2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Dusk Till Dawn: Self-supervised Nighttime Stereo Depth Estimation using Visual Foundation Models

Madhu Vankadari, Samuel Hodgson, Sangyun Shin*, Kaichen Zhou*, Andrew Markham, and Niki Trigoni

***Abstract*— Self-supervised depth estimation algorithms rely heavily on frame-warping relationships, exhibiting substantial performance degradation when applied in challenging circumstances, such as low-visibility and nighttime scenarios with varying illumination conditions. Addressing this challenge, we introduce an algorithm designed to achieve accurate self-supervised stereo depth estimation focusing on nighttime conditions. Specifically, we use pretrained visual foundation models to extract generalised features across challenging scenes and present an efficient method for matching and integrating these features from stereo frames. Moreover, to prevent pixels violating photometric consistency assumption from negatively affecting the depth predictions, we propose a novel masking approach designed to filter out such pixels. Lastly, addressing weaknesses in the evaluation of current depth estimation algorithms, we present novel evaluation metrics. Our experiments, conducted on challenging datasets including Oxford RobotCar and Multi-Spectral Stereo, demonstrate the robust improvements realized by our approach.**

I. Introduction

Depth estimation is a pivotal subject within computer vision, with wide-ranging implications for applications such as autonomous driving, augmented and virtual reality, and robotics [1], [2]. Despite the accomplishments of supervised depth estimation algorithms, these methods typically depend on high-resolution ground truth data - a challenge that requires substantial computational resources, costly 3D LiDAR sensors, and heavy computational requirements [3], [4].

Addressing the need for ground-truth data, recent research has shown interest in self-supervised depth estimation methods [5], [6]. Such approaches typically warp the source frame to the target frame using the learned depth information. A photometric loss is found between the reconstructed and actual target images to constrain the learning process. While monocular depth estimation algorithms are widely applicable, they often lack scale information [7] and exhibit limited generalizability [8]. In contrast, self-supervised stereo depth estimation algorithms that use the correspondence between left and right frames yield more robust performance [9].

Self-supervised stereo depth estimation, however, relies on photometric consistency assumptions and conventional warping relationships, which are constrained by favorable lighting conditions. These assumptions break down in nighttime scenarios, characterized by low texture and fluctuating illumination [11]. Traditional approaches primarily depend on pretrained classification networks, such as [12], for feature extraction. The utility of these networks is confined to relatively small datasets due to the need for labeled data, and their results are domain-specific. Contemporary self-supervised feature extractors offer enhanced robustness and clarity in feature mapping, as they do not rely on labeled data [13], [14]. Conventional stereo depth estimation techniques also tend to combine stereo features indiscriminately, without addressing regions that have correspondence issues [9], [15].

Fig. 1: A comparison of the estimated disparity using our method (Ours) with a SOTA stereo-matching method, IGEV-Stereo [10]. Note how the sky (blue rectangle) is incorrectly estimated by IGEV-Stereo as being very near. Similarly, there is a lack of detail showing the edge of the wall and the lamp-post (green rectangle). In comparison, our method is able to accurately estimate these depths.

In light of this, our approach concentrates on obtaining accurate self-supervised stereo depth estimation at night. The contributions of our paper can be summarized as follows:

- We present an architecture capable of efficient self-supervised stereo depth estimation at night, using visual foundation models and a photometric loss function.
- We introduce a feature-level mask to mitigate the impact of pixels violating illumination assumptions.
- We propose a distance regularizer aimed at enhancing the feature descriptions to estimate accurate depth maps.
- We provide a more rational set of evaluation metrics to assess the performance of depth estimation.

II. Related Work

Photometric consistency assumes the scene to be static, free of any noise or occlusions, lambertian and temporally illumination invariant. Various changes in the appearance of scenes during nighttime cause systems trained only for daytime datasets to fail. Approaches are required to deal with the scene lighting issues, such as the lack of the sun as a primary source of light.

A. Nighttime Self-supervised Depth Estimation

Loosely, we classify previous work as being based either on domain adaptation or image enhancement. Adaptation-

∗ refers to equal contribution
All authors are with the University of Oxford, Oxford, UK

based approaches seek to overcome the domain shift between day and night using synthesised data to align day/night or clear/inclement weather image encodings [16], [11]. They aim to obtain the same features across different conditions. Synthesised data is created using standard datasets (i.e. [17]), and a Generative Adversarial Network (GAN). [18] uses a Monodepth2 [4] architecture with an adversarially-trained nighttime image encoder. Features are aligned using a GAN to generate day images from night, and a discriminator to enforce similarity. [19] emulates this approach, but trains in the output space as well as the feature space. [20] uses the GAN outputs to train two feature extractors, a day-night invariant feature extractor that forms the backbone for depth prediction, as well as night and day style extractors for training. Outside of our broader categorisation, [21] uses a feature space (rather than image space) contrastive loss to improve domain generalisation beyond that of [4]. Enhancement-based approaches aim to improve performance over day models by taking greater account of scene lighting, for example, by isolating illumination information within the image. [22] uses a learned image enhancement and adapted masking from [4]. [23] estimates the illumination change between the consecutive nighttime images to relax assumptions of photometric loss for better depth estimation.

B. All-day Self-supervised Depth Estimation

Concurrent state-of-the-art approaches [16] and [11] also consider weather conditions in their approaches. Both adapt [4] for their all-day unified networks, using GANs to augment for both weather and lighting conditions. [16] uses a semi-augmented warping to minimise GAN-induced inconsistency between consecutive frames, and uses raw inputs for pose estimation to minimise error propagation. Similarly, [11] uses daytime depth estimation to distill depth knowledge to nighttime using image translation networks. [24] uses a complex, partially adversarial architecture, with a learned image enhancement to estimate illumination and uncertainty, which is then masked from the loss.

C. Supervised Stereo Depth Estimation

All approaches mentioned so far are monocular, even if stereo images are used during training. In terms of supervised stereo depth estimation, [25] uses a transformer and 4D correlation feature matrix, including an iterative refinement inspired by [26], to derive depth, flow, and disparity. [27] uses a coarse-to-fine approach with a hierarchical network and adaptive group correlation for getting fine disparities. [28] builds an attention-based cost volume to suppress redundant information and enhance matching-related information. [10] unifies stereo and optical flow approaches based on 2D convolution, avoiding the memory cost of 3D convolution. [29] trains a model using day datasets that have ground truth depth and further uses domain adaptation to work on day to night. Our method does not use any daytime data, domain adaptation, or ground truth depth for training.

III. Proposed Method

Given a pair of stereo images (I_l, I_r), we aim to estimate the per-pixel depth map $\mathbf{d}$ using self-supervised learning. The proposed framework is depicted in Fig 2. Our method is composed of 4 main components, namely the feature extractor, projection head, stereo matcher, and upsampler.

A. Feature Extractor

In contrast to existing approaches, we use visual foundation models, DINO [13] and DINOv2 [14], as feature extractors for the input images. The models are trained on the ImageNet [30] and Facebook LVD-142M [14] datasets, respectively. In our experiments, we use the pretrained *small*-models of DINO and DINOv2, with patch sizes of 8 and 14, respectively. For the rest of the paper, we refer to DINO ViT-S/8 (patch size 8) as the encoder unless otherwise stated. Our encoder has a *conv*-layer followed by a series of 12 *transformer* layers. Given an image $I \in \mathbb{R}^{H\times W\times 3}$ with height H and width W, the conv-layer converts the image into overlapping 8×8 patches with stride 4, resulting in a feature-map $\hat{f} \in \mathbb{R}^{h'\times w'\times 384}$ where $h' = H/4$, $w' = W/4$. After processing $\hat{f}$ through the first 6 transformer layers, the estimated output feature-map $\tilde{f} \in \mathbb{R}^{h'\times w'\times 384}$ is subsampled with stride 2 in the height and width dimensions, giving an $8\times$ downscaled feature map that is further processed by 6 more transformer layers resulting $f \in \mathbb{R}^{h\times w\times 384}$, where $h = H/8$, $w = W/8$.

Selection of the transformer layers is based on [31]. They observe that various encoder layers act similarly to hierarchical CNN layers in capturing local-to-global information as the depth of the network increases. Deeper layers capture semantics, and shallow layers capture local details (including positions). Middle layers tend to carry both. For accurate disparity estimation, we use both deep and middle layer features, $(f, \tilde{f})$, for stereo matching.

B. Projection Head

The projection head takes feature vectors $f, \tilde{f}$ from the encoder and projects them into lower dimensional space $\mathbb{R}^D$ where $D < 384$. This is done because (1) when PCA is performed on the 384 dimensions, we observe the 10 first principal components to explain more than 50% of the total variance, suggesting it is safe to project to lower dimensional space and (2) the computational complexity of the stereo matcher during inference and training will be reduced.

The projection head consists of two *conv*-layers with kernel size as 1×1, and ReLU as activation function in the middle. We use $D = 128$ as the output description dimension and the same projection head for both left and right features.

C. Stereo Matcher

The stereo matcher takes the multi-scale left and right features $\{(f_l, f_r), (\tilde{f}_l, \tilde{f}_r)\}$ as input and estimates the disparity map $d_r \in \mathbb{R}^{h'\times w'\times 1}$ in three stages. First, features are enhanced with cross-image feature context using a transformer module. Then, a disparity map d_g is estimated at $1/8$ scale using the (f_l, f_r) features and global matching. Lastly, the

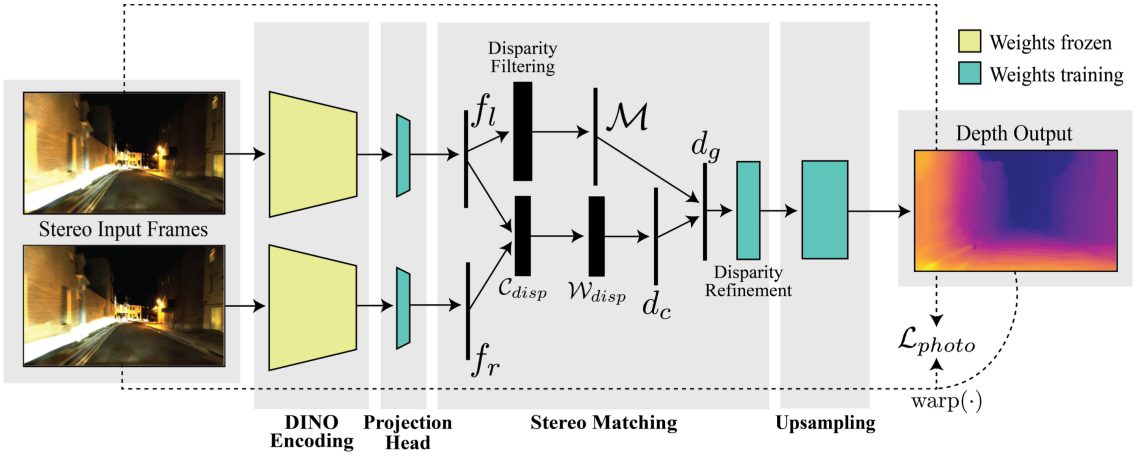

Fig. 2: Our approach consists of four main elements. Features are encoded independently for each input using DINO [13], a learnable projection head adapts these features and reduces their dimension, giving f_l and f_r. Stereo matching of the features then takes place, with disparity filtering yielding the mask $\mathcal{M}$, and the combination of f_l and f_r providing the correspondence volume $\mathcal{C}_{disp}$. $\mathcal{W}_{disp}$ is found by using softmax on $\mathcal{C}_{disp}$, which is used to find coarse disparity d_c. Coarse disparity and the mask combine to give global disparity d_g, which is refined and upsampled to give final depth.

disparity is bilinearly interpolated by $2\times$ and corrected for the interpolation artifacts with a residual disparity estimated using $(\tilde{f}_l, \tilde{f}_r)$ and local matching. These modules are explained in detail in the following sections.

Transformer: Stereo images are processed through the encoder and projection head individually, meaning no cross-image interaction or information exchange. In previous works [25], [32], [33], it is suggested that adding cross-image information results in better matching accuracy. We therefore use a transformer module similar to [25] with fixed positional encoding after the projection head. To further reduce computational complexity, we only aggregate features that are on the respective epipolar line for any given feature. We use rectified stereo-images that make the search problem 1D, as the epipolar lines are strictly straight. Given f_l and f_r, the transformer module outputs $\mathbf{f}_l$ and $\mathbf{f}_r$ in $R^{h\times w\times D}$.

Coarse Disparity Estimation: The output features from the transformer module are used to compute the dense correspondence volume C_{disp} using normalized feature correlation (i.e, global matching) with a simple matrix multiplication

$$\mathcal{C}_{disp} = \frac{\mathbf{f}_l \cdot \mathbf{f}_r^{\top}}{\sqrt{D}} \in \mathbb{R}^{h\times w\times w}. \tag{1}$$

We obtain the matching distribution $\mathcal{W}_{disp}$ using a softmax over the last dimension of $\mathcal{C}_{disp}$. This is then multiplied with a 1D pixel grid $\mathcal{P}_{1D} \in \mathbb{R}^{w}$ to obtain the corresponding pixel locations $\mathcal{G}_{1D} \in \mathbb{R}^{h\times w}$. Finally, the coarse disparity d_c can be computed as the difference between $\mathcal{P}_{1D}$ and $\mathcal{G}_{1D}$. Formally, this can be written as:

$$\begin{aligned} \mathcal{W}_{disp} = \mathrm{softmax}(C_{disp}), \text{ and } \mathcal{G}_{1D} = \mathcal{W}_{disp}\mathcal{P}_{1D}, \\ d_c = \mathrm{ReLU}(\mathcal{G}_{1D} - \mathcal{P}_{1D}), \end{aligned} \tag{2}$$

where ReLU ensures the disparity is always positive.

Disparity Filtering: One observation made during our experiments is that features belonging to areas of low texture, particularly the sky, resulted in very noisy disparity estimates similar to [34], [10] as shown in Fig. 1. We explain this as being due to reduced camera sensitivity in low-light causing an accumulation of noise, resulting in erroneous and noisy features. Such features exacerbate incorrect feature matches, causing noisy estimates, as we are extracting disparity as a byproduct, rather than estimating it directly as in [29]. To address this, we propose a simple yet effective solution based on intra-image feature description distances, using them to mask the noisy areas. We conjecture that features belonging to noisy areas tend to have a lower minimum distance from their nearest neighbors, compared to the features that belong to well-lit areas. In a set of n features, given a feature $f_i \in f$ with $i \in \{1, 2, 3..., n\}$: (1) we normalize to have unit length; (2) we estimate one to all cosine similarity to find the nearest neighbor f_j, with $j \neq i$; and (3) we calculate l_2 distance p_i between f_i and f_j. Formally, this is, $f'_j = \mathrm{argmax}_j(f'_i \cdot f'^{\top}_j), j \neq i$ where $f' = \frac{f}{||f||}$ and $p_i = ||f_i - f_j||$. Estimated distances are then used to filter good disparity values from noisy ones by estimating a mask $\mathcal{M}$, which is used to find masked disparity d_m as:

$$d_m = \mathcal{M} \cdot d_c, \quad \mathcal{M} = \begin{cases} 1, & \text{if } p_i > \zeta, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where ζ is set to 0.2 in our experiments.

Disparity Propagation Some poorly lit regions that contain useful information may get lost during masking. This creates holes in the estimated disparity. We, therefore, propagate the disparity at the valid pixels to the masked-out pixels by measuring self-similarity, as in [25]. This is done using an attention layer to find global disparity, d_g:

$$d_g = \mathrm{softmax}\left(\frac{f_l \cdot f_l^{\top}}{\sqrt{D}}\right) \cdot d_m \tag{4}$$

Disparity Refinement: The current global disparity d_g is at $1/8\times$ resolution. To increase the resolution, we upsample d_g using bilinear interpolation by $2\times$. Doing so inevitably introduces interpolation artifacts, so we use fine-level features

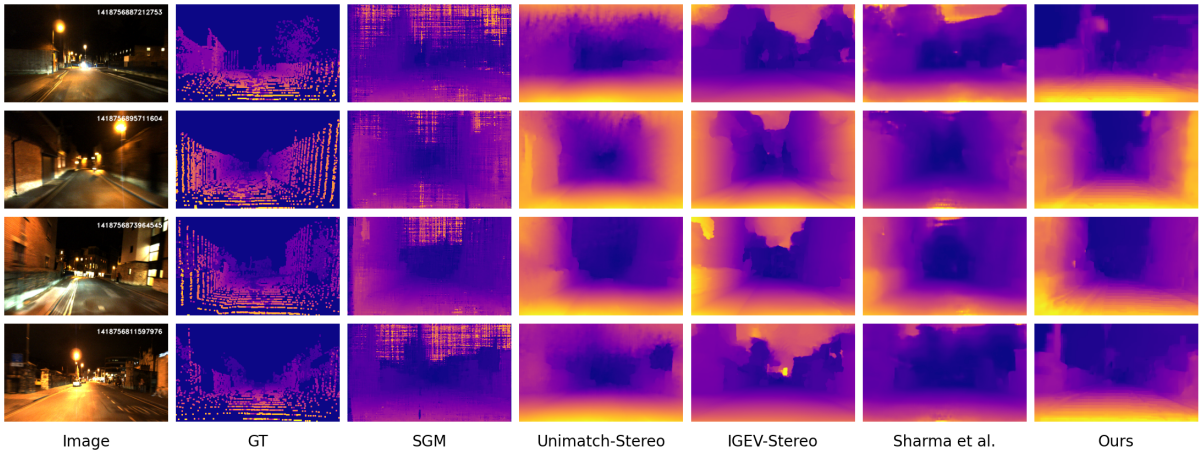

Fig. 3: The qualitative comparison of the proposed method with SGM [34] and the state-of-the-art supervised methods Unimatch-Stereo [25], IGEV-Stereo [10], and Sharma et al. [29]. The brighter the pixel is, the closer it is to the camera.

$\tilde{f}_l, \tilde{f}_r$ to account for this. We first warp the right-features $\tilde{f}_r$ onto the left frame using the upsampled disparity d_g^{up}. The original left-features $\tilde{f}_l$ and the warped-features $\tilde{f}_l'$ are then passed through the transformer, disparity estimation, and filtering to estimate residual disparity ∇d. Using localised attention on subsampled features in the refinement process is suggested in [25] to improve accuracy through local feature interactions. The residual disparity is added to the upsampled disparity, giving $d_r = d_g^{up} + \nabla d$. Finally, the refined disparity d_r is correlated with $\tilde{f}_l$ to output disparity $\tilde{d}_r \in \mathbb{R}^{h' \times w' \times 1}$.

D. Upsampling

The disparity output $\tilde{d}_r$ is upsampled to give final output disparity $d \in \mathbb{R}^{H \times W \times 1}$. Instead of bilinear interpolation, which can result in blurry borders, we use the learnable convex upsampling method proposed by RAFT [26]. In it, two (3×3) convolutional layers are used to predict a mask and perform softmax over the 9 neighbours of a given pixel.

E. Training Losses

Photometric Loss: Estimated disparity is used to reconstruct the left image from the right using bilinear interpolation. The reconstructed image is compared with the original left image to calculate the photometric loss. Given the left and right-images (I_l, I_r) and estimated disparity d, the photometric loss L_{photo} is:

$$\mathbf{d} = \frac{b\lambda}{d}, \quad \hat{I}_l = \text{warp}(I_r, \mathbf{d}, K, T_{lr}), \tag{5}$$

$$L_{\text{photo}} = \alpha |I_l - \hat{I}_l| + (1 - \alpha)\text{SSIM}(I_l, \hat{I}_l), \tag{6}$$

where $\mathbf{d}$ is the output depth map, b is the baseline distance between the left and right cameras, λ is the focal length, K is the camera calibration matrix, $T_{lr} \in SE(3)$ is extrinsics of the stereo-rig, α is the convex combination weight between L_1 and $SSIM$ losses, and is set to 0.15. The $\text{warp}(\cdot)$ function warps the left from the right image using $\mathbf{d}, K$ and T_{lr}. More details of this loss can be found in [4].

Distance Regularizer: We also encourage all of the features to maximise the minimum distance from their nearest neighbour using a regularization loss inspired by [35]. This allows the features from poorly lit areas to improve. However, there is an imbalance between well and poorly-lit areas in the majority of our training split images, similar to the class imbalance problem from classification literature. This is addressed by using a modulation factor, γ, to reduce the concentration of loss on features that already have higher minimum distance with their nearest neighbour, and to focus more on small feature distances. Formally:

$$L_{\text{reg}} = -\frac{1}{n}\sum_{i=0}^{n}(1 - p_i)^{\gamma} \log(p_i), \tag{7}$$

where γ is a modulation factor similar to focal-loss in [36], used here to focus more on features that have low minimum distances. We set $\gamma = 2$ in our experiments.

In order to make the estimated disparity spatially smooth while preserving the edges, the common edge-aware Disparity Smoothness Loss L_{smooth} from [3] is used. Finally, the total loss is $L_{\text{total}} = L_{\text{photo}} + \beta_1 L_{\text{reg}} + \beta_2 L_{\text{smooth}}$, where β_1 balances how much we spread the features on the unit-sphere and is set as $\beta_1 = 1$, and we choose $\beta_2 = 0.1$.

IV. Experiments

A. Datasets

Throughout our experiments, we train on the RobotCar Dataset [17] and test on both the Robotcar [17] and MS2 [37] datasets. Details of each are given below.

Oxford RobotCar: The Oxford RobotCar Dataset [17] is an autonomous driving dataset collected on the same route over a year in Oxford, UK. We follow the data splits proposed in [23] to exclude the geographical overlaps between training and test splits. We use the six sequences from the traverse on 2014-12-16-18-44-24 for our experiments, providing 19612 images for training, 4559 images for validation, and 709 images for testing. Ground truth depth data for evaluation is generated by projecting the LiDAR data from several nearby frames into the test frame using the available RTK [1] data.

[1]Quantitative results may change when other forms of pose data such as VO or INS is used to generate ground truth depth

Metric	*Method*	type	*Abs. Rel.* (↓)	*Sq. Rel.* (↓)	*RMSE* (↓)	*Log RMSE* (↓)	$\delta < 1.25(\uparrow)$	$\delta < 1.25^2(\uparrow)$	$\delta < 1.25^3(\uparrow)$
U	SGM [34]	-	0.237	3.453	8.393	0.358	0.689	0.862	0.924
	UniMatch-Stereo [25]	Sup	0.207	2.521	9.087	0.373	0.588	0.793	0.906
	IGEV-Stereo [10]	Sup	**0.147**	**1.655**	7.092	0.312	**0.782**	0.888	0.934
	Sharma et al. [29]	Sup	0.225	1.728	**6.489**	0.278	0.669	**0.920**	**0.963**
	Ours	Self-Sup	0.177	1.970	7.077	**0.274**	0.744	0.900	0.951
W	SGM [34]	-	0.246	3.711	9.313	0.374	0.630	0.825	0.900
	Unimatch-Stereo [25]	Sup	0.278	4.379	10.237	0.426	0.422	0.660	0.828
	IGEV-Stereo [10]	Sup	**0.184**	2.649	7.433	0.327	**0.703**	0.830	0.894
	Sharma et al. [29]	Sup	0.229	**2.113**	**6.750**	**0.284**	0.639	**0.892**	**0.945**
	Ours	Self-Sup	0.192	2.427	7.100	**0.275**	0.703	0.870	0.931

TABLE I: Quantitative evaluation of our proposed method against the SOTA. This evaluation is carried out with a maximum depth range of 50 meters. U refers to the unweighted metric, and W to the proposed weighted metric. In the type-column, "sup" refers to supervised training, and "self-sup" refers to self-supervised training. **Bold** shows the best performance and underline refers to second best results.

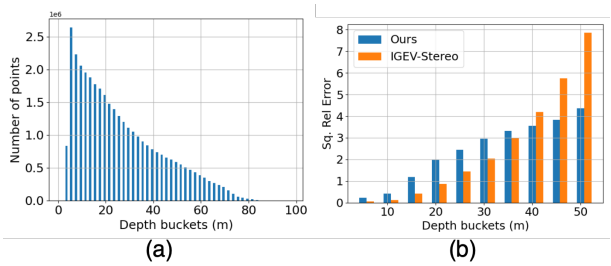

Fig. 4: The visualization of (a) the ground truth depth distribution of the Robotcar test split, and (b) square relative error calculated at different depth-bins using the proposed weighted metric.

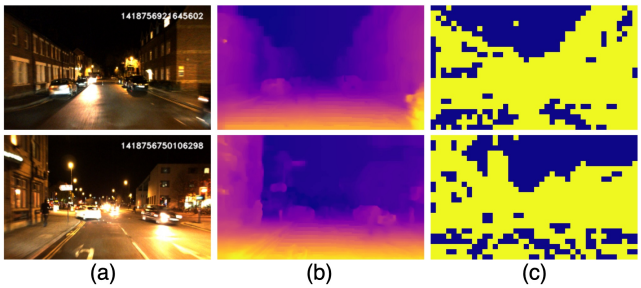

Fig. 5: Visualization of the estimated masks in (c), with their input-images (Left camera) in (a), and the estimated disparity-maps in (b).

Multi-Spectral Stereo (MS2) Dataset: The MS2 dataset [37] contains 184K pairs taken from multi-spectral sensors on a vehicle in Daejeon, South Korea. The sequences include various lighting, weather, and traffic conditions. Following the evaluation split proposed in [37], we use the *Road*, *City*, and *Campus* nighttime sequences, further sub-sampling them with 5m distance between consecutive test images. This gave 1,470 pairs for evaluation. We use the (filtered) ground truth depth data released with the dataset for the evaluation.

B. Training details

The framework is trained using the Robotcar dataset for 20 epochs with an input image resolution of 192×320. We used a batch size of 8 and the Adam [38] optimizer with a constant learning rate of $1e-4$.

C. Baselines:

To the best of our knowledge, there is no self-supervised system that estimates depth from night-time stereo-images. We therefore compare our method with a classical method, Semi-Global Matching (SGM) [34], and 3 state-of-the-art supervised methods: UniMatch-Stereo[2] [25], IGEV-Stereo [10], and Sharma et al. [29]. Note that these methods are trained end-to-end only for the purpose of stereo-matching, with large amounts of ground truth data. Also, we found that the disparity estimation of Unimatch-Stereo and IGEV-Stereo drops drastically when tested at the same resolution as ours. Therefore, we use $2\times$ more resolution while reporting their results.

[2]We used their in-the-wild use stereo-matching with refinement model from GitHub during the evaluation

D. Depth Evaluation

The depth-estimation performance is evaluated qualitatively and quantitatively on both the Robotcar [17] and MS2 [37] datasets. Quantitative evaluation is carried out using the metrics proposed in [39]. For every test image, existing methods compute their metrics as the mean of all valid pixels up to a given depth range (50m in our evaluation), usually only with a sparse set of LiDAR points. Taking the overall mean is sensible when pixels are uniformly distributed over the depth range, however, this is not the case for the RobotCar dataset [17] as shown in Fig 4(a). The same effect is observed for most autonomous driving datasets, including the MS2 dataset [37]. This is due to the fact that the majority of pixels in an image are occupied by points that are very close to the camera. Evaluating performance on this kind of data skews the evaluation, as the performance on nearby points outweighs the performance on the far-away points. To mitigate this, we propose the use of depth bins splitting the given depth range into M bins. Letting x be a metric, unweighted (U) and weighted (W) evaluation metrics for a given test image can be written as:

$$\text{x}_{\text{unweighted}} = \frac{1}{Z}\sum_{i=1}^{Z}\text{x}_i, \quad \text{x}_{\text{weighted}} = \frac{1}{M}\sum_{N=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\text{x}_{ij}, \quad (8)$$

where Z is the total number of valid pixels, M is the total number of bins and N_i is the total number of valid pixels in the i-th depth-bin. We set $M = 10$, i.e., each bin covers 5m depth. We reported the numbers from both metrics, and we did not use eigen or garg crops [4] during our evaluation.

RobotCar: We give a qualitative comparison depicted in Fig 3. Our method is able to extract crisp details, even in poorly illuminated regions. The effect of the masking

Metric	*Method*	*Abs. Rel.*	*RMSE*	$\delta < 1.25$	$\delta < 1.25^3$
U	SGM [34]	0.185	6.106	0.773	0.968
	Unimatch-Stereo [25]	0.095	3.396	0.910	0.994
	IGEV-Stereo [10]	0.099	3.918	0.894	0.985
	Sharma et al. [29]	0.193	5.077	0.713	0.990
	Ours	0.182	5.838	0.740	0.977
W	SGM [34]	0.183	6.482	0.764	0.969
	Unimatch-Stereo [25]	0.112	4.182	0.860	0.991
	IGEV-Stereo [10]	0.125	4.977	0.815	0.976
	Sharma et al. [29]	0.180	5.489	0.733	0.987
	Ours	0.180	6.162	0.716	0.978

TABLE II: Quantitative evaluation of our proposed method (trained on RobotCar) against the SOTA on MS2 Dataset

Method	*Abs. Rel.* (↓)	*RMSE* (↓)	$\delta < 1.25$(↑)	$\delta < 1.25^3$(↑)
Base model	0.201	7.734	0.724	0.943
w/ Mask	0.214	7.644	0.712	0.948
w/ Mask + Reg	0.188	7.409	0.744	0.946
w/ DINO-V2	0.204	7.871	0.711	0.944

TABLE III: Ablation study showing the importance of different modules in our system. This evaluation is carried out using the unweighted metrics with 50 meters as the maximum depth.

is clear when looking at the disparity estimated for the sky pixels. We also visualize the estimated masks and the respective filtered disparity maps in Fig. 5. Plausible masks are generated even for noisy low-illumination areas such as the sky. The original ground truth depth images are very sparse, making comparison hard. Hence, we dilated them for visualization. Quantitative results are given in Table I. Our method performs on par with the baselines in the majority of metrics across both variations despite being self-supervised. Fig 4(b) shows the mean squared relative error of IGEV-Stereo [10] and Ours for the test set. Per the unweighted metrics in Table I, IGEV-Stereo [10] performs better. One can see, however, the performance clearly degrades as the depth range increases compared to ours. This effect is much better captured in the weighted metric. Similar observations are made for other metrics as well. Lastly, note that our outstanding performance for challenging regions, such as the sky, is not taken into account during the quantitative evaluation due to the absence of the ground truth. As one can see in Fig. 3, all other methods estimate valid depth (brighter pixels) for the sky where they are supposed to be darker, as seen in the ground truth.

MS2: To further evaluate the generalisability of our method, we evaluated the model trained on RobotCar dataset using the test split of the MS2 dataset with the same 50m maximum depth range. Despite differences in geographic locations and lighting, and being trained on a relatively small dataset, our model still estimates very plausible disparity maps and pixel masks as shown in Fig 6. Quantitatively, our method performs better than SGM [34] and is comparable to other methods as shown in Table II.

E. Ablation Studies

We performed various ablation studies on RobotCar dataset to understand the impact of different modules. The results are shown in Table III. **Base model** uses DINO-V1 as the encoder, with stereo-matching and upsampling, trained using L_{photo} alone. **W/ mask** had features masked before stereo-matching, also trained with L_{photo} alone. **W/ mask + reg** used the regularization loss L_{reg} with L_{photo}. The improvement in both error and accuracy metrics explains the importance of the proposed masking and regularization loss for accurate depth estimation.

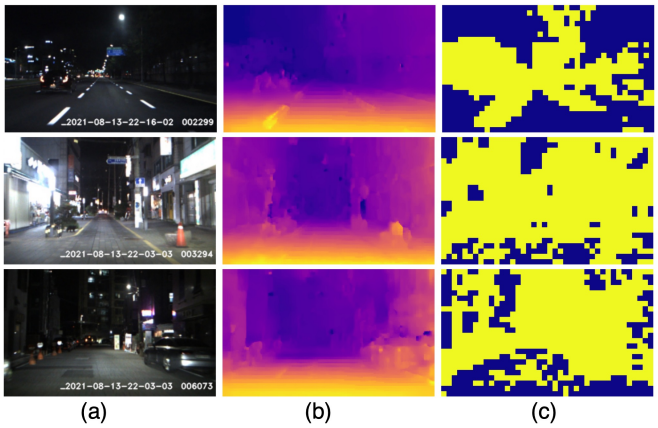

Fig. 6: Visualization of the estimated disparities in (b) and the masks in (c) with their input-images (Left camera) in (a) when tested on the unseen MS2 dataset.

F. Failure Cases:

Interestingly, training with a DINOv2 [14] encoder yielded a performance drop despite being pretrained on a larger amount of data than DINO [13] as shown in the last row of Table III. Also, the overexposed areas and lane markings create undesirable edges in the estimated disparity maps (can be observed in Fig. 5, Fig. 6). This can be a limitation of the commonly used edge-aware disparity smoothness loss L_{smooth}. We currently leave these issues for future investigation.

V. Future Work and Conclusions

We introduce an algorithm achieving precise self-supervised stereo depth estimation for nighttime conditions, leveraging visual foundation models. We present an efficient masking method and distance regularizer to enhance the accuracy of depth estimation, and novel, weighted evaluation metrics that provide more accurate evaluation given the non-uniform ground truth depth distributions. Our approach shows effective performance across a range of challenging scenarios and generalizes well to unseen datasets.

VI. Acknowledgments

This work was supported by AWS via the Oxford-Singapore Human-Machine Collaboration Programme, and EPSRC via ACE-OPS (EP/S030832/1). The authors would also like to thank the anonymous reviewers for their helpful comments.

References

[1] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[2] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[3] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[5] Y. Wang, Y. Liang, H. Xu, S. Jiao, and H. Yu, "Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation," *arXiv preprint arXiv:2309.00526*, 2023.

[6] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 560–15 569.

[7] K. Zhou, L. Hong, C. Chen, H. Xu, C. Ye, Q. Hu, and Z. Li, "Devnet: Self-supervised monocular depth learning via density volume construction," in *European Conference on Computer Vision*. Springer, 2022, pp. 125–142.

[8] Z. Feng, L. Yang, L. Jing, H. Wang, Y. Tian, and B. Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in *European Conference on Computer Vision*. Springer, 2022, pp. 228–244.

[9] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8071–8081.

[10] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.

[11] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, "Robust monocular depth estimation under challenging conditions," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[12] R. Garg, V. K. B.G., G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 740–756.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[15] I. Fang, H.-C. Wen, C.-L. Hsu, P.-C. Jen, P.-Y. Chen, Y.-S. Chen, *et al.*, "Es3net: Accurate and efficient edge-based self-supervised stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4471–4480.

[16] K. Saunders, G. Vogiatzis, and L. Manso, "Self-supervised monocular depth estimation: Let's talk about the weather," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[17] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, 2017.

[18] M. B. Vankadari, S. Garg, A. Majumder, S. Kumar, and A. Behera, "Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222133092

[19] C. Zhao, Y. Tang, and Q. Sun, "Unsupervised monocular depth estimation in highly complex environments," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, pp. 1237–1246, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236469578

[20] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[21] J. Spencer, R. Bowden, and S. Hadfield, "Defeat-net: General monocular depth via simultaneous unsupervised representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun 2020, pp. 14 390–14 401. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01441

[22] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[23] M. Vankadari, S. Golodetz, S. Garg, S. Shin, A. Markham, and N. Trigoni, "When the sun goes down: Repairing photometric losses for all-day depth estimation," 2022.

[24] Y. Zheng, C. Zhong, P. Li, H. ang Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang, and D. Zhao, "Steps: Joint self-supervised nighttime image enhancement and depth estimation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[25] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[27] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 263–16 272.

[28] G. Xu, Y. Wang, J. Cheng, J. Tang, and X. Yang, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[29] A. Sharma, L.-F. Cheong, L. Heng, and R. T. Tan, "Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 23–31.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[31] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep vit features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, vol. 2, no. 3, p. 4, 2021.

[32] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8922–8931.

[33] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947.

[34] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via semi-global matching on the gpu," *Procedia Computer Science*, vol. 80, pp. 143–153, 2016.

[35] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou, "Spreading vectors for similarity search," *arXiv preprint arXiv:1806.03198*, 2018.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[37] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1043–1053.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[39] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.