VGGT-Segmentor: Geometry-Enhanced Cross-View Segmentation

Yulu Gao*
Hangzhou International Innovation Institute of Beihang University
Hangzhou, China
gyl97@buaa.edu.cn

Bohao Zhang*
Beihang University
Beijing, China
zbbhhh@buaa.edu.cn

Zongheng Tang
Hangzhou International Innovation Institute of Beihang University
Hangzhou, China
tzhhhh123@buaa.edu.cn

Jitong Liao
Beihang University
Beijing, China
jitongliao@buaa.edu.cn

Wenjun Wu
Beihang University
Beijing, China
wwj09315@buaa.edu.cn

Si Liu†
Beihang University
Beijing, China
liusi@buaa.edu.cn

Abstract

Instance-level object segmentation across disparate egocentric and exocentric views is a fundamental challenge in visual understanding, critical for applications in embodied AI and remote collaboration. This task is exceptionally difficult due to severe changes in scale, perspective, and occlusion, which destabilize direct pixel-level matching. While recent geometry-aware models like VGGT provide a strong foundation for feature alignment, we find they often fail at dense prediction tasks due to significant pixel-level projection drift, even when their internal object-level attention remains consistent. To bridge this gap, we introduce VGGT-Segmentor (VGGT-S), a framework that unifies robust geometric modeling with pixel-accurate semantic segmentation. VGGT-S leverages VGGT's powerful cross-view feature representation and introduces a novel Union Segmentation Head. This head operates in three stages: mask prompt fusion, point-guided prediction, and iterative mask refinement, effectively translating high-level feature alignment into a precise segmentation mask. Furthermore, we propose a single-image self-supervised training strategy that eliminates the need for paired annotations and enables strong generalization. On the Ego–Exo4D benchmark, VGGT-S sets a new state-of-the-art, achieving 67.7% and 68.0% average IoU for Ego→Exo and Exo→Ego tasks, respectively, significantly outperforming prior methods. Notably, our correspondence-free pretrained model surpasses most fully-supervised baselines, demonstrating the effectiveness and scalability of our approach.

1. Introduction

Achieving instance-level correspondence across vastly different viewpoints is a key challenge in multi-view visual understanding, driving applications in embodied AI [14, 26] and remote collaboration systems [4, 23]. While traditional multi-view methods such as multi-view stereo [18, 24, 42] have significantly advanced scene geometry and keypoint correspondence, instance-level cross-view semantic correspondence, which concerns finding and segmenting the same physical object in two separate views, remains a largely underexplored frontier.

With the release of the large-scale Ego–Exo4D dataset [21], researchers can now systematically investigate the ego–exo object correspondence task. Given an object mask as a query in one view, the goal is to locate and segment the same physical entity in another view. This capability is crucial for embodied intelligence and remote collaboration systems, as it enables the observation of key manipulated objects from an external viewpoint and provides real-time guidance or prompts in the first-person view.

The task is highly challenging due to the significant differences in scale, perspective, and occlusion between the two views. The ego camera is positioned close to the operator's hands, while the exo camera is often farther away or at a different height, causing the same object to appear differently in each view. Ego frames are frequently occluded by hands and tools, whereas exo frames contain numerous distractor objects and complex backgrounds, making pixel-level matching unstable.

Early works often rely on semantic consistency [31] or the contextual understanding provided by large language models [17, 53], but they tend to overlook geometric structures and spatial relationships. VGGT [45] offers a novel

*These authors contributed equally.
†Corresponding author.

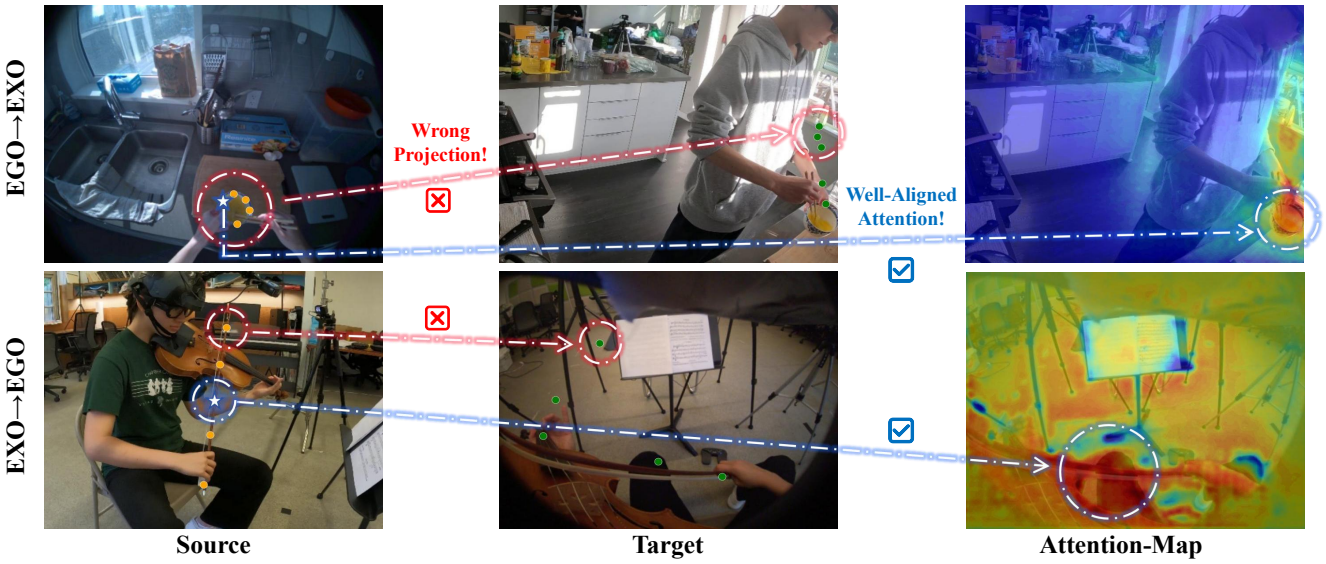

Figure 1. **Visualizing VGGT Cross-View Correspondence.** Left: source image. Middle: target image with the projections of source-sampled points obtained by directly applying VGGT, which exhibit the systematic drift and misalignment. Right: star markers in the source image with the corresponding attention map on the target image, illustrating VGGT's instance-consistent object alignment across views.

perspective. As a large transformer driven by visual geometry, VGGT jointly infers scene depth, camera parameters, and point maps across multiple views, enabling consistent modeling of both geometry and appearance. This provides a robust foundation for cross-view feature alignment.

However, our study reveals a critical challenge in applying VGGT directly to dense segmentation: in ego–exo scenarios, severe occlusion and large viewpoint changes can cause its pixel-level point projections to drift, as illustrated in Figure 1. Notably, while the raw point tracking shows instability, VGGT's internal feature alignment remains consistently reliable, successfully focusing on the approximate object region.

To this end, we propose VGGT-Segmentor (VGGT-S). The model leverages VGGT's strengths in cross-view feature modeling and introduces an object-level union segmentation head, which integrates the object mask as an explicit query into the cross-view reasoning process. The pipeline consists of three stages. The first is Mask Prompt Fusion, where two-view images are encoded by VGGT and then fused with the source-view object mask feature. This is followed by Point-Guided Prediction, where VGGT tracks points from the source mask and outputs a set of coarsely projected points in the target view to guide the fused features. The final stage is Mask Refinement, which refines the predicted mask by iteratively optimizing object boundaries and filling occluded regions. Additionally, we propose a Single-Image Self-Supervised Training strategy that enables training without costly paired annotations, leading to powerful generalization.

On the Ego–Exo4D benchmark, VGGT-S achieves state-of-the-art average IoU scores of 67.7% (Ego→Exo) and 68.0% (Exo→Ego), outperforming the previous best methods by 18.0% and 12.8%, respectively. Remarkably, even our correspondence-free pretrained VGGT-S variant surpasses prior fully-supervised baselines, highlighting its potential for scalable cross-view understanding without paired annotations.

Our key contributions are as follows:

- We introduce VGGT-S, a geometry-enhanced cross-view segmentation framework that fully exploits VGGT's multi-view geometric representations.
- We design the Union Segmentation Head, which comprises three coordinated stages including Mask Prompt Fusion, Point-Guided Prediction, and Mask Refinement, enabling robust cross-view segmentation.
- We propose a Single-Image Self-Supervised Training strategy that reduces the need for paired annotations while enabling superior generalization for both Ego→Exo and Exo→Ego cross-view segmentation.
- We achieve state-of-the-art results on the Ego–Exo4D benchmark, significantly surpassing previous methods.

2. Related Work

2.1. Cross-View Modeling

Cross-view alignment and multi-view modeling are key directions in 3D vision. Classical structure-from-motion [1, 6, 35, 41, 43, 44, 47, 50] and multi-view stereo methods [16, 18, 19, 36–38, 52] rely on keypoint matching

and geometric constraints to accurately reconstruct camera parameters and dense geometry in static scenes, but they are computationally demanding and struggle with non-rigid motion and large baselines. End-to-end neural methods have gradually reduced the need for traditional geometric optimization. VGGT [45] employs a large transformer in a feed-forward manner to jointly predict camera parameters, depth, and point maps, delivering efficient and accurate reconstruction without complex post-processing and serving as a geometry-consistent backbone for downstream tasks. Methods such as DUSt3R [46] and MASt3R [29] are related but often still depend on post-optimization. Given the substantial viewpoint differences in the ego–exo setting, pure reconstruction or two-view matching does not transfer directly to instance-level correspondence, motivating a unified approach that combines geometric structure and contextual semantics for instance-level correspondence. SegMASt3R [25] is a successful example of cross-view object segmentation that leverages 3D geometric priors to establish correspondences.

2.2. Visual Object Correspondence

Instance-level correspondence aims to establish matches for object instances across different views [21]. Some previous studies work on cross-view person matching [2, 15, 48, 49]. In the ego–exo setting, this problem is referred to as object correspondence. XSegTx [21] adapts a cross-image transformer architecture, conditioning on a query mask to perform mutual attention between egocentric and exocentric frames for joint mask prediction. XView-XMem [21] enhances tracking across interleaved ego-exo sequences by integrating embeddings from XSegTx into a working memory module to mitigate track drift. PSALM [53] combines a segmentation model with a large language model to tackle this task in a zero-shot manner. ObjectRelator [17] enhances PSALM by fusing language descriptions with visual queries and explicitly aligning object representations across different views to improve consistency. DOMR [31] proposes a Dense Object Matching framework that pairs objects across views by jointly modeling visual, spatial, and semantic cues, modeling the contextual relationships among multiple objects simultaneously to suppress ambiguous matches.

2.3. Segmentation Models

Segmentation is fundamental to visual understanding, including semantic segmentation [8–11], instance segmentation [5, 22, 32], and panoptic segmentation [27]. Recent unified segmentation models like Mask2Former [12], along with multimodal promptable approaches such as SEEM [54] and large-scale promptable models like SAM [28] and SAM 2 [40], have demonstrated strong generalization on large datasets. However, most existing segmentation methods are single-view and lack cross-view alignment mechanisms. MASA [30] leverages SAM's rich segmentation outputs to establish instance-level correspondences through extensive data transformations. Its core innovation lies in a self-training strategy that bootstraps instance associations from unlabeled images by applying geometric transformations to create pixel-level correspondences. These are then lifted by SAM to the instance level for contrastive similarity learning, enabling robust zero-shot tracking.

3. Method

3.1. Overview

VGGT [45] is a vision model for multi-view geometric consistency, using a unified encoder with integrated tracking and feature interaction to model dense features. As illustrated in Figure 2(A), VGGT-S augments the VGGT encoder with a lightweight **Union Segmentation Head** that converts cross-view geometric cues into target-view masks. Given a source–target image pair (I_s, I_t) (e.g., Exo→Ego), the VGGT encoder produces dense feature maps F_s and F_t. The source mask M_s is encoded and integrated into cross-view feature interactions. A compact set of representative points sampled from M_s is tracked to the target frame via the VGGT's track head, generating P_t. These point prompts guide the prediction of the target mask $\hat{M}_t$ on F_t. During training, the VGGT encoder remains frozen and only the Union Segmentation Head is optimized, keeping the framework end-to-end while minimizing computational and memory overhead.

3.2. VGGT Encoder

Following VGGT, each image is patchified by a DINO-style [7] stem, which refers to a ViT-based patch embedding approach that splits images into patches and embeds them as tokens. They are then processed together through alternating frame-wise and global self-attention layers. A DPT-style [39] head, which is a decoder for dense prediction that upsamples and fuses tokens into spatial feature maps, transforms tokens into dense feature maps geometrically aligned with depth, point, and tracking information. We extract these maps as inputs to our head:

$$x_s = \mathrm{Stem}(I_s), \quad x_t = \mathrm{Stem}(I_t), \tag{1}$$

$$h_s, h_t = \mathrm{VGGT}(x_s, x_t), \tag{2}$$

$$F_s, F_t = \mathrm{DPT}(h_s, h_t). \tag{3}$$

The resulting geometry-aware features F_s and F_t are fed into the Union Segmentation Head.

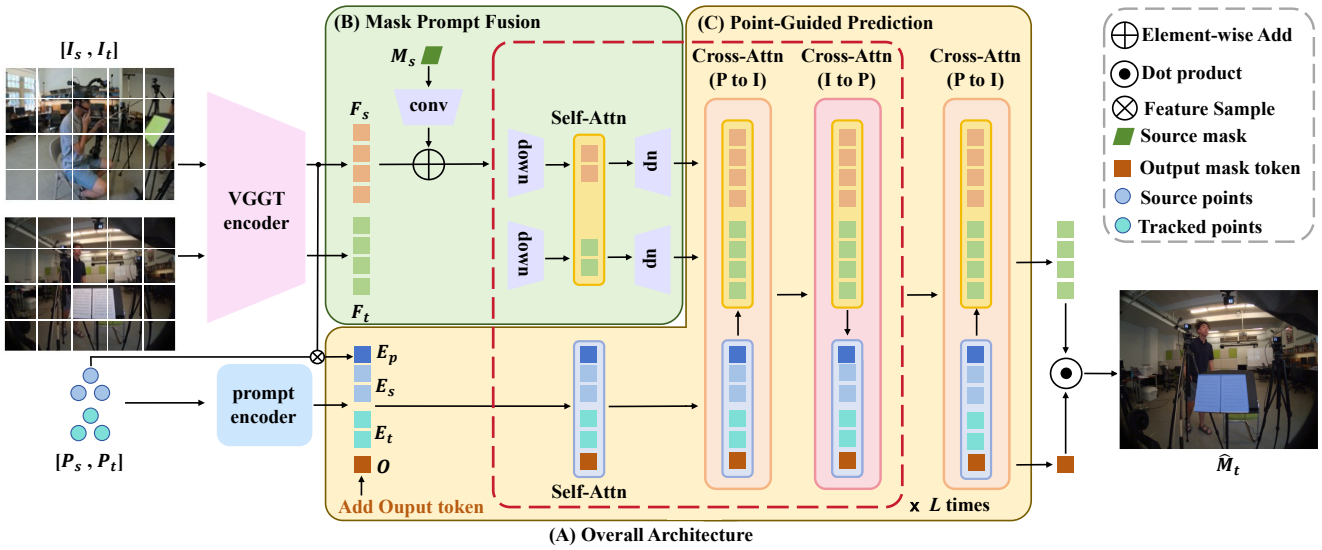

Figure 2. **(A) Overall Architecture of VGGT-S**, which integrates the original VGGT encoder with our **Union Segmentation Head**. **(B) Mask Prompt Fusion** stage, which injects the source mask M_s into source feature map F_s and target feature map F_t via convolutional fusion and a **Bottleneck Fusion** module. **(C) Point-Guided Prediction** stage, which uses point sets (P_s, P_t) to guide target mask prediction through bidirectional interactions between point embeddings and image features.

3.3. Union Segmentation Head

The Union Segmentation Head consists of three stages, Mask Prompt Fusion, Point-Guided Prediction and Mask Refinement.

Mask Prompt Fusion. As shown in Figure 2(B), we first encode the source mask M_s into a high-dimensional embedding that captures its spatial layout and identity:

$$E_m = \mathrm{Conv}(M_s). \tag{4}$$

This embedding E_m is added to the source features F_s directly:

$$F_s' = F_s + E_m. \tag{5}$$

Although M_s is now fused into F_s', it has not yet interacted sufficiently with F_t. Therefore, we introduce a **Bottleneck Fusion** module that integrates self-attention (SelfAttn), feed-forward network (FFN) as well as downsampling D_r and upsampling U_r (ratio r):

$$\tilde{F}_s = \mathrm{D}_r(F_s'), \quad \tilde{F}_t = \mathrm{D}_r(F_t), \tag{6}$$

$$\dot{F}_s, \dot{F}_t = \mathrm{FFN}\big(\mathrm{SelfAttn}([\tilde{F}_s\ ,\ \tilde{F}_t])\big), \tag{7}$$

$$F_s^\star = \mathrm{U}_r(\dot{F}_s), \quad F_t^\star = \mathrm{U}_r(\dot{F}_t). \tag{8}$$

Here $[\cdot\ ,\ \cdot]$ denotes concatenation. The resulting $F^\star = [F_s^\star, F_t^\star]$ is a compact yet expressive representation containing both geometric and semantic cues from two views.

Point-Guided Prediction. We next generate point prompts from the source mask. Let the foreground pixel set be

$$\Omega = \{(x, y) \mid M_s(x, y) = 1\}. \tag{9}$$

We sample K_{pt} representative points using K-Means algorithm [33]:

$$P_s = \mathrm{kmeans}(\Omega, K_{\mathrm{pt}}). \tag{10}$$

VGGT's track head $\mathcal{T}$ projects them to the target frame:

$$P_t = \mathcal{T}(P_s; I_s, I_t). \tag{11}$$

A prompt encoder ψ maps points to embeddings, and a learnable output mask token O together with source point features sampled from F_s are appended:

$$E_p = \mathcal{G}(P_s, F_s), \quad E_s = \psi(P_s), \quad E_t = \psi(P_t), \tag{12}$$

$$Q_0 = [E_p, E_s, E_t, O], \tag{13}$$

where Q_0 denotes the prompt queries.

As shown in Figure 2(C), we apply L-layer lightweight decoder blocks, each consisting of self-attention among prompts, followed by point-to-image and image-to-point cross-attention (CrossAttn):

$$\bar{Q}_\ell = \mathrm{SelfAttn}(Q_{\ell-1}), \tag{14}$$

$$Q_\ell = \mathrm{CrossAttn}_{P\to I}(\bar{Q}_\ell, F_\ell^\star), \tag{15}$$

$$H_\ell = \mathrm{CrossAttn}_{I\to P}(F_\ell^\star, Q_\ell), \quad \ell = 1, \ldots, L. \tag{16}$$

where $F_\ell^\star$ denotes the output of the Bottleneck Fusion module within the ℓ-th block, and H_ℓ represents the resulting fused image features produced by the same block.

Finally, we perform an additional point-to-image cross-attention using the refined output mask token O_L, and generate an initial mask through per-pixel dot products on H_t,

which corresponds to the target-view component of the final fused image features H_L:

$$\tilde{O} = \mathrm{CrossAttn}_{P\to I}(O_L, H_t), \tag{17}$$

$$z(x,y) = \big(W\tilde{O} + b\big)^{\top} \mathbf{f}_t(x,y), \tag{18}$$

$$\hat{M}_t^{(0)}(x,y) = \sigma\big(z(x,y)\big), \tag{19}$$

where W and b denote the weights and bias of an MLP, $\mathbf{f}_t(x,y)$ is the feature vector at pixel position (x,y) on H_t and $\sigma(\cdot)$ is the sigmoid function.

Mask Refinement. To sharpen boundaries and handle occlusions, we adopt an iterative refinement module. At iteration k,

$$\hat{M}_t^{(k+1)} = \Psi\big(F_s,\, M_s,\, F_t,\, \hat{M}_t^{(k)},\, Q\big), \tag{20}$$

where Ψ denotes our lightweight mask decoder, Q denotes the refined prompt queries.

During training, we perform refinement iterations and backpropagate gradients only through the final iteration and half of the samples in each batch undergo refinement, while the other half do not. This process progressively sharpens object boundaries, fills occluded regions, and improves cross-view segmentation quality. More details are in the Supplementary Material.

3.4. Single-Image Self-Supervised Training

To reduce reliance on paired annotations and enhance generalization, we introduce a Single-Image Self-Supervised Training strategy inspired by the augmentation methods of MASA [30]. Given any image I, we generate an augmented view I' and obtain a pseudo mask M from an offline segmentor [28]. The model is required to predict the same object's mask $\hat{M}'$ on I'.

The training strategy employs dynamic augmentations from two families: (1) VGGT-adaptive (e.g., scaling, mild rotations, cropping), which preserve VGGT's point mapping. In this case, both views are processed through the VGGT encoder, and the VGGT's track head provides point prompts on the target view. (2) VGGT-non-adaptive (e.g., large rotations, horizontal flips), which heavily disrupt cross-view alignment and cause VGGT to fail in maintaining effective correspondence. Here, the two views are processed independently by VGGT encoder, and we perturb target ground-truth points to synthesize prompts. By mixing these two families, the model learns a cross-view mask head well aligned with VGGT features. It can recover target masks under substantial viewpoint changes, enabling robust Ego→Exo and Exo→Ego transfer without paired annotations.

Specifically, we train the model on a $1/20$ subset of the SA-1B dataset [40] to obtain a correspondence-free pretrained variant. When evaluated on the Ego-Exo4D dataset, this variant still delivers competitive results.

4. Experiments

4.1. Setup and Implementation Details

Dataset. We use the ego–exo correspondence benchmark from the Ego-Exo4D dataset [21], which contains synchronized first-person and third-person videos of professional skill demonstrations across various domains. The dataset includes 1,335 annotated takes and 5,566 target objects. It provides 1.8 million masks sampled at 1 FPS, of which 742K are egocentric and 1.1 million are exocentric. On average, each video consists of approximately 5.5 objects and 173 frames per track. The annotations cover a wide range of objects, including tools, relevant environmental items, and human body parts. We use the official train/validation split for our experiments, and the evaluation metric is the mean Intersection over Union (IoU) between predicted and ground-truth masks.

Implementation Details. We adopt the official VGGT encoder settings, using an image patch size of 14. In the Mask Prompt Fusion stage, we downsample the source mask through a convolution layer, reducing its size to half of the original resolution. This ensures consistency with the feature map of the image output by VGGT. In the Point-Guided Prediction stage, we apply the K-Means algorithm [33], setting the number of clusters to 5 to match the number of sampled points. Clustering is refined only once to save training time. Following SAM [28], we supervise the model's predictions using a linear combination of focal and dice losses with a weight ratio of $20:1$. For optimization, we use AdamW [34], with an initial learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. The model is trained for 12 epochs, with the learning rate reduced by a factor of 0.1 after 8 and 11 epochs. To prevent gradient explosion, we clip the L_2 norm of all gradients to 1.0. All experiments are conducted on 4×NVIDIA RTX 4090 GPUs, with a batch size of 8 during training. For inference speed, we run 100 forward passes on a single image using a single GPU and report the average time. In the Ego→Exo task, the remapping strategy introduces an additional mapping step, which is omitted in the subsequent time measurements. We also adopt a cropping strategy. Both are detailed in the Supplementary Material.

4.2. Main Results

We evaluate our method on the Ego-Exo4D benchmark and report the results in Table 1. Our approach achieves 67.7% IoU on Ego→Exo and 68.0% IoU on Exo→Ego, surpassing the previous state-of-the-art method, DOMR, by 18.0% and 12.8%, respectively. Compared to the LLM-based ObjectRelator, our method outperforms it by 22.3% and 17.1% in the two directions, while also demonstrating significantly higher efficiency during inference.

In the zero-shot setting, our model achieves 54.1% IoU

Table 1. Comparison with prior methods on Ego-Exo4D dataset. "ZSL" denotes the zero-shot learning results. "Type S" denotes spatial-only modeling, while "Type ST" denotes spatio-temporal modeling. Our VGGT-S provides both supervised and zero-shot learning results.

Ego→Exo				Exo→Ego			
Method	**ZSL**	**Type**	**IoU ↑**	**Method**	**ZSL**	**Type**	**IoU ↑**
XSegTx [21]	✓	S	0.3	XSegTx [21]	✓	S	1.3
SEEM [54]	✓	S	1.1	SEEM [54]	✓	S	4.1
XSegTx [21]	×	S	6.2	XSegTx [21]	×	S	30.2
CMX [51]	×	S	6.8	CMX [51]	×	S	12.0
PSALM [53]	✓	S	7.9	PSALM [53]	✓	S	9.6
XView-XMem [21]	✓	ST	16.2	XView-XMem [21]	✓	ST	13.5
XView-XMem [21]	×	ST	17.7	XView-XMem [21]	×	ST	20.7
XView-XMem + XSegTx [21]	×	ST	36.9	XView-XMem + XSegTx [21]	×	ST	36.1
SSCC [3]	✓	S	38.4	SSCC [3]	✓	S	43.7
SSCC [3]	×	S	39.1	SSCC [3]	×	S	47.1
PSALM [53]	×	S	41.3	PSALM [53]	×	S	47.3
ObjectRelator [17]	×	S	45.4	ObjectRelator [17]	×	S	50.9
DOMR [31]	×	S	49.7	DOMR [31]	×	S	55.2
VGGT-S (Ours)	✓	S	**54.1**	**VGGT-S**	✓	S	**58.4**
VGGT-S (Ours)	×	S	**67.7**	**VGGT-S**	×	S	**68.0**

Table 2. Comparison with prior methods on MvMHAT dataset.

Method	**AP**
MvMHAT [20]	63.8
DOMR [31]	71.1
VGGT-S (Ours)	80.7

on Ego→Exo and 58.4% IoU on Exo→Ego. We improve over PSALM by 46.2% and 48.8%, and over XView-XMem by 37.9% and 44.9%, respectively. Notably, XView-XMem leverages spatiotemporal cues, whereas our method relies solely on image-level features and still outperforms it. Our correspondence-free pretrained variant also surpasses the supervised method, DOMR, on both tasks, with gains of 4.4% and 3.2%, demonstrating strong generalization to unseen objects and scenes.

To further validate the generalizability of VGGT-S, we finetune the correspondence-free pretrained model on the MvMHAT dataset [20] for 1 epoch. Surprisingly, the resulting AP reaches 80.7%, surpassing DOMR by 9.6% and the method in [20] by 16.9%, as Table 2 shows. These results demonstrate the strong generalization capability of our VGGT-S model.

4.3. Ablation Studies

Component Analysis. A step-by-step ablation of the proposed components is provided in Table 3. We begin with a Plain Head that encodes the source view mask and predicts the target mask using an output mask token, establishing a direct baseline. In the next step, adding Bottleneck Fusion leads to clear improvements, demonstrating that cross-view feature aggregation is crucial for viewpoint transfer, as target features gain spatial prior information from the source object. Introducing Point-Guided Prediction results in a significant increase in IoU by incorporating sparse, geometry-aware anchors, which are robust to perspective and scale changes. Finally, the Mask Refinement module consistently boosts IoU with minimal computational overhead by refining boundaries and correcting small misalignments. The full model, incorporating all components, achieves an overall improvement of 32.2% on the Ego→Exo task and 30.9% on the Exo→Ego task over the Plain Head setting, validating the effectiveness of the geometry-enhanced design.

Table 3. Component analysis. "BF" denotes the Bottleneck Fusion module in Mask Prompt Fusion stage. "PGP" denotes the Point-Guided Prediction. "MR" denotes Mask Refinement stage.

Method	**IoU ↑**		**Time (ms)**
	Ego→Exo	**Exo→Ego**	
Plain Head	35.5	37.1	105.8
+ BF	50.2	52.3	107.4
+ PGP	62.2	63.5	153.2
+ MR	67.7	68.0	161.4

Effect of Bottleneck Fusion Resolution. We investigate the impact of fusion resolution in the Bottleneck Fusion module at spatial sizes of 37×37, 74×74, and 518×518, as summarized in Table 4. Increasing the resolution from

Table 4. Effect of Bottleneck Fusion resolution.

Fusion Size	IoU ↑		Time (ms)
	Ego→Exo	Exo→Ego	
37×37	67.7	68.0	161.4
74×74	68.4	68.5	180.9
518×518	*OOM*	*OOM*	*OOM*

Table 5. Effect of the number of points used in Point-Guided Prediction.

#Points	IoU ↑		Time (ms)
	Ego→Exo	Exo→Ego	
1	61.5	63.4	160.1
5	67.7	68.0	161.4
9	68.3	68.5	162.9

Table 6. Effect of iterations in Mask Refinement.

#Refine Iters	IoU ↑		Time (ms)
	Ego→Exo	Exo→Ego	
0	62.2	63.5	153.2
1	66.3	67.5	157.7
2	67.7	68.0	161.4
3	67.9	68.4	165.3

Table 7. Effect of input image size.

Image Size	IoU ↑		Time (ms)
	Ego→Exo	Exo→Ego	
420×420	66.1	66.3	136.8
518×518	67.7	68.0	161.4
700×700	68.5	68.9	225.4

Table 8. Effect of the number of decoder blocks.

#Blocks	IoU ↑		Time (ms)
	Ego→Exo	Exo→Ego	
1	65.1	65.5	158.2
2	67.7	68.0	161.4
3	68.4	68.7	165.4
6	68.8	69.3	176.8

37×37 to 74×74 results in improvements of 0.7% and 0.5% IoU for the two tasks, respectively. However, this also increases latency due to the quadratic complexity of self-attention at higher spatial resolutions. Further scaling to 518×518 causes out-of-memory (*OOM*) issues during training. Balancing both accuracy and efficiency, we adopt 37×37 as the default resolution for mask and image fusion in our main experiments, which retains most of the benefits of cross-view coupling while maintaining inference efficiency.

Effect of the Number of Points. Table 5 analyzes the impact of the number of points used in Point-Guided Prediction. Increasing the number of sampled points from 1 to 5 improves the IoU by 6.2% and 4.6% on the Ego→Exo and Exo→Ego tasks, respectively. Further increasing the number of points from 5 to 9 results in only marginal gains of 0.6% and 0.5% for the two tasks. We adopt 5 points for all final results. These experiments demonstrate that sparse points provide an effective and efficient guidance signal for cross-view segmentation.

Effect of Mask Refinement Iterations. We vary the number of Mask Refinement iterations in Table 6. As the number of iterations increases from 0 to 3, IoU improves from 62.2% to 67.9% on the Ego→Exo task and from 63.5% to 68.4% on the Exo→Ego task, resulting in total gains of +5.7% and +4.9%, respectively. Since each iteration re-invokes the mask head, the computational cost scales approximately linearly with the number of iterations. With our lightweight head, two iterations provide an optimal trade-off, delivering significant improvements over a single pass with minimal additional latency, while further iterations result in only marginal gains.

Effect of Input Image Size. Table 7 evaluates the impact of input resolutions 420×420, 518×518, and 700×700. While higher input resolutions lead to monotonic improvements in IoU, they also increase computational and memory requirements, resulting in higher latency and reduced throughput during inference. This trade-off is consistently observed across both Ego→Exo and Exo→Ego settings. Therefore, we adopt 518×518 as the default resolution, as it strikes a good balance between accuracy and efficiency for both directions, and aligns with our training time configuration and hardware profile.

Effect of the Number of Decoder Blocks. Table 8 ablates the number of decoder blocks. Performance improves steadily from 1 to 6 blocks, suggesting that deeper cross-view fusion enhances alignment and refines mask details. To maintain a compact and efficient model, we use 2 blocks by default in all reported results. This configuration captures most of the benefits from iterative point and image interactions without introducing noticeable slowdowns.

4.4. Qualitative Results

Visualization of VGGT-S vs. DOMR. Figure 3 compares VGGT-S with DOMR on both Ego→Exo and Exo→Ego tasks. Leveraging geometry-enhanced cues, VGGT-S demonstrates clear advantages in spatial localization. Even

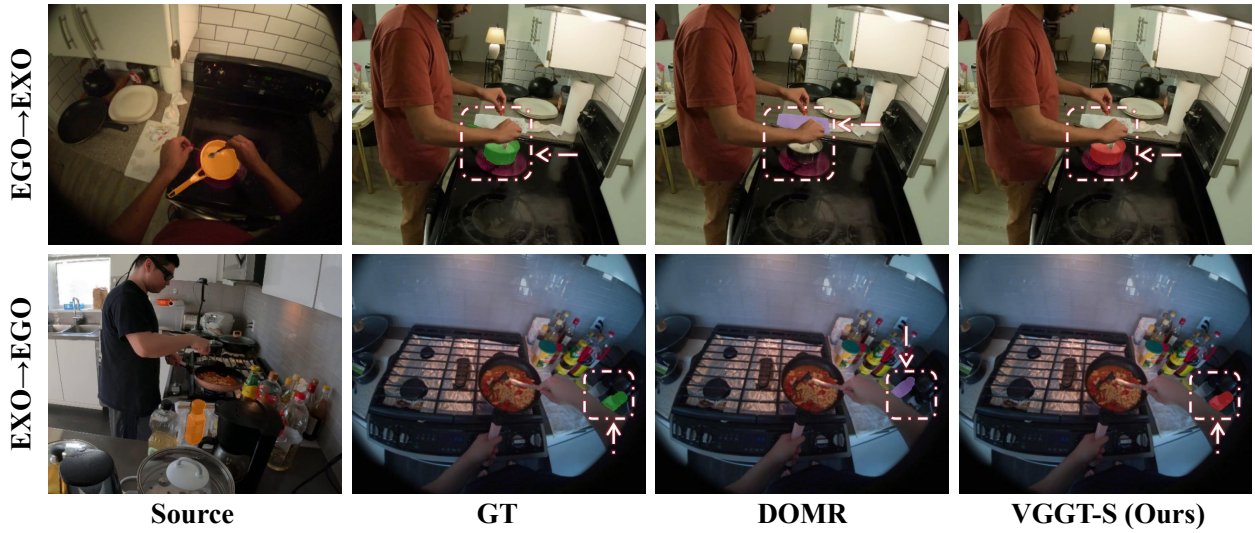

Figure 3. **Visualization of VGGT-S vs. DOMR.** The first row shows the Ego→Exo task. DOMR incorrectly takes the chopping board as the predicted result, while VGGT-S correctly identifies the pot. The second row illustrates the Exo→Ego task. Two similar bottles are nearby. Due to a lack of geometric information, DOMR mistakenly confuses them, whereas VGGT-S continues to make accurate predictions.

under significant viewpoint changes and in the presence of visually similar distractors, our method effectively restricts the correspondence search to geometrically reasonable regions, ensuring consistent alignment between views. This geometric constraint reduces ambiguity during matching. As a result, VGGT-S more reliably identifies the correct target among multiple confusing proposals, producing cleaner and better-aligned masks with sharper boundaries, whereas DOMR tends to drift towards nearby look-alike objects, exhibits unstable correspondences, and often leads to noticeable boundary misalignment.

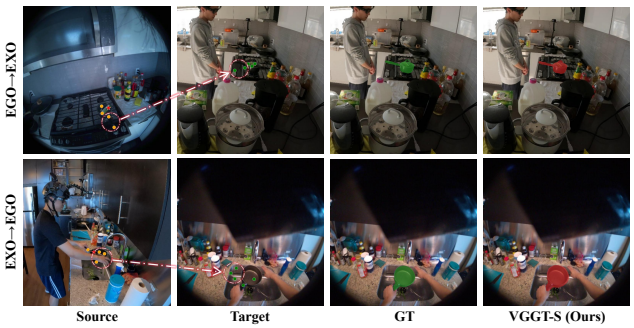

Figure 4. **Visualization of the Effect of the Union Segmentation Head.** Although VGGT projects points to incorrect locations, our Union Segmentation Head adjusts the predicted mask to geometrically consistent positions. Zooming in provides better results.

Visualization of the Effect of the Union Segmentation Head. To evaluate the effect of the Union Segmentation Head, we visualize predictions in Figure 4. The Union Segmentation Head explicitly aggregates contextual information while addressing the VGGT point projection bias. When raw VGGT point reprojections experience slight drift or local misalignment, the Union Segmentation Head corrects these inconsistencies through feature fusion and spatial consensus, pulling masks back to geometrically consistent locations. This results in improved alignment with the scene structure.

Test on Outdoor Datasets. We further assess the generalization of our correspondence-free pretrained VGGT-S on MAVREC dataset [13]. Details and visualization can be found in the Supplementary Material.

5. Conclusion

We introduced VGGT-Segmentor (VGGT-S), a geometry-enhanced framework for cross-view instance-level segmentation between egocentric and exocentric perspectives. By leveraging VGGT's geometry-consistent representations and incorporating a Union Segmentation Head with Mask Prompt Fusion, Point-Guided Prediction, and Mask Refinement, our method effectively transfers object masks across large viewpoint and scale variations. Additionally, the proposed Single-Image Self-Supervised Training strategy enables training without paired annotations, supporting Ego–Exo transfer without correspondence supervision. Extensive experiments on the Ego–Exo4D benchmark demonstrate that VGGT-S achieves state-of-the-art performance, strong generalization, offering a simple yet scalable solution for cross-view object segmentation.

Acknowledgments

This research is supported in part by the Postdoctoral Research Funding of Hangzhou International Innovation Institute of Beihang University (Grant No. 2026BKZ008), National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (No. 62461160308, No. 62576024, U23B2010), "the Fundamental Research Funds for the Central Universities" (No. 501RCQD2025), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161), Beijing Natural Science Foundation (QY25227), Ningbo Science and Technology Innovation 2025 Major Project (2025Z034), NSFCRGC Project (N CUHK498/24).

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 2

[2] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016. 3

[3] Alan Baade and Changan Chen. Self-supervised cross-view correspondence with predictive cycle consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16753–16763, 2025. 6

[4] Allison Bayro, Hongju Moon, Yalda Ghasemi, Heejin Jeong, and Jae Yeol Lee. Object manipulation in physically constrained workplaces: remote collaboration with extended reality. *IISE Transactions on Occupational Ergonomics and Human Factors*, 13(3):177–190, 2025. 1

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 3

[6] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 778–792. Springer, 2010. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[13] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22678–22690, 2024. 8

[14] Chrisantus Eze and Christopher Crick. Learning by watching: A review of video-based learning approaches for robot manipulation. *IEEE Access*, 2025. 1

[15] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 3

[16] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2

[17] Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6530–6540, 2025. 1, 3, 6

[18] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[19] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 2

[20] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *ACM MM*, 2021. 6

[21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19383–19400, 2024. 1, 3, 5, 6

[22] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. 3

[23] Yuping He, Yifei Huang, Guo Chen, Lidong Lu, Baoqi Pei, Jilan Xu, Tong Lu, and Yoichi Sato. Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. *International Journal of Computer Vision*, 134(2):62, 2026. 1

[24] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018. 1

[25] Rohit Jayanti, Swayam Agrawal, Vansh Garg, Siddharth Tourani, Muhammad Haris Khan, Sourav Garg, and Madhava Krishna. Segmast3r: Geometry grounded segment matching. *arXiv preprint arXiv:2510.05051*, 2025. 3

[26] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025. 1

[27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 3

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3, 5

[29] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3

[30] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18973, 2024. 3, 5

[31] Jitong Liao, Yulu Gao, Shaofei Huang, Jialin Gao, Jie Lei, Ronghua Liang, and Si Liu. Domr: Establishing cross-view segmentation via dense object matching. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 412–421, 2025. 1, 3, 6

[32] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 3

[33] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4, 5

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[36] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *European Conference on Computer Vision*, pages 734–750. Springer, 2022. 2

[37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3504–3515, 2020.

[38] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8645–8654, 2022. 2

[39] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 5

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[42] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35(2):151–173, 1999. 1

[43] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12517–12526, 2022. 2

[44] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 2

[45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 3

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[47] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *European conference on computer vision*, pages 230–247. Springer, 2020. 2

[48] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 3446–3455, 2021. 3

[49] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018. 3

[50] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016. 2

[51] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694, 2023. 6

[52] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21508–21518, 2023. 2

[53] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 1, 3, 6

[54] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 3, 6