UNIDISP: LEARNING UNIFIED HIERARCHICAL STEREO REPRESENTATIONS FOR PROGRESSIVE DISPARITY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep stereo disparity estimation has long followed staged pipelines of feature extraction, correspondence modeling, and disparity refinement. Although effective, such pipelines distribute stereo representation learning and disparity inference across multiple modules, while global ambiguous correspondences can remain difficult to resolve from local matching evidence alone. We propose **UniDisp**, a unified backbone that estimates disparity from stereo representations. This is motivated by a simple hypothesis: disparity can be treated as an **intrinsic geometric attribute** encoded in stereo representations, rather than a quantity derived from matching within a modular stereo pipeline. We instantiate this idea with a hierarchical stereo encoder that jointly models cross-view displacement cues across spatial scales, a prediction head that maps multi-level stereo representations directly to disparity, organized into a two-stage framework with structurally identical yet functionally complementary encoder–head stages. The first stage performs initial estimation from the raw stereo input, while the second uses implicit warping to construct disparity-aware inputs for residual refinement. This design unifies stereo representation learning, disparity prediction, and refinement into a compact backbone, enabling disparity to be inferred from the learned stereo representations themselves. Extensive experiments demonstrate consistent cross-domain transfer, reaching state-of-the-art performance on several real-world benchmarks.

1 INTRODUCTION

Stereo matching aims to estimate pixel-level disparity from rectified stereo images, providing dense geometric information for a wide range of applications, including 3D perception, 3D reconstruction, autonomous driving, and robot vision Li et al. (2019); Seitz et al. (2006); Murray & Little (2000). Modern deep stereo matching methods have long relied on explicit correspondence modeling: encoding left and right views, constructing matching representations that encompass candidate disparities, and obtaining final predictions through cost aggregation, disparity regression, or iterative updates Kendall et al. (2017); Chang & Chen (2018); Lipson et al. (2021); Xu et al. (2023). While effective at modeling local correspondences, this paradigm separates stereo representation learning, correspondence modeling, and disparity inference into distinct modules. As matching and refinement modules become increasingly complex, the overall pipeline relies more heavily on intermediate representations and cascaded processing, potentially making global matching ambiguities harder to resolve from local matching evidence alone.

This motivates us to reconsider the role of disparity in stereo vision. Our core hypothesis is that disparity can be treated as an intrinsic geometric attribute encoded in the binocular relationship and should therefore be inferred from a unified stereo representation. Early work such as DispNetS showed that disparity can be directly regressed from channel-wise concatenated stereo images Mayer et al. (2016), while recent studies have further explored cost-volume-free or regression-centric stereo estimation Wang & Deng (2026); Guan et al. (2026). These studies suggest that an explicit matching space is not indispensable for disparity estimation. Rather than asking whether direct disparity regression is possible, we ask a more fundamental question: how should stereo representations be organized so that binocular geometry can be encoded and disparity can be inferred from them? To

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

this end, this paper propose UniDisp, a unified stereo disparity estimation backbone based on fully convolutional hierarchical stereo representations.

A key challenge is to represent binocular displacement across different spatial scales. In rectified stereo images, disparity appears as horizontal cross-view displacement, while its magnitude and spatial structure vary across a scene. We therefore design a Hierarchical Stereo Representation (HSR) Encoder that directly encodes the joint stereo input through a convolutional hierarchy. At high-resolution stages, the local inductive bias of convolution preserves fine-grained cross-view displacement cues. As spatial resolution decreases, the effective receptive field expands and the same image-space displacement occupies fewer feature coordinates, making larger displacement structures easier to represent. In this way, the encoder organizes binocular displacement into a hierarchy of stereo representations, from fine-grained local offsets to broader structural displacements.

Given these hierarchical stereo representations, the next question is how pixel-level disparity can be inferred from them. We design a Hierarchical Disparity Prediction (HDP) Head that reads out disparity from the full representation hierarchy and progressively integrates multi-scale geometric cues in a coarse-to-fine manner. Residual propagation carries coarse displacement estimates toward finer stages, allowing high-level structural information and fine-grained spatial cues to jointly contribute to the final prediction. UniDisp further employs two encoder–head stages that share the same architecture but serve complementary roles. The first stage predicts an initial disparity, while the second re-encodes the implicitly warped stereo input to estimate a residual correction. Within the iterative refinement process, the same hierarchical representation–prediction formulation is reused for residual estimation on disparity-aware stereo inputs.

UniDisp is trained from scratch using only synthetic data and demonstrates strong cross-domain generalization with competitive performance across real-world scenarios. More fundamentally, UniDisp approaches stereo matching from the perspective of how stereo representations should be organized, treating disparity as an intrinsic property of the learned binocular representation. Accordingly, the HSR Encoder hierarchically encodes stereo geometry, while the HDP Head directly infers disparity from the resulting representation, with the same design used for both initialization and refinement. This perspective offers a new way to think about stereo representation learning and provides a simpler backbone-level approach to disparity estimation.

2 RELATED WORK

Stereo Matching with Explicit Correspondence Modeling. Deep neural networks have become the foundation of modern stereo matching. Early work learned local matching costs from image patches Žbontar & LeCun (2016), while DispNet Mayer et al. (2016) demonstrated end-to-end disparity estimation. GC-Net Kendall et al. (2017) and PSMNet Chang & Chen (2018) subsequently established the widely adopted pipeline of feature extraction, cost-volume construction, cost aggregation, and disparity regression. Later studies improved this paradigm through group-wise correlation Guo et al. (2019), guided aggregation Zhang et al. (2019), adaptive cost aggregation Xu & Zhang (2020), cascaded and fused cost volumes Shen et al. (2021), and attention-enhanced matching representations Xu et al. (2022). More recent methods further strengthen correspondence modeling through geometry encoding Xu et al. (2023), motif-aware correlation Chen et al. (2024), and learned disparity hypotheses Guan et al. (2024).

Recent stereo methods also increasingly incorporate monocular geometric knowledge. MonSter Cheng et al. (2025) jointly exploits stereo and monocular depth cues, while DEFOM-Stereo Jiang et al. (2025) and BridgeDepth Guan et al. (2025) introduce representations or geometric priors from pretrained monocular foundation models. Such approaches provide stronger semantic or geometric priors for difficult regions, but generally retain explicit correspondence representations or rely on pretrained monocular knowledge.

Progressive Disparity Refinement. RAFT-Stereo Lipson et al. (2021) introduced recurrent disparity refinement to stereo matching by repeatedly querying multi-level correlation features and updating disparity with convolutional GRUs. CREStereo Li et al. (2022) combines recurrent refinement with coarse-to-fine local correlation, while IGEV-Stereo Xu et al. (2023) augments iterative updates with a geometry encoding volume. Subsequent methods enrich the information available during refinement through high-frequency cues in DLNR Zhao et al. (2023), multi-frequency hidden-state ag-

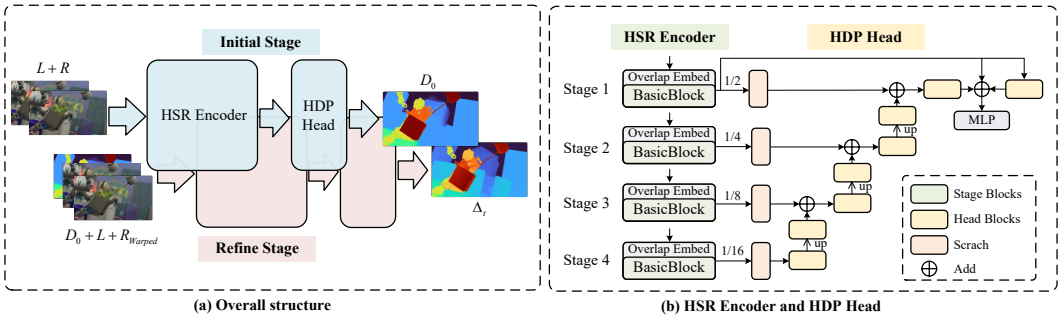

Figure 1: **Overview of UniDisp.** UniDisp consists of structurally identical initialization and refinement stages, each built with an HSR Encoder and an HDP Head to progressively infer disparity from hierarchical stereo representations.

gregation in Selective-Stereo Wang et al. (2024), and multi-range geometry encoding in IGEV++ Xu et al. (2025). These methods demonstrate the effectiveness of progressively correcting disparity estimates using matching evidence accumulated across iterations.

UniDisp follows the idea of progressive refinement from a different representation perspective. Rather than organizing stereo inference around explicit matching volumes, we treat disparity as an intrinsic geometric attribute of the learned stereo representation and directly infer it from a hierarchical stereo backbone. The same representation principle is shared by initialization and refinement, without relying on explicit cost volumes or external monocular priors.

3 METHOD

Given a pair of rectified stereo images $\mathbf{I}_L, \mathbf{I}_R \in \mathbb{R}^{3\times H\times W}$, our goal is to estimate a dense disparity map $\hat{\mathbf{D}} \in \mathbb{R}^{H\times W}$ for the left view. Instead of separately extracting features and explicitly constructing correspondence volumes, UniDisp directly learns hierarchical stereo representations from the binocular input and infers disparity from the resulting feature hierarchy. As illustrated in Fig. 1, UniDisp consists of an HSR Encoder for hierarchical stereo representation learning and an HDP Head for progressive disparity prediction. These components are organized into two structurally consistent stages: the first estimates an initial disparity, while the second uses disparity-guided warping to expose residual misalignment and predicts a correction. In this way, stereo representation learning, disparity prediction, and refinement are carried out under a unified hierarchical encoding–prediction architecture.

3.1 HIERARCHICAL STEREO REPRESENTATION ENCODER

Stereo disparity exhibits geometric structures over different spatial ranges: fine-resolution representations are important for preserving small displacement variations and object boundaries, while representations with larger receptive fields provide broader spatial support for resolving larger-scale displacement structures. Hierarchical and multi-resolution representations have therefore been widely adopted in modern stereo matching Chang & Chen (2018); Tankovich et al. (2021); Lipson et al. (2021); Xu et al. (2023). Following this principle, we construct the HSR Encoder as a four-stage hierarchy that progressively transforms joint stereo observations into multi-level stereo representations.

A key aspect of our design is to explicitly decouple spatial transition from intra-stage representation learning. Inspired by the stage-wise formulation of hierarchical vision encoders such as SegFormer Xie et al. (2021), each stage first performs an overlapping spatial transition and then applies a sequence of representation blocks at a fixed resolution. Formally, the i-th stage is defined as

$$\begin{aligned}\mathbf{F}^i &= \mathcal{L}_i\left(\mathrm{OPE}_i\left(\mathbf{F}^{i-1}\right)\right)\\ &= \left(\mathcal{B}_i^{D_i}\circ\cdots\circ\mathcal{B}_i^1\right)\left[\mathrm{Conv}_{3\times3}^{s=2,p=1}\left(\mathbf{F}^{i-1}\right)\right], \qquad i=1,\ldots,4,\end{aligned} \tag{1}$$

where $\mathbf{F}^i \in \mathbb{R}^{H_i \times W_i \times C_i}$ denotes the representation at stage i, with H_i and W_i reduced by a factor of two relative to the preceding stage. OPE_i performs spatial downsampling and channel projection through an overlapping 3×3 convolution, while $\mathcal{L}_i$ consists of D_i representation blocks $\{\mathcal{B}_i^j\}_{j=1}^{D_i}$ that preserve the spatial resolution and channel dimension within the stage.

Separating these two operations provides a clean hierarchical formulation in which the spatial topology of the encoder is independent of the internal representation operator. Consequently, different convolutional or attention-based blocks can be evaluated under the same resolution, channel, and depth configuration, allowing the representation mechanism itself to be studied without altering the hierarchical structure. Our controlled comparisons in Sec. 4.2 show that a fully convolutional configuration provides the strongest stereo representation within this framework. We therefore instantiate $\mathcal{L}_i$ entirely with convolutional blocks, yielding a simple hierarchical encoder with strong local inductive bias while progressively enlarging the effective receptive field across stages.

3.2 Hierarchical Disparity Prediction Head

The HSR Encoder progressively encodes stereo geometry into a hierarchy of representations with different spatial resolutions and receptive fields. Rather than predicting disparity solely from the deepest representation, we introduce a Hierarchical Disparity Prediction Head (HDP Head) that progressively integrates the entire feature hierarchy for disparity inference. Inspired by Dense Prediction Transformer Ranftl et al. (2021), encoder representations are first projected to a common feature dimension and then successively fused from low to high resolution.

Formally, the hierarchical prediction process is defined as

$$\begin{aligned} \mathbf{G}^4 &= \phi_4\big(\mathbf{F}^4\big)\,, \\ \mathbf{G}^i &= \mathcal{R}_i\left(\phi_i\big(\mathbf{F}^i\big)\,, \mathrm{Up}\big(\mathbf{G}^{i+1}\big)\right), \qquad i = 3, 2, 1, \\ \mathbf{H} &= \mathcal{R}_0\left(\psi\big(\mathbf{F}^1\big)\,, \mathbf{G}^1\right), \end{aligned} \tag{2}$$

where $\phi_i(\cdot)$ projects the encoder representations to a common feature dimension, $\mathrm{Up}(\cdot)$ denotes spatial upsampling, and $\mathcal{R}_i$ denotes the fusion transformation. Each $\mathcal{R}_i$ adopts the same convolutional representation operator used in the HSR Encoder, maintaining a consistent transformation mechanism across encoding and prediction. The additional transformation $\psi(\cdot)$ provides a direct pathway from the highest-resolution encoder representation to the final fused representation $\mathbf{H}$.

This progressive fusion propagates large-scale geometric context from deeper stages toward finer resolutions while recovering local structures from higher-resolution representations. The resulting representation $\mathbf{H}$ is directly mapped to disparity prediction without constructing an additional multi-scale hierarchy. The same HDP Head is employed in both stages of UniDisp: the initialization stage predicts the initial disparity through a discrete disparity distribution, whereas the refinement stage predicts a residual correction from disparity-aware representations. The two stages therefore share the same hierarchical prediction mechanism while performing complementary estimation and refinement.

3.3 Unified Progressive Disparity Estimation

UniDisp employ the HSR Encoder and HDP Head within a two-stage disparity estimation framework following the classification-before-regression formulation of WAFT-Stereo Wang & Deng (2026). In the initial stage, the left and right images are jointly encoded into hierarchical stereo representations, from which the HDP Head predicts a discrete disparity distribution. The initial disparity $\hat{\mathbf{D}}_0$ is obtained by expectation and restored to the target resolution using convex upsampling.

The initial estimate is then used to warp the right image toward the left view. Unlike WAFT-Stereo, we apply pixel unshuffle before refinement to retain local image information at a reduced spatial resolution. The left view, warped right view, and current disparity are subsequently processed by a second HSR–HDP network to predict a residual correction. The overall progressive estimation can

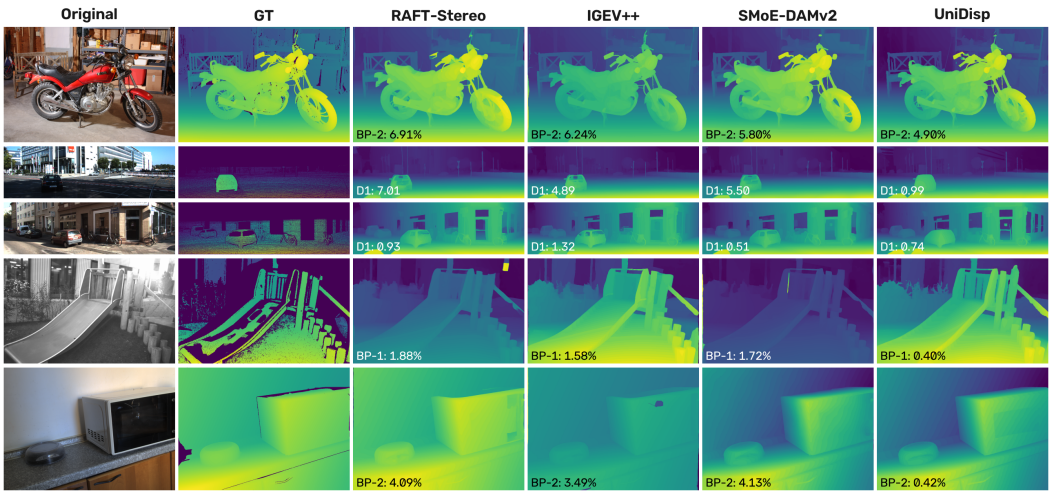

Figure 2: Cross-Domain generalization on real-world scenarios. Comparison with RAFT-Stereo, IGEV++, and SMoE across five datasets. Our UniDisp is more robust to global matching ambiguities, particularly in challenging specular regions on Booster (bottom row).

be summarized as

$$\begin{aligned} \hat{\mathbf{D}}_0 &= \mathrm{HDP}_0\left(\mathrm{HSR}_0(\mathbf{I}_L, \mathbf{I}_R)\right), \\ \widetilde{\mathbf{I}}_R &= \mathcal{W}(\mathbf{I}_R, \hat{\mathbf{D}}_0), \\ \Delta\mathbf{D} &= \mathrm{HDP}_1\left(\mathrm{HSR}_1(\mathbf{I}_L, \widetilde{\mathbf{I}}_R, \hat{\mathbf{D}}_0)\right), \\ \hat{\mathbf{D}} &= \hat{\mathbf{D}}_0 + \Delta\mathbf{D}. \end{aligned} \tag{3}$$

Warping transforms the original large displacement into a smaller residual misalignment, allowing the second stage to focus on correcting the initial estimate. Although initialization and refinement operate under different geometric conditions, both follow the same hierarchical stereo representation and disparity prediction principle.

Training objective. We follow the supervision strategy of WAFT-Stereo. The initial disparity distribution is supervised using a soft classification loss constructed from the ground-truth disparity, while refinement predictions are optimized using the Mixture-of-Laplace loss with exponentially decayed weights:

$$\mathcal{L} = \mathcal{L}_{\text{init}} + \sum_{t=1}^{T} \gamma^{T-t} \mathcal{L}^t_{\text{MoL}}, \qquad \gamma = 0.9. \tag{4}$$

This two-stage design retains the coarse-to-fine estimation strategy of WAFT-Stereo while unifying initialization and refinement under the same HSR–HDP architecture.

3.4 Discussion

A central perspective of this work is to reconsider stereo matching from how stereo representations should be organized for disparity estimation. We hypothesize that disparity can be encoded as an intrinsic geometric attribute of jointly learned stereo representations, motivating the HSR Encoder to hierarchically model stereo geometry and the HDP Head to progressively decode disparity. The same HSR–HDP design is shared by initialization and refinement, while our experiments favor fully convolutional representation operators for accuracy and cross-domain generalization. Together, these designs establish a streamlined representation–prediction paradigm without explicit matching construction or external visual foundation models.

4 EXPERIMENTS

This section presents the implementation details and evaluation protocol of our method. We then compare the pretrained model with state-of-the-art stereo methods, including approaches with and without depth priors. Finally, we conduct ablation studies to analyze the stereo representation operators of the proposed model.

Implementation details. Our main model, denoted as UniDisp-L, contains four hierarchical stages with widths of $[64, 128, 320, 512]$ and depths of $[3, 6, 40, 3]$. For controlled architectural ablations, we use a smaller variant, UniDisp-S, with the same stage widths and reduced depths of $[3, 4, 6, 3]$. UniDisp-S retains the complete hierarchical encoder, disparity prediction head, and progressive estimation pipeline, while substantially reducing the cost of repeated ablation experiments. For other parameters, we follow the WAFT-Stereo Wang & Deng (2026) settings: initial disparity bins = 40, MAX DISP = 800, and 4 iterations for disparity refinement. **Training.** We train UniDisp on synthetic stereo data with crop size 480×640. Our training pool consists of seven datasets: Scene Flow Mayer et al. (2016), Spring Mehl et al. (2023), FallingThings Tremblay et al. (2018), UnrealStereo4K Tosi et al. (2021), CarlaHighRes Yang et al. (2019), Sintel Stereo Butler et al. (2012), and Virtual KITTI 2 Cabon et al. (2020), together with the indoor subset of WMGStereo Yan et al. (2026). UniDisp is trained for 400K iterations on the seven-dataset mixture with batch size 8, followed by 150K iterations on FlyingThings3D and WMGStereo-Indoor with batch size 2.

Evaluation protocol. We evaluate cross-domain generalization capabilities across five widely used real-world scenarios: KITTI 2012 Geiger et al. (2012), KITTI 2015 Menze & Geiger (2015), ETH3D Schöps et al. (2017), Middlebury V3 Scharstein et al. (2014), and Booster Zama Ramirez et al. (2022). These datasets span substantially different acquisition conditions, including outdoor driving scenes, high-resolution indoor environments, and scenes containing challenging surface properties. The same synthetically trained UniDisp checkpoint is evaluated on all five benchmarks, providing a consistent measure of cross-domain transfer without target-domain adaptation. Furthermore, we validated the in-domain robustness on the Scene Flow dataset and evaluated performance on the KITTI dataset-specific benchmark. We follow the standard evaluation practice used in stereo matching benchmarks Mayer et al. (2016); Yan et al. (2026). All metrics are computed on valid pixels. We report End-Point Error (EPE), which measures the average disparity error in pixels, and Bad-Pixel Rate (BP-X), which counts the fraction of pixels with an error larger than X pixels. For KITTI 2015, we additionally report D1, the official outlier rate where a pixel is considered erroneous if its disparity error is both larger than 3 pixels and greater than 5% of the ground-truth disparity.

4.1 COMPARISON WITH STATE OF THE ART

We evaluate UniDisp against both mainstream stereo methods and recent methods that leverage monocular depth priors. UniDisp achieves state-of-the-art accuracy under the classical stereo setting and remains highly competitive with depth-prior-based methods in both accuracy and cross-domain robustness.

Cross-domain generalization evaluation. We compare UniDisp-L with representative stereo matching methods, including conventional models such as RAFT-Stereo Lipson et al. (2021), IGEV Xu et al. (2023), IGEV++ Xu et al. (2025), and NMRF Guan et al. (2024), as well as recent approaches that leverage pretrained visual foundation models (VFM), including DEFOM-Stereo Jiang et al. (2025), MonSter Cheng et al. (2025), BridgeDepth Guan et al. (2025), SMoE-DAV2 Wang et al. (2025), StereoFlow Wang et al. (2026), and DispViT+ Guan et al. (2026). All methods are re-evaluated under a unified protocol, with unreleased pre-trained weights results marked as "–". As shown in Table 1, UniDisp demonstrates strong and consistent cross-domain transfer across the five real-world datasets. Without VFM initialization or target-domain adaptation, it achieves the best results on KITTI 2012 (2.54 D1), ETH3D (0.97 BP-1), and Booster (9.63 BP-2). On KITTI 2015, UniDisp obtains a D1 of 3.24, nearly matching the best result of 3.23 achieved by MonSter. It also remains competitive on Middlebury-Q, achieving a BP-2 of 2.59 compared with 2.40 for DispViT+.

Fighre 2. presents a qualitative comparison between UniDisp and representative conventional stereo matching methods. UniDisp shows a clear advantage in resolving global matching ambiguities and achieves lower bad-pixel rates in most of the presented examples. These results are particularly notable in comparison with recent VFM-assisted methods, which leverage large-scale pretrained

Table 1: **Cross-domain evaluation on real-world scenarios.** We follow standard benchmark metrics. Results are reported over all valid pixels on KITTI-12 and KITTI-15, and over non-occluded pixels on Middlebury and ETH3D.

Methods	VFM	KITTI-12 D1	KITTI-15 D1	ETH3D BP-1	Middlebury_Q BP-2	Booster BP-2
RAFT-Stereo	No	4.06	5.43	2.29	7.43	17.75
IGEV	No	4.73	5.79	3.64	6.21	17.62
IGEV++	No	5.75	6.14	4.14	6.27	17.85
NMRF	No	3.92	5.29	3.48	5.06	26.14
DEFOM-Stereo	Yes	3.43	4.77	2.17	3.51	11.34
MonSter	Yes	2.90	**3.23**	1.10	3.87	16.06
BridgeDepth	Yes	3.33	4.52	1.13	2.85	16.67
SMoE-DAv2	Yes	3.55	4.80	2.08	5.69	23.21
STEREOFLOW	Yes	3.50	3.90	1.70	4.10	-
DISPVIT+	Yes	3.20	3.50	2.30	**2.40**	-
UniDisp	No	**2.54**	3.24	**0.97**	2.59	**9.63**

visual or monocular geometric priors, whereas UniDisp learns transferable stereo representations solely from stereo supervision. Importantly, the same source-domain-trained model is evaluated across all target datasets without dataset-specific checkpoint or model selection. This consistent cross-domain performance suggests that our stereo backbone can capture broadly transferable stereo geometry, even without external foundation-model priors.

In-domain evaluation on Scene Flow. We further evaluate UniDisp on the FlyingThings3D test set under the in-domain setting, as shown in Table 2. UniDisp achieves an EPE of 0.59 pixels and a BP-1 of 3.44%, with the latter being the lowest among the compared methods. This indicates that most predictions remain within one-pixel error, while a small number of difficult regions still produce relatively large deviations and increase the mean EPE. Therefore, UniDisp does not uniformly outperform existing methods on Scene Flow, but exhibits a favorable low-error distribution with room to reduce large residual errors. More importantly, Scene Flow performance does not directly determine cross-domain generalization. Several methods with lower in-domain EPE degrade more noticeably on real-world datasets, whereas UniDisp maintains competitive performance across the five real-world scenarios in Table 1.

KITTI benchmark. We fine-tune the model for 150k iterations on a mixture of KITTI 2012 and KITTI 2015 with a batch size of 2. Table 3 compares UniDisp with representative stereo matching methods on the official KITTI 2012 and KITTI 2015 test sets. On KITTI 2012, UniDisp achieves 0.92/1.27% Out-3 on non-occluded/all pixels, respectively, outperforming a broad range of established matching-based methods. Although DispViT+ and WAFT-Stereo achieve lower errors, UniDisp remains highly competitive without an explicit cost volume or external visual foundation representations. On KITTI 2015, UniDisp also performs strongly on foreground regions, achieving 2.27/2.44% D1-fg, with the lowest non-occluded foreground error among the compared methods and an all-pixel result comparable to the best-performing approaches. Overall, these results demonstrate that unified hierarchical stereo representations with progressive refinement are particularly effective for challenging foreground structures and disparity discontinuities, while precise background matching remains an important direction for further improvement.

4.2 Ablation Study

We conduct comprehensive ablation studies to examine the architectural design of UniDisp, with all experiments performed on UniDisp-S under the same training configuration. Specifically, we investigate how residual convolutional designs and kernel sizes affect hierarchical stereo representation, whether advanced convolutional blocks further enhance representation capacity and cross-domain generalization, and whether global attention in deeper stages can strengthen long-range stereo modeling.

Residual convolutional designs. We first study the effect of residual formulation on hierarchical stereo representation learning by comparing three closely related designs in Table 4. BasicBlock

Table 2: **In-domain evaluation.** We report EPE and BP-3 over all valid pixels on SceneFlow. The maximum disparity is set to the commonly used value of 192.

Method	VFM	EPE	BP-3
RAFT-Stereo	No	0.72	7.48
IGEV	No	0.48	5.39
IGEV++	No	0.47	4.46
NMRF	No	0.45	4.50
DEFOM-Stereo	Yes	0.46	5.30
MonSter	Yes	0.38	4.32
BridgeDepth	Yes	0.37	3.67
SMoE-DAv2	Yes	0.73	7.54
STEREOFLOW	Yes	**0.32**	–
DISPVIT+	Yes	0.34	3.50
UniDisp (Ours)	No	0.59	**3.44**

Table 3: **KITTI benchmark comparison.** We report Out-3 on KITTI 2012 and foreground (FG) error on KITTI 2015.

Method	KITTI 2012		KITTI 2015	
	Noc	All	Noc	All
IGEV-Stereo	1.12	1.44	2.62	2.67
Selective-IGEV	1.07	1.38	2.55	2.61
NMRF-Stereo	1.01	1.35	2.90	3.07
MoCha-Stereo	1.06	1.36	2.42	**2.43**
MC-Stereo	1.04	1.34	2.55	2.51
IGEV++	1.04	1.36	2.54	2.54
DS-Stereo	0.98	1.34	2.54	2.64
GREAT-IGEV (Solid)	0.95	1.32	2.53	2.56
DispViT+	<u>0.82</u>	<u>1.02</u>	3.10	3.26
WAFT-Stereo	**0.77**	**0.95**	<u>2.34</u>	2.47
UniDisp (Ours)	0.92	1.27	**2.27**	<u>2.44</u>

adopts the standard post-activation residual design, and PreActBasic He et al. (2016) moves normalization and activation ahead of convolution. Among them, BasicBlock-3 is the most stable across both source and target domains, achieving the lowest Things EPE/BP-1 (1.2/7.4) together with the strongest Middlebury and Booster results (6.9/20.3). Enlarging the kernel to 5×5 does not improve KITTI or ETH3D and slightly weakens the remaining benchmarks. More importantly, removing the residual path consistently degrades performance, while pre-activation only partially recovers the loss. This suggests that, within our hierarchical encoder, the residual organization is more consequential than simply enlarging the local receptive field.

Advanced convolutional blocks. We further investigate whether more advanced convolutional designs can enhance stereo representation learning by comparing BasicBlock with several representative modern convolutional units in Table 4: Bottleneck He et al. (2016)compresses and restores channels through 1×1 projections, ConvNeXt Liu et al. (2022)combines depthwise convolution with channel-wise MLPs, and RepVGG Ding et al. (2021) aggregates parallel convolutional and identity branches. RepVGG is the strongest of these alternatives, reaching the best KITTI-12 D1 (3.4) and ETH3D BP-1 (3.6) in the table. However, this gain does not transfer uniformly, as BasicBlock-3 remains clearly better on Things, Middlebury, and Booster. Bottleneck and ConvNeXt show an even larger drop on several datasets. The results therefore indicate that increasing block sophistication is not necessary for stronger stereo representations; in our setting, the simpler BasicBlock provides the more reliable cross-domain behavior.

Representation operators across hierarchical stages. The decoupled stage design of the HSR Encoder allows us to study different representation operators while keeping the hierarchical topology unchanged. We therefore compare convolutional and hybrid convolution–attention configurations under the same stage resolutions, channel dimensions, depths, and transition operators. Specifically, *C* and *A* denote convolutional and attention-based blocks, respectively, and the four characters indicate the operators used from the first to the fourth stage.

As shown in Table 5, the fully convolutional configuration consistently performs best across both synthetic and real-world benchmarks. Replacing the last stage with attention increases the Things EPE from 1.18 to 1.61 and substantially degrades Middlebury BP-2 from 6.90 to 11.38. Extending attention to the last two stages further increases the error to 1.94 on Things and 13.91 on Middlebury. A similar, although smaller, degradation is observed on KITTI-15.

These results indicate that, within our hierarchical stereo representation framework, introducing attention into the deeper stages does not provide an accuracy advantage. Instead, retaining convolutional operators throughout the hierarchy yields more effective representations and noticeably stronger cross-domain performance. We attribute this behavior to the suitability of convolutional local inductive bias for progressively encoding stereo displacement structures, while the hierarchical reduction in spatial resolution naturally enlarges the effective receptive field at deeper stages. Based on these observations, we adopt the fully convolutional CCCC configuration for the HSR Encoder.

Table 4: **Ablation of convolutional representation operators.** We compare residual formulations, kernel sizes, and representative modern convolutional blocks in UniDisp-S. Results are rounded to one decimal place for readability. ETH3D and Middlebury are evaluated over non-occluded pixels.

Block	Kernel	In-Domain		Cross-Domain				
		Things EPE ↓	Things BP-1 ↓	KITTI-12 D1 ↓	KITTI-15 D1 ↓	ETH3D BP-1 ↓	Middlebury BP-2 ↓	Booster BP-2 ↓
BasicBlock	3×3	1.2	7.4	3.5	3.9	3.8	6.9	20.3
	5×5	1.4	8.6	3.5	3.9	3.8	8.3	21.3
PlainBasic	3×3	1.6	10.1	4.1	4.4	4.7	8.1	28.8
	5×5	1.8	11.5	4.4	4.9	6.6	9.2	35.2
PreActBasic	3×3	1.4	8.7	3.7	4.1	5.0	8.1	21.7
Modern Convolutional Network Blocks								
Bottleneck	3×3	2.0	11.7	3.5	4.2	5.0	9.8	28.6
ConvNeXt	3×3	1.7	9.3	4.0	4.2	14.8	8.3	22.6
RepVGG	3×3	1.4	9.1	3.4	4.0	3.6	9.0	23.7

These ablation studies reveal key architectural principles underlying effective unified stereo representation learning. The results consistently suggest that carefully structured convolutional inductive biases and hierarchical feature propagation are more important than increased receptive fields or architectural complexity. Together, these lightweight designs enable UniDisp to encode stereo geometry directly within the representation hierarchy, further supporting unified representation-based disparity inference as a simple and effective alternative to increasingly specialized matching pipelines.

Table 5: **Ablation of hierarchical representation operators.** *C* and *A* denote convolutional and attention-based stages,respectively.

Pattern	Things EPE	Things BP-1	KITTI-15 D1	Middlebury BP-2
CCCC	**1.19**	**7.38**	**3.89**	**6.90**
CCCA	1.61	9.53	4.25	11.38
CCAA	1.94	10.72	4.29	13.91

5 Conclusion and Future Work

In this work, we presented UniDisp, a unified stereo backbone that treats disparity as an intrinsic geometric attribute of hierarchical stereo representations and directly infers it from the learned representation hierarchy. By combining the HSR Encoder, HDP Head, and structurally unified initialization and refinement stages, UniDisp simplifies the conventional stereo pipeline while achieving strong in-domain accuracy and consistent cross-domain generalization without relying on visual foundation models or target-domain adaptation. Despite these results, long-tail errors remain in challenging regions involving weak textures, complex boundaries, severe ambiguities, and large disparities, where broader contextual reasoning is often required. Future work will therefore investigate more effective global stereo modeling, particularly attention-based mechanisms for capturing long-range cross-view dependencies, together with stronger progressive refinement strategies to further improve robustness in complex real-world scenarios.

AI use statement

In this work, we used generative AI tools to assist with language polishing in the methodology section, as well as debugging code and LaTeX formatting. Generative AI was not used to generate experimental results, modify benchmark data, or make research conclusions on behalf of the authors. All AI-assisted content and code were reviewed and verified by the authors. We take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (ed.), *European Conference on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012. doi: 10.1007/978-3-642-33783-3_44.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. URL https://arxiv.org/abs/2001.10773.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yong Qin, and Jia Wu. MoCha-Stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27768–27777, 2024.

Junda Cheng, Li Liu, Gangwei Xu, Xianqi Wang, Ziyang Zhang, Yiming Deng, Jin Zang, Yufeng Chen, Zhaopeng Cai, and Xin Yang. MonSter: Marry monodepth to stereo unleashes power. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6273–6282, 2025.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13733–13742, 2021.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Tongfan Guan, Chen Wang, and Yung-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5459–5469, 2024.

Tongfan Guan, Jiaming Guo, Chen Wang, and Yung-Hui Liu. BridgeDepth: Bridging monocular and stereo reasoning with latent alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27681–27691, 2025.

Tongfan Guan, Jiaxin Guo, Tianyu Huang, Jinhu Dong, Chen Wang, and Yun-Hui Liu. DispViT: Direct stereo disparity regression with a single-stream vision transformer. In *International Conference on Learning Representations*, 2026.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3273–3282, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, volume 9908 of *Lecture Notes in Computer Science*, pp. 630–645. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46493-0_38.

Hongbo Jiang, Zhaoyang Lou, Lihe Ding, Rui Xu, Mingkui Tan, Wei Jiang, and Rui Huang. DEFOM-Stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21857–21867, 2025.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75, 2017.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16242–16251. IEEE, 2022.

Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7644–7652, 2019.

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision*, pp. 218–227, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Nikolaus Mayer, Eddy Ilg, Philip Häusser, 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*, pp. 4040–4048, 2016.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991, June 2023.

Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061–3070, 2015. doi: 10.1109/CVPR.2015.7298925.

Don Murray and James J. Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8(2):161–171, 2000. doi: 10.1023/A:1008987612352.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch (eds.), *Pattern Recognition*, volume 8753 of *Lecture Notes in Computer Science*, pp. 31–42. Springer International Publishing, 2014. doi: 10.1007/978-3-319-11752-2_3.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, 2017. doi: 10.1109/CVPR.2017.272.

Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 519–528, 2006. doi: 10.1109/CVPR.2006.19.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13906–13915, 2021.

Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14362–14372, 2021.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8942–8952, June 2021.

Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2038–2041, June 2018. doi: 10.1109/CVPRW.2018.00275.

Hao Wang, Haoran Geng, Xiaotong Yang, Jing Tang, Songlin Wei, Linlong Lang, Yeying Jin, Zheng Zhu, Zhaoxin Fan, and Biao Leng. STEREOFLOW: Progressive stereo matching with stereodit and transition flow matching. *arXiv preprint arXiv:2607.19986*, 2026.

Xianqi Wang, Gangwei Xu, Haokun Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19701–19710, 2024.

Yifan Wang and Jia Deng. WAFT-Stereo: Warping-alone field transforms for stereo matching. *arXiv preprint arXiv:2603.24836*, 2026.

Yun Wang, Longguang Wang, Chenghao Zhang, Yongjian Zhang, Zhanjie Zhang, Ao Ma, Chenyou Fan, Tin Lun Lam, and Junjie Hu. Learning robust stereo matching in the wild with selective mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21276–21287, 2025. doi: 10.1109/ICCV51701.2025.01976.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12981–12990, 2022.

Gangwei Xu, Xianqi Wang, Xuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21919–21928, 2023.

Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):7108–7122, 2025. doi: 10.1109/TPAMI.2025.3569218.

Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1959–1968, 2020.

David Yan, Alexander Raistrick, and Jia Deng. What makes good synthetic training data for zero-shot stereo matching? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34171–34180, 2026.

Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5515–5524, June 2019. doi: 10.1109/CVPR.2019.00566.

Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21168–21178, 2022.

Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65):1–32, 2016.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H. S. Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 185–194, 2019.

Haoliang Zhao, Huizhou Zhou, Yao Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1327–1336, 2023.