BaseBoostDepth: Exploiting Larger Baselines For Self-supervised Monocular Depth Estimation

Kieran Saunders[1]
190229315@aston.ac.uk
Luis J. Manso[1]
l.manso@aston.ac.uk
George Vogiatzis[2]
g.vogiatzis@lboro.ac.uk

[1] Aston University
Birmingham, UK

[2] Loughborough University
Leicestershire, UK

Abstract

In the domain of multi-baseline stereo, the conventional understanding is that, in general, increasing baseline separation substantially enhances the accuracy of depth estimation. However, prevailing self-supervised depth estimation architectures primarily use minimal frame separation and a constrained stereo baseline. Larger frame separations can be employed; however, we show this to result in diminished depth quality due to various factors, including significant changes in brightness, and increased areas of occlusion. In response to these challenges, our proposed method, **BaseBoostDepth**, incorporates a curriculum learning-inspired optimization strategy to effectively leverage larger frame separations. However, we show that our curriculum learning-inspired strategy alone does not suffice, as larger baselines still cause pose estimation drifts. Therefore, we introduce incremental pose estimation to enhance the accuracy of pose estimations, resulting in significant improvements across all depth metrics. Additionally, to improve the robustness of the model, we introduce error-induced reconstructions, which optimize reconstructions with added error to the pose estimations. Ultimately, our final depth network achieves state-of-the-art performance on KITTI and SYNS-patches datasets across image-based, edge-based, and point cloud-based metrics without increasing computational complexity at test time. The project website can be found at https://kieran514.github.io/BaseBoostDepth-Project/.

1 Introduction

For decades, depth estimation has stood as a fundamental element in the domain of computer vision, finding diverse applications in areas like self-driving, virtual reality, robotics, and scene reconstruction. While the principles of multiple view geometry have long been understood, the rise of deep learning has made single-view depth prediction feasible.

Most self-supervised approaches to monocular depth estimation use photometric loss to evaluate view-synthesis between consecutive video frames, deviating from traditional supervised learning that relies on significant ground truth depth data obtained from expensive sen-

Figure 1: When comparing **BaseBoostDepth** with the baseline Monodepth2, we observe significant improvements in edge-based depth estimation metrics.

sors like LiDAR. Self-supervised methods have attracted attention for their cost efficiency, as they remove the need for ground truth data. Consequently, they can be trained on larger datasets owing to the abundance of available video data, leading to enhanced generalizability as shown in prior research [11], compared to their supervised counterparts.

However, the significance of baseline width in self-supervised methods has not been explored to the same extent as it has in the field of multi-baseline stereo. In multi-baseline stereo, a consistent trend is known: narrower baselines pose an easier pixel matching problem but result in poorer depth estimates.

Despite the potential accuracy advantages of wider baselines, current self-supervised monocular depth (SSMD) methods, such as Monodepth2 (MD2) [9], use narrower baselines in their reconstruction processes. MD2 does this using source images which consist of one subsequent and one preceding consecutive frame to reconstruct the target image. Additionally, it leverages narrow stereo frames in relation to the target image to aid in the reconstruction process. While it is possible to use larger monocular baselines, research conducted by Lokender *et al*. [25] has suggested that employing wider baselines over a larger temporal window introduces challenges such as brightness inconsistencies and increased occlusions, thus making the use of larger baselines a complex problem.

One might consider a straightforward approach: combining large and small baselines and updating the depth estimation based on the most accurate image reconstruction. However, as demonstrated in Section 4.2, this approach introduces a significant bias in favor of smaller baselines, as depth inaccuracies in those images yield lower photometric errors.

Brightness-contrast cues [10, 22, 24], which play a crucial role in our method, rely on the fact that objects closer to the camera tend to appear brighter than those farther away. Additionally, while traditional image-based metrics have proven useful, we aim to bolster the case for wider baselines by also examining edge-based metrics [13], providing a more accurate depiction of how humans perceive depth from two-dimensional images. Furthermore, we analyze point cloud metrics [17] to validate the suitability of our depth estimations for use in 3D applications.

In this work, we leverage wide monocular baselines to achieve state-of-the-art (SotA) depth predictions, as depicted in Figure 1. Our proposed method, **BaseBoostDepth**, outperforms MD2 in terms of image and edge-based metrics. Distinctively, our approach exhibits a stronger reliance on brightness-contrast cues extracted from the input image. These cues significantly enhance boundary definition in our depth estimations without any edge-based supervision. To our knowledge, we are the first to observe the significance of brightness-contrast cues in SSMD estimation.

To accomplish this, we put forward four main contributions:

- **Curriculum-Learning-Inspired Optimization Strategy (3.2)** – This strategy involves a gradual transition from smaller to wider monocular baselines through two stages of training: warmup and boosting.
- **Tri-Minimization (3.3)** – Inspired by multi-baseline stereo, we minimize errors by reconstructing the target image (center frame) from triplets of future and past frames, effectively using multiple reconstructions from different baselines.
- **Incremental Pose Estimations (3.4)** – To address significant drift in pose estimation over larger baselines, which tends towards underestimation, we introduce incremental pose estimation. This technique involves breaking down the pose estimation process into smaller increments within larger intervals.
- **Error-Induced Reconstructions (3.5)** – In addition to using incremental pose estimation, we optimize reconstructions by applying controlled error to the pose estimates. This approach is motivated by our observation that incremental pose estimation does not benefit all reconstructions.

To systematically evaluate each contribution, we conduct an ablation study in Section 4.2 and show SotA performance on both the KITTI and SYNS datasets.

2 Related Work

Self-supervised Depth: Garg *et al*. [8] pioneered self-supervised learning for stereo depth via view synthesis between stereo pairs. Subsequently, Monodepth [9] used photometric loss, combining L_1 loss and SSIM [29], to enforce left-right consistency in reconstructed images. Our focus is on monocular cameras due to their inherent simplicity, which contrasts with the cost constraints and spatial limitations associated with stereo setups. SfM-Learner [37] was the first to utilize view synthesis for monocular depth estimation. Unlike traditional stereo methods, this and subsequent approaches leverage a depth network along with pose estimations to warp images, thereby maximizing photometric uniformity. MD2 [9] introduced per-pixel minimization of photometric error to address occlusion issues, incorporating auto-masking for textureless regions, stationary pixels and dynamic objects. Improving upon MD2, some methods have proposed better depth network architectures [28, 32, 33, 34, 36], or introduced cost volumes to utilize multiple frames as input [16, 30, 31]. Other methods have focused on improving the robustness of monocular depth estimation [15, 21, 22, 35, 38], or on handling the rigid scene assumption [0, 5, 14, 20].

Wider Baselines & Brightness-contrast Cues: Lokender *et al*. [25] were among the first to propose the use of larger frame separations to enhance depth accuracy. However, concerns were raised that larger frame separations would introduce challenges due to increased occlusion and brightness inconsistencies. Madhu *et al*. [26] tackled brightness inconsistencies by implementing per-pixel neural intensity transformation, allowing for a two-frame separation instead of one. Their findings suggested that this increased separation led to improved depth, particularly when addressing brightness inconsistencies. Notably, unlike our method, their exploration did not extend to larger frame separations.

3 Method:

Overview: We present **BaseBoostDepth**, which uses a curriculum-learning-inspired optimization strategy divided into warm-up and boosting stages. Our approach is capable of accurately estimating depth with clearly defined object boundaries. Unlike previous methods, we effectively exploit wider baselines and observe a greater effect of brightness-contrast cues, resulting in SotA depth estimations. Our method is depth backbone-agnostic, allowing any pre-trained or from-scratch depth network to be boosted and achieve enhanced object boundary definition. An overview of the overall framework is depicted in Figure 2.

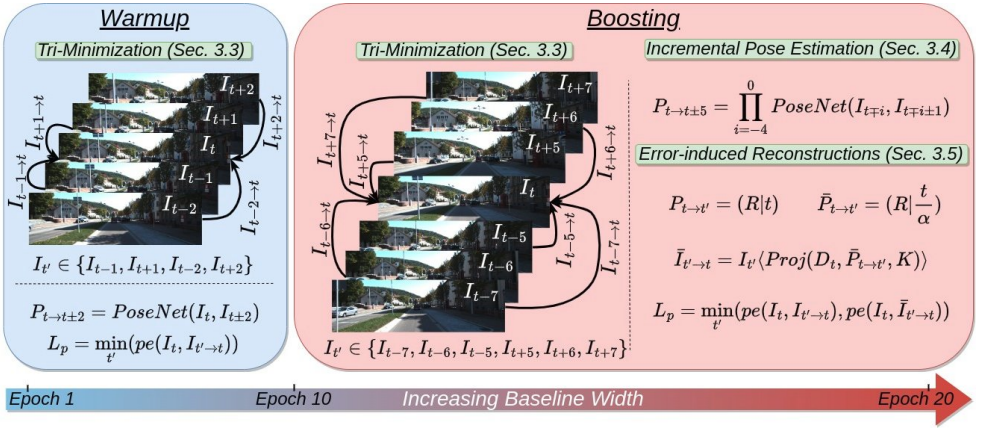

Figure 2: We progressively increase the baseline width used for training the depth and pose networks (3.2). We employ tri-minimization in both warmup and boosting stages (3.3), incorporating incremental pose estimation (3.4) and error-induced reconstructions (3.5) exclusively during the boosting stage.

3.1 Preliminaries

Adhering to the methodology outlined by Zhou *et al.* [37], we concurrently train both an ego-motion network and a depth network to facilitate view synthesis between successive frames. Our approach entails using the target depth estimation $D_t = DepthNet(I_t)$ and camera pose estimations $P_{t\to t'} = PoseNet(I_t, I_{t'})$ to synthesize the target image, where $I_{t'} \in \{I_{t-1}, I_{t+1}\}$, relying solely on source frames. Here, PoseNet and DepthNet denote the pose and depth estimation networks, respectively. The synthesized projection is obtained through inverse warping, as illustrated in the equation below.

$$I_{t'\to t} = I_{t'}\langle Proj(D_t, P_{t\to t'}, K)\rangle \tag{1}$$

The Proj() function yields the resulting 2D coordinates of the depths after projecting into the camera of frame $I_{t'}$, with $\langle\rangle$ denoting the sampling operator. Furthermore, following from [9], our methodology incorporates photometric loss (pe), which is defined as follows:

$$pe(I_a, I_b) = \frac{0.85}{2}\Big(1 - SSIM(I_a, I_b)\Big) + (1 - 0.85)||I_a - I_b||. \tag{2}$$

In MD2, the per-pixel photometric loss employed for training the pose and depth network uses minimum aggregation, as depicted below:

$$L_p = \min_{t'}(pe(I_t, I_{t'\to t})). \tag{3}$$

For each pixel, a determination is made regarding whether to use the next or the previous frames for reprojection based on the minimization of reprojection errors.

3.2 Curriculum-learning-inspired Optimization Strategy

Inspired by curriculum learning's gradual progression from easier to harder samples during training, we categorize image reconstructions by baseline widths. "Easy" samples have smaller baselines, while "harder" samples have larger baselines. Increasing frame separation can yield larger baselines, but baseline widths vary significantly. For example, with a two-frame separation, baselines range from approximately 0.1 to 0.8 distance units, with 0.1 representing roughly 0.54m (stereo images baseline, I_s).

We select frame separations for reconstruction based on predetermined distance estimation using a pre-trained pose network (MD2) between each frame and its subsequent frame in the training dataset. The L_2 norm of the estimated translation matrix determines the baseline for each frame separation, denoted as b for one frame of separation. To calculate the baseline between the target and potential source images, we multiply the predetermined baseline (b) by the number of frames (k) separating them. This can be represented as follows:

$$G(t, t+k) = b \times k. \tag{4}$$

Where $G(t,s) = 0.1$ for stereo frames. This assumes approximately constant velocity over small frame separations.

We select the source frames index denoted by $\hat{x}^+$ to reconstruct the target image I_t, adhering to the following equation:

$$\hat{x}^+ = \arg\max_x \left(G(t,x) \mid x \in \Omega, G(t,x) <= \tau\right) \tag{5}$$

Here, we choose the source image index ($\hat{x}^+$) with the highest baseline relative to the target image ($G(t,x)$), provided it falls below the threshold (τ), from a predefined set of potential source images (Ω). We set τ and Ω to $0.1 + (0.04 \times \text{epoch})$ and $\{s, t+1, t+2\}$ respectively for the warm-up stage and to $(0.1 \times \text{epoch}) - 0.4$ and $\{s, t+1, t+2, t+3, t+4, t+5\}$ respectively for the boosting stage. To mitigate external influences like lighting variations and dynamic objects, we limit our model to a maximum of two frames during the warm-up stage.

We use the positive frame indices as shown in Eq. 5, but we also include the corresponding negative versions of the monocular images. This leads to an updated set of source images $I_{t'} \in \{I_{\hat{x}^+}, I_{\hat{x}^-}\}$ when a monocular frame index is chosen, where $\hat{x}^+ = t+k$ and $\hat{x}^- = t-k$. If a stereo index is selected as the source frame, then $I_{t'} \in \{I_{\hat{x}^+}\}$.

In summary, for each target frame in a batch, we select the maximum frame separation within the bounds of Eq. 5, which leads to a varied set of source images $I_{t'}$ across the batch.

3.3 Tri-minimization:

We introduce tri-minimization, a technique to reconstruct the target image (center frame) from triplets of future and past frames with different baselines to address occlusion, mitigate brightness inconsistencies, and reduce the impact of dynamic objects.

To achieve tri-minimization we attempt to extend the set of source image to include three future frames and three previous frames, including the selected source frame from Eq. 5. The formal equation is shown below:

$$I_{t'} \in \begin{cases} \{I_{\hat{x}^+}\}, & \text{if } \hat{x}^+ = s \\ \{I_{\hat{x}^+}, I_{\hat{x}^-}, I_s\}, & \text{if } \hat{x}^+ = t+1 \\ \{I_{\hat{x}^+}, I_{\hat{x}^+-1}, I_{\hat{x}^-}, I_{\hat{x}^-+1}, I_s\}, & \text{if } \hat{x}^+ = t+2 \\ \{I_{\hat{x}^+}, I_{\hat{x}^+-1}, I_{\hat{x}^+-2}, I_{\hat{x}^-}, I_{\hat{x}^-+1}, I_{\hat{x}^-+2}\}, & \text{otherwise.} \end{cases} \tag{6}$$

Given that a monocular frame is selected (i.e., $\hat{x}^+ \neq s$), then $\hat{x}^- + 1 = t-k+1$ and $\hat{x}^+ - 1 = t+k-1$. When using tri-minimization, we encourage larger baseline widths to counteract the preference for smaller baselines in minimization aggregation. This is achieved by using more widely separated potential source frames (Ω) and a more aggressive τ threshold in the boosting stage ($\Omega = \{s, t+1, t+2, t+3, t+4, t+5, t+6, t+7\}$ and $\tau = (0.15 \times \text{epoch}) - 0.9$).

3.4 Incremental Pose Estimation

When training with larger frame separations and assuming constant velocity within short time intervals (less than one second), consecutive frame translations are expected to remain approximately linear. However, upon inspection, we observed pose estimation drift with increased frame separations, indicating better performance with smaller separations and worse with larger ones. Therefore, we propose incremental pose estimation as follows:

$$P_{t \to t \pm n} = \prod_{i=-(n-1)}^{0} PoseNet(I_{t \mp i}, I_{t \mp i \pm 1}). \tag{7}$$

Equation 7 represents the matrix multiplication of incremental pose estimations, leading to a refined pose estimation over larger frame separations. For further support, see the supplementary materials.

During tri-minimization, we discovered that using incremental pose estimations was beneficial only for the smallest frame separation to the target image within the set of source images $I_{t'}$. However, we found that rotation estimations from incremental pose estimations were beneficial for all reconstructions.

3.5 Error-induced Reconstructions

Based on the discovery that partial incremental pose results in better image-based and edge-based performance than a full incremental pose, we propose that adding a fixed error to the pose network could lead to improved performance. Integrating reconstructions based on pose estimations with a fixed error in translations empirically enables the depth network to better understand the influence of pose estimations on reconstruction accuracy. By incorporating these reconstructions, we expand the solution space, where the introduced pose errors act as a form of perturbation to guide the depth network towards exploring alternative solutions and enhance its ability to generalize. This observation is well-supported by experimentation. Refer to Section 4.2 for more details.

Using the incremental pose estimations, we define the rotation and translation as $(R|t) = P_{t\to t'}$, then the error-induced pose is defined as $\bar{P}_{t\to t'} = (R|\frac{t}{\alpha})$. Then;

$$\bar{I}_{t'\to t} = I_{t'}\langle Proj(D_t, \bar{P}_{t\to t'}, K)\rangle. \tag{8}$$

Note that we do not change the corrected rotation estimations and that the error-induced pose gradients are cut off during backpropagation. Finally, we minimize between the standard reconstructions $pe(I_t, I_{t'\to t})$ which use incremental pose and the error-induced reconstructions $pe(I_t, \bar{I}_{t'\to t})$ which use the error-induced pose:

$$L_p = \min_{t'}(pe(I_t, I_{t'\to t}), pe(I_t, \bar{I}_{t'\to t})). \tag{9}$$

The final loss, incorporating photometric loss with automasking (μ) from MD2 [9] and per-pixel smoothness loss from [19], is defined as $L = \mu L_p + \lambda L_s$, and this combined loss is averaged across each pixel, scale, and batch.

4 Results:

Experimental set-up: For training **BaseBoostDepth**, we utilize pretrained ImageNet weights [8] with PyTorch [18] on an NVIDIA A6000 GPU. We employ the Adam optimizer [12] for 20 epochs, using an input size of 640×192 and a multi-step learning rate strategy. The learning rate starts from 1e-4 and is progressively reduced at epochs 11, 13, 15, 16, 17, 18, and 19 by a factor of 0.4. Hyperparameters ω, β, and γ are set to 0.01, 0.01, and 0.001, respectively, with a smoothing loss parameter λ of 0.001. Through empirical testing, the warm-up stage spans the initial 10 epochs using 4 resolution scales, while the boosting stage covers the subsequent 10 epochs with only the largest resolution scale. For variations like **BaseBoostDepth**$_{pre}$ and **BaseBoostDepth**$^{\dagger}_{pre}$ in Table 2 and Table 3, we train exclusively with the boosting stage starting from pretrained weights and depth backbones from MD2 (ResNet-18 encoder) and MonoViT, respectively.

4.1 Datasets

KITTI [9]: For validation, we use the official Zhou split with 4,424 images and train on the full set of 39,810 images. Testing is done on 697 images from the Eigen *et al*. test set [4]. Due to ground truth accuracy limitations, our evaluation focuses on image-based metrics for depth estimate assessment. All models are trained exclusively on the KITTI dataset.
SYNS-Patches [1]: This dataset comprises 1,438 outdoor images with accurately measured ground truth depth information. We adopt edge-based metrics from Koch *et al*. [13] and point cloud-based metrics from Örnek *et al*. [17], following the methodology outlined by Spencer *et al*. [23]. Note that the exact steps for evaluating this dataset are not provided; therefore, we have created our own version of the evaluation, which is released with the code.

4.2 Ablation Study:

This subsection investigates the impact of each contribution on the baseline model, as shown in Table 1. We primarily analyze the KITTI test dataset using image-based metrics and evaluate edge-based metrics for each contribution using the SYNS test set.

Ablation	Contributions						KITTI							SYNS	
	Skip	Pre	Tri.	Incri. Pose	Part. Incri.	Err. Rec.	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Acc	Comp
Monodepth2 [D]	1						0.106	0.818	4.750	0.196	0.874	0.957	0.979	2.516	17.193
Monodepth2 [D]	4						0.107	0.832	4.723	0.186	0.887	0.961	0.982	2.512	14.856
Monodepth2 [D]	[4]						0.146	1.164	5.289	0.221	0.813	0.940	0.975	2.465	5.278
BaseBoostDepth	*C*	✗					0.115	0.916	4.856	0.190	0.877	0.960	**0.983**	2.442	5.518
BaseBoostDepth	*C*	✗	✓				0.112	0.867	4.762	0.187	0.879	0.962	**0.983**	**2.417**	**3.433**
BaseBoostDepth	*C*	✗	✓	✓			0.109	0.868	4.767	0.186	0.883	0.961	0.982	2.489	6.547
BaseBoostDepth	*C*	✗	✓	✓	✓		0.107	0.799	4.656	0.184	0.884	**0.963**	**0.983**	2.450	4.290
BaseBoostDepth	*C*	✗	✓	✓	✓	✓	0.106	0.736	4.584	0.184	0.883	**0.963**	**0.983**	2.453	3.810
BaseBoostDepth$_{pre}$	*C*	✓	✓	✓	✓	✓	**0.104**	**0.738**	**4.544**	**0.183**	**0.888**	**0.963**	**0.983**	2.432	4.763

Table 1: **Ablation Study:** Here, we present all contributions of our work. **Bold** represents the best results for the metric and underscore is the second best.

Baseline (Row 1 & Row 2): The first row depicts MD2 using a one-frame separation, trained with monocular and stereo images. In contrast, the second row involves reconstructions using up to 4-frames of separation, but shows no significant quantitative improvements in image-based or edge-based depth metrics. This highlights the challenge of using wider baselines with standard minimum aggregation, where smaller baselines produce more accurate reconstructions despite potentially less accurate depth estimations. As a result, wider baselines are often overlooked in optimization steps, a challenge addressed by **BaseBoostDepth**.

Wide Baseline (Row 3): Results are shown after training with a 4-frame separation ($I_{t'} \in I_{t-4}, I_{t+4}$). The network faces challenges with larger baselines due to increased occlusion, brightness changes, and dynamic objects. While image-based metrics decline, edge-based metrics improve significantly, indicating better edge definition with larger baselines.

Curriculum-learning-inspired Optimization (Row 4): Our curriculum-learning-inspired optimization strategy (*C*) yields worse image-based metrics compared to row 1 due to larger baselines introduced during the boosting phase, leading to unstable optimization. However, we maintain respectable image-based metrics and achieve significant edge-based improvements similar to row 3, thanks to the gradual introduction of larger baselines.

Tri-minimization (Row 5): Tri-minimization yields improvements in image-based metrics and notably achieves impressive edge-based metrics surpassing those of row 3. This shows our ability to achieve greater edge-based metrics while improving image-based metrics.

Incremental Pose Estimation (Row 6 & Row 7): We hypothesize that many observed errors were due to drifted pose estimations. To tackle this, in row 6 we introduce incremental pose estimation, resulting in another decline in image-based metrics but an increase in edge-based metrics. In row 7, we apply incremental pose solely to the smallest frame separation in tri-minimization, while consistently using incremental pose estimations for rotation (referred to as "partial incremental pose"). This approach again leads to significant reductions in both image-based and edge-based metrics.

Error-induced Reconstructions (Row 8): We observe another notable decrease in both edge-based and image-based metrics when using error-induced reconstructions with an α value of 5.5. Fine-tuning details are provided in the supplementary materials.

Pre-trained with MD2 (Row 9): Finally, we initialize our model with Monodepth2 weights and then apply our boosting stage. Combining contributions, we exceed the image-based metrics set by Monodepth2 and achieve significant improvements in edge-based performance.

4.3 Comparison with SotA

In Table 2, we compare the previous SotA results with different variations of **BaseBoostDepth**. The version trained from scratch achieved performance comparable to MD2. However, significant performance gains were observed when applying the boosting phase to pre-trained depth networks. Our method consistently outperforms the original approach by leveraging all contributions from the boosting phase, and **BaseBoostDepth**$^{\dagger}_{pre}$ establishes a new SotA benchmark for the given resolution using the MonoViT depth backbone and pre-trained weights.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [0]	0.106	0.818	4.750	0.196	0.874	0.957	0.979
CADepth [CA]	0.102	0.752	4.504	0.181	0.894	0.964	0.983
DIFFNet [DN]	0.101	0.749	4.445	0.179	0.898	0.965	0.983
MonoViT [MV]	0.098	0.683	4.333	0.174	0.904	0.967	0.984
BaseBoostDepth	0.106	0.736	4.584	0.184	0.883	0.963	0.983
BaseBoostDepth$_{pre}$	0.104	0.738	4.544	0.183	0.888	0.963	0.983
BaseBoostDepth$^{\dagger}_{pre}$	**0.096**	**0.648**	**4.201**	**0.170**	**0.906**	**0.968**	**0.985**

Table 2: **Quantitative Results for the KITTI Eigen Test Dataset.** Note that we do not use any test-time refinement processes. Our depth estimation inference relies on a single frame, and we train at a resolution of 640 × 192 without edge supervision. Models subscript with *pre* are trained only with the boosting stage from pretrained weights, while † indicates the use of MonoViT's depth backbone; otherwise, a ResNet18 backbone is used.

4.4 Evaluation of Edge and Point Cloud Performance

Method	Image-Based					Edge-Based		Point Cloud-Based	
	Abs Rel	MAE	Sq Rel	RMSE	RMSE log	Acc	Comp	F-Score	IoU
Monodepth2 [0]	0.334	6.901	5.285	12.089	0.405	2.516	17.193	0.242	0.149
CADepth [CA]	0.363	8.787	5.548	13.512	0.546	2.473	19.045	0.022	0.012
DIFFNet [DN]	0.311	6.554	4.690	11.610	0.383	2.411	12.116	0.258	0.161
MonoViT [MV]	0.287	6.195	4.399	11.124	0.354	2.443	15.672	0.264	0.164
BaseBoostDepth	0.334	6.878	4.854	11.847	0.409	2.453	**3.810**	0.275	0.174
BaseBoostDepth$_{pre}$	0.328	6.752	4.815	11.752	0.405	2.432	4.763	0.268	0.168
BaseBoostDepth$^{\dagger}_{pre}$	**0.278**	**5.951**	**3.795**	**10.575**	**0.351**	**2.409**	5.314	**0.300**	**0.191**

Table 3: **Quantitative Results for the SYNS Test Dataset.** Here we show test evaluations done on the SYNS dataset using image, edge and point cloud based metrics.

To truly show the benefits of using large baselines, we evaluate our depth estimates against the SYNS-patches dataset to compare with other SotA models in Table 3.

Our analysis shows that leveraging our boosting phase improves performance across image, edge, and point cloud-based metrics on the SYNS dataset, *regardless of the depth backbone used*. Comparing Monodepth2 with **BaseBoostDepth** trained using pre-trained ImageNet weights, we maintain similar image-based metric performance but achieve significant improvements in edge-based metrics and excel in point cloud metrics, demonstrating accuracy in 3D space.

In summary, **BaseBoostDepth**$^{\dagger}_{pre}$ leads on the SYNS dataset, slightly behind in edge composition compared to **BaseBoostDepth**. Additionally, CNN-based architectures generally outperform vision transformers in edge composition, likely due to the texture-focused

nature of CNNs versus the shape-focused approach of transformers. Ultimately, the choice of depth backbone depends on specific user requirements.

Finally, we present qualitative results in Figure 3, which illustrate the advantages of our depth network over prior methods, owing to our superior depth edge accuracy.

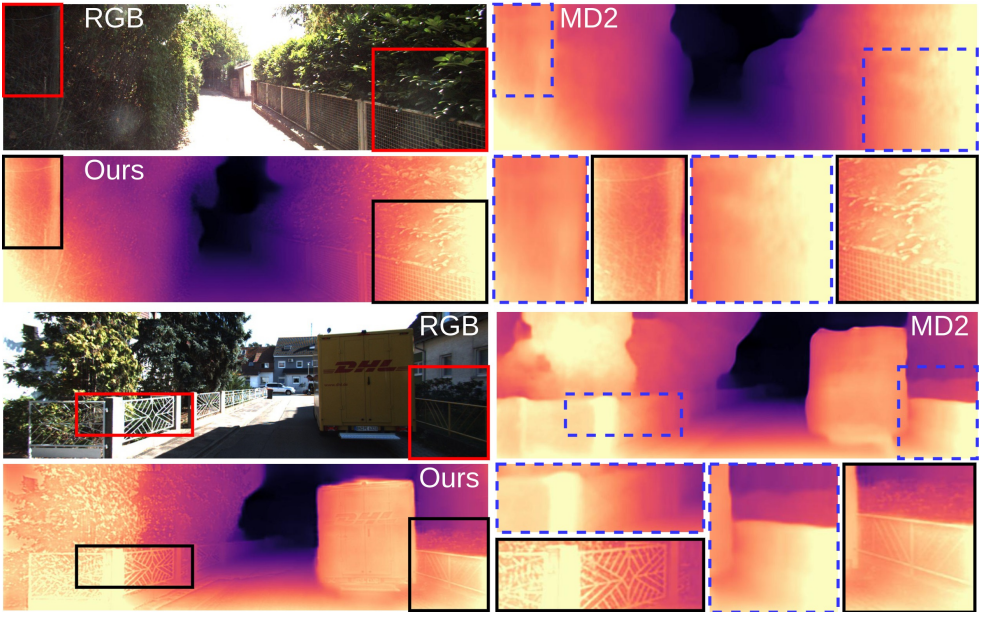

Figure 3: Here, we show clear examples where BaseBoostDepth (black solid border) outperforms Monodepth2 (blue dashed border) in capturing sharp details around fine regions. Thin mesh railings are visible with our depth network, and leaf structures emerge more clearly.

5 Conclusion

Our study demonstrates leveraging wider baselines for improved self-supervised monocular depth estimation, enhancing image, edge, and point cloud metrics. Traditionally, wider baselines were avoided in depth estimation due to perceived limitations and the oversight of wider baselines when using minimum aggregation. However, by implementing our curriculum-learning-inspired strategy, and carefully guiding the pose estimations, we extract great benefits for edge-based depth improvements. Our boosting strategy is depth backbone-agnostic and can be initialized from the warm-up phase or with pre-trained weights. Additionally, our improvements do not result in any increase in computational cost at test time. We anticipate that our findings will advance research toward more refined detail adaptation in the future.

6 Acknowledgement

This research was funded and supported by the EPSRC's DTP, Grant EP/W524566/1. Most experiments were run on Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

[1] Wendy J Adams, James H Elder, Erich W Graf, Julian Leyland, Arthur J Lugtigheid, and Alexander Muryy. The southampton-york natural scenes (syns) dataset: statistics of surface attitude. *Scientific reports*, 6(1):35805, 2016.

[2] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[5] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *European Conference on Computer Vision*, pages 228–244. Springer, 2022.

[6] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

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

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[9] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[10] Paul B Hibbard, Ross Goutcher, Rebecca L Hornsey, David W Hunter, and Peter Scarfe. Luminance contrast provides metric depth information. *Royal Society Open Science*, 10(2):220567, 2023.

[11] Hanjiang Hu, Baoquan Yang, Zhijian Qiao, Ding Zhao, and Hesheng Wang. Seasondepth: Cross-season monocular depth prediction dataset and benchmark under multiple environments. *arXiv preprint arXiv:2011.04408*, 2020.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[14] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1863–1872, 2021.

[15] Yifan Mao, Jian Liu, and Xianming Liu. Stealing stable diffusion prior for robust monocular depth estimation. *arXiv preprint arXiv:2403.05056*, 2024.

[16] Xingyu Miao, Yang Bai, Haoran Duan, Yawen Huang, Fan Wan, Xinxing Xu, Yang Long, and Yefeng Zheng. Ds-depth: Dynamic and static depth estimation via a fusion cost volume. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] Evin Pınar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. *arXiv preprint arXiv:2203.08122*, 2022.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learni pdf.

[19] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019.

[20] Kieran Saunders, George Vogiatzis, and Luis J Manso. Dyna-dm: Dynamic object-aware self-supervised monocular depth maps. In *2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 10–16. IEEE, 2023.

[21] Kieran Saunders, George Vogiatzis, and Luis J Manso. Self-supervised monocular depth estimation: Let's talk about the weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8907–8917, 2023.

[22] Barry J Schwartz and George Sperling. Luminance controls the perceived 3-d structure of dynamic 2-d displays. *Bulletin of the psychonomic society*, 21(6):456–458, 1983.

[23] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *arXiv preprint arXiv:2208.01489*, 2022.

[24] Nan-Ching Tai and Mehlika Inanici. Luminance contrast as depth cue: Investigation and design applications. *Computer-aided design and applications*, 9(5):691–705, 2012.

[25] Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo rgb-d for self-improving monocular slam and depth prediction. In *European conference on computer vision*, pages 437–455. Springer, 2020.

[26] Madhu Vankadari, Stuart Golodetz, Sourav Garg, Sangyun Shin, Andrew Markham, and Niki Trigoni. When the sun goes down: Repairing photometric losses for all-day depth estimation. In *Conference on Robot Learning*, pages 1992–2003. PMLR, 2023.

[27] Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. *arXiv preprint arXiv:2404.09831*, 2024.

[28] Youhong Wang, Yunji Liang, Hao Xu, Shaohui Jiao, and Hongkai Yu. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. *arXiv preprint arXiv:2309.00526*, 2023.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[30] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021.

[31] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6112–6122, 2021.

[32] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2021.

[33] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023.

[34] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. *arXiv preprint arXiv:2208.03543*, 2022.

[35] Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3d adversarial attacks against monocular depth estimation in autonomous driving. *arXiv preprint arXiv:2403.17301*, 2024.

[36] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021.

[37] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[38] Ruijie Zhu, Ziyang Song, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Ec-depth: Exploring the consistency of self-supervised monocular depth estimation under challenging scenes. *arXiv preprint arXiv:2310.08044*, 2023.