

# ADAADepth: Adapting Data Augmentation and Attention for Self-Supervised Monocular Depth Estimation

Vinay Kaushik<sup>1</sup>, Kartik Jindgar<sup>2</sup> and Brejesh Lall<sup>3</sup>

*Abstract*—Self-supervised learning of depth has been a highly studied topic of research as it alleviates the requirement of having ground truth annotations for predicting depth. Depth is learnt as an intermediate solution to the task of view synthesis, utilising warped photometric consistency. Although it gives good results when trained using stereo data, the predicted depth is still sensitive to noise, illumination changes and specular reflections. Also, occlusion can be tackled better by learning depth from a single camera. We propose ADAADepth, utilising depth augmentation as depth supervision for learning accurate and robust depth. We propose a relational self-attention module that learns rich contextual features and further enhances depth results. We also optimize the auto-masking strategy across all losses by enforcing  $L_1$  regularisation over mask. Our novel progressive training strategy first learns depth at a lower resolution and then progresses to the original resolution with slight training. We utilise a ResNet18 encoder, learning features for prediction of both depth and pose. We evaluate our predicted depth on the standard KITTI driving dataset and achieve state-of-the-art results for monocular depth estimation whilst having significantly lower number of trainable parameters in our deep learning framework. We also evaluate our model on Make3D dataset showing better generalization than other methods.

## I. INTRODUCTION

Depth from a single image has been of utmost importance in computer vision community with the advent of deep learning. Depth prediction provides solutions for several applications including smart mobility [1], smartphone AR [2], 3D zooming [3], face anti-spoofing [4], image dehazing [5], etc. Humans are able to perceive depth in the visible world by utilising cues like occlusion, texture differences, relative scale of neighbouring objects, lighting and shading variations along with object semantics.

Multi-view and stereo methods are computationally expensive and have high memory overheads. Depth from single image drastically reduces these complexities and is favourable for real-time systems. Deep learning provides the tools to predict depth from a single image by transforming the task into a learning problem[6], [7], given the ground truth depth annotations. However, capturing vast amount of ground truth data in different scenarios is a formidable task.

<sup>1</sup>Vinay Kaushik and <sup>3</sup>Brejesh Lall are with the Department of Electrical Engineering, IIT Delhi, India.<sup>2</sup>Kartik Jindgar is with Manipal University, Jaipur. {Vinay.Kaushik,brejesh}@ee.iitd.ac.in k.jindgar@gmail.com

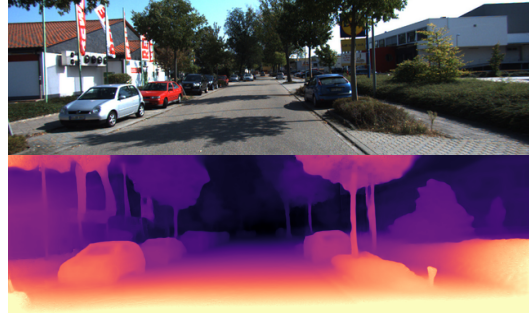


Fig. 1. Depth predicted from our network

Self-supervision for computing depth eliminates this limitation by utilising photometric warp for learning depth[8], [9].

Learning from a monocular sequence is challenging due to scale ambiguity and unknown camera pose. Thus, there's an explicit need to compute camera egomotion[10], [11]. The necessity of joint learning for depth and egomotion means that the quality of depth is highly dependent on the correctness of camera pose. Also, static scene assumption in self-supervised learning paradigm leads to holes and aberrations in pixels belonging to a moving object in the scene. Occlusions at image boundaries makes it difficult to learn depth near boundary regions (bottom image region in a forward moving camera). Although there have been innovations in deep learning architectures[12], [13], loss functions[8], [10], masking strategies [14], [15], [16], [10], there is still a huge scope of improvement to bridge the gap between self-supervised and supervised methods. This paper aims to reduce that gap by incorporating novel relational self-attention and data augmentation utilising learnt depth.

We utilise a ResNet18 encoder for our ablation and quantitative analysis and show substantial improvements in learning depth. Our main contributions are as follows:

- We introduce data augmentation as a supervisory loss, improving depth at occluded edges and image boundaries while making the model more robust to illumination changes and image noise.
- Our self attention module learns optimal feature relations that drastically improve our depth prediction.
- We show that our novel progressive learning strat-

egy learns robust scale-invariant features leading to significant improvements in depth prediction while saving huge computational overhead of training a high resolution model from scratch.

Our network can predict state-of-the-art depth while having significantly lower number of parameters.

## II. RELATED WORK

Depth estimation from a single colored image is a challenging task due to the obscure nature of this problem. A single depth map can be mapped to innumerable possible colored images. Over the last few years learning models have proven to be successful in effectively learning and exploiting this relationship between color images and their corresponding depths.

### A. Supervised Depth Estimation

Eigen[6] was one of the first ones to explore end to end supervised learning of depth from a single colored image using a multi-scale deep neural network. He trained a model to learn directly from raw colored images and their corresponding depths. Several different approaches have been proposed since then. [17] introduced a patch-based model which generated super-pixels to combine local information. [18] used a non-parametric scene sampling pipeline where candidate images from the dataset were matched with target image using high level image and optical flow features.

Acquiring large amounts of ground truth data in the real world is a challenge and this creates large overheads, both in terms of cost and time as it requires use of lasers like LIDAR. This is the reason that supervised models, despite their superior performance, are not universally applicable. As a result several works have turned to unsupervised or weakly supervised models and use of synthetically generated data.

[19] used real world size of objects to compute depth maps. They used geometric relations to calculate depth maps which were then refined using energy function optimization. [20] used relative depth annotation instead of actual ground truth depth data. They learned to estimate metric depth using relative depth annotations. These works however, still require supervision signals in the form of additional set of depths or other annotations. Generating large amounts of realistic synthetic data that includes several types of variations found in the real world is not a superficial task as well.

### B. Self-supervised Depth Estimation

A more promising substitute for supervised and weakly supervised models is the self supervised approach. Either stereo or monocular inputs are used for these models. Depth, hallucinated by the model, is used to warp the source image into the target frame. The difference between the reconstructed and reference frame is penalised and added as a reconstruction loss to provide a supervisory signal to the model.

1) *Self-supervised Stereo Training*: For self-supervised stereo depth estimation, synchronized stereo image pairs are fed into the model. The model estimates disparity or inverse depth between the two frames and in the process learns to predict the depth of single images. Garg [9] presented an approach that reconstructed left images by inverse warping the right images using the predicted depth and known camera extrinsics. The photometric error between the reconstructed image and the original images was used to train the encoder. [8] incorporated a left-right consistency term amongst other losses. [21] utilised stereo matching to provide sparse supervision in form of depth hints to predict depth. Since then several works have refined self-supervised stereo training of depth. However, some problems still plague stereo estimation. Occlusion drastically affects stereo frames due to the fixed baseline between cameras. Also, wide baseline stereo data might not be available in all real world scenarios e.g. mobile phone camera.

2) *Self-supervised Monocular Training*: Self supervised monocular depth estimation is naturally unimpeded by a lot of these restraints. In monocular training, temporally consecutive frames are fed into the model instead of stereo pairs. The model has to also learn pose in addition to depth due to the unknown and varying baseline. Zhou et al[15] provided one of the initial works in this domain where they used an end to end learning approach with supervision provided by view synthesis. They used two separate networks for learning depth and pose. [10] used a minimum reprojection loss to handle occlusion and prevent the network from learning erroneously from occluded pixels. They computed an automasking framework to prevent learning depth from stationary pixels (static camera). Several works have also incorporated optical flow estimation in their pipelines and tried to exploit relationships between depth, pose and optical flow to achieve more accurate results. [22] proposed a cross-task consistency loss, [23] performed motion segmentation, [11] decomposed motion into rigid and non rigid components and used a residual flow learning module to handle non rigid cases, [24] used losses that ensured 3D structural consistency and enforced geometric constraints,  $S^3$ Net[25] fuses semantic constraints into depth framework, Shu[26] introduces a feature metric loss computed from FeatureNet to improve depth. Huynh [27] formulates a depth attention volume for guiding monocular depth. Xian [28] constructs a structure guided ranking loss for self-supervised learning of depth.

3) *Self-Attention in Deep learning*: Wang[29] introduced self-attention as a non-local operation by correlating response at a spatial position as weighted sum of features at all positions. Building on the same framework, Zhang[30] utilised self-attention in GANs for image generation tasks. Fu[31] formulated a dual

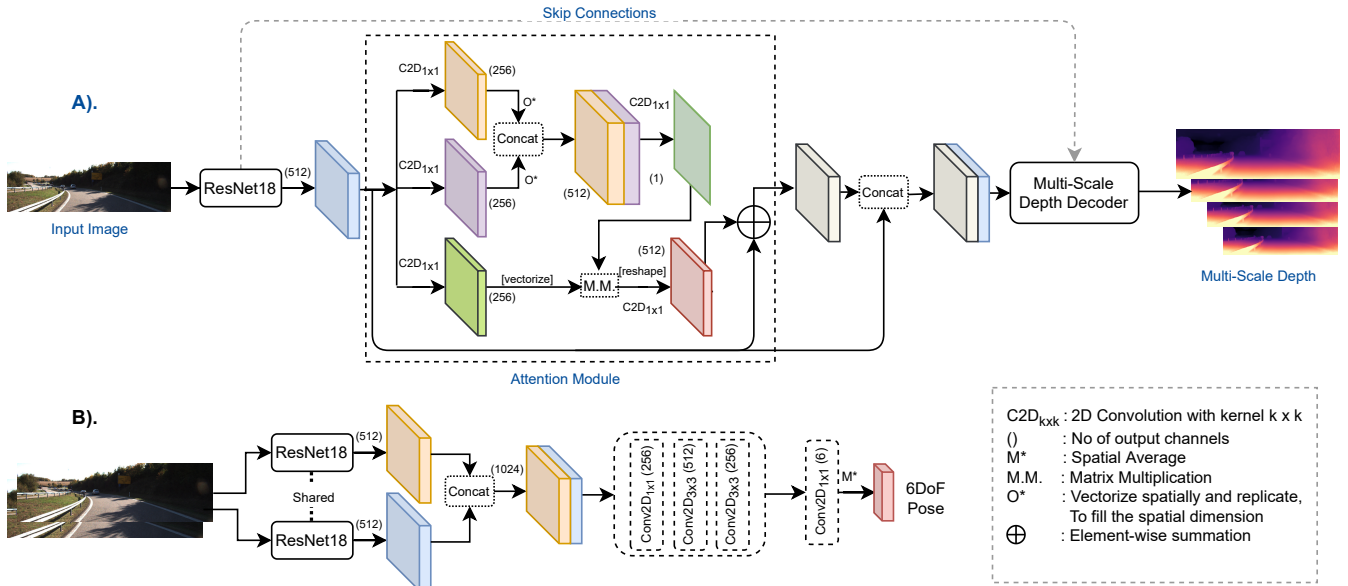


Fig. 2. Architecture diagram

attention network for semantic segmentation that unlike traditional works which focus on multi-scale feature fusion, focused on self-attention to integrate local features with their global dependencies adaptively. Since then, self-attention has already been utilised in medical applications[32], video recognition, semantic segmentation[33], object detection[34] and video understanding[35]. Unlike a convolutional operation, self-attention provides the ability to learn features and dependencies in non-contiguous regions making it an important building block of deep learning frameworks. We formulate a relational self-attention mechanism, learning from relational reasoning[36] to embed better context in the self-attention framework. Our model achieves better accuracy without learning for optical flow or motion segmentation by encompassing robust geometric constraints, a relational self-attention framework and utilising augmentation for depth supervision along with our progressive learning strategy.

### III. METHODOLOGY

Self-supervised learning utilising photometric consistency has become the de-facto standard for learning depth without ground truth data. The problem of depth prediction is transformed into a problem of view synthesis, where the goal is to use predicted depth of the input image to find per pixel correspondence for reconstructing the input image from another view. By solving for view synthesis, we can train our network to predict depth. We utilise the same approach while incorporating multiple novel data-driven and geometric constraints. Here, we describe a model that jointly learns to predict depth and pose. The network comprises of a shared VGG encoder, depth decoder and pose sub-network. The encoder takes an RGB image as input and extracts it's features that are utilised by both depth decoder and pose sub-network. For training our

network we use a 3 frame sequence, where the middle frame is target image  $I_t$  and the remaining two frames are source images  $I_{S1}, I_{S2}$ . We predict target depth  $D_t$ , source depth  $D_{S1}$  and  $D_{S2}$ , pose  $P_1$  and pose  $P_2$ , where pose  $P_i$  is the 6DoF transformation from target to the  $i^{th}$  source.

We first outline our training model architecture along with the necessary notations required in formulating losses for training our model then describe in detail the geometric constraints of depth prediction. We describe in detail the augmentation loss framework and the self-attention module for and then delineate each loss along with it's significance in our algorithm.

#### A. Training Model Architecture

As shown in 2, our model consists of a ResNet18 encoder [8] taking an RGB image as input. Features extracted from the source  $I_s$  and target  $I_t$  images are concatenated and fed to the pose sub-network to compute the 6x1 egomotion vector. Our depth decoder takes in feature of the target image  $I_t$  to predict depth  $D_t$  of that image. The encoder-decoder framework is similar to the U-Net architecture introduced by [38], that enables us to encapsulate both global as well as local features while predicting depth at 4 scales. The relational attention module takes input as encoder's features and generates attention maps that are concatenated to the original features and fed to the depth decoder as in Figure2. The pose network comprises of 4 convolutional layers to get a 6x1 output vector[39] containing rotation(3x1) and translation (3x1) information as shown in Figure2. We use Sigmoid activation at depth outputs and ELU activation everywhere else[8]. The target image  $I_t$  and it's corresponding predicted depth  $D_t$  is then processed by the augmentation pipeline to get transformed augmented image  $I_{aug}$  and true augmented depth  $D_{aug}^{true}$ .  $I_{aug}$  is then fed to

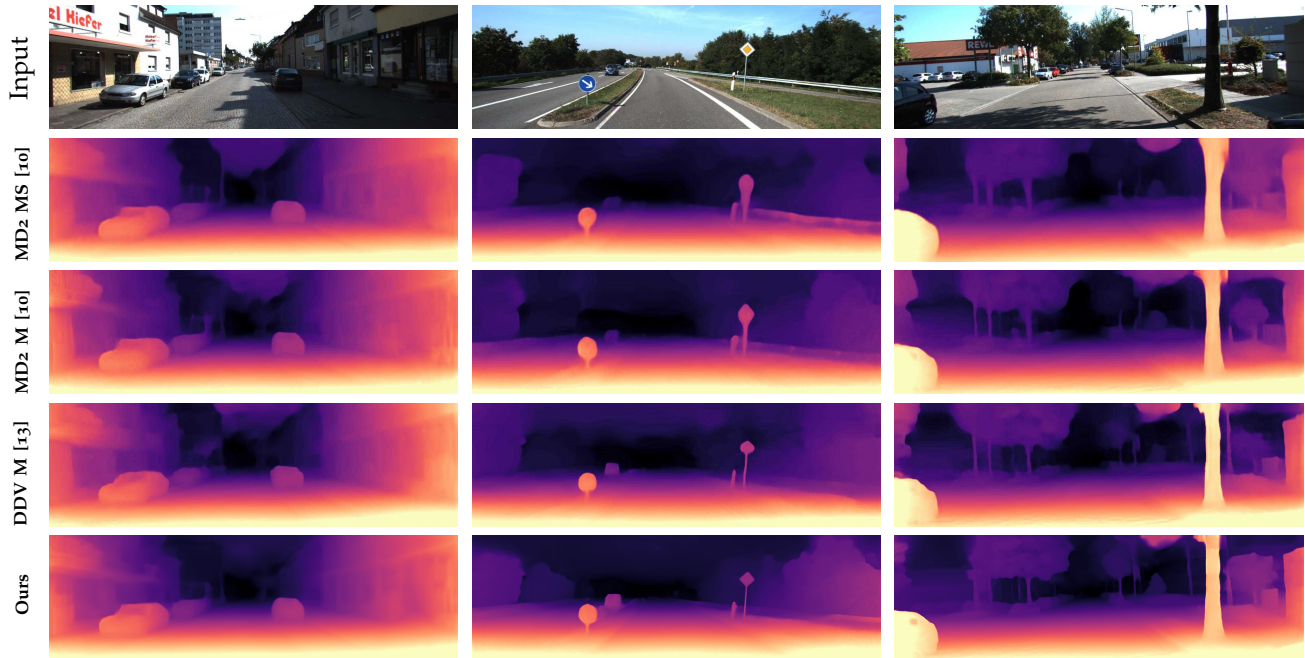


Fig. 3. **Qualitative results on the KITTI Eigen split [37] test set.** Our models perform better on thinner objects such as trees, signs and bollards, as well as being better at delineating difficult object boundaries. The depth of far objects including sky is further improved.

the network to predict output augmented depth  $D_{aug}^{out}$ . The model returns  $D_t, D_s, P, D_{aug}^{true}, D_{aug}^{out}$  for computing training losses. Target depth  $D_t$  warps source image  $I_s$  to compute synthetic image  $I_t^s$  by using bi-linear sampling for sampling source images. While testing, the network can simply compute  $D_t$  from  $I_t$ .

### B. Constraints for depth prediction

In this section, we describe the formulation of various loss functions used in our network for self-supervised learning of depth and pose.

1) *Minimum Photometric Loss*: As described by [10], this loss is a slight variation from the normal photometric loss. Instead of taking per pixel average of photometric loss for all sources, we compute minimum of photometric loss for all sources. This successfully tackles scenarios where a target pixel is visible in one source image but not visible in the other source image due to occlusion and only back-propagates the minimum error, thereby ignoring the erroneous one.

$$L_p = \min_{s \in \{1,2\}} pe(I_t, I_t^s) \quad (1)$$

Here, photometric error  $pe$  is defined by a weighted combination of L1 loss and Structural Similarity (SSIM) [46], similar to [10][24][23].

$$pe(I_t, I_s) = \frac{\alpha}{2}(1 - SSIM(I_t, I_s)) + (1 - \alpha)||I_t - I_s|| \quad (2)$$

,where  $\alpha = 0.85$  Similar to [10], we apply a per pixel binary mask  $\lambda$  to the computed losses. The mask  $\lambda$  is generated by comparing the photometric error between source and target frames with that between

the synthesised source and target frames.

$$\lambda = [\min_s pe(I_t, I_t^s) < \min_s pe(I_t, I_s)] \quad (3)$$

This eliminates static pixels from corrupting the loss and the network skips learning depth altogether if the camera isn't moving. We observe that although this improves depth prediction drastically, it leads to random white noise around static regions and makes the learning of depth more sensitive to noisy images. This happens because the mask doesn't consider neighbouring pixels while comparing photometric errors and simply takes a threshold of per pixel values. To alleviate this problem, we enforce a L1 regularisation over inverse of  $\lambda$ , thereby motivating the mask to be positive for those sparse number of pixels.

$$L_r = |1 - \lambda| \quad (4)$$

We compute first order gradient smoothness loss  $L_s$ [8] over mean normalized inverse depth  $d_t$ [39] to ensure that the predicted depth is locally smooth as well as consistent in textured regions.

$$L_s = |\delta_x d_t| e^{-|\delta_x I_t|} + |\delta_y d_t| e^{-|\delta_y I_t|} \quad (5)$$

where  $\delta_x$  and  $\delta_y$  are gradients in horizontal and vertical direction respectively.

2) *Data augmentation for depth supervision*: Several works have utilised data augmentation[10], [47], [13] in their deep learning pipeline to make their networks more robust to challenging scenarios and invariant to changes in noise, brightness and contrast that are common in the real world. Traditionally, pipelines performed data augmentation at the data loading stage

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al[15]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang et al[40]	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al[41]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Geonet[11]	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO[39]	0.151	0.125	5.583	0.228	0.81	0.936	0.974
LEGO[42]	0.162	1.352	6.276	0.252	-	-	-
DF-Net[43]	0.150	0.124	5.507	0.223	0.806	0.933	0.973
Ranjan et al[23]	0.148	0.149	5.464	0.226	0.815	0.935	0.973
EPC++[16]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2Depth(M)[44]	0.141	1.025	5.290	0.215	0.816	0.945	0.979
Monodepth2[10]	0.115	0.882	4.701	0.190	0.879	0.961	0.982
DDV[13]	0.106	0.861	4.699	0.185	0.889	0.962	0.982
<b>Proposed Approach</b>	<b>0.108</b>	<b>0.745</b>	<b>4.436</b>	<b>0.181</b>	<b>0.889</b>	<b>0.966</b>	<b>0.984</b>

TABLE I

SELF-SUPERVISED DEPTH PREDICTION RESULTS ON KITTI DATASET [45] TRAINED AT 1024 x 384 RESOLUTION. RESULTS ON EIGEN SPLIT [6] FOR DEPTHS AT CAP 80M, AS DESCRIBED IN [6].

and the augmented data was fed into the network for training. We utilise augmentation for generating augmented inputs and outputs that are used to train the network in a semi-supervised manner. We incorporate an augmentation loss, in the form of a depth supervision, that improves the predicted depth. While training, in the first forward pass, the network takes  $I_t$  as input, giving depth  $D_t$  as network output. We pass the pair  $(I_t, D_t)$  to the augmentation pipeline, applying identical random image cropping, flipping, skewing and scaling and affine transformations to both. Additionally, we perform random changes to brightness, jitter, gamma and saturation to the input image  $I_t$ . We also add random gaussian noise to  $I_t$ . The augmentation pipeline returns augmented image  $I_{aug}$  and true augmented depth  $D_{aug}^{true}$ . While training, in the second forward pass,  $I_{aug}$  is fed to the network generating augmented predicted depth  $D_{aug}^{out}$ . Augmented depth maps generated in first pass serve as ground truth for depth maps generated in the second pass. The augmentation loss minimises the difference between the output augmented depth and the true augmented depth, enforcing both depths to be consistent with each other.

$$L_a = \|D_{aug}^{true} - D_{aug}^{out}\|_1 \quad (6)$$

Due to camera egomotion, occlusion is present at certain image boundaries. Rescaling and crop transformations randomly remove boundary regions from the image, while ensuring that it's size remains the same. Thus, the boundaries of the augmented depth are more accurate due to lower probability of occlusion.

3) *Relational Self-Attention*: The relational self-attention block takes input the features  $X$  from the ResNet18 encoder and computes self-attention  $Y$  that is added as a residual connection to the input feature  $X$  to compute the output feature. The operation can be summarised as follows:

$$y_i = \frac{1}{N} \sum_{\forall j} W_f^T[\theta(x_i), \phi(x_j)]g(x_j) \quad (7)$$

Here,  $W_f$  is a weight factor that projects the concatenated vector to the scalar by performing a convolution

with single output channel,  $[\cdot, \cdot]$  denotes concatenation,  $N$  defines the number of positions in  $X$ . Also, the functions  $\theta$ ,  $\phi$  and  $g$  are defined by  $1 \times 1$  2D convolution operations as shown in Figure 2. The input  $X$  generates the projection, query, key and value embeddings as

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{W}_f \mathbf{X}, \\ g(\mathbf{X}) &= \mathbf{W}_g \mathbf{X}, \\ \theta(\mathbf{X}) &= \mathbf{W}_\theta \mathbf{X}, \\ \phi(\mathbf{X}) &= \mathbf{W}_\phi \mathbf{X}, \end{aligned} \quad (8)$$

where  $W_f, W_g, W_\theta, W_\phi \in \mathbb{R}^2$  are weight matrices to be learnt. The pairwise relation between query  $\theta$  and key  $\phi$  is projected by  $W_f$  and multiplied by value  $g(X)$  to compute our relational self-attention that is then element-wise added to input  $X$  to give output of our attention block.

The output is then concatenated with the encoder's features and utilised by the decoder to computed multi-scale depth.

4) *Final Training Loss*: We combine photometric and smoothness losses with our data augmentation loss along with  $L_1$  regularization over mask to obtain our final objective.

$$L_f = \alpha_p L_p + \alpha_s L_s + \alpha_a L_a + \alpha_r L_r \quad (9)$$

All our losses are computed per-pixel and averaged over entire image, scales and batch.

	Type	Abs Rel	Sq Rel	RMSE	$\log_{10}$
Karsch [48]	D	0.428	5.079	8.389	0.149
Liu [49]	D	0.475	6.562	10.05	0.165
Laina [50]	D	<b>0.204</b>	<b>1.840</b>	<b>5.683</b>	<b>0.084</b>
Monodepth [8]	S	0.544	10.94	11.760	0.193
Zhou [15]	M	0.383	5.321	10.470	0.478
DDVO [51]	M	0.387	4.720	8.090	0.204
Monodepth2 [10]	M	0.322	3.589	7.417	0.163
DDV [13]	M	0.297	2.902	7.013	0.158
<b>Proposed Approach</b>	<b>M</b>	<b>0.289</b>	<b>2.552</b>	<b>6.869</b>	<b>0.155</b>

TABLE II

**Make3D[52] results.** ALL SELF-SUPERVISED MONOCULAR (M) METHODS USE MEDIAN SCALING.

## IV. EXPERIMENTS AND RESULTS

This section introduces the dataset and describes the training details. We describe in detail various comparative qualitative and quantitative studies along with an ablation study undertaken for validation and show that our method surpasses all other existing related methods.

### A. Dataset

Our model was trained on KITTI 2015 dataset[45]. This dataset comprises videos captured by a camera mounted on a car moving through the German city of Karlsruhe and is widely recognized and often used for tasks like estimation of depth, optical flow and car’s egomotion. We used the Eigen test split[6] of this dataset and tested our model using the ground truth labels present in it. The test set consists of 697 images and it is ensured that frames that are similar to those present in the test set are removed from the training set. We also test our trained model on the 134 images in Make3D dataset[52].

### B. Parameter Settings

Similar to other self-supervised models[10][23][24], we use ImageNet weights for initialising our network and train our model using a single NVIDIA 2080Ti GPU. Three temporally consecutive images are fed into the model and Adam optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  is used. Initial learning rate is set to  $10^{-4}$  and batch size to 12. While optimizing our network we set weights of different loss terms to  $\alpha_p = 1.0, \alpha_s = 0.001, \alpha_a = 0.1$  and  $\alpha_r = 0.001$ . While preparing training data, static frames are removed from the dataset as proposed by Zhou et al[15]. Basic augmentation in the form of random cropping, color jittering, resizing and flipping is also performed as part of our data preparation pipeline. We train our network over two phases in a progressive manner. In the first phase, images of 640x192 resolution are fed into the network. After training it for 50 epochs, in the second phase, we freeze the pose encoder and feed the higher resolution 1080x384 images to the depth network and train our model for 5 epochs with batch size 2. Progressive training aids in further improvement and faster convergence in our depth prediction model at a higher resolution as shown in Table IV.

### C. Main Results

We compare our results with other recent models in Table I. These results show that our monocular model is able to comprehensively outperform all existing state of the art self-supervised monocular methods. Our model is even able to surpass methods that incorporate optical flow prediction into their pipeline[11], [24],[23], while having lower number training parameters. During evaluation, as common practice[6], we cap depth to 80m. Table IV shows the comparison with DDV[13]

and moonodepth2[10] trained using features from same ResNet18 encoder. Our method has better RMSE, Sq Rel, Abs Rel than other similar methods. This shows that our model performs significantly better in all metrics on Eigen split of KITTI 2015 dataset.

### D. Qualitative Analysis

Figure 3 displays qualitative improvements in our method over baseline Monodepth2(MD2)[10] and DDV[13]. Our algorithm retains structural details in objects like poles, sign boards and trees while learning smooth depth over entire scene. We also have the least noise in disparity values of the infinitely distant sky.

### E. Make3D

Table II shows results of our model that is trained on KITTI dataset and tested on the Make3D dataset[52]. We use the crop defined by [15] and apply depth median scaling for fair comparison. The table shows our method’s superior performance than other self-supervised methods while bridging the gap between supervised ones[50].

### F. Ablation Study of Losses

We also undertake exhaustive quantitative comparison of all the losses to analyze the impact of each loss component. Table III shows different combinations of losses applied and the corresponding results achieved by our model. It is evident from the table, that with just the inclusion of augmentation loss, we get significant gains over the baseline. The augmentation loss makes the model more robust to variation in brightness, contrast and image noise. Supervising the network in form of augmentation loss utilising the true augmented depth drastically improves the depth prediction at occluded regions including image boundaries. Similarly, appearance and color based transforms help the network in learning to predict more consistent and robust depth which is less affected by noise and illumination changes. We observe that adding reflection padding to our network doesn’t have noticeable effect on the depth prediction results as the augmentation loss already improves depth at image boundaries. We also observe that attention improves *AbsRel* more than Augmentation and the combination of both losses have a multi-fold improvement over baseline. We also tried replacing skip connections by attention module but the added complexity was drastically high with no significant improvement in depth prediction. As depicted in Table V, concatenating encoder feature to the attention gave better results than simply passing attention block’s output to the decoder. This tells us that attention though significant isn’t sufficient to achieve optimal result. Also, increasing the augmentation loss weight  $\alpha_a > 0.1$  induced texture copy artifacts and decreasing it led to minimal improvement in accuracy.

Aug Loss	Attention	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
No	No	0.115	0.919	4.854	0.194	0.877	0.958	0.980
Yes	No	0.113	0.837	4.726	0.189	0.879	0.961	0.982
No	Yes	0.111	0.827	4.742	0.189	0.878	0.960	0.982
<b>Yes</b>	<b>Yes</b>	<b>0.111</b>	<b>0.817</b>	<b>4.685</b>	<b>0.188</b>	<b>0.883</b>	<b>0.961</b>	<b>0.982</b>

TABLE III

ABLATION STUDY FOR DEPTH PREDICTION AT 640X192 IMAGE RESOLUTION USING RESNET18 ENCODER ON EIGEN SPLIT[6]. WE OBSERVE THAT THE COMBINATION OF AUGMENTATION LOSS AND OUR ATTENTION FRAMEWORK GIVES US THE BEST DEPTH RESULTS.

Method	Backbone	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [10]	ResNet18	0.115	0.902	4.847	0.193	0.877	0.960	0.981
DDV [13]	ResNet18	0.111	0.941	4.817	0.189	0.885	0.961	0.981
<b>Proposed Approach</b>	<b>ResNet18</b>	<b>0.111</b>	<b>0.817</b>	<b>4.685</b>	<b>0.188</b>	<b>0.883</b>	<b>0.961</b>	<b>0.982</b>
<b>Proposed Approach 1024x384</b>	<b>ResNet18</b>	<b>0.108</b>	<b>0.745</b>	<b>4.436</b>	<b>0.181</b>	<b>0.889</b>	<b>0.966</b>	<b>0.984</b>

TABLE IV

COMPARING OUR METHOD AT 640X192 RESOLUTION WITH OTHER METHODS UTILISING SAME NETWORK BACKBONE.

Depth Decoder's Input	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Attention + No feature concat	0.113	0.879	4.777	0.190	0.880	0.959	0.981
Attention in all skip connections + Feature concat	0.112	0.856	4.699	0.188	0.880	0.961	0.982
Attention + feature concat in all skip connections	0.112	0.866	4.742	0.189	0.879	0.960	0.982
<b>Attention + Feature concat</b>	<b>0.111</b>	<b>0.817</b>	<b>4.685</b>	<b>0.188</b>	<b>0.883</b>	<b>0.961</b>	<b>0.982</b>

TABLE V

COMPARING OUR METHOD AT 640X192 RESOLUTION WITH MULTIPLE VARIATIONS OF SELF-ATTENTION FEATURES. AUGMENTATION LOSS IS APPLIED DURING TRAINING. FEATURE IS THE RESNET18 ENCODER'S OUTPUT FEATURE.

As observed by Monodepth2[10], it is necessary to handle static frames, i.e. frames where either the camera is stationary or regions such as sky that does not change across consecutive frames. Automasking masks out these areas and prevents the model from learning erroneous depth. To enforce mask to be consistent and smooth, and eliminating noisy values, we apply L1 regularisation over inverse of our mask. This slightly increases the number pixels to be evaluated and reduces artifacts over still regions like sky. Our methods predicts superior results both qualitatively and quantitatively when compared to other self-supervised monocular depth prediction methods.

## V. CONCLUSION

We propose a self-supervised model which utilises relational self-attention for jointly learning depth and camera egomotion. The model is able to predict accurate and sharp depth estimates by incorporating data augmentation as depth supervision. Our algorithm predicts state-of-the-art depth on the KITTI benchmark[45]. In future, we shall utilise optical flow for motion segmentation, pretrained models and semantic cues for further strengthening depth of moving objects. Architectural innovations in deep learning such as vision transformers along with cues like optical flow and semantic information present in the scene can further optimize robustness and consistency in predicting depth.

## REFERENCES

- [1] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Bouteau, J.-Y. Ertaud, and X. Savatier, "Deep learning for real-time 3d multi-object detection, localisation, and tracking: Application to smart mobility," *Sensors (Basel, Switzerland)*, vol. 20, 2020.
- [2] J. P. C. Valentin, A. Kowdle, J. T. Barron, N. Wadhwa, M. Dzitsiuk, M. Schoenberg, V. Verma, A. Csaszar, E. Turner, I. Dryanovski, J. Afonso, J. Pascoal, K. Tsotsos, M. Leung, M. Schmidt, O. G. Guleryuz, S. Khamis, V. Tankovich, S. R. Fanello, S. Izadi, and C. Rhemann, "Depth from motion for smartphone ar," *ACM Transactions on Graphics (TOG)*, vol. 37, pp. 1 – 19, 2018.
- [3] J. L. G. Bello and M. Kim, "Deep 3d-zoom net: Unsupervised learning of photo-realistic 3d-zoom," *ArXiv*, vol. abs/1909.09349, 2019.
- [4] Y. Liu, Y. Tai, J.-L. Li, S. Ding, C. Wang, F. Huang, D. Li, W. Qi, and R. Ji, "Aurora guard: Real-time face anti-spoofing via light reflection," *ArXiv*, vol. abs/1902.10311, 2019.
- [5] Y.-J. Kim and C. Yim, "Image dehaze method using depth map estimation network based on atmospheric scattering model," *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, pp. 1–3, 2020.
- [6] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.
- [7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.
- [8] 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.
- [9] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*. Springer, 2016, pp. 740–756.
- [10] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in

- Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838.
- [11] Z. Yin and J. Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.
  - [12] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3d packing for self-supervised monocular depth estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.
  - [13] A. Johnston and G. Carneiro, “Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4756–4765.
  - [14] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, “Sfm-net: Learning of structure and motion from video,” *arXiv preprint arXiv:1704.07804*, 2017.
  - [15] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.
  - [16] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, “Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding,” *arXiv preprint arXiv:1810.06125*, 2018.
  - [17] A. Saxena, M. Sun, and A. Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, p. 824–840, May 2009.
  - [18] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, p. 2144–2158, Nov 2014.
  - [19] Y. Wu, S. Ying, and L. Zheng, “Size-to-depth: A new perspective for single image depth estimation,” 2018.
  - [20] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-image depth perception in the wild,” 2016.
  - [21] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, “Self-supervised monocular depth hints,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
  - [22] Y. Zou, Z. Luo, and J.-B. Huang, “Df-net: Unsupervised joint learning of depth and flow using cross-task consistency,” *Lecture Notes in Computer Science*, p. 38–55, 2018.
  - [23] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, “Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 12 240–12 249.
  - [24] Y. Chen, C. Schmid, and C. Sminchisescu, “Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.
  - [25] B. Cheng, I. S. Saggü, R. Shah, G. Bansal, and D. Bharadia, “s<sup>3</sup> net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data,” in *European Conference on Computer Vision*. Springer, 2020, pp. 52–69.
  - [26] C. Shu, K. Yu, Z. Duan, and K. Yang, “Feature-metric loss for self-supervised learning of depth and egomotion,” in *European Conference on Computer Vision*. Springer, 2020, pp. 572–588.
  - [27] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkilä, “Guiding monocular depth estimation using depth-attention volume,” in *European Conference on Computer Vision*. Springer, 2020, pp. 581–597.
  - [28] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, “Structure-guided ranking loss for single image depth prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 611–620.
  - [29] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
  - [30] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.
  - [31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.
  - [32] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
  - [33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnnet: Criss-cross attention for semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.
  - [34] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra r-cnn: Towards balanced learning for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 821–830.
  - [35] J. Lin, C. Gan, and S. Han, “Tsm: Temporal shift module for efficient video understanding,” *arXiv preprint arXiv:1706.01427*, 2017.
  - [36] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” *arXiv preprint arXiv:1706.01427*, 2017.
  - [37] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *ICCV*, 2015.
  - [38] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
  - [39] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey, “Learning depth from monocular videos using direct methods,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2022–2030.
  - [40] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, “Unsupervised learning of geometry with edge-aware depth-normal consistency,” 2017.
  - [41] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints,” 2018.
  - [42] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, “Lego: Learning edge with geometry all at once by watching videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 225–234.
  - [43] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, “Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.
  - [44] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 8001–8008, Jul 2019.
  - [45] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
  - [46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
  - [47] S. Pillai, R. Ambrus, and A. Gaidon, “Superdepth: Self-supervised, super-resolved monocular depth estimation,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9250–9256.
  - [48] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *PAMI*, 2014.
  - [49] M. Liu, M. Salzmann, and X. He, “Discrete-continuous depth estimation from a single image,” in *CVPR*, 2014.
  - [50] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *3DV*, 2016.
  - [51] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, “Learning depth from monocular videos using direct methods,” in *CVPR*, 2018.
  - [52] A. Saxena, M. Sun, and A. Ng, “Make3d: Learning 3d scene structure from a single still image,” *PAMI*, 2009.