Stereo Depth Estimation with Temporal Consistency and Foundation Model Priors

Yuanhan Chen
University of Toronto
yuanhan.chen@mail.utoronto.ca

Abstract

Stereo depth estimation from video is enhanced by combining temporal consistency cues with learned monocular depth priors. We propose a novel stereo matching framework that integrates Temporally Consistent Stereo (TC-Stereo) [16] with a Depth Foundation Model (DEFOM) [3] based prior. Our model propagates disparity forward through time and leverages a pre-trained monocular depth network to initialize and regularize disparity in ill-posed regions. A hybrid CNN+ViT encoder fuses stereo and monocular features, yielding a robust cost volume for matching. We obtain an initial disparity for each frame by temporal disparity propagation from the previous frame and monocular depth initialization from the foundation model, which are merged via a completion network. Disparities are then refined iteratively with a dual-space ConvGRU – alternating between disparity and disparity-gradient domains – and a scale adjustment module that corrects the depth scale drift from the monocular prior. Experiments on challenging stereo video sequences (e.g. TartanAir) demonstrate that our integrated approach slightly outperforms state-of-the-art stereo models in both accuracy and temporal stability. In particular, it achieves lower end-point error (EPE) and reduces 1-pixel and 3-pixel disparity error rates compared to TC-Stereo alone. The proposed method produces sharper, more consistent depth maps across time, highlighting the benefits of fusing temporal information and foundation model priors in stereo matching. The code is available at https://github.com/oscarchen178/StereoProject

1. Introduction

Recovering dense 3D structure from stereo image pairs is fundamental in computer vision, essential for applications like autonomous driving, robotics, and AR/VR. Many such applications operate on stereo video streams, where maintaining temporal consistency of disparity is as critical as per-frame accuracy [16] [8]. Temporal inconsistencies can degrade user experience in AR or impair robot navigation [8]. However, achieving both accuracy and temporal stability in stereo video is challenging due to photometric ambiguities and scene dynamics. Regions lacking texture or containing repetitive patterns lead to ambiguous disparities, causing jitter when processed independently per frame [4]. Additionally, dynamic scenes introduce occlusions and newly visible areas, which naive per-frame approaches handle poorly, resulting in abrupt disparity changes.

Recent works address these issues by explicitly enforcing temporal consistency. Methods like Temporally Consistent Stereo Matching [16] propagate disparities from previous frames to guide current-frame estimation, significantly reducing flicker. Similarly, CODD integrates learned per-pixel motion transformations to warp previous depth maps, yielding smoother depth transitions [8]. Transformer-based approaches like DynamicStereo leverage multi-frame attention to stabilize disparities across sequences [4]. Despite these advancements, purely temporal methods struggle in persistently ambiguous regions.

Alternatively, monocular depth networks provide strong priors, particularly effective in areas where stereo geometry alone fails. Methods like MonSter integrate monocular depth maps with stereo pairs by adjusting monocular predictions to match stereo-derived scales [2]. DEFOM-Stereo further leverages powerful pre-trained vision transformer (ViT)-based monocular models to enhance stereo matching, achieving superior performance in challenging conditions [3].

Inspired by these complementary strengths, we propose a model TC-DEFOM with unified approach integrating temporal consistency and monocular depth priors. Building upon TC-Stereo [16], our model augments temporal disparity propagation with monocular guidance from a robust pre-trained depth model. The propagated disparity from the previous frame provides a sparse, confident prior, while the monocular network offers a dense depth estimation for the current image. A completion module merges these cues, filling new or ambiguous regions and retain-

ing reliable propagated disparities. This initial map undergoes iterative refinement via a dual-phase approach: global scale adjustment to reconcile monocular-stereo metric differences and dual-space refinement enforcing local smoothness. Our approach significantly reduces temporal inconsistency and improves accuracy in challenging regions, as demonstrated on datasets like TartanAir [14].

Our contributions include: (1) We present the first integration of temporal stereo matching with a depth foundation model prior, demonstrating a unified architecture that leverages both past frames and learned monocular cues. (2) We propose a scale-aware disparity initialization strategy that combines propagated disparity from the last frame with a monocular depth estimate for the current frame, merged via a novel completion network to provide a scale-calibrated, temporally coherent starting disparity. (3) We incorporate a dual-phase iterative refinement: a Scale Update module to adjust global disparity scale, followed by a Dual-Space refinement module that alternates updates in disparity and gradient space to enforce smoothness and temporal stability in ill-posed regions . (5) Our integrated approach is comparable to state-of-the-art models on stereo video benchmarks.

2. Related works

Deep learning has significantly advanced stereo matching, introducing diverse architectures over the past decade. Early approaches such as GC-Net [5] and PSMNet [1] implemented the traditional stereo pipeline end-to-end, featuring dense 4D cost volumes processed by 3D CNNs. GC-Net pioneered learnable cost volumes, while PSMNet enhanced multi-scale context via pyramid pooling. Despite their effectiveness, these methods were memory-intensive and limited in featureless or large-disparity scenarios.

Recent models, like RAFT-Stereo [9], shifted to iterative refinement strategies, avoiding costly 3D convolutions. RAFT-Stereo uses correlation volumes and recurrent ConvGRUs for iterative disparity refinement, significantly boosting efficiency and accuracy. This paradigm inspired further improvements seen in models such as IGEV-Stereo [15] and CREStereo [6].

Transformers have also emerged, offering enhanced long-range dependency modeling. STTR [7] employs Transformers for direct feature matching, addressing repetitive pattern issues. Hybrid models like ViTAStereo [10] integrate pre-trained ViTs with traditional CNN cost volumes, achieving state-of-the-art results on KITTI benchmarks [12]. Overall, contemporary stereo networks increasingly leverage hybrid architectures, combining robust convolutional cost computation, iterative refinement, and Transformer-based attention to handle challenging textureless and occluded regions.

Temporal Stereo: While most stereo research focuses on single pairs, there is growing interest in methods that exploit temporal continuity in stereo videos. Simply running a stereo network on each frame independently often produces temporarily inconsistent results – depth flicker or jumps – even if each frame is accurate, as highlighted by Li et al. [8]. Initial attempts to address this applied post-filtering on disparity sequences (e.g. temporal smoothing filters or Kalman filtering), but these can lag behind fast motion or fail when new occlusions appear. Recent methods incorporate temporal reasoning within the network to achieve true temporally consistent stereo.

TC-Stereo [16] is a prime example. It augments a RAFT-Stereo framework with modules for temporal propagation and dual-space refinement. TC-Stereo warps the previous frame's disparity into the current frame using the estimated camera motion, obtaining a sparse prior disparity map. A temporal disparity completion network then fills in missing regions to produce a full initial disparity for the current frame. After that, a dual-space iterative refinement is performed: a ConvGRU module refines the disparity in both the original disparity space and the disparity gradient space. The insight is that disparity gradients are more stable to estimate in ambiguous regions ; by enforcing consistency in the gradient domain, the model encourages the disparity to vary smoothly over time on each surface. TC-Stereo was shown to markedly reduce flicker and errors in challenging video scenarios, since it carries over reliable depth estimates and only needs to update local changes.

Another approach, DynamicStereo [4], uses spatio-temporal Transformers to process a window of video frames jointly. By attending across frames, it reinforces consistent disparities and was demonstrated on dynamic scenes. DynamicStereo achieves efficiency via a divided attention scheme and produces smooth depth even for non-rigid motion. Beyond propagation and attention, integrating motion estimation directly is another line of work. In CODD [8], the system includes a separate motion network to predict a 3D rigid transformation for each pixel's depth. This warps the previous depth map more accurately into the current view before fusing it with the stereo network's prediction. By explicitly handling object motion and not just camera motion, CODD can maintain temporal consistency even when independent objects move. Across these methods, the consensus is that temporal context can greatly improve stereo video results. The main challenges are avoiding drift or error accumulation over time and handling dis-occlusions. Most approaches therefore still rely on a strong per-frame stereo backbone and treat temporal modules as refinements on top.

Monocular-Guided Stereo and Depth Priors: A complementary research thread combines monocular depth estimation with stereo to exploit their respective strengths. Monocular depth networks (e.g. MiDaS , DPT [13]) are

trained on large image datasets and can often estimate relative depth from a single image surprisingly well, even for scenes very different from the training set. However, monocular predictions lack a fixed scale: they produce depth maps that are relative (up to an unknown scale and shift). Stereo, on the other hand, yields metric depth via triangulation but fails when the images lack texture or overlap (e.g. one camera sees an object that the other does not). Merging these two modalities can yield the best of both: monocular inference fills in ambiguous regions, and stereo ensures geometric consistency and correct scaling. MonSter [2] is a notable effort along these lines. It runs a monocular depth network on one image to get an initial depth, then feeds that into a stereo matching module that learns to predict per-pixel scale and shift corrections. Essentially, MonSter trusts the monocular network for the overall shape of the depth map, but uses the stereo pair to fine-tune each pixel's value so that it aligns with the actual disparity evidence. This addresses the scale ambiguity by allowing a spatially varying rescaling of the monocular depth.

More recently, DEFOM-Stereo [3] takes a different approach: rather than post-adjusting monocular output, it injects a pre-trained monocular depth model into the stereo network architecture itself. DEFOM-Stereo uses a large Vision Transformer (ViT) based depth model (termed a Depth Foundation Model) to extract features from the input image and also predict an initial relative depth. These monocular features are fused with traditional stereo CNN features in a unified encoder, enriching the stereo cost volume with high-level context. The monocular depth prediction is converted to an initial disparity and used to seed the iterative stereo refinement process. By integrating the monocular prior, the stereo network benefits from the prior's broad scene understanding while still ensuring the final output obeys stereo geometry. DEFOM-Stereo achieved leading accuracy on several benchmarks, even ranking 1st on KITTI 2015 [12] among published methods at the time. This underscores how powerful foundation model priors can be for stereo. But the network must learn to balance the monocular prior against stereo cues. If it relies too much on the prior, it might ignore stereo evidence. Proper training (and sometimes freezing the monocular model) is required to get the best results.

Our work sits at the intersection of these areas. We extend the concept of monocular-stereo fusion into the temporal domain: not only do we use a monocular prior to aid stereo, but we also propagate information through time for consistency. By unifying TC-Stereo with DEFOM-Stereo, we aim to produce a stereo video depth estimator that is robust to ambiguity (thanks to the learned prior) and stable over time (thanks to temporal propagation). To our knowledge, this is the first stereo model to integrate a foundation depth model and temporal cues simultaneously. Next, we detail the proposed model's architecture and components.

3. Technical section

The proposed model for this project is developed based on TCStereo and integrated DEFOMStereo. Let's first discuss technical details of these two model.

3.1. Comparing technical details of TCStereo and DEFOMStereo

TCStereo (Temporally Consistent Stereo) and DEFOM-Stereo (Depth Foundation Model Stereo) represent advanced stereo matching frameworks derived from the RAFT-Stereo paradigm. Despite sharing core iterative refinement principles, their design goals diverge significantly: TCStereo prioritizes temporal consistency for stereo video sequences, whereas DEFOMStereo integrates monocular depth priors for robust single-pair stereo estimation.

Feature Extraction: TCStereo employs conventional CNN-based encoders, producing matching and context features at 1/4 scale, analogous to RAFT-Stereo. It enhances temporal coherence by caching and shifting features across consecutive frames, thereby propagating useful context forward without redundant computations.

In contrast, DEFOMStereo adopts a hybrid architecture combining CNN encoders with a ViT-based backbone derived from Depth Anything V2, a pre-trained monocular depth model. This integration provides semantically richer feature representations. The fusion of CNN and ViT features occurs through simple channel alignment and summation, yielding combined matching features and multi-scale context features at 1/4, 1/8, and 1/16 resolutions.

Cost Volume Construction: Both models rely on cost volumes built from feature correlations. TCStereo constructs a dense 3D cost volume at 1/4 resolution using cosine similarity. It employs a winner-take-all (WTA) strategy to generate a semi-dense initial disparity map, relying on a confidence threshold to identify reliable matches. The cost volume is precomputed and queried during iterative refinement, promoting efficiency.

DEFOMStereo similarly utilizes an all-pairs correlation approach, structured into a multi-scale correlation pyramid facilitating both coarse and fine matching. It introduces specialized lookups: a Scale Lookup (SL) for global disparity scaling adjustments, and a Pyramid Lookup (PL) for local refinement around current disparity estimates.

Initial Disparity Estimation: The methods diverge notably in disparity initialization strategies. TCStereo initializes disparity using temporal information. For the first frame, it depends on a semi-dense WTA disparity from the cost volume. Subsequent frames use temporal propagation, projecting disparities from previous frames into the current viewpoint via camera motion estimation, creating sparse

disparity maps completed by a dedicated Temporal Disparity Completion (TDC) module.

DEFOMStereo, however, leverages monocular depth priors directly from Depth Anything V2. It converts monocular depth estimates into disparity maps through normalization and scaling, ensuring dense initial guesses that address ambiguities present in textureless or occluded areas. Unlike TCStereo, it requires no dedicated completion module, directly entering the refinement stage with a dense initial disparity map.

Iterative Refinement Modules: Refinement processes in both models follow RAFT-style iterative updates but differ substantially in their implementation:

TCStereo's Dual-Space Refinement: This method uniquely updates disparities in both disparity and gradient spaces within each iteration. The disparity-space refinement performs conventional disparity adjustments, whereas the gradient-space refinement manages disparities indirectly, ensuring smoothness and consistency, especially in reflective and occluded regions. TCStereo integrates a ConvGRU-based hidden state fused with temporal information from prior frames, maintaining continuity and stability across iterations.

DEFOMStereo's Scale-and-Delta Refinement: DEFOMStereo introduces a sequential refinement scheme. Initially, a Scale Update (SU) module iteratively corrects the global scale misalignment of monocular priors through Scale Lookup operations. Following scale correction, a Delta Update (DU) module conducts fine-grained disparity refinement utilizing local Pyramid Lookups around current disparity estimates. This two-step approach effectively separates global disparity scaling from local matching adjustments.

Implementation Structure: Both models follow similar high-level structures but introduce specific functional blocks aligned with their innovations:

TCStereo integrates specialized modules like Temporal Disparity Completion (TDC) and Temporal State Fusion (TSF), enabling coherent disparity propagation across frames. Its refinement leverages dual-space updates within iterative GRU loops.

DEFOMStereo incorporates ViT and CNN feature fusion at an early stage, alongside distinct modules for Scale Lookup and Pyramid Lookup operations. Its refinement structure explicitly divides global scale correction (SU) from detailed local refinement (DU).

Output Generation: The final output for both approaches is a refined disparity map at full resolution, upsampled from 1/4 resolution outputs of the iterative refinement process. TCStereo uniquely ensures temporal coherence by projecting the final disparities and hidden states forward to subsequent frames, whereas DEFOMStereo independently processes stereo pairs without maintaining temporal state.

Unique Innovations Summary:

- TCStereo stands out for its temporal disparity completion strategy, dual-space iterative refinement, and explicit state fusion across frames, effectively reducing temporal jitter and ensuring stable sequential disparities.
- DEFOMStereo excels by leveraging a pre-trained monocular depth foundation model, introducing innovative scale adjustment modules, and enhancing generalization and robustness across diverse domains.

In conclusion, TCStereo's temporal continuity mechanisms and DEFOMStereo's integration of monocular priors represent complementary advancements in stereo vision. These distinct technical strengths inform the subsequent design of our integrated stereo depth estimation model, which aims to leverage the temporal consistency of TCStereo alongside the robust, monocular-driven generalization of DEFOMStereo.

3.2. TC-DEFOM: Model Implementation Detailsn

TC-DEFOM Stereo Architecture: We introduce a new stereo model, termed TC-DEFOM (Temporal-Consistent Depth Foundation Model Stereo), which merges the temporal consistency mechanisms of TC-Stereo with the robustness of DEFOM-Stereo's monocular depth priors. The overall design follows the RAFT-Stereo style recurrent framework inherited from TC-Stereo – including disparity completion, temporal state fusion, and dual-space refinement – and augments it with a pre-trained depth foundation model and a scale-alignment module as proposed in DEFOM-Stereo . Below we detail each component, highlighting which ideas stem from TC-Stereo and which from DEFOM-Stereo. Feature Extraction

We reuse TCStereo's CNN encoders and incorporate DEFOM features. A MultiBasicEncoder (from TCStereo) processes the left image to produce multi-scale context features for hidden-state and refinement networks. A BasicEncoder (TCS) processes the right image (or, if a shared backbone is used, a small 2-layer conv net upsamples left features to serve as right features). In parallel, a frozen DefomEncoder (DEFOMStereo) – a ViT/DPT-based network – processes both images and outputs deep features plus a monocular inverse-depth map. The Defom encoder's feature maps are aligned to the CNN feature dimensions via a 1×1 convolution and added to the corresponding left/right CNN features. This fused feature pair is used for subsequent matching.

- Context encoder: MultiBasicEncoder (TCS) extracts hierarchical context features from the left image.
- Feature encoder: BasicEncoder (TCS) extracts features from the right image (or a 2-layer conv substitutes if sharing backbone).
- Monocular depth encoder: DefomEncoder (DEFOM) produces high-level features for both images and a coarse

inverse-depth map.
- Feature fusion: Defom features are aligned (1×1 conv) and summed with the corresponding CNN features.

Cost Volume: We build cost volumes by correlating the fused left/right features. Two forms of cost are used:
- Static cost volume (TCStereo): We compute full all-pairs correlation via the original CorrBlock1D, producing a 4D cost of shape (B, W_2, H, W_1). We then zero out any costs where the right-image coordinate exceeds the left (to enforce non-negative disparity). This cost is used for supervised loss computation (endpoint/contrastive loss).
- Iterative cost features (DEFOMStereo): For refinement, we use NewCorrBlock1D, which builds a pyramid of correlations over multiple downsampled levels and supports scale search. At each GRU iteration, given the current disparity estimate, NewCorrBlock1D samples correlation values around the predicted position using multiple scale hypotheses (from a predefined scale list). The resulting concatenated cost features (over all levels and scales) are fed to the update GRUs. This provides robust matching cues across scales.

Initial Disparity: Disparity initialization depends on whether we have a previous frame:
- Monocular init (DEFOMStereo): For the first frame (no history), we convert the frozen-encoder inverse-depth map $\hat{z}$ to disparity. We normalize $\hat{z}$ by its maximum and multiply by a factor η and the image width: $d_{\text{init}} = (\hat{z}/\max \hat{z}), \eta W$, with $\eta = 0.5$ as in DEFOMStereo. This yields a coarse initial disparity map (all pixels valid) and a dummy zero cost.
- Temporal init (TCStereo): For subsequent frames (video mode), we warp the previous disparity into the current frame using the known camera motion (K, baseline, relative pose). This produces a sparse disparity estimate and validity mask from the last frame's disparity (unaltered core logic from TCStereo).

These steps yield the starting disparity field d and mask for iterative refinement.

Iterative Refinement: We refine disparity using a cascade of GRU update blocks, following the RAFT paradigm (default 5 iterations). The first K iterations (where K=scale_iters) use a scale-aware GRU (ScaleBasicMultiUpdateBlock, from DEFOMStereo). This block outputs a disparity update and a learned scale factor s. We update the normalized pixel coordinates via a contraction:

$$\text{if scale-iter:} \quad \text{coords} \leftarrow \text{coords}_0 - s \cdot (\text{coords}_0 - \text{coords})$$

where coords_0 is the reference coordinate grid. For remaining iterations, we use the standard BasicMultiUpdateBlock (TCStereo), which outputs a displacement Δ and updates $\text{coords} \leftarrow \text{coords} + \Delta$. In either case, the current disparity is $\text{d} = \text{coords}_0 - \text{coords}$.

After each update, we apply a small CNN to the disparity gradient (via DispGradPredictor) and then a final upsampling block (DispRefine) to produce a refined disparity increment. The residual $(\text{refined} - \text{current})$ is fused into the GRU hidden state using a gated HiddenstateUpdater (TCS). At the end of all iterations, we upsample the disparity to full resolution.

Training Setup: The model was trained end-to-end with supervised disparity losses. The loss combines multi-scale L1 disparity loss (upsampled to full resolution) and a disparity-gradient loss at each output, mirroring TCStereo's regimen. We use the AdamW optimizer (learning rate 2×10^{-4}, weight decay 1×10^{-5}) for $\sim$100k iterations on a mix of synthetic and real stereo datasets (e.g. TartanAir) with batch size 4 and random crops (e.g. 480x640). The network uses 5 update iterations per forward pass (with both scale-aware and standard updates as described). When training in temporal mode, we enable warping for initialization; otherwise only the monocular DEFOM init is used.

4. Experimental results

We evaluated the proposed TC-DEFOM model on a very limited stereo dataset derived from the TartanAir benchmark, using only the *Abandoned Factory* scene. This small subset (a single environment) provides a constrained testbed to validate the model but offers minimal diversity in imagery. We trained the model with the same hyperparameter settings as the baseline TCStereo network (e.g., learning rate, optimizer, loss weights), except we drastically limited the training duration. In particular, we conducted experiments at only 10, 100, and 300 training steps (in contrast to the thousands of iterations typical in full training) to observe the model's learning progress under extremely curtailed training. All models were trained from scratch on the *Abandoned Factory* subset with no additional pre-training, and evaluated on held-out frames from the same scene.

Model_epoch	EPE	D1	D3
TC-DEFOM_10	16.803540	83.463899	80.718340
TCStereo_10	21.854693	99.122010	97.318627
TC-DEFOM_100	7.377263	78.479445	61.564768
TCStereo_100	7.334006	76.911067	58.205395
TC-DEFOM_300	7.232800	80.543986	63.236010
TCStereo_300	7.218970	78.115603	58.799200

Table 1. Comparison of performance metrics for different models and steps.

We report performance using standard stereo depth metrics: the end-point error (EPE) in disparity (average pixel disparity error), and the D1 and D3 error rates (percentage of pixels with disparity error greater than 1 and 3 pixels, re-

spectively). Table X summarizes the quantitative results for the TC-DEFOM model compared to the baseline TCStereo at the three training step milestones. At the very early stage of training (only 10 steps), the integrated TC-DEFOM model already achieves a noticeably lower EPE (approximately 16.8 pixels) than the baseline TCStereo (about 21.9 pixels). Similarly, the outlier rates are substantially better with TC-DEFOM at 10 steps (D3 $\approx$ 80.7% vs. 97.3% for TCStereo), indicating that even with minimal training the combined model produces more accurate disparity estimates on many pixels. This suggests that the DEFOM component may be providing useful monocular cues or regularization that "jump-start" learning. After 100 training steps, both models improve significantly – EPE drops to $\sim$7.3 for both, and D3 outlier rates fall to around 58–62%. At this point, the baseline slightly edges out the TC-DEFOM model in terms of D1/D3 (for example, D3 is 58.2% for TCStereo vs. 61.6% for TC-DEFOM at 100 steps), while their EPE values are nearly identical. Extending training to 300 steps yields little further improvement in EPE (both models plateau around 7.2), and in fact the outlier rates slightly worsen for the TC-DEFOM model (D3 rises to 63.2%) whereas the baseline remains roughly steady (D3 58.8%). This plateau and minor regression in error rates beyond 100 steps suggest that with such a small training set, the models quickly reach their capacity and may begin to overfit the scene specifics. Overall, on this limited dataset the TC-DEFOM architecture did not yet demonstrate a clear performance gain over the original TCStereo baseline – apart from an initial faster convergence at 10 steps, its final accuracy after 300 steps was essentially on par with, or slightly behind, the baseline.

5. Discussion

These experimental results highlight several challenges and limitations of our current approach. First, the dataset itself is a significant limiting factor. Using only a single TartanAir scene (*Abandoned Factory*) means the training data lack diversity. The model is learning from a very narrow distribution of imagery, which constrains its ability to generalize and accelerates overfitting. Indeed, the slight increase in outlier rate from 100 to 300 steps for TC-DEFOM (and the plateau in error reduction) can be seen as a symptom of overfitting to the small training set. Also any benefit from the integrated monocular module might require varied scenes and longer training to manifest. In summary, the experimental scope was too narrow, and a larger, more diverse dataset would be needed to fairly evaluate the new model's capabilities without encountering an early performance ceiling.

Second, we did not perform any hyperparameter tuning or optimization specific to the new TC-DEFOM architecture. We carried over the training parameters from the TCStereo baseline (learning rate schedule, regularization strength, etc.) without adjustment. The lack of tuning means the TC-DEFOM model might have been trained suboptimally. For example, a more complex model might benefit from a smaller learning rate or different weight initialization. In essence, the experiment treated the new model as a drop-in replacement with identical training strategy, which may not be ideal; a tailored training regimen could yield better results and was not attempted in this initial study.

Finally, and most critically, the way in which we integrated the Temporal Consistency (TC) stereo module with the DEFOM module was fairly naive, leading to potential redundancy and conflicts within the model. Both the original TCStereo and the DEFOM approach employ recurrent neural units (GRUs) but for different purposes. The update blocks serves different purpos. The TCStereo baseline's GRU is designed to propagate and refine disparity estimates over time, while the DEFOM model's GRU was developed in the context of monocular depth prediction. In the current TC-DEFOM implementation, we essentially combined the two architectures outright. Such architectures can make the training dynamics unnecessarily complex. The result can be inefficient learning and diminished returns from what should have been a more powerful combined model.

Moreover, the DEFOM component's approach to scale handling, which was a strength for monocular depth, may not translate well to a stereo-based system. The scale shift for mono depth and past frame depth is different. Therefore, the scale-adjustment could be detrimental if left unchecked: it might, for example, introduce drift or inconsistency where there should be none, or bias the network to doubt the reliable stereo geometry. In our TC-DEFOM integration, we did not modify this aspect, meaning the model might be trying to "solve" a non-existent scale ambiguity problem. This misalignment between DEFOM's original design assumptions and the stereo application could partially explain why TC-DEFOM did not outperform the baseline. In summary, simply fusing the two models without redesigning their interaction led to an ineffective use of the DEFOM features – the experiment revealed that a more clever integration is required so that the monocular-derived cues constructively augment the stereo temporal pipeline, rather than interfere with it.

6. Concluding remarks

This initial test of the TC-DEFOM model—which fuses a temporal stereo pipeline with a monocular depth module—showed only a slight early learning boost and no clear advantage over the strong TCStereo baseline. This outcome suggests that naïvely combining two systems isn't enough, but it does highlight precisely where integration and training need refinement.

To strengthen TC-DEFOM, we will first vastly expand

our datasets. Instead of one small scene, we'll train on large, varied sources—synthetic suites like SceneFlow [11] plus real-world benchmarks such as the full KITTI stereo set [12] and diverse TartanAir scenes [14]. This broader data exposure should reduce overfitting and reveal the true value of combining stereo and monocular cues.

Next, we'll overhaul how the stereo and monocular streams interact. We'll explore a unified recurrent unit that ingests monocular predictions alongside stereo features each time step, or we'll merge feature maps early so the network learns a single coherent depth output. This principled fusion aims to eliminate conflicting processes and let the modules genuinely support one another.

We'll also prune unnecessary complexity. By designing a more suitable temporal updater, removing features that don't boost accuracy, and performing targeted hyperparameter and curriculum training (for instance, pretraining the monocular branch), we expect a leaner, more stable model that's easier to optimize.

In summary, our modest proof-of-concept underscores the need for richer data and smarter architecture when uniting stereo and monocular methods. By scaling up training, refining module fusion, and streamlining the design, we're confident the next TC-DEFOM iteration will deliver significantly more accurate and robust depth estimates in stereo video.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2

[2] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. *arXiv preprint arXiv:2501.08643*, 2025. 1, 3

[3] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3

[4] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13229–13239, 2023. 1, 2

[5] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 2

[6] 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)*, pages 16242–16251, 2022. 2

[7] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6186, 2021. 2

[8] Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X. Creighton, Russell H. Taylor, Ganesh Venkatesh, and Mathias Unberath. Temporally consistent online depth estimation in dynamic scenes. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3017–3026, 2023. 1, 2

[9] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227, 2021. 2

[10] Chuang-Wei Liu, Qijun Chen, and Rui Fan. Playing to vision foundation model's strengths in stereo matching. *IEEE Transactions on Intelligent Vehicles*, pages 1–12, 2024. 2

[11] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 7

[12] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 3, 7

[13] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 2

[14] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020. 2, 7

[15] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21919–21928, 2023. 2

[16] Jiaxi Zeng, Chengtang Yao, Yuwei Wu, and Yunde Jia. Temporally consistent stereo matching. 2024. 1, 2