Monocular Depth Estimation Methods using CNN and Transformer

Anonymous CVPR submission

Paper ID ****

Abstract

Depth information is an important attribute in a variety of domains including but not limited to self-driving, augmented reality, face unlocking etc. A variety of deep learning methods exist that aim to solve this problem using a single camera (monocular). Moreover, methods based on transformers dominate these methods in terms of acccuracy currently. We aim to present and analyse different methods to solve monocular depth estimation problem. We also aim to compare each method focusing especially on CNN-based and transformer-based methods. We show on a public dataset NYUDepth v2 that combining CNN and transformer gives the best performance. We propose two approaches in this direction. First, an end-to-end training of combined CNN and Transformer network. Second uses knowledge distillation of large transformer models to smaller CNN-based networks.

1. Introduction

Monocular Depth Estimation (MDE) is a widely studied problem that aims to estimate the depth of each pixel in the image using only one camera image (monocular) at a given time. This is quite natural for humans as we utilize additional information in the scene like the relative sizes of other known objects in the scene, the appearance of objects in varying lighting, shading and occlusions, surface textures and focal fields etc. However computing a depth estimation model is an ill-posed problem fundamentally because any 2D image could have been generated from an infinite range of 3D scenes. Computational models for Monocular depth estimation have traditionally used auxiliary information like object sizes and location, interaction of objects with occlusion and perspective and texture variations to name a few to estimate depth.

Monocular depth estimation It is important especially when stereo images are not available or other sensors like Lidar are impractical or costly. Depth estimation helps to understand the environment and helps to make better decisions. It has wide applications in augmented reality to give realistic views of the digital objects. Also, self-driving vehicles uses depth information to detect objects and avoid collisions.

There has been a lot of deep learning methods based on CNN that tries to solve this problem. Recently, transformer based models show promising accuracy in this domain. But transformer-based model face a lot of challenges too. For instance, they are difficult to train and tune. They require a lot of data to converge. On the contrary, they provide a global receptive field unlike CNNs. CNNs are generally easy to train and converge faster. We aim to explore and experiment with methods that use both CNN and transformer based features simultaneously. Also, transformer based models are many times big networks trained on large amount of data to get state-of-the-art accuracies. But it may not be practical to deploy such models on embedded devices for real-time execution. Hence, we also propose an approach to use knowledge distillation to transfer transformer based learning to a smaller CNN-based network. We experiment and analyse the accuracies by using such methods.

We base our experiments on a widely used public dataset called NYU-Depth V2 [16] dataset. The dataset comprises of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It features 1449 densely labeled pairs of aligned RGB and depth images, 464 new scenes taken from 3 cities and 407,024 new unlabeled frames.

We propose two such approaches and show that they outperform methods trained on single features. Our contributions can be summed up as follows:

1. We show that training an end-to-end network with CNN and transformer achieves better performance than doing a multi-stage training.
2. In cases, where end-to-end training is not practical, we show knowledge distillation can be a viable method to distill information from transformer model to small cnn based networks.

2. Related Work

2.1. CNN for monocular depth estimation

A lot of the initial MDE research was based on using CNNs. One of the seminal works in this field was by Eigen et al. [2]. They were the first to utilize a few fundamental components in the single image depth estimation pipeline which were later followed by many others. In addition to proposing the concept of directly regressing over each pixel, they had also proposed the approach of spliting the estimation into two: one which estimates the global structure of the input scene and ther other which refines this structure with local information. The introduction of scale invariant loss was also proposed to handle scale dependency of estimation error.

The ideas proposed by Eigen et.al were later taken up by Xu et al. [15] and Li et al[8]. with the addition of fusion of multiple semantic layers of the CNN within a Conditional Random Field (CRF) framework. Li et al. also proposed the use of multi-scale CRFs and a cascade of CRFs, one for each level.

2.2. Transformers for monocular depth estimation

One of the first attempts to use attention models for estimating depth from monocular images were by Xu et al.[18] where the structure from (Xu) [15] along with the use of a structured attention model where the information exchanged between embedding of different scales were controlled by the attention model. This particular approach operates on the feature-level and fuses features from different scales and enforces structure.

Another use of attention based models was in Yuru et al. [7] where a supervised attention-based Context Aggregation Network (ACAN) was proposed to estimate depth maps. The method uses deep residual architecture, dilated layer and self-attention modules for scale control with dense prediction. The use of the self-attention module helps in mapping the relation between every pixel which translate to the attention weights. Also another key feature is the use of image-pooling which combines image level information for the depth estimation.

Iterative approaches in MDE by Ranftl et al. [14][13] have used 3D movies as data source to learn from dataset with varying parameters of environments like scale, range of depth, aspect ratios etc. This had enabled the model to do zero-shot cross-dataset transfer learning. They went to propose hybrid and vision transformer based dense prediction of depth.

2.3. Combining CNN and Vision Transformer for monocular depth estimation

We further explored the idea proposed by Ranftl et al. in combining CNNs with transformers. Specifically in the hybrid model, non-overlapping patches of the input RGB image is converted into tokens by passing the patches through a ResNet-50 feature extractor. These embeddings along with the positional embedding are they passed through multiple transformer stages. These tokens are then reassembled at different resolutions into an image-like representation. The hybrid model extracts features at $\frac{1}{16}$ scale of input resolution, which is a much deeper resolution than most methods with convolutional backbones.

Our reasoning for the use of CNNs and transformers in this way is the presence of inductive bias and low sample complexity in CNNs. CNNs would manage to produce a good feature representation of the data with very few samples. These features when passed into a transformer and trained with supervision should be able to match the accuracy of a purely transformer model with fewer data samples.

2.4. Teacher Student Network for supervised learning tasks

Lowering model complexity and computation while having highly accurate outputs from all deep learning models have been under exploration for a long time. Some of the ways of simplifying the model have been model pruning[4] and quantization [12]. We explore another strategy called knowledge distillation proposed by Hinton et al. [5], which aims to transfer knowledge from a heavier, more accurate teacher network onto a lighter student network. This method was initially proposed for image classification and have since then been diversified into other tasks like semantic segmentation, object detection and depth prediction[11]. Initially distillation strategies revolved around distiling the class probability distribution for each pixel. Shen et al. [9] later propsed pair-wise and holistic distillation, which is a more generic, structured knowledge distillation framework for dense prediction. The holistic distillation consists of conditional adversarial learning and the use of a discriminator.

2.5. Datasets

Here we highlight some of the major datasets we came across for training model for MDE. Such datasets have been created using ground truth depth estimation mechanisms like disparity, LiDAR, structured light among others.

NYU-v2 dataset was introduced in [16], containing 1449 RGB images with dense depth labels. It contains a total of 407K frames of 464 scenes. KITTI dataset[3], has two versions with 394 road scenes having RGB stereo sets and GT depth maps. They have been captured using the Velodyne laser scanner. Pandora[1] contains 250K full resolution RGB and depth images. SceneFlow[10] in one of the first large-scale synthetic datsets having 39K stereo images with corresponding disparity, depth , optical flow and segmentation masks. Additional datasets have been listed in

Table 1.

Dataset	Labelled images	Annotation
NYU-v2	1449	RGBD + segmentation
KITTI	94K	RGBD + optical flow
Pandora	250K	RGBD
SceneFlow	39K	RGBD + segmentation
DIML Indoor[6]	220K	RGBD
ReDWeb[17]	3600	stereo

Table 1. Some of the datasets for monocular depth estimation

3. Methods

3.1. Combining CNN and Transformer

We investigate the effectiveness of combining CNN and Transformer architecture for learning the task of monocular depth estimation in two ways. We investigate that this would leverage the deep representational learning of CNN and attention modules of Transformer to encode global context. We investigate two methods for fusion. In the former method, we train both the encoder and the transformer separately. Our hypothesis is that since the Vision Transformer is applied to sequential patches of image, it is better to keep the two components independent. The second architecture is the End to End architecture designed for dense prediction task where we fine-tune our CNN model and transformer model in an end to end fashion. Further details can be found in the experiment section.

3.2. Teacher Student Network

Knowledge Distillation is an approach to train smaller networks (student) using models trained on large amount of data (teacher). It has been observed that the model converges faster than if trained from scratch using the large amount of data. The reason for such behaviour is that the smaller model may not have enough representation capability to learn from the full data. Hence, it is helpful to transfer the learning using soft labels. Since, it may be impractical to train CNN-Transformed based models end-to-end, we propose knowledge distillation to capture the power of both transformer and CNN features. We try to distill information from a large transformer based model to a small CNN based model with significantly less number of parameters. This can also solve another problem that we would like to highlight here. The ground truth depth data has errors in the depth value as captured from the Kinect camera. This is due to the inherent sensor noises, limited range issues etc. Due to this , we can see black patches in the ground truth itself (figure). Hence, it becomes difficult for the model to learn that missing information especially with a smaller network. The larger network is able to learn this with large amounts of data. Hence distilling that information can help to achieve better quality outputs. The loss function that we used is as follows:

$$\mathcal{L}_{\text{depth}} = \alpha \sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}\left(\sum_i g_i\right)^2}$$

where $g_i = \log \hat{d}_i - \log d_i$ with the ground truth depth d_i and the predicted depth $\hat{d}_i$. We set λ and α to 0.85 and 10, same as [cite transdepth]

4. Experiments

4.1. Combining CNN and Transformer

The input image is first resized into 224 x 224 x 3. In our results, we finetune the encoder ResNeXt-101 backbone on NYU depth dataset. We found that using higher capacity encoder like the one we have used performs significantly better than the same encoder that was only trained on ImageNet. For combining the output of CNN with vision transformer, we use the output from a CNN model (ResNet backbone) and feed it into a vision transformer capable of predicting dense output. The patch embedding layer is applied to final feature output of the CNN. This patch embedding's kernel should be pxp, which means that input sequence is obtained by simply flattening the spatial dimensions of the feature map and projecting to the Transformer's dimension. The only difference between the two proposed methods is that in the first one, the output of the encoder is trained separately to predict the depth whereas the second trains the combined network end to end.

4.2. Teacher Student Network

We design the following experiments to implement and analyse this approach.

1. First, we fine-tune a small CNN based network [cite] using 1449 labelled samples of NYUDepth v2 dataset. This acts as the baseline for comparison.

2. Second, we take a pre-trained transformer network [13] and fine-tune the pretrained small CNN network using ground truth and the output of transformer network. We follow the architecture as described in the [figure]. We use two loss terms to help the model learn from ground-truth as well as the output of transformer network. We try different loss functions to train the models.

3. Third, we also experiment with filling the missing information in the ground-truth relying on the output of the transformer network and then fine tuning on the dense data obtained.

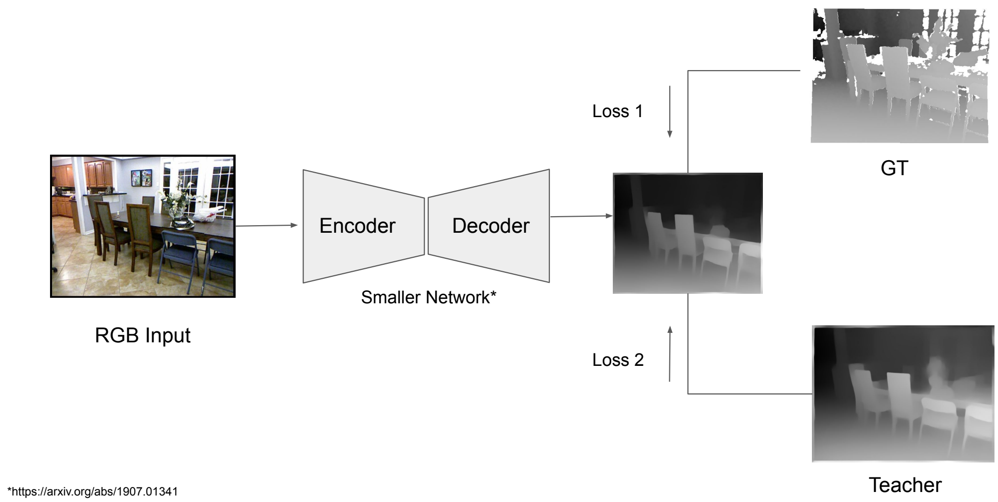
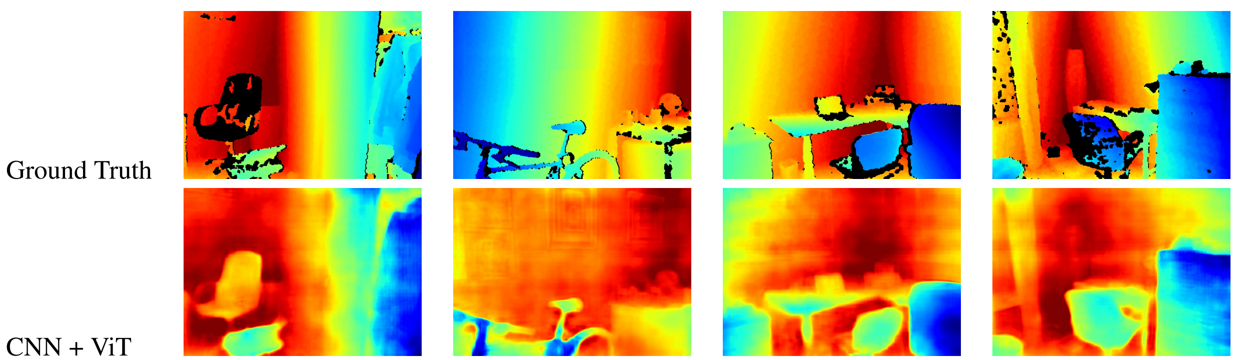


Figure 1. Teacher Student Network Architecture



4.3. Results and Discussion

4.3.1 CNN + Transformer method

In Figure 3.2, we can see that quite a lot of the ground truth have a series of black patches. This happens when for a certain region Kinect's speckle pattern emitter doesn't get back to the Kinect's IR camera. This can be due to a number of factors including the surface being too reflective or that surface not visible to the IR camera and pattern emitter at the same time. The CNN + ViT output shows that the model is able to predict a smooth output directly from the RGB image. Most of the surfaces in the images have the correct level of depth and has learnt to group close-by pixels in the RGB image, similar depth values. Although, this is a valid mapping for most surfaces and objects, it does not hold true always. For instance, the bike in the second example has different levels of depth at different parts of it which the models fails to assign.

we evaluate End to End training and Multi-Stage training across all metrics. For metrics which rely on relative distance between ground truth and predicted pixel values like LOG_RMS and SILOG (Scale invariant logarithmic error) [2] given by:

$$D\left(y, y^*\right)=\frac{1}{n} \sum_i d_i^2-\frac{1}{n^2}\left(\sum_i d_i\right)^2 \quad d_i=\log y_i-\log y_i^* \tag{1}$$

where smaller value is better. On the other hand, metrics which measure accuracy with threshold t: percentage (%) of d_i^* subject to 2,

$$\max \left(\frac{d_i^*}{d_i}, \frac{\tilde{d}_i}{d_i^*}\right)=\delta<t\left(t \in\left[1.25,1.25^2, 1.25^3\right]\right) \tag{2}$$

where a higher value is better. In Figure 2, we can see that End to End training outperforms Multi-Stage architecture in all metrics.

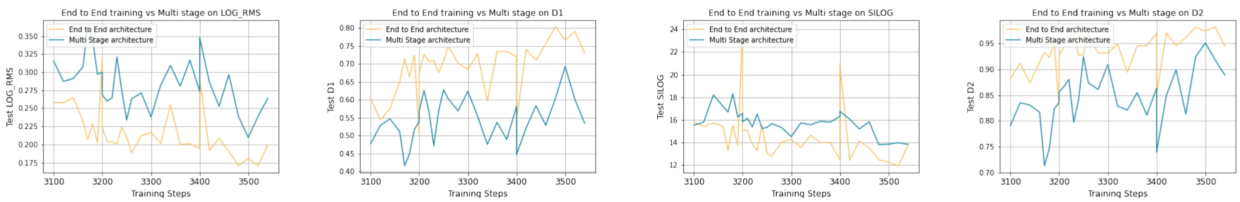


Figure 2. Quantative results comparing End to End architecture vs Multi Stage architecture

4.3.2 Teacher Student Network

The results can be seen in table 2. The Baseline Finetuned model is the Midas based CNN network finetuned on NYUDepth v2 dataset. We see that it is able to achieve RMSE of 0.619 and Delta1 of 0.701. This acts as the baseline for the other experiments that we wish to perform. The second model ”Finetuned w/ Teacher” is the model trained with Teacher supervision on the same CNN network. We used the loss function as described in the Section 3.2 and architecture as described in 1. We tried different weights to add both the losses. The best is shown in table as RMSE of 1.151 and Delta1 of 0.417. Ideally, we hoped to achieve better performance using knowledge distillation but the results are contrary to that. We believe that this could happen because of the less data used for fine-tuning the model. Since, the loss is changed than the pretrained model was trained on, it might take a little more data and time to converge. Some of the qualitative results were promising showing that the model did learn something. But more research and tuning might be required to make it better than the baseline. As we described earlier, the ground-truth contains error in depth values especially missing values. Hence, the models may not be able to learn those values if loss only calculates error with ground-truth. Hence, we tried another way to train the model. We filled the missing values with the output of the transformer based model. The intuition was to give explicit guidance to the network and use only one loss term. We hoped that it would make it easy for the model to learn slightly better. But as can be seen in the table, the model could only reach at the same level as using 2 loss terms. This suggests that that the benefit of filling the values was outweighed by not giving the full output of the transformer model to learn using an extra loss term.

Model	RMSE	Delta1	MAE
Baseline fine-tuned	0.619	0.701	0.470
Fine-tuned w/ Teacher	1.151	0.417	0.817
Filling unknown	1.181	0.407	0.907

Table 2. Results comparison among baseline and Teacher supervision on NYUDepth-v2

4.4. Conclusion

We analyse the effect of using CNN and transformer together in the domain of Monocular Depth Estimation. We show qualitatively and quantitatively that end-to-end learning using CNN+Trasnformer achieves better performance. We also show analysis of knowledge distillation using two approaches. We distill the information of a larger transformer network to a smaller CNN network and compare the performance with the model finetuned directly on the data. We also fill the missing values in the ground-truth with the transformer output values and finetune the network. Ideally, we hoped to show better performance but due to time, data and GPU constraints, we were not able to beat the baseline. We aim to use larger data to finetune the network with Teacher supervision. We also aim to experiment with more knowledge distillation approaches apart from adding loss terms. For instance, learning from multi-layer outputs of the bigger model and not just the final output.

References

[1] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4661–4670, 2017. 2

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014. 2, 4

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[4] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017. 2

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[6] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE transactions on Image Processing*, 27(8):4131–4144, 2018. 3

[7] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar

guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[8] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 2

[9] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[10] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2

[11] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019. 2

[12] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 2

[13] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 2, 3

[14] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 2

[15] Elisa Ricci, Wanli Ouyang, Xiaogang Wang, Nicu Sebe, et al. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1426–1440, 2018. 2

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1, 2

[17] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 3

[18] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3917–3925, 2018. 2