. Geometry-Derived Structural Slimming retains important pairs using structural and geometric evidence, while Prior-Free Geometry Distillation transfers a prior-conditioned teacher to a compact RGB-only student without conditional adapters or a learned point head. Across driving benchmarks, GeoX-2 achieves strong geometry accuracy, reaches 1003.1 FPS offline KITTI throughput, and leads multi-view inference efficiency across varying view counts, establishing a scalable route to high-throughput driving geometry.

## GENERATIVE AI USE STATEMENT

Codex was used for language polishing, LaTeX editing, and assistance with paper illustrations. The research ideas, code, experiments, data analysis, and conclusions were produced and verified by the authors.

## REFERENCES

Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022.

OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Vancouver, Canada*, pp. 18–22, 2023.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13728–13737. IEEE, 2021.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. Vgg-$t^3$: Offline feed-forward 3d reconstruction at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36464–36474, June 2026.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2485–2494, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pp. 409–418. PMLR, 2021.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232. IEEE, 2023.

Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X 3 kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13343–13353. IEEE, 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9098. IEEE, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024.

Jiawei Lian, Haoyi Sun, Yang Wu, Lifu Mu, Siyuan Wang, Le Hui, Ning Mao, Tao Wei, Pan Zhou, Kun Zhan, et al. Liauto-geox: Efficient grounded driving transformer. *arXiv preprint arXiv:2606.05774*, 2026.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in neural information processing systems*, 35:10421–10434, 2022.

Aoming Liu, Reuben Tan, Boqing Gong, and Bryan A Plummer. Fine-grained token allocation via operation pruning for efficient mllms. *arXiv preprint arXiv:2507.02909*, 2025.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025.

Lifu Mu, Shuai Chen, Wen Zheng, Haoyi Sun, Xueyang Fu, Pengfei Yu, Ning Mao, Tao Wei, and Zhou Pan. Liauto-mindvit: A hybrid vision backbone with adaptive bidirectional mamba. *arXiv preprint arXiv:2609.24337*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, et al. Omnivggt: Omni-modality driven visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36485–36497, 2026.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, et al. Litevggt: Boosting vanilla vggt via geometry-aware cached token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36422–36432, 2026.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in neural information processing systems*, 36:64318–64330, 2023.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 32–42. IEEE, 2021.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-$\omega$. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Zeyu Wang, Dingfu Li, Chenxu Luo, Chao Xie, and Xiaoming Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, 2023.

Zipeng Wang and Dan Xu. Flashvggt: Efficient and scalable visual geometry transformers with compressed descriptor attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21826–21835, 2026.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21672–21683. IEEE, 2023.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph E Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Where do transformers really belong in vision models? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 599–609, 2021.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pp. 3095–3101. IEEE, 2021.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision*, pp. 785–801. Springer, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21791–21801. IEEE, 2023.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5116–5125. IEEE, 2023.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. Dvgt: Driving visual geometry transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14658–14668, 2026.