2T-UNET: A Two-Tower UNet with Depth Clues for Robust Stereo Depth Estimation

Mansi Sharma[1,2,*] Rohit Choudhary[2,*] Rithvik Anil[2]
[1] Thapar Institute of Engineering & Technology, Patiala, Punjab, India
[2] Indian Institute of Technology Madras, India
mansi.sharma@thapar.edu, mansisharma@ee.iitm.ac.in, ee20s002@smail.iitm.ac.in,
rithvik.anil@gmail.com

Abstract

Stereo correspondence matching is an essential part of the multi-step stereo depth estimation process. This paper revisits the depth estimation problem, avoiding the explicit stereo-matching step using a simple two-tower convolutional neural network. The proposed algorithm is entitled as 2T-UNet. The idea behind 2T-UNet is to replace cost volume construction with twin convolution towers. These towers have an allowance for different weights between them. Additionally, the input for twin encoders in 2T-UNet are different compared to the existing stereo methods. Generally, a stereo network takes a right and left image pair as input to determine the scene geometry. However, in the 2T-UNet model, the right stereo image is taken as one input and the left stereo image along with its monocular depth clue information, is taken as the other input. Depth clues provide complementary suggestions that help enhance the quality of predicted scene geometry. The 2T-UNet surpasses state-of-the-art monocular and stereo depth estimation methods on the challenging Scene flow dataset, both quantitatively and qualitatively. The architecture performs incredibly well on complex natural scenes, highlighting its usefulness for various real-time applications. Pretrained weights and code will be made readily available.

1. Introduction

Depth estimation from RGB images has been a crucial research topic in computer vision and computer graphics. It plays a critical role in the domain of 3D reconstruction, virtual reality, augmented reality [1, 2], mapping and localization, image refocusing [3], image segmentation [4], robotics navigation [5], autonomous driving [6], novel view synthesis [7], free-viewpoint TV [8, 9] and 3D displays [10–12]. In general, stereo depth estimation is a multi-step process where the steps can be classified into four categories: 1) feature generation, 2) feature matching, 3) disparity estimation, and 4) refinement of the disparity. The feature generation step in a standard stereo depth estimation network compute features for both views using a shared-weight CNN. The next step, *i.e.*, feature matching, uses a cost volume to calculate how close the feature maps are at various levels of disparity. Patch correspondence of features extracted from the stereo pair is performed in this step [13–17]. For example, a patch in the left feature map centered around (i, j) is matched with multiple patches centered around $(i, j - d)$ in the right feature map, where d is the disparity level. A similarity rating is computed for each of the pairs at various disparity levels. Current methods consider different metrics for similarity computation, such as L_1 norm, L_2 norm, cosine similarity, correlation, etc [16, 18–20]. Some methods construct a 4D cost volume by simply appending features at different levels of disparity to the corresponding patch from the left feature map. This cost volume is regressed through several convolution layers to the estimated depth map and refined further to the desired output [15, 17, 21–23].

The differentiating attribute of our 2T-UNet architecture is that there is no need for explicit cost volume construction. The cost volume moves the features at several incremental levels of disparity to compute the similarity scores [15, 17, 19, 23]. Alternatively, the twin encoders in 2T-UNet implicitly capture the disparity or depth information by shifting its weights. Removing cost volume by transferring its functions to the encoders offers our network flexibility in deciding the range of disparity variation that should be considered for a given scene. Based on the training data, standard stereo-based depth prediction networks set the highest disparity value. However, our method allows the network to infer the highest disparity value on its own [24]. This helps us achieve robust depth predictions on tricky scenes with fine details and clear object boundaries in the depth map. Inspired by the work of Watson et al. [25], 2T-

*These authors contributed equally to this work

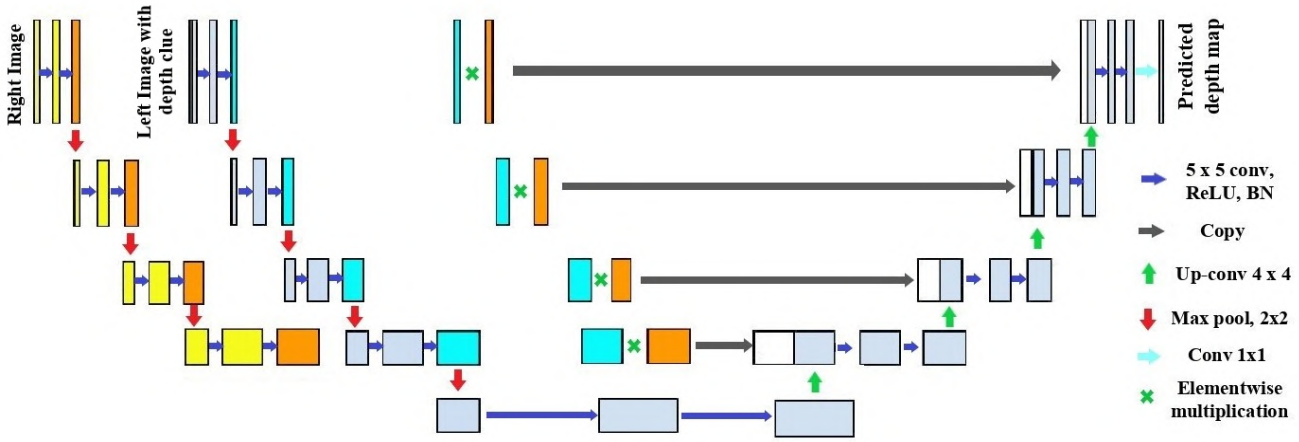


Figure 1. Overview of proposed 2T-UNet architecture: The network, inspired by UNet, employs a novel secondary encoder to eliminate the cost volume construction. A novel fusion strategy is introduced, where twin features are fused from the encoder before sending the information to the decoder.

UNet is provided with a monocular depth clue, a complementary depth suggestion from a pre-trained network. The depth hints are used in the domain of self-supervised learning, wherein the network predicts disparities of the scene using an RGB stereo pair alone and no depth labels. The stereo pair is reconstructed using predicted disparity in the self-supervised depth estimation approach. The reconstruction loss with the target views (stereo pair) is critical for better convergence of their network. The depth hints are computed using standard heuristically designed stereo methods [25].

In the proposed method, the depth clue is obtained using the left view of the stereo pair and passing it through an off-the-shelf monocular depth estimation algorithm. The depth clue is provided to the network by concatenating monocular depth information with the left stereo view. The secondary encoder takes input of the right view alone. The off-the-shelf algorithm is not trained on the same dataset as our network. Our method differs from Watson et al. [25] work in two crucial aspects:

Application domain: Our method is on stereo depth estimation domain. However, Watson et al. [25] depth hints work is in the domain of self-supervised monocular depth estimation.

Location of application in the network: Our method provides a coarse depth clue to the network, whereas the depth hint in Watson et al. [25] is used during the loss calculation.

The 2T-UNet is composed of following four novel ideas:

• Our architecture uses differently weighted [24] twin encoder. This is contrary to the conventional stereo algorithms that generally make use of encoders with shared weights followed by cost volume construction [15, 19, 21, 22].

• In addition, monocular depth clues are used to aid our network in predicting a high-quality depth map. We computed coarse geometric information of the scene using the algorithm of Ranftl et al. [26]. The depth clue is provided as complementary guidance to our network. The depth clues assist the proposed network to better preserve depth discontinuities and retain much better feature definitions in the depth map predictions. We also make sure that depth is learned from features derived from the stereo pair and not directly from the given depth clues. The coarse depth clues used in our proposed 2T-UNet network are depicted in Fig. 2 (d).

• Another distinguishing characteristic of 2T-UNet is that it uses a simple fusion strategy which is a crucial step in replacing the cost volume construction. The features are bitwise multiplied before being sent from encoder to decoder, *i.e.*, each corresponding element of the feature maps are multiplied before concatenating them in the decoder [24].

A standard stereo depth estimation network [15, 17, 21–23], on the other hand, performs the stereo matching step using cost volume. Standard methods achieve this by using an additional parameter d_{max}, which depicts the maximum extent to which the stereo feature must be shifted disparity-wise in order to perform stereo matching. This maximum disparity parameter is set explicitly in the model definition based on the variation in the dataset [19]. To enable our network to match features without explicitly specifying the disparity limit parameter, the feature matching process is re-imagined.

To make 2T-UNet network perform the feature matching without explicitly setting the disparity limit parameter, we re-imagine the feature matching step. Instead of shifting

the features within the cost volume, we use a functionally identical shift in weights at the encoders to capture the disparity. Thus, to capture disparity information, it is necessary that shift in the weight in either of the encoders influences the other. By multiplying the features elementwise, we are ensuring that the weights are updated in a dependent fashion. Backpropagation through the multiplication operation ensures that weight update happens in tandem between the two encoders[24]. Furthermore, in conventional stereo networks, the cost volume covers feature shift at all possible disparity values, and therefore contains large amounts of information. This large information bundle cannot be captured by shifting the weights in a single convolution block. Therefore, in the proposed 2T-UNet, we split workload among a sequence of weight shifts along with the twin encoders with periodic element-wise multiplication operations.

• Convolutional operations are an integral part of a CNN. This operation captures information about a point and its neighbours. For network to capture the disparity between views, we should extend the information extraction to both views. This awareness extension is achieved by multiplying the features from the encoders elementwise, *i.e.*, fusing them into one feature map, before sending them to the decoder. The fusion operation ensures that a feature point is aware of its neighbours and its disparity with the other view [24]. In addition, 2T-UNet executes feature fusion at various resolutions, capturing disparity information at different scales. The spread of disparity capture allows depth map to retain minute details and well-preserved object boundaries.

To summarise, 2T-UNet is inspired by SDE-DualENet (SDE-DENet) [24], but differs significantly in its architecture and certain core ideas. The SDE-DENet [24] is based on EfficientNets [36], while 2T-UNet is built on top of the UNet architecture [37]. Both techniques differ significantly in the way in which information from the encoder is added to the decoder. In SDE-DENet, the encoder feature is added elementwise to the decoder feature, resulting in some information loss. In 2T-UNet, however, the encoder features are concatenated to the decoder feature, carrying more information. The distinguishing attribute that sets 2T-UNet apart from SDE-DENet is the usage of monocular depth clues. These clues aid the network converge better and give results with clear object boundaries, while maintaining minute details. There are five main sections that make up the rest of this article. Section 2 covers various depth estimation algorithms. In Section 3, the proposed CNN architecture is thoroughly explained. Section 4 explains the implementation details of the experiment. In Section 5, we elaborate our experiment by describing the evaluations and detailed comparative analysis of the experiment. In Section 6, we perform an ablation study of our architecture. The proposed scheme is concluded in Section 7, which includes comprehensive findings and recommendations for further research.

2. Related work

In this section, we provide a brief summary of studies in the domain of image-based depth estimation.

Monocular depth estimation: Alhashim et al. [29] showed that a detailed high-resolution depth maps how, even for a very simple decoder, our method can achieve detailed high-resolution depth maps. The issue of ambiguous reprojections in depth prediction using stereo-based self-supervision is examined by Watson et al. [25]. They included Depth Hints to counteract such negative impacts. Laina et al. [30] proposed a robust, single-scale CNN architecture approach that uses residual learning. Ranftl et al. [26] recommended using principled multi-objective learning to incorporate data from various sources to provide a robust training goal invariant to changes in depth range and scale, and emphasize the significance of pretraining encoders on auxiliary tasks. By using two deep network stacks, Eigen et al. [32] proposed a novel method that provides more accurate local predictions while producing a rough global prediction based on the entire image.

Cantrell et al. [31] combined the advantages of both semantic segmentation and transfer learning for better depth estimation precision. Hu et al. [33] suggested changes in the loss functions and in the method of combining features acquired at various scales. Bhat et al. [27] suggested an architectural block based on transformers that separates the depth range into bins with adaptively determined centres per image. Yan et al. [28] suggested two significant contributions. Firstly, the structure perception module uses the self-attention mechanism to capture long-range dependencies, aggregating discriminative features in channel dimensions. Secondly, the detail emphasis module re-calibrates channel-wise feature maps and selectively emphasises the informative features, resulting in more accurate and sharper depth prediction.

Stereo depth estimation: Duggal et al. [23] demonstrated the case where the cost volume of each pixel is pruned in portions without the need to assess the matching score fully. They constructed a differentiable PatchMatch module to achieve the same. High-resolution imagery processing is a challenge for many state-of-the-art (SOTA) methods due to processing performance limits. Yang et al.[34] provided an end-to-end framework that gradually looks for correspondences throughout a coarse-to-fine hierarchy. They accurately predicted disparity for near-range structures with low latency. Komatsu et al. [14] suggested plane-sweeping strategy to creat two cost volumes using high and low spatial frequency characteristics. Chang et al. [15] proposed a pyramid stereo matching network consisting of two main modules: the spatial pyramid pooling module for creating cost volume using global context infor-

Table 1. Comparison of proposed 2T-UNet scheme with different monocular *(top)* and stereo *(below)* based depth estimation methods on Scene flow dataset. Lower values are better for abs_rel, sq_rel, log_{10} and RMSE. Higher values indicate better quality for σ_1, σ_2, σ_3 and SSIM measures. Best method per metric is highlighted in bold. Second best method per metric is underlined.

Method	$abs\_rel \downarrow$	$sq\_rel \downarrow$	$log_{10} \downarrow$	RMSE $\downarrow$	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	SSIM $\uparrow$
AdaBins [27]	1.222	0.414	0.283	0.264	0.240	0.411	0.569	0.644
CADepth [28]	0.456	0.059	0.142	0.087	0.522	0.760	0.867	0.826
DenseDepth [29]	1.976	0.738	0.377	0.309	0.142	0.288	0.444	0.555
FCRN [30]	1.143	0.288	0.280	0.219	0.208	0.394	0.599	0.658
Depth Hints [25]	0.788	0.136	0.208	0.136	0.338	0.635	0.763	0.769
SerialUNet [31]	0.861	0.165	0.216	0.151	0.344	0.576	0.735	0.713
MSDN [32]	0.856	0.174	0.234	0.189	0.241	0.457	0.702	0.715
SIDE [33]	0.958	0.218	0.239	0.169	0.325	0.540	0.707	0.726
MiDaS [26]	0.338	0.031	0.146	0.072	0.550	0.782	0.879	0.840
OctDPSNet [14]	1.552	0.498	0.422	0.319	0.139	0.275	0.407	0.509
PSMNet [15]	0.317	0.022	0.226	0.062	0.501	0.665	0.773	0.781
DeepROB [23]	0.245	0.016	0.155	0.049	0.622	0.766	0.840	0.816
HSMNet [34]	0.485	0.106	0.286	0.164	0.383	0.517	0.600	0.685
STTR [35]	1.016	0.342	2.341	0.410	0.003	0.005	0.008	0.018
SDE-DENet [24]	**0.166**	**0.008**	0.097	**0.031**	**0.740**	0.858	0.903	0.872
2T-UNet (ours)	0.218	0.013	**0.084**	0.037	0.736	**0.880**	**0.935**	**0.886**

mation and the 3D CNN module for regularizing the cost volume. Li et al. [35] revisited the issue from a sequence-to-sequence correspondence viewpoint in order to replace the costly pixel-by-pixel building of the cost volume with dense pixel matching that makes use of location information and attention. Anil et al. [24] eliminated cost volume construction by learning to match correspondence between pixels with various sets of weights in the dual towers.

3. Proposed Architecture

We propose 2T-UNet, a two-tower UNet with identical contracting paths (encoder), but with different weights and inputs. Fig. 1 illustrates the architecture of our 2T-UNet architecture. This network is inspired by the UNet architecture [37]. 2T-UNet differs from standard UNet in three key aspects:

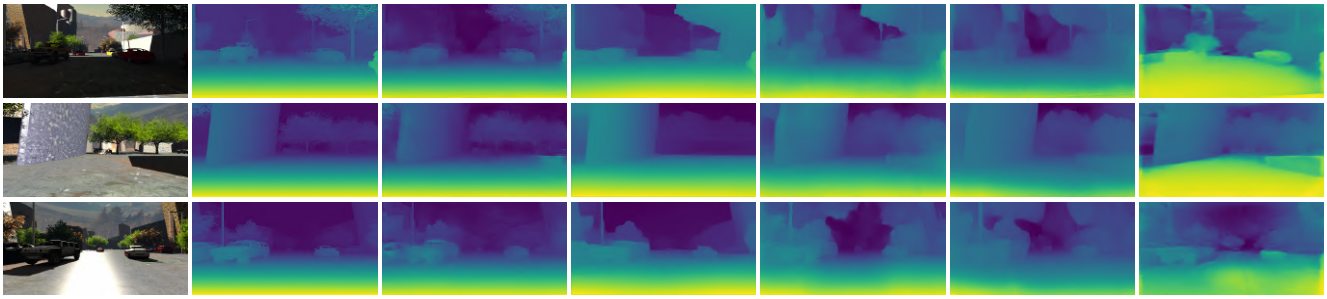
- The standard UNet has a singular contracting (encoder) and expanding (decoder) path, whereas in our 2T-UNet, there exists twin encoders and one decoder.
- We assist stereo-based depth estimation process by providing the network with monocular depth clues of the left view and the stereo image pair.
- The features from the twin encoders are multiplied element-wise to make up the fusion strategy [24]. The fusion is performed before concatenating the features from encoders with the upsampled feature map in the decoder.

The contracting paths consist of the repeated application of two 5×5 convolutions. These convolutions are applied with two padding to preserve the input feature resolution. This is followed by downsampling, which is done by using a 2×2 max-pool layer with stride equals two. At each downsampling step, we double the channels while reducing the spatial resolution by half. The expanding path is composed of repeated units of 4×4 up-convolution. Reducing the feature channels to half and doubling the spatial resolution.

A standard UNet is composed of skip connections between the encoder and decoder, which aids in information transfer, thus retaining the spatial resolution and object boundaries in the prediction [37]. In standard UNet, the encoder feature map is transferred as a whole to the decoder to concatenate with the decoder feature map [37]. In our 2T-UNet model, however, there are two encoders and a single decoder. The model architecture necessitates the use of a fusion strategy to combine two encoder features into a single feature map. This fusion is accomplished by multiplying the features of the two encoders element by element before sending them to the decoder for concatenation. Two 5×5 convolutions with two padding are applied on the concatenated feature maps. Finally, the expansion path outputs a depth map that matches the input's spatial resolution. In 2T-UNet, the primary encoder receives the left image concatenated with the monocular depth clue as input, while the secondary encoder receives the right image alone. The primary encoder is the only one that is directly connected to the expanding path. The secondary encoder terminates at the last skip connection in the architecture.

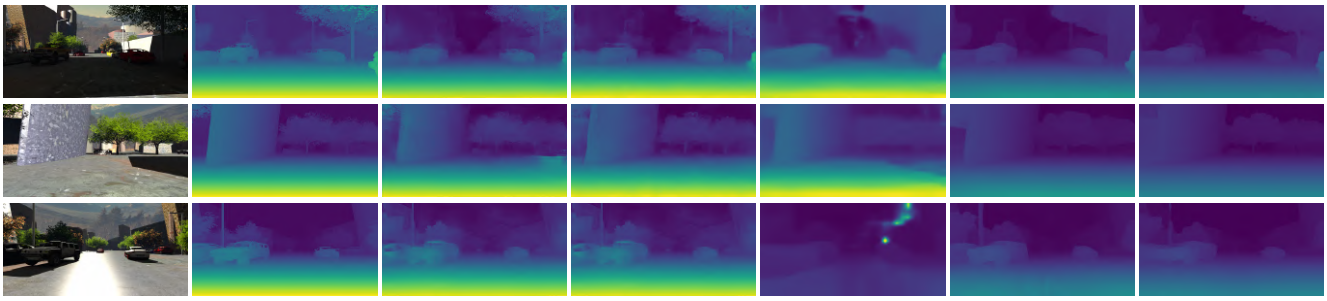
4. Implementation Details

The 2T-UNet is implemented using PyTorch. The model is trained and evaluated on a high-end Tesla K80 GPU on Google Colaboratory. The training took about 15 hours to complete. Inference time per stereo pair is around 0.22 seconds. We trained our model on synthetic driving dataset



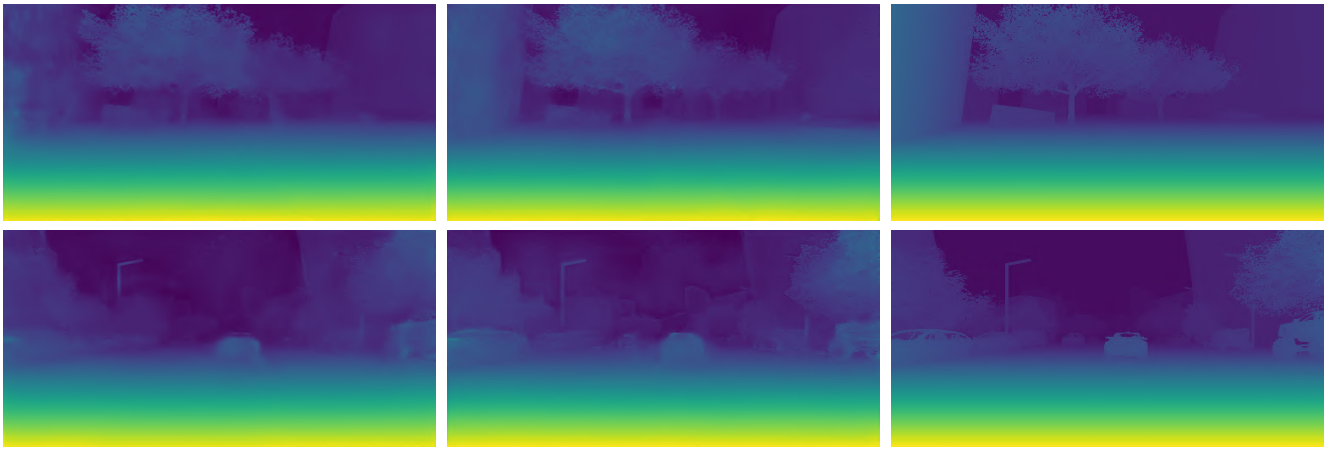
(a) Left image (b) Ground truth (c) 2T-UNet (ours) (d) MiDaS[26] (e) CADepth[28] (f) Depth Hints[25] (g) AdaBins[27]

Figure 2. Visual comparison results of proposed 2T-UNet method with state-of-the-art monocular depth estimation algorithms.



(a) Left image (b) Ground truth (c) 2T-UNet (ours) (d) SDE-DENet[24] (e) HSMNet[34] (f) DeepROB[23] (g) PSMNet[15]

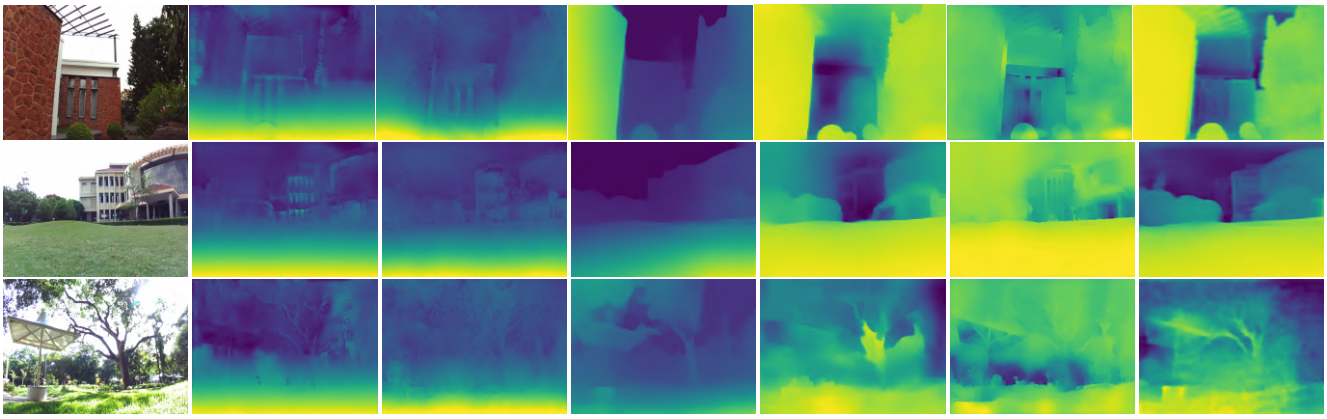
Figure 3. Visual comparison results of proposed 2T-UNet method with state-of-the-art stereo based depth estimation algorithms.



(a) Predicted Depth Map without Depth clue (b) Predicted Depth Map with Depth clue (c) Ground Truth

Figure 4. A visual comparison of estimated depth maps using proposed 2T-UNet model considering without depth clue (a) and with depth clue (b). (row1) The depth estimate and feature definitions of building in the foreground is more accurate in (b); (row2) The features of the street lights and smaller vehicles on both sides of (b) are much closer to ground truth (c).

from the Scene flow database [39]. The dataset is composed of dynamic natural driving scenarios from the first-person perspective of the driver. The scenes contain highly detailed objects such as cars, trees, warehouses, and streetlights. The dataset has 4400 pairs of stereo images. The network is trained for 15 epochs using a standard 90:10 train-test split. It is interesting to notice the computational complexity of 2T-UNet model with and without considering depth clues. The total number of trainable parameters remains the same for both these models, while the total parameters increase by 15% for the model that utilizes depth clues. The inference time of the 2T-UNet model that uses depth clue is 218 ms, while that of the model without depth clue is 217 ms, a difference of approximately 1 ms is noticed. It is safe to say



(a) Left image (b) 2T-UNet(ours) (c) SDE-DENet[24] (d) MiDaS[26] (e) SIDE[33] (f) DenseDepth[29] (g) AdaBins[27]

Figure 5. Visual comparison results of proposed 2T-UNet architecture trained on Scene flow dataset and tested on natural stereoscopic scenes. We compare with state-of-the-art depth estimation algorithms on $House1$, $Himalaya$, $Canopy3$ scenes *(top to bottom)* from complex natural dataset [38].

that the computational complexity of the two models is not significantly different.

5. Evaluation and Comparative Analysis

The proposed 2T-UNet is compared with monocular and stereo-based depth prediction algorithms. We compare the performance of our architecture with the existing state-of-the-art monocular depth prediction algorithm: Depth Hints [25], DenseDepth [29], MiDaS [26], FCRN [30], SIDE [33], SerialUNet [31], MSDN [32], AdaBins [27], CADepth [28]. In addition, stereo-based depth prediction algorithms are also used for performance comparison: PSMNet [15], OctDPSNet [14], DeepROB [23], HSMNet [34], STTR [35] and SDE-DENet [24]. We use publicly available pre-trained models for evaluating the baseline methods. We separately computed depth maps of right and left images to compare the performance with monocular-based algorithms. Furthermore, a stereo image pair is provided as input for comparison with stereo-based depth prediction methods, with one image serving as a source (reference) and the other as a target. In our model, the stereo image pair is provided as input to the network.

The proposed architecture delivers encouraging results on the Scene flow dataset with high-quality depth maps. In Table 1, quantitative comparison results of 2T-UNet architecture with other baseline methods on the Scene flow dataset are reported. For quantitative comparisons, we use standard error metrics: absolute relative error (abs_rel), squared relative error (sq_rel), root mean square error (RMSE), average (log_{10}) error, threshold accuracy (σ_i) [29, 40]. We also compute mSSIM, which is the mean structural similarity score [41]. The error metrics for a predicted depth image and its corresponding ground truth are determined in the following manner:

Absolute relative error :

$$abs\_rel = \frac{1}{|T|}\sum_{p\epsilon T}\frac{|y_p - y_p^*|}{y_p^*} \tag{1}$$

Squared relative error :

$$sq\_rel = \frac{1}{|T|}\sum_{p\epsilon T}\frac{||y_p - y_p^*||^2}{y_p^*} \tag{2}$$

Root mean square error :

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{p\epsilon T}||y_p - y_p^*||^2} \tag{3}$$

Average log error :

$$log_{10} = \sqrt{\frac{1}{|T|}\sum_{p\epsilon T}||logy_p - logy_p^*||^2} \tag{4}$$

Threshold accuracy : percentage of y_p such that

$$max(\frac{y_p}{y_p^*}, \frac{y_p^*}{y_p}) = \sigma_i < thres \tag{5}$$

for $thres = 1.25, 1.25^2, 1.25^3$.

Figures 2 and 3 show the visual comparison of the predicted depth maps of the selected networks. We select three scenes from the Scene flow driving synthetic dataset which include tree, car, road, buildings. The scenes include ill-posed areas such as reflective glass, walls and road surfaces. Our proposed algorithm consistently outperforms stereo-based and monocular-based baseline methods, both quantitatively and qualitatively. The 2T-UNet architecture achieves comparable result with SDE-DENet [24] method.

Table 2. Comparison between the proposed 2T-UNet with and without the inclusion of monocular depth clues in the pipeline.

Method	$abs\_rel \downarrow$	$sq\_rel \downarrow$	$log_{10} \downarrow$	RMSE $\downarrow$	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	SSIM $\uparrow$
2T-UNet (without Depth Clue)	0.328	0.059	0.083	0.041	0.648	0.741	0.852	0.805
2T-UNet (with Depth Clue)	0.218	0.013	0.084	0.037	0.736	0.880	0.935	0.886

Our method provides robust depth estimation results particularly in the regions of car windows and wall. The thin structures and object boundaries are clearly retained in our depth estimates.

In a natural scene with intricate movements, lighting change, and illumination, estimating depth can be difficult. We conduct a visual comparison of our method against existing approaches on challenging nature scenes [38], as shown in Figure 1, to demonstrate its efficacy. For this objective, we employ 2T-UNet architecture that has been trained on the Scene flow dataset. Due to a lack of ground truth data, the quantitative analysis for intricate natural scenes [38] is not carried out.

6. Ablation study

We conduct an ablation study on our architecture to explore the significance of integrating monocular depth clues into the depth estimation process. Table 2 presents a comparative analysis of the 2T-UNet model with and without the inclusion of depth clues as an additional input. The model that incorporates depth clues alongside RGB images exhibits superior performance across all evaluation metrics The predicted depth maps with and without the use of depth clues are compared visually in Fig. 4. The estimated depth map with depth clues has clear depth discontinuity and object structures. The depth structure and fine details are well preserved in scenes with large variations in depth.

7. Conclusion

In this paper, we explore the issue of stereo depth estimation using a simple CNN and propose an end-to-end Two Tower UNet architecture. The suggested architecture bypasses cost volume construction and, as a result, avoids specifying the disparity range of a scene. Furthermore, the network learns better object boundaries with clarity in depth discontinuity by using depth clues. The proposed 2T-UNet model gives quantitatively and qualitatively better depth predictions at inherently ill-posed regions despite eliminating cost volume construction. We experimentally demonstrated the quality of results using the complex natural scenes and Scene flow dataset. As an extension of the current study, we plan to investigate other feature fusion techniques. Additionally, we would also wish to expand 2T-UNet as a complementary network to additional methods based upon stereo depth estimation. We will further explore the possibilities of extending our work to free-viewpoint rendering and in realistic integration of virtual scenes in augmented reality and 3D environments.

References

[1] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, 2011. doi: 10.1109/ICCV.2011.6126513. 1

[2] P. Avinash and M. Sharma. Predicting forward backward facial depth maps from a single rgb image for mobile 3d ar application. In *IC3D*, pages 1–8, Dec 2019. doi: 10.1109/IC3D48390.2019.8975899. 1

[3] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. Active refocusing of images and videos. *ACM Trans. Graph.*, 26(3): 67–es, jul 2007. ISSN 0730-0301. doi: 10.1145/1276377.1276461. 1

[4] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision – ACCV*. Springer International Publishing, 2017. 1

[5] G.N. Desouza and A.C. Kak. Vision for mobile robot navigation: a survey. *IEEE TPAMI*, 24(2):237–267, 2002. doi: 10.1109/34.982903. 1

[6] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE CVPR*, 2015. doi: 10.1109/CVPR.2015.7298925. 1

[7] Vineet Thumuluri and Mansi Sharma. A unified deep learning approach for foveated rendering & novel view synthesis from sparse rgb-d light fields. In *IC3D*, pages 1–8, 2020. doi: 10.1109/IC3D51119.2020.9376340. 1

[8] Mansi Sharma and G. Ragavan. A novel image fusion scheme for ftv view synthesis based on layered depth scene representation and scale periodic transform. In *International Conference on 3D Immersion*, 2019. doi: 10.1109/IC3D48390.2019.8975902. 1

[9] Mansi Sharma, Gowtham Ragavan, and B Arathi. A novel algebaric variety based model for high quality free-viewpoint view synthesis on a krylov subspace. In *IC3D*, pages 1–8, 2019. doi: 10.1109/IC3D48390.2019.8975992. 1

[10] Mansi Sharma, M. Venkatesh, Gowtham Ragavan, and Rohan Lal. A novel approach for multi-view 3d hdr content generation via depth adaptive cross trilateral tone mapping. In *IC3D*, pages 1–8, 12 2019. doi: 10.1109/IC3D48390.2019.8975988. 1

[11] Joshitha Ravishankar and Mansi Sharma. A hierarchical coding scheme for glasses-free 3d displays based on scalable hybrid layered representation of real-world light fields. *CoRR*, abs/2104.09378, 2021. URL https://arxiv.org/abs/2104.09378.

[12] M. Sharma, S. Chaudhury, B. Lall, and M.S. Venkatesh. A flexible architecture for multi-view 3dtv based on uncali-

brated cameras. *JVCIR*, 25(4):599–621, 2014. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2013.07.012. 1

[13] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2020.3032602. 1

[14] R. Komatsu, H. Fujii, Y. Tamura, A. Yamashita, and H. Asama. Octave deep plane-sweeping network: Reducing spatial redundancy for learning-based plane-sweeping stereo. *IEEE Access*, 7:150306–150317, 2019. doi: 10.1109/ACCESS.2019.2947195. 3, 4, 6

[15] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF CVPR*, pages 5410–5418, 2018. doi: 10.1109/CVPR.2018.00567. 1, 2, 3, 4, 5, 6

[16] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. Di Stefano. Real-time self-adaptive deep stereo. In *IEEE CVPR*, 2019. 1

[17] Sizhang Dai and Weibing Huang. A-tvsnet: Aggregated two-view stereo network for multi-view stereo depth estimation. *arXiv preprint arXiv:2003.00711*, 2020. 1, 2

[18] Yue Zhao, Y. Xiong, and D. Lin. Recognize actions by disentangling components of dynamics. In *IEEE CVPR*, June 2018. 1

[19] Y Wang, Z Lai, G Huang, B H Wang, L van D Maaten, M Campbell, and K Q Weinberger. Anytime stereo image depth estimation on mobile devices. *arXiv:1810.11408*, 2018. 1, 2

[20] Z. Yin, T. Darrell, and F. Yu. Hierarchical discrete distribution decomposition for match density estimation. In *IEEE CVPR*, June 2019. 1

[21] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. In *ICLR*, 2019. 1, 2

[22] F. Zhang, V. Prisacariu, R. Yang, and P. H.S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE CVPR*, 2019. 2

[23] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[24] Rithvik Anil, Mansi Sharma, and Rohit Choudhary. Sdedualenet: A novel dual efficient convolutional neural network for robust stereo depth estimation. In *VCIP*, 2021. doi: 10.1109/VCIP53242.2021.9675391. 1, 2, 3, 4, 5, 6

[25] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[26] R. Ranftl, K. Lasinger, D. Hafner, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. doi: 10.1109/TPAMI.2020.3019967. 2, 3, 4, 5, 6

[27] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE CVPR*, 2021. doi: 10.1109/CVPR46437.2021.00400. 3, 4, 5, 6

[28] J. Yan, H. Zhao, P. Bu, and Y. Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *3DV*, 2021. doi: 10.1109/3DV53792.2021.00056. 3, 4, 5, 6

[29] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv e-prints*, abs/1812.11941:arXiv:1812.11941, 2018. URL https://arxiv.org/abs/1812.11941. 3, 4, 6

[30] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. *3DV*, pages 239–248, 2016. doi: 10.1109/3DV.2016.32. 3, 4, 6

[31] Kyle J. Cantrell, C. D. Miller, and C. W. Morato. Practical depth estimation with image segmentation and serial u-nets. *VEHITS*, 2020. ISSN 2184-495X. doi: 10.5220/0009781804060414. 3, 4, 6

[32] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 3, 4, 6

[33] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *IEEE WACV*, 2019. 3, 4, 6

[34] G. Yang, J. Manela, M. Happold, and D. Ramanan. Hierarchical deep stereo matching on high-resolution images. In *IEEE CVPR*, June 2019. 3, 4, 5, 6

[35] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. *arXiv:2011.02910*, 2020. 4, 6

[36] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. ISBN 978-3-319-24574-4. 3, 4

[38] A Wadaskar, M Sharma, and R Lal. A rich stereoscopic 3d high dynamic range image & video database of natural scenes. In *IC3D*, 2019. doi: 10.1109/IC3D48390.2019.8975903. 6, 7

[39] N. Mayer, Eddy Ilg, Philip Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *IEEE CVPR*, 2016. 5

[40] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Trans. Graph.*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356528. URL https://doi.org/10.1145/3355089.3356528. 6

[41] M. Sharma, A. Sharma, K. R. Tushar, and A. Panneer. A novel 3d-unet deep learning framework based on high-dimensional bilateral grid for edge consistent single image depth estimation. In *IC3D*, 2020. doi: 10.1109/IC3D51119.2020.9376327. 6