Rethinking Dense Optical Flow without Test-Time Scaling

Praroop Chanda[1,3] Suryansh Kumar[1,2,3,4,*]
Visual and Spatial AI Lab[1], VCCM Section
College of PVFA[2], Department of ECEN[3], Department of CSCE[4],
Texas A&M University, College Station, Texas, USA

Abstract

*Recent progress in dense optical flow has been driven by increasingly complex architectures and multi-step refinement for test-time scaling. While these approaches achieve strong benchmark performance, they also require substantial computation during inference. This raises a fundamental question: Is scaling test-time computation the only way to improve dense optical flow accuracy? We argue that it is not. Instead, powerful visual semantic and geometric priors encoded in modern foundation models can reduce, if not overcome, the need for computationally expensive iterative refinement at test-time. In this paper, we present a framework that estimates dense optical flow in a single forward pass, leveraging pretrained foundation representations, while avoiding iterative refinement and additional inference-time computation, thus offering an alternative to test-time scaling. Our method extracts visual semantic features from a frozen DINO-v2 backbone and combines them with geometric cues from a monocular depth foundation model. We fuse these complementary priors into a unified representation and apply a global matching formulation to estimate dense correspondences without recurrent updates or test-time optimization. Despite avoiding iterative refinement, our approach achieves strong cross-dataset generalization across challenging benchmarks. On Sintel Final, we obtain **2.81 EPE** without refinement, significantly improving over state-of-the-art (SOTA) SEA-RAFT under comparable training conditions and outperforming RAFT, GMFlow (without refinement), and recent FlowSeek in the same setting. These results suggest that strong foundation priors can substitute for test-time scaling, offering a computationally efficient alternative to refinement-heavy pipelines.*

1. Introduction

Dense optical flow is a fundamental problem in computer vision, robotic perception, and machine vision. The task seeks to estimate a optical flow vector for every pixel between two consecutive images [33]. Accurate estimation of optical flow enables a wide range of downstream applications, including video understanding [5], saliency detection [20], action recognition [29, 36], dense 3D reconstruction [3, 17, 19, 32], shape deformation modeling [14–16, 18], and motion modeling [7]. Over the past decade, deep learning methods have dramatically improved optical flow accuracy. Architectures such as RAFT [37] and its variants [39] achieve near-saturating performance on standard benchmarks. However, these improvements come at an increasing computational cost. Modern pipelines rely heavily on large annotated datasets, extensive training schedules, and most critically multi-step test-time refinement, which requires substantial inference-time compute.

This trend naturally raises a fundamental question: Is allocating more computation at time-time the only way forward to improve optical flow performance? Recent developments in foundation models for visual representation suggest an alternative viewpoint. Large-scale pretrained models trained once on internet-scale datasets have demonstrated strong transferability across diverse visual tasks, including object detection [45], segmentation [44], depth estimation [42, 43], and geometric reasoning [23, 42]. Rather than repeatedly training specialized architectures or increasing inference-time computation, these models show that powerful pretrained representations can generalize across domains and tasks. This observation raises an intriguing possibility for dense optical flow, i.e., instead of relying on increasingly expensive test-time refinement unlike SOTA [30, 39], we can leverage strong visual semantic and geometric priors from foundation models to perform accurate motion estimation directly.

Surprisingly, dense optical flow has remained largely disconnected from such paradigm despite the rapid progress of foundation models in other vision domains. Current state-of-the-art approaches [30, 37, 39] continue to emphasize architectural innovations, dataset scaling, and iterative update mechanisms, treating optical flow as a task that must be explicitly learned and repeatedly corrected during in-

*Corresponding Author: Suryansh Kumar

ference. While these strategies improve benchmark performance, they also reinforce the assumption that accurate optical flow estimation requires task-specific feature encoders and expensive refinement procedures. Nevertheless, as we know, dense optical flow is fundamentally a correspondence problem, where the key challenge is to learn representations that reliably match pixels across frames while respecting scene structure and motion boundaries. Modern vision foundation models already encode many of these properties, including strong semantic discrimination and boundary-aware geometric cues. This observation suggests that dense optical flow may benefit more from reusing powerful pretrained representations than from further increasing architectural complexity or test-time computation.

In this work, we revisit the dense optical flow problem from a foundation model perspective. We argue that strong semantic and geometric priors contained in modern pretrained models can significantly reduce the reliance on test-time scaling through iterative refinement. Instead of allocating additional computation during inference, we investigate whether carefully designed representations can enable accurate motion estimation in a single forward pass. Our central hypothesis is that dense optical flow can emerge from sufficiently powerful visual semantic and geometric representations without repeated correction through iterative updates.

Concretely, we introduce a dense optical flow framework that leverages two complementary foundation priors. First, we extract semantic visual features from the self-supervised DINO-v2 model [27], which provides spatially coherent representations capable of capturing fine-grained visual correspondence. Second, we incorporate geometric structure using a monocular depth foundation model [43]. Although depth prediction is learned from data rather than derived from explicit geometry, modern depth prediction models produce high-frequency structural cues and sharp boundaries that are particularly informative for correspondence estimation near occlusions and motion discontinuities. By combining these semantic and geometric signals, we construct a unified representation suitable for dense matching.

The proposed framework fuses DINO-v2 features and depth foundation representations within a global matching formulation inspired by GMFlow [41]. Unlike RAFT [37], SEA-RAFT [39], and very recently proposed FlowSeek [30], our model performs optical flow estimation in a single forward pass and does not require iterative refinement to achieve competitive performance. Empirically, our approach demonstrates strong cross-dataset generalization across challenging benchmarks. On the Sintel Final benchmark, we achieve 2.81 EPE without refinement, substantially improving over SEA-RAFT (4.32 EPE) under comparable training conditions and outperforming RAFT, GMFlow (without refinement), and FlowSeek in the same setting. These results suggest that strong foundation priors can partially substitute for test-time scaling in dense optical flow estimation. More broadly, our work highlights the potential of foundation-model-driven inference pipelines to reduce the reliance on refinement-heavy architectures and large task-specific training pipelines. Our **contributions** are summarized as follows:

- We propose a dense optical flow estimation framework that operates without test-time refinement and leverages frozen pretrained foundation models instead of training task-specific encoders.
- We show that combining DINO-v2 visual semantic features [27] with depth foundation representations [43] provides a powerful prior for dense optical flow correspondence estimation.
- Our work demonstrate that strong pretrained representations can reduce reliance on iterative inference procedures. Empirically, our framework achieves competitive and often superior performance compared to refinement-based SOTA approaches such as RAFT [37], GMFlow [41], SEA-RAFT [39], and FlowSeek [30], including **2.81 EPE** on Sintel Final without refinement.

2. Related Work

***(i)* Learning-Based Optical Flow.** Deep learning has greatly advanced the accuracy of dense optical flow estimation. Early convolutional architectures such as FlowNet [4, 9] and PWC-Net [34] positioned optical flow as a supervised regression problem over learned cost volumes, while subsequent approaches introduced multi-scale processing and feature warping strategies to improve efficiency and robustness [10, 35]. A major change occurred with RAFT [37], which reformulated optical flow estimation as iterative refinement over an all-pairs correlation field. RAFT dramatically improved accuracy yet introduced recurrent update operators that require multiple refinement steps at test time. SEA-RAFT [39] further improved this paradigm by simplifying the architecture and proposing probabilistic supervision and rigid-motion pretraining, showing a strong balance between efficiency and accuracy. Despite these improvements, current SOTA pipelines still rely heavily on multi-step test-time iterative refinement to gain performance.

Transformer-based approaches have also been explored to improve global correspondence reasoning. Methods such as GMFlow [41] and FlowFormer [8] leverage attention mechanisms to capture long-range dependencies and handle large displacements. GMFlow reformulates optical flow estimation as a global matching problem using attention and softmax normalization, removing recurrent updates but still learning flow-specific feature representations through supervised training. FlowFormer extends this idea by proposing cost-volume transformers to better model long-range dependencies. While these models improve global reasoning, they depend on task-specific feature learning and often

benefit from additional refinement.

***(ii)* Geometric Priors in Optical Flow.** Incorporating geometric cues into optical flow has long been explored as a way to improve correspondence estimation [33]. Earlier work often relied on piecewise rigid motion assumptions or geometric regularization [38]. More recently, learning-based approaches have begun integrating depth priors [21, 22] into optical flow pipelines. FlowSeek [30] represents a very recent example of this direction. It injects monocular depth predictions from foundation models [43] into a RAFT-style architecture and introduces motion bases to improve cross-dataset generalization. However, FlowSeek still trains a flow-specific backbone and relies heavily on iterative refinement to realize its performance gains. In contrast, our work treats depth predictions as representation priors rather than refinement guidance. Instead of embedding depth within a recurrent refinement pipeline, we directly fuse depth-aware features with semantic representations to enable single-pass correspondence estimation.

***(iii)* Foundation Models as Visual Representation.** Vision foundation models have recently demonstrated remarkable transferability across tasks. Self-supervised models such as DINO [2] and DINO-v2 [27] learn semantically meaningful feature representations that generalize well across domains without task-specific supervision. In parallel, depth foundation models such as Depth Anything [42, 43] have shown strong cross-dataset generalization while producing high-frequency geometric cues and sharp boundary predictions. Despite the success of foundation models across many visual tasks, their role in dense optical flow remains largely unexplored. Existing optical flow pipelines typically retain task-specific encoders even when incorporating geometric priors. In this work, we instead treat foundation models as frozen representation priors and focus on designing an inference pipeline that exploits their semantic and geometric cues. By combining DINO-v2 features with depth foundation representations and performing global correspondence estimation without iterative refinement, our approach aligns dense optical flow with the broader paradigm of foundation-driven inference, reducing reliance on task-specific training and test-time scaling.

Overall, our work differs from existing dense optical flow models in three key aspects. First, unlike refinement-based methods such as RAFT [37] and its extension SEA-RAFT [39], which rely on recurrent update operators and multi-step test-time refinement, our framework estimates optical flow in a single forward pass without allocating additional inference-time computation. Second, although our matching formulation follows the global correspondence strategy introduced in GMFlow [41], we depart from GMFlow by removing flow-specific feature learning and instead relying on frozen foundation representations to drive correspondence estimation. Third, while FlowSeek [30] integrates depth foundation models into a refinement-based architecture, it still trains a dedicated flow backbone and depends on iterative updates to achieve its performance gains. In contrast, we utilize both vision and geometric foundation models as fixed representation priors and focus on designing a lightweight inference pipeline that leverages them for dense optical flow correspondence estimation. Taken together, our approach shifts the emphasis from increasing architectural complexity or test-time scaling toward representation-driven optical flow, where strong pretrained semantic and geometric features enable accurate motion estimation without iterative refinement.

3. Method

Given two consecutive RGB frames $\mathbf{I}_1 \in \mathbb{R}^{H\times W\times 3}, \mathbf{I}_2 \in \mathbb{R}^{H\times W\times 3}$, our goal is to estimate dense optical flow $\mathbf{V} \in \mathbb{R}^{H\times W\times 2}$, where H, W denote the height and width of the image, respectively. We first extract dense visual features using a vision foundation model (Sec. 3.1), and then incorporate monocular depth priors to encode geometric structure (Sec. 3.2). These complementary cues are integrated through a joint cross-modal feature fusion strategy (Sec. 3.3), producing a unified representation for dense correspondence estimation. Finally, we apply a transformer-based global matching and flow propagation pipeline (Sec. 3.4) to infer optical flow in a single-pass, without iterative refinement or test-time optimization. In contrast to refinement-based optical flow pipelines, our framework estimates dense correspondences in a single forward pass by leveraging pretrained semantic and geometric priors. The overall loss function is provided in Sec. 3.5.

3.1. Foundation Visual Feature Extraction

Rather than learning a task-specific feature encoder as in prior optical flow methods [30, 39], we reuse pretrained semantic representations from vision foundation models. We begin by extracting dense semantic visual representations from each input frame using a frozen vision foundation model [27]. Specifically, given $\mathbf{I}_1, \mathbf{I}_2$, we compute feature embeddings via a pretrained DINOv2-S (small) [27] encoder:

$$\mathbf{F}_i^D = \Phi_{\text{DINO}}(\mathbf{I}_i), \quad i \in \{1, 2\}, \tag{1}$$

where, Φ_{DINO} denotes the DINOv2 backbone trained with large-scale self-supervised learning. The resulting feature maps are extracted at a spatial resolution of $\frac{1}{8}$ relative to the input image:

$$\mathbf{F}_i^D \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_D}, \tag{2}$$

where C_D denotes the channel dimensionality of the DINOv2 [27] features. DINOv2 representations provide dense, spatially coherent embeddings that encode both semantic consistency and fine-grained structural information.

Unlike task-specific optical flow encoders, these features are learned from large-scale image collections without motion supervision, and therefore capture visual regularities that generalize robustly across domains, appearance variations, and scene types. This property is particularly valuable in the zero-shot optical flow setting, where no dataset-specific adaptation is permitted.

Importantly, we keep the DINOv2 [27] backbone **frozen** at train and test time. This design choice serves two purposes. First, it preserves the rich visual priors acquired during large-scale pre-training thereby preventing overfitting to motion-specific biases. Second, freezing the backbone stabilizes optimization and ensures that optical flow estimation is performed purely via inference over fixed foundation representations, rather than through task-driven representation learning. As a result, our method decouples dense correspondence estimation from optical-flow-specific feature training, thus forming a key component of our approach.

3.2. Depth Prior Feature Encoding

To incorporate geometric context into dense optical flow correspondence estimation, we extract depth-aware representations from each input frame using a pretrained monocular depth foundation model. These geometric cues are particularly informative because motion discontinuities often align with depth boundaries. Concretely, given input images $\mathbf{I}_1, \mathbf{I}_2$, we compute depth features as

$$\mathbf{F}_i^Z = \Phi_{\text{Depth}}(\mathbf{I}_i), \quad i \in \{1, 2\}, \tag{3}$$

where Φ_{Depth} denotes the frozen Depth Anything V2-B (base) encoder [43]. Rather than relying on scalar depth predictions alone, we utilize intermediate feature representations produced by the depth decoder. These features encode rich geometric structure, including depth discontinuities, object boundaries, and spatial layout cues, while implicitly capturing uncertainty in regions such as occlusions and reflective surfaces [11] [12]. Prior works [31, 40] has shown that such intermediate representations often provide more informative and transferable geometric signals than final depth, particularly in downstream tasks requiring dense spatial reasoning. Since the native resolution and dimensionality of the depth features differ from those of the DINOv2 visual embeddings, we align them through a learnable projection module, i.e.,

$$\tilde{\mathbf{F}}_i^Z = \Psi_{\text{proj}}(\mathbf{F}_i^Z), \tag{4}$$

where, Ψ_{proj} is a lightweight convolutional downsampling network that maps the depth features to the same spatial resolution and channel dimensionality as $\mathbf{F}_i^D$. This projection enables easy integration of geometric and semantic information in subsequent processing stages, while remaining agnostic to camera intrinsics or explicit 3D reconstruction assumptions.

We emphasize that the depth encoder is kept frozen at train and test time. In doing so, we treat monocular depth estimation as a source of reusable geometric prior rather than a task to be optimized jointly with optical flow. This design ensures that our method preserves the strong generalization properties of depth foundation models and adheres to a strictly zero-shot setting, where dense optical flow is inferred entirely from fixed, pretrained representations.

3.3. Cross-Modal Feature Fusion

As alluded to above, the semantic representations extracted by DINOv2 and the geometric representations derived from the depth foundation model encode complementary, yet inherently different information. While DINOv2 features emphasize semantic consistency and visual appearance, depth features encode scene structure, boundaries, and geometric discontinuities. To effectively leverage both modalities, we construct a unified joint representation that integrates semantic and geometric cues at the feature level. Specifically, for each input frame, we first concatenate the DINOv2 features and the projected depth-aware features along the channel dimension:

$$\mathbf{F}_i^C = \text{Concat}\left(\mathbf{F}_i^D, \tilde{\mathbf{F}}_i^Z\right), \tag{5}$$

where, $\mathbf{F}_i^D$ denotes the semantic visual features extracted by the DINOv2 backbone, and $\tilde{\mathbf{F}}_i^Z$ denotes the depth-aware features aligned in resolution and dimensionality as described in Sec. 3.2. Rather than relying on direct concatenation alone, we pass the combined representation through a learnable cross-modal fusion network Ψ_{fusion}. Mathematically,

$$\hat{\mathbf{F}}_i = \Psi_{\text{fusion}}\left(\mathbf{F}_i^C\right), \tag{6}$$

where, Ψ_{fusion} consists of lightweight convolutional layers with residual connections. This fusion network is designed to explicitly model interactions between semantic appearance cues and geometric structure. This enables the network to reweight, suppress, or reinforce features across modalities in a data-driven manner.

Our framework differs from FlowSeek [30], where depth features are injected into a recurrent refinement rather than shaping the correspondence representation itself. Crucially, this fusion is performed before any correspondence matching or motion estimation. By integrating semantic and geometric information early, the resulting representation encodes both appearance similarity and structural consistency, which is essential for resolving ambiguities in challenging regions such as low-texture surfaces, repetitive patterns, motion boundaries, and illumination changes. Unlike approaches that inject geometric priors at later stages or through iterative refinement, our fusion strategy produces a single, coherent representation that the subsequent global matching and propagation pipeline can directly consume.

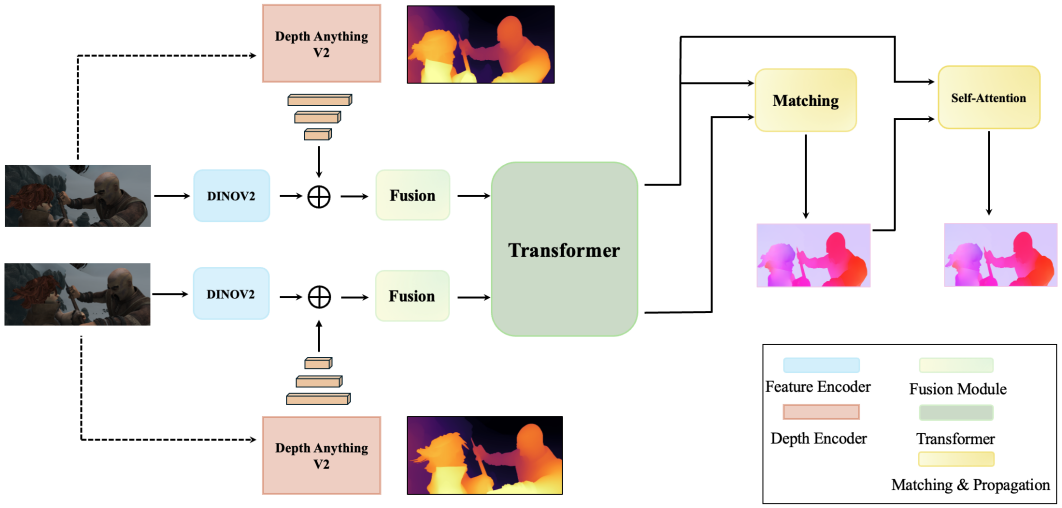

Figure 1. **Overview**. The conventional CNN-based feature encoder is replaced with DINOv2 [27] to provide semantically rich, large-scale self-supervised visual features, while original transformer-based feature interaction, global matching, and flow propagation modules remain unchanged. In addition, monocular depth estimates from Depth Anything V2 [43] are introduced as a geometric prior to improve feature conditioning and correspondence estimation.

Finally, we emphasize that the fusion network operates entirely on frozen foundation features and introduces only a modest number of learnable parameters. This design preserves the strong cross-dataset generalization of pretrained foundation representations while allowing sufficient flexibility to reconcile modality-specific biases, ensuring that dense optical flow estimation remains an inference problem over fixed, pretrained representations rather than a task requiring domain-specific feature learning.

3.4. Global Matching and Propagation

The fused feature representations $\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2$ are first processed through a transformer encoder to produce $\mathbf{F}_1$ and $\mathbf{F}_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D}$ for the two input frames, we then estimate dense optical flow via a transformer-inspired global matching and propagation pipeline. This design follows the global matching formulation introduced in prior work [41], but is employed here strictly as a single-pass inference mechanism operating on fixed foundation representations, without iterative refinement or test-time optimization.

Global Matching. We first compute an all-pairs correlation volume that measures the similarity between every spatial location in the 1st frame and every location in the second frame:

$$\mathbf{C}_{\text{flow}} = \frac{\mathbf{F}_1\mathbf{F}_2^\top}{\sqrt{D}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}}, \tag{7}$$

where, D denotes the feature dimensionality. We convert the correlation scores into a probabilistic matching distribution via a softmax over all candidate correspondences:

$$\mathbf{M}_{\text{flow}} = \text{softmax}\left(\mathbf{C}_{\text{flow}}\right) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}}. \tag{8}$$

Let $\mathbf{G}_{\text{2D}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2}$ denote the 2D coordinate grid of the second frame. The expected correspondence for each pixel in the first frame is then obtained as the expectation under the matching distribution

$$\hat{\mathbf{G}}_{\text{2D}} = \mathbf{M}_{\text{flow}}\,\mathbf{G}_{\text{2D}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2}. \tag{9}$$

The initial dense optical flow field is computed as the displacement between corresponding coordinates

$$\hat{\mathbf{V}}_{\text{flow}} = \hat{\mathbf{G}}_{\text{2D}} - \mathbf{G}_{\text{2D}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2}. \tag{10}$$

This formulation yields sub-pixel accurate correspondences and enables global reasoning over large displacements without relying on local search windows or recurrent updates. We intentionally keep the matching operator unchanged to isolate the effect of foundation-driven representations.

Flow Propagation. This step effectively spreads reliable optical flow estimates across semantically and geometrically consistent regions. The softmax-based matching formulation assumes that reliable correspondences exist for all pixels, which may not hold in occluded, textureless, or out-of-boundary regions. To address this, we propagate flow estimates using feature self-similarity within the first frame. Specifically, we compute an attention matrix based on intra-frame feature affinity as

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{F}_1\mathbf{F}_1^\top}{\sqrt{D}}\right) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}}, \tag{11}$$

which captures structural similarity between pixels in the reference frame. The final optical flow is obtained by propagating reliable flow estimates across similar features:

$$\mathbf{V} = \mathbf{A}\,\hat{\mathbf{V}}_{\text{flow}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2}. \tag{12}$$

Method	#refine	Things (val, clean)	Sintel (train, clean)				Sintel (train, final)			
		EPE	EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}
RAFT [37]	32	4.25	1.43	0.33	1.54	9.04	2.71	0.51	2.98	17.62
GMFlow [41]	0	3.48	1.50	0.46	1.77	8.26	2.96	0.72	3.45	17.70
GMFlow [41]	1	2.80	1.08	0.30	1.25	6.26	2.48	0.51	2.81	15.76
SEA-RAFT (S) [39]	4	-	1.27	-	-	-	4.32	-	-	-
FlowSeek (T) [30]	4	3.94	1.16	0.25	1.31	7.26	2.48	0.43	2.63	16.60
Ours	**0**	**3.02**	**1.46**	0.39	2.05	7.74	**2.81**	0.64	3.36	16.99

Table 1. **Cross-dataset generalization** after training on Chairs and Things. No target-domain fine-tuning is applied. Lower is better.

This proposed propagation allows information from confidently matched regions to inform ambiguous or occluded areas, leveraging the structural coherence encoded in the fused foundation features.

Importantly, overall matching and propagation pipeline is executed in a single forward pass, without recurrence, iterative refinement, or test-time optimization. Combined with the frozen foundation representations described in the previous sections, this design ensures that dense optical flow estimation is performed purely as inference over pretrained semantic and geometric priors, aligning with the zero-shot formulation of our approach. Figure 1 provides the overall pipeline of the proposed framework.

3.5. Training Loss

We supervise flow predictions using an ℓ_1 regression loss between the predicted flow and the ground-truth flow field. The ℓ_1 loss provides robustness to outliers and aligns with the endpoint error metric used during evaluation. The loss L is applied to both intermediate and final flow predictions, with higher weight assigned to the final prediction.

$$L = \sum_{i=1}^{N} \gamma^{N-i} \left\| \mathbf{v}^{(i)} - \mathbf{v}_{gt} \right\|_1, \tag{13}$$

where, N denotes the total number of flow predictions, $\mathbf{v}^{(i)}$ denotes the predicted flow at stage i, and γ controls the relative weighting between intermediate and final predictions.

4. Experiments

Implementation details. We implement our approach in PyTorch [28]. Training is conducted on two NVIDIA RTX 6000 GPUs using the AdamW optimizer [24]. Unless otherwise stated, we keep all hyperparameters fixed across experiments. We freeze DINOv2 visual backbone and the Depth Anything V2 depth backbone throughout training and inference. Only the lightweight projection, fusion, and matching modules are optimized. This design keeps the number of trainable parameters small while isolating the contributions of the pretrained semantic and geometric priors.

Datasets and evaluation setup. We follow the standard optical flow training and evaluation protocol established in [37, 39]. We train on synthetic datasets and evaluate generalization on real-world benchmarks without target-domain fine-tuning. Specifically, we train on FlyingChairs [4] and FlyingThings3D [25], and then evaluate cross-dataset generalization on the Sintel [1] and KITTI [26] training splits. For completeness and comparison with prior SOTA methods, we further report results after domain-specific fine-tuning on Sintel and KITTI 2015. These fine-tuned results are presented separately and do not support our main claims about generalization without target-domain adaptation.

Metrics. We evaluate optical flow accuracy using the standard endpoint error (EPE), defined as the average ℓ_2 distance between predicted and ground-truth flow vectors over all pixels. To provide a more fine-grained analysis across motion regimes, we report EPE stratified by ground-truth flow magnitude: $s_{0\text{–}10}$, $s_{10\text{–}40}$, and s_{40+}, corresponding to pixel displacements of 0–10, 10–40, and greater than 40 pixels, respectively. For Sintel, we report both overall EPE and these motion-stratified metrics. For KITTI 2015, we also report F1-all, which measures the percentage of outlier pixels under the official KITTI evaluation protocol.

Training protocol. Training proceeds in stages. We first train on FlyingChairs for 200k iterations using a batch size of 16, a learning rate of 4×10^{-4}, and random crops of size 384×512. We then continue training on FlyingThings3D for 800k iterations with a reduced learning rate of 2×10^{-4} and crop size 384×768. For fine-tuning experiments, we adopt a mixed training set consisting of KITTI [6], HD1K [13], FlyingThings3D, and Sintel (denoted as TSKH), and train for an additional 200k iterations using crops of size 320×896. Finally, for the KITTI 2015 evaluation, we fine-tune the model for 90k iterations with a batch size of 8, a learning rate of 2×10^{-4}, and a crop size of 320×1152.

4.1. Cross-Dataset Generalization

Table 1 summarizes cross-dataset generalization results after training exclusively on FlyingChairs and FlyingThings3D, with no target-domain fine-tuning at test time. This setting is deliberately challenging, as it probes whether a

model trained on synthetic data can transfer to domains with substantially different appearance, motion statistics, and rendering characteristics. Overall, our method generalizes strongly across synthetic and real benchmarks. On the FlyingThings3D validation set, our approach achieves an EPE of 3.02, outperforming RAFT and GMFlow without refinement, while competitive with methods that rely on iterative updates. Notably, this performance is obtained without recurrent refinement or test-time optimization.

On Sintel, our method shows a clear advantage on more challenging Final pass. In particular, we achieve an EPE of **2.81** on Sintel Final, substantially improving over SEA-RAFT (4.32 EPE) despite SEA-RAFT employing multiple refinement steps. Compared to GMFlow without refinement, our method also achieves lower overall error on Sintel Final while operating in a strictly single-pass inference regime. When compared to FlowSeek [30], our approach attains comparable performance on both Sintel passes, despite FlowSeek benefiting from additional pretraining on TartanAir. Taken together, these results support our central hypothesis: stronger semantic and geometric priors can partially substitute for increased inference-time computation in dense optical flow estimation.

4.2. Results on Sintel and KITTI

Results on Sintel. Table 2 reports performance on Sintel train set after training on Chairs, Things, and the mixed TSKH dataset. Overall, our method achieves competitive performance on both Clean and Final passes, despite operating in a single-pass setting without iterative refinement or additional large-scale pretraining. On the more challenging Sintel Final split, our approach outperforms RAFT, GMFlow without refinement, and FlowSeek, while remaining close to refinement-based GMFlow. This trend is notable because refinement-based methods explicitly revisit and correct flow estimates via multiple update iterations, whereas our model relies on a single forward pass over frozen foundation representations. These results shows that the use of visual semantic and geometric representation priors captures sufficient global context to handle the appearance changes and motion patterns present in Sintel Final.

SEA-RAFT [39] achieves the strongest performance on both Sintel splits. As shown in Table 2, this gap likely reflects differences in both training scale and auxiliary pretraining. In particular, SEA-RAFT is trained using substantially larger compute resources, e.g., $8\times$ NVIDIA L40 GPUs compared to $2\times$ RTX 6000 GPUs in our setup, corresponding to approximately $4\times$ higher effective training compute. This difference is also reflected in larger effective batch sizes and higher-resolution training crops (batch size 32 with 432×960 crops for SEA-RAFT versus batch size 8 with 320×896 crops in our training). Moreover, both SEA-RAFT and FlowSeek benefit from additional pretraining on TartanAir, which introduces greater scene diversity and motion variability beyond the TSKH mixture. Despite these advantages, our approach remains competitive, highlighting the effectiveness of foundation-model-driven representations for dense optical flow estimation.

Method	Extra Data	#refine	Sintel Clean (EPE)	Sintel Final (EPE)
RAFT [37]	–	32	0.768	1.217
GMFlow [41]	–	0	0.947	1.276
GMFlow [41]	–	1	0.762	1.110
SEA-RAFT (S) [39]	TartanAir	4	0.546	0.782
FlowSeek (T) [30]	TartanAir	4	0.71	1.28
Ours	**–**	**0**	**0.847**	**1.140**

Table 2. **Performance on the Sintel train set** after training on Chairs, Things, and mixed datasets (TSKH). Our method outperforms RAFT, GMFlow (wo refinement), and FlowSeek on Sintel Final. SEA-RAFT performs best on both splits, likely due to differences in training scale and extra pretraining, including substantially larger effective batch sizes (**32 with crop size** 432×960 **for SEA-RAFT versus 8 with crop size** 320×896 **in our training**).

Results on KITTI. Table 3 presents results on the KITTI train set after fine-tuning on KITTI following training on Chairs, Things, and TSKH. Our framework achieves performance comparable to GMFlow without refinement, while using frozen foundation priors rather than learned flow-specific encoders. Although refinement-based methods such as SEA-RAFT and FlowSeek obtain lower EPE and F1-all scores, they benefit from both recurrent update mechanisms and pretraining on TartanAir, which provides optical flow priors and scene structures closely aligned with KITTI.

KITTI poses additional challenges due to frequent occlusions, thin structures, and sharp motion boundaries. Iterative refinement is particularly effective in such scenarios as it allows a model to correct local errors and enforce spatial consistency. In contrast, our framework deliberately avoids refinement to maintain a simple, scalable inference pipeline based on frozen foundation model priors. The observed performance gap, therefore, reflects a trade-off between architectural simplicity and fine-grained local accuracy.

For completeness, we note that FlowSeek [30] does not release a checkpoint trained exclusively on KITTI; consequently, we evaluate it using the publicly available model trained under a mixed-dataset setting. Overall, these results indicate that while refinement and additional pretraining remain beneficial for achieving peak performance on KITTI, foundation-model-driven representations already provide a strong baseline. Incorporating lightweight refinement mechanisms or larger-scale pretraining remains a promising direction for improving performance further without abandoning the foundation-based paradigm.

5. Ablation

1. Performance analysis of the proposed fusion module. We analyzed the impact of the proposed fusion mod-

RAFT [37] GMFlow [41] SEA-RAFT [39] FlowSeek [30] **Ours**

Figure 2. **Qualitative comparison on Sintel (Final).** This benchmark contains severe motion blur, illumination variation, and large displacements. Despite operating without iterative refinement, our method preserves sharp motion boundaries and produces accurate flow estimates, performing comparably to—and in some cases better than—refinement-based approaches.

Method	Extra Data	#refine	KITTI EPE	KITTI F1-all
RAFT [37]	–	32	0.63	1.47
GMFlow [41]	–	0	2.06	7.57
GMFlow [41]	–	1	1.36	5.17
SEA-RAFT (S) [39]	TartanAir	4	0.93	2.65
FlowSeek (T) [30]	TartanAir	4	1.26	3.90
Ours	**–**	**0**	1.99	7.40

Table 3. **Performance on the KITTI train set** after fine-tuning on KITTI, following training on Chairs, Things, and the mixed TSKH dataset.

ule, which integrates visual semantic features from DINOv2 with geometric cues from the depth foundation model. As shown in Table 4, enabling the fusion module consistently improves cross-dataset generalization across popular benchmarks. On FlyingThings3D, fusion reduces the EPE from 3.52 to 3.02, indicating that combining visual and geometric priors improves correspondence estimation even on synthetic validation data. The improvement is more pronounced on the Sintel benchmark. On the Clean split, fusion reduces EPE from 1.575 to 1.46, while on the more challenging Final split, the error decreases from 3.12 to 2.81, corresponding to approximately a 10% relative improvement. A motion-stratified study further highlights the benefits of fusion across displacement regimes. In particular, the largest gain appears in the large-motion regime (s_{40+}) on Sintel Final, where EPE decreases from 19.37 to 16.99.

Method	Things (val, clean)	Sintel (train, clean)				Sintel (train, final)			
	EPE	EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}
(w/o) Fusion	3.52	1.575	0.45	2.01	8.53	3.12	0.685	3.57	19.37
(w/) Fusion	3.02	1.46	0.39	2.05	7.74	2.81	0.64	3.36	16.99

Table 4. Ablation study evaluating the contribution of the proposed fusion module.

2. Usefulness of depth features. To analyse the usefulness of depth feature, we train the depth model for 100k iterations on the FlyingChairs dataset with and without depth features. For evaluating the variant without depth, we also exclude the cross-modal fusion module used for feature integration. We observed that incorporating depth features leads to a significant improvement in dense optical flow estimation, reducing the EPE from **1.77** to **0.87**.

6. Conclusion and Limitations

In this paper, we present a dense optical flow estimation framework that utilizes pretrained complementary visual priors from foundation models. By using frozen representations from DINOv2 [27] and a monocular depth from [43] foundation model and integrating them via a lightweight fusion and global matching pipeline, our approach estimates optical flow in a single forward pass without task-specific backbone training, iterative refinement, or test-time optimization. Experimental evaluations show strong cross-dataset generalization. Moreover, our ablation show that strong foundation model representations have the potential to substitute for test-time scaling in dense optical flow.

Despite encouraging results, our work has a few limitations. First, our proposed approach may yield lower accuracy in challenging scenarios involving heavy occlusions, thin structures, or fine motion boundaries. Next, our approach depends on the availability and quality of large pretrained foundation models, whose training costs lie outside the scope of this work and whose representations may carry biases inherited from large-scale datasets. Addressing these limitations opens promising research directions.

Acknowledgment. The authors thank High Performance Research Computing (HPRC) at Texas A&M University, Texas, USA for providing us with the startup credits for utilizing the GPU-server facility.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012.

[2] 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] Weirong Chen, Suryansh Kumar, and Fisher Yu. Uncertainty-driven dense two-view structure from motion. *IEEE Robotics and Automation Letters*, 8(3):1763–1770, 2023.

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, pages 713–729. Springer, 2020.

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[7] Yuxiang Huang, Yuhao Chen, and John Zelek. Zero-shot monocular motion segmentation in the wild by combining deep learning with geometric motion model fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2733–2743, 2024.

[8] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022.

[9] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[10] Azin Jahedi, Maximilian Luz, Marc Rivinius, Lukas Mehl, and Andrés Bruhn. Ms-raft+: high resolution multi-scale raft. *International Journal of Computer Vision*, 132(5): 1835–1856, 2024.

[11] Nishant Jain, Suryansh Kumar, and Luc Van Gool. Enhanced stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2023.

[12] Nishant Jain, Suryansh Kumar, and Luc Van Gool. Learning robust multi-scale representation for neural radiance fields from unposed images. *International Journal of Computer Vision*, 132(4):1310–1335, 2024.

[13] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.

[14] Suryansh Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2019.

[15] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 51–60, 2020.

[16] Suryansh Kumar and Luc Van Gool. Organic priors in non-rigid structure from motion. In *European Conference on Computer Vision*, pages 71–88. Springer, 2022.

[17] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *Proceedings of the IEEE international conference on computer vision*, pages 4649–4657, 2017.

[18] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2018.

[19] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1705–1717, 2019.

[20] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7274–7283, 2019.

[21] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. In *The Eleventh International Conference on Learning Representations*.

[22] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Single image depth prediction made better: A multivariate gaussian take. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17346–17356, 2023.

[23] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36:37193–37229, 2023.

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

[25] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[26] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] 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.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[29] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9945–9953, 2019.

[30] Matteo Poggi and Fabio Tosi. Flowseek: Optical flow made easier with depth foundation models and motion bases. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

[31] Qiyang Qian, Hansheng Chen, Masayoshi Tomizuka, Kurt Keutzer, Qianqian Wang, and Chenfeng Xu. Bridging viewpoint gaps: Geometric reasoning boosts semantic correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11579–11589, 2025.

[32] Shenhan Qian, Ganlin Zhang, Shangzhe Wu, and Daniel Cremers. Flow4r: Unifying 4d reconstruction and tracking with scene flow. *arXiv preprint arXiv:2602.14021*, 2026.

[33] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2432–2439. IEEE, 2010.

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1408–1423, 2019.

[36] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[38] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1384, 2013.

[39] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024.

[40] Shangbo Wu, Yu-an Tan, Ruinan Ma, Wencong Ma, Dehua Zhu, and Yuanzhang Li. Boosting generative adversarial transferability with self-supervised vision transformer features. *arXiv preprint arXiv:2506.21046*, 2025.

[41] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[43] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[44] Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23385–23395, 2024.

[45] Weiming Zhuang, Chen Chen, Zhizhong Li, Sina Sajadmanesh, Jingtao Li, Jiabo Huang, Vikash Sehwag, Vivek Sharma, Hirotaka Shinozaki, Felan Carlo Garcia, et al. Argus: A compact and versatile foundation model for vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4418–4429, 2025.