Lightweight single image dehazing utilizing relative depth information

Panagiotis Frasiolas
CERTH
6th km
Charilaou-Thermi
road
Greece 57001,
Thessaloniki
frasiolas@iti.gr

Asterios Reppas
CERTH
6th km
Charilaou-Thermi
road
Greece 57001,
Thessaloniki
asterisreppas@iti.gr

Konstantinos
Konstantoudakis
CERTH
6th km
Charilaou-Thermi
road
Greece 57001,
Thessaloniki
k.konstantoudakis@iti.gr

Dimitrios Zarpalas
CERTH
6th km
Charilaou-Thermi
road
Greece 57001,
Thessaloniki
zarpalas@iti.gr

ABSTRACT

Considering the need for lightweight and fast implementations, this paper presents an architecture based on a MobileVit encoder for efficiency and speed, introducing a fully convolutional lightweight decoder with skip connections for feature extraction. The main purpose of this network is to address the problem of single image dehazing. Recognizing the critical role of depth information in assisting the above task, the merging of these two tasks into a single network was performed in a supervised manner. Taking into account that there is a shortage of datasets that provide both dehazing and relative depth estimation ground truths, Depth Anything was utilized to extract the relative depth values of the images, which is the SOTA network in this task.

Keywords

Lightweight, Vision Transformers, relative depth, dehazing

1 INTRODUCTION

Despite the recent advancements in computer vision research, scene understanding remains a fundamental problem. Monocular depth estimation provides a deeper insight to the scene, capturing depth information and transforming perception from a two-dimensional representation to a richer three-dimensional understanding. It has a potential to revolutionize applications such as autonomous navigation [4], augmented reality, and scene understanding.

In the presence of haze, because it has a strong effect on visual clarity and detail, comprehending a scene becomes really challenging. Single image dehazing aims to mitigate the adverse effects of atmospheric scattering, enhancing the visibility and fidelity of images captured in hazy or foggy conditions and aims to restore the true radiance of objects obscured by haze or fog. In essence, both depth estimation and image dehazing share a common goal: the recovery of a more faithful representation of the scene.

Recent studies of both monocular depth estimation and single image dehazing methods, have introduced Vision Transformers [25, 1, 8] as a fundamental component for a global understanding of the scene, unlike traditional methods which rely on convolutional neural networks [17], [7], [5] with limited receptive field.

Tasks like self-driving cars require real-time processing, because it directly impacts user experience and safety. Single image dehazing is essential in scenarios where visibility is compromised due to adverse weather conditions. Real-time dehazing can enhance image clarity and enable immediate responses. Real-time processing ensures that the information provided is current, allowing systems to react swiftly and effectively to changing environments and unforeseen obstacles. That is the reason why a lightweight model has been developed in this paper.

The proposed model is based on an encoder-decoder architecture. MobileVit [22], a Vision Transformer with a low complexity, is employed in the encoder, aiming to minimize the model's parameter count. The decoder is a fully-convolutional neural network. In the earlier stages of the decoder both dehazing and depth estimation are learned simultaneously and in later stages these tasks are separated into 2 branches. Skip-connections from the encoder to the decoder result in an efficient and

effective feature processing pipeline and enable the network to leverage both fine and coarse-grained details. The combination of these components allows the decoder to generate accurate pixel-wise predictions. This approach contributes to the ability of the network to efficiently and accurately process visual data.

The main contributions in this paper are summarized below:

- Proposition of a lightweight network for single image dehazing utilizing MobileVit [22] in the encoder
- Utilization of the relative depth values in order to help the image dehazing tasks and also extract a relative depth image from the input hazy image.
- Utilization of a fully-convolutional, fast and accurate decoder with skip connections that requires few parameters.

2 RELATED WORK

2.1 Monocular depth estimation

An early work on monocular depth estimation is that of Eigen et al. [7], which uses two CNNs. The first is used to predict a coarse global depth, and the second to refine the prediction locally. Jung et al. [13] proposed a solution for poor boundary localization and spurious regions by using a two-stage convolutional network as a generator. Their approach employs a deep adversarial learning framework, with an adversarial discriminator training criterion aiming to effectively tell real and synthetically generated depth images apart. A lightweight model was proposed by Wofk et al. [31], utilizing MobileNet [12] as the decoder in order to have a low count of parameters. They also incorporated skip-connections between the encoder and the decoder. Rudolph et al. [28] used a Guided Upsampling Block (GUB) for building the decoder. GUB relies on the image to guide the decoder in upsampling the feature representation and the depth map reconstruction, achieving high resolution results with fine-grained details. Lee et al. [16] proposed a token sharing transformer that utilizes global token sharing, which enables the model to obtain an accurate depth prediction with high throughput in embedded devices. The model used for the ground truth relative depth images is Depth Anything [33], a practical solution for robust monocular depth estimation, focusing on simplicity and effectiveness. By scaling up the dataset to approximately 62 million unlabeled images and employing data augmentation techniques and auxiliary supervision, the method achieves impressive generalization across various datasets.

2.2 Single image dehazing

To describe the formation of a hazy image, the atmospheric scattering model was first proposed by McCartney [21]. The equation of this model can be written as

$$I(x) = J(x) \cdot t(x) + a \cdot (1 - t(x)) \quad (1)$$

where I(x) is a hazy image, J(x) is the real scene to be recovered, t(x) is the medium transmission, a is the global atmospheric light.

DCP Net [9] is a simple but effective image prior - dark channel prior to remove haze from a single input image. DehazeNet [2], which is one of the earliest deep learning works, uses a CNN with specialized Maxout layers for haze-related feature extraction and introduces the Bilateral Rectified Linear Unit (BReLU) activation function to enhance haze-free image quality. Ren et al. [27] used an encoder-decoder architecture and adopted a novel fusion-based strategy which derives three inputs from an original hazy image by applying white balance, contrast enhancement, and gamma correction. Dong et al. [5] presented a Multi-Scale Boosted Dehazing Network using the U-Net framework, which is designed based on two principles: boosting and error feedback. The model incorporates the Strengthen-Operate-Subtract boosting strategy in the decoder, gradually enhancing the haze-free image. They introduced a dense feature fusion module with back-projection feedback in the U-Net architecture to maintain spatial information. Hong et al. [11] introduced a knowledge distillation-based dehazing network that employs process-oriented learning with the student network mimicking image reconstruction. Wu et al. [32] introduced a novel regularization technique that utilizes contrastive learning. CR leverages hazy images as negatives and clear images as positives, guiding the restored image closer to clear images and away from hazy ones in the representation space. Cui et al. [3] was inspired by the consistent degradation of various regions in corrupted images, and suggested a shift towards prioritizing essential areas for reconstruction. In the latter approach, they introduced a dual-domain selection mechanism to accentuate critical information for restoration, including elements like edge signals and challenging regions. FFA-NET [24] is an end-to-end feature fusion attention network, consisting of three key components.1) Channel Attention with Pixel Attention mechanism, 2) Local Residual Learning, and 3) An Attention-based different levels Feature Fusion (FFA) structure, that performs especially outstanding in regions with thick haze and rich texture details. AOD-Net [18] is designed based on a re-formulated atmospheric scattering model and directly generates clean images through a lightweight CNN, making it easily embeddable into other deep models. MSCNN [26] is a multi-scale CNN consisting of a coarse-scale

net that predicts a holistic transmission map based on the entire image, and a fine-scale net that refines results locally. LightDehazeNet [29] jointly estimates both the transmission map and atmospheric light using a transformed atmospheric scattering model. There are not many lightweight models to perform image dehazing delivering good results. The proposed method targets these two goals concurrently.

2.3 Vision Transformers

Vision Transformer (ViT) [6] adapts the transformer architecture used in natural language processing to extract multiscale information from images by breaking them down into smaller patches. The most important part is the self-attention mechanism which helps to encode relationships between the patches. ViT-based models have achieved remarkable results in tasks like image classification and segmentation, depth estimation, and single-image dehazing. Depth estimation works like Ada-Bins [1] proposed a transformer-based architecture block that divides the depth range into bins whose center value is estimated adaptively per image. The final depth values are estimated as linear combinations of the bin centers. Ranft et al. [25] gathered tokens from different stages in the vision transformer to create representations that resemble images at various resolutions. These representations are gradually fused to produce full-resolution predictions using a convolutional decoder. In another encoder-decoder architecture, Kim et al. [14] deployed a hierarchical transformer-based encoder to capture the global information in an image, and a lightweight decoder to generate an estimated depth-map, while also considering local connectivity. Vision transformers are also used in single-image dehazing. Guo et al. [8] proposed a novel transmission-aware 3D position embedding to involve haze density-related prior information into the vision transformer. Lu et al. [20] created two modules, one for handling both fine textures and large hazy areas, and another for addressing uneven haze distribution in image dehazing. The first module uses parallel dilated convolutions with large receptive fields, while the second efficiently extracts global and local information in parallel to improve dehazing results. Zhao et al. [34] combined intrinsic image decomposition and image dehazing, enhancing the generation of high-quality haze-free images. The Complementary Feature Selection Module (CFSM) was used to effectively fuse complementary features, thereby boosting feature aggregation. In the scope of this research paper, vision transformers are adopted as the encoder of the network. The reason for this selection relies on the proven capacity of vision transformers to produce rapid and precise results. Vision transformers are favored for their adeptness in managing visual data, detecting patterns, and comprehending the content of images.

3 METHOD

An image dehazing network is trained Fig.1 which aims to predict the dehazed image $Y \in \Re^{H\times W\times 3}$ and the relative depth map $D \in \Re^{H\times W\times 1}$ from an RGB hazy image $X \in \Re^{H\times W\times 3}$. The primary focus lies on the dehazing output, with the relative depth output serving a supplementary role. The relative depth information primarily aids the image dehazing task, given its inherent inclusion within the haze-scattering model. Eq.1.

To accomplish that, a model based on an encoder-decoder architecture was implemented. In the encoder, a pre-trained MobileVit [22] was used, and in the decoder a fully convolutional network. MobileViT is designed to bring together the strengths of CNNs and vision transformers to create a lightweight and fast-to-evaluate network for mobile vision tasks. It offers a new perspective on how to process visual information efficiently using transformer-based approaches in the context of mobile devices. Most of the standard encoders like ResNet [10] are fully convolutional and do not utilize the benefits of Vision Transformers. The resulting feature map is further upscaled and integrated with the MobileVit layer outputs. To the best of our knowledge, this is the first paper that uses MobileVit as the encoder for the single image dehazing problem. The model has a total of 2.29 million parameters, positioning it as a lightweight solution without compromising its performance.

3.1 Encoder

The encoder extracts the feature map from the input image. For this extraction to be possible the classification layer of MobileVit was deleted. MobileViT block combines CNN (local information) and transformers (global information). It uses convolutions for local details, then transforms patches to capture relationships between image parts. Four intermediate blocks and the output of the Encoder are used for feature extraction, each capturing different aspects of the input image. As shown in Fig.1 each of these blocks (light blue color), has a different width, height and channel values.

- Block 1: $\frac{1}{2} \times \frac{1}{2} \times C1$
- Block 2: $\frac{1}{4} \times \frac{1}{4} \times C2$
- Block 3: $\frac{1}{8} \times \frac{1}{8} \times C3$
- Block 4: $\frac{1}{16} \times \frac{1}{16} \times C4$
- Out: $\frac{1}{32} \times \frac{1}{32} \times C5$

As these stages progress, the feature maps become smaller but contain richer information. This helps in the understanding of the relationships between different parts of the image.

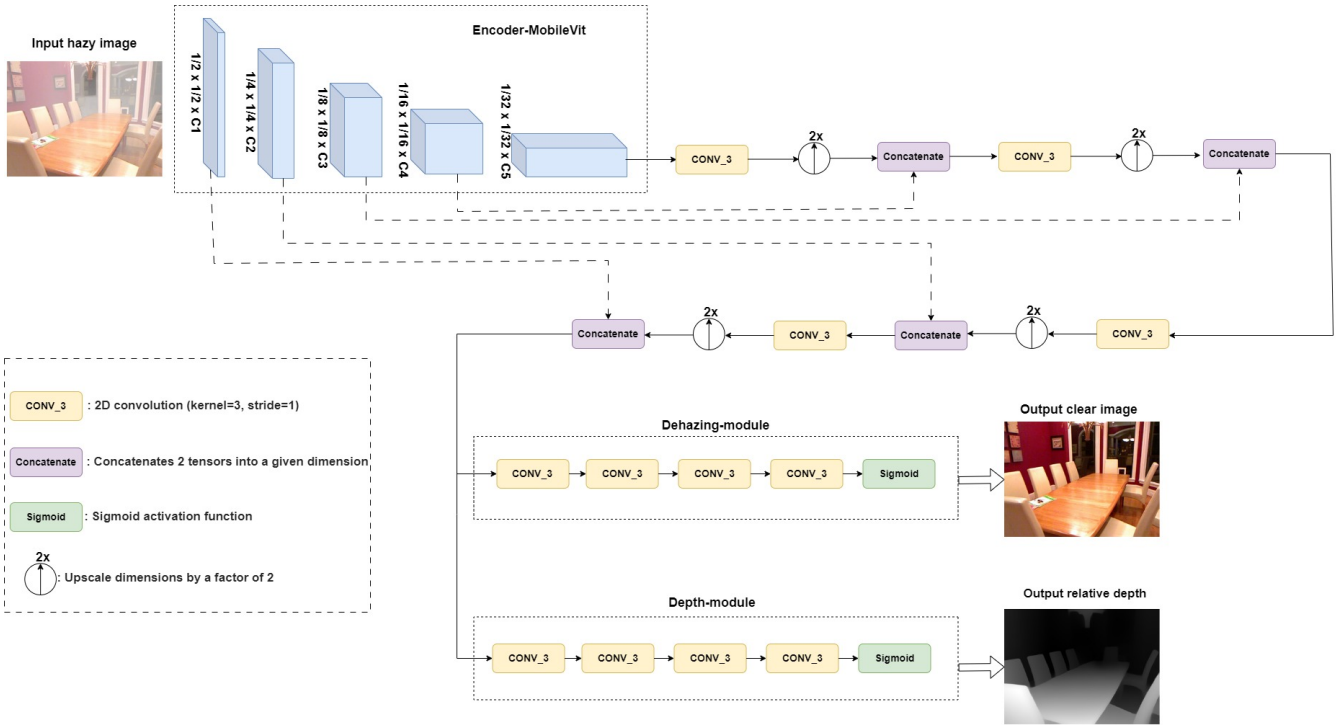


Figure 1: Model used for image dehazing utilizing relative depth. Each feature extraction block of MobileVit is concatenated with the output at different stages of the decoder. All the convolutions have kernel size=3 and stride=1.

3.2 Decoder

The decoder is the part of the network where the information provided from the features extracted from the encoder is used to reconstruct both the clear image and the relative depth map. The features from the MobileVit blocks are passed through a series of convolution, upsampling, and concatenation. As depicted in Fig.1 the first input of the decoder is the output of MobileVit with dimensions: $\frac{1}{32} \cdot H \times \frac{1}{32} \cdot W \times C5$. After undergoing a 3×3 convolution, the output is upsampled using Bilinear Interpolation. This enables the concatenation of features with those extracted from Block 4, as they share the same dimensions, specifically $\frac{1}{16} \cdot H \times \frac{1}{16} \cdot W$ of the original input. This process iterates through all blocks until the output matches the dimensions of the input hazy image ($H \times W$). The rationale behind this approach lies in the simultaneous presence of dehazing and depth feature information, which mutually reinforce each other, thereby enhancing overall performance.

After that, the decoder is split into 2 modules, a dehazing-module and a depth-module. The dehazing module utilizes this mixed information to refine its image reconstruction process, ensuring that the final output is visually coherent and faithful to the input. Similarly, the depth module benefits from this combined information to achieve more discernible depth cues.

3.3 Loss Function

Let y_{gt} (ground truth) be the clear image, y_p the predicted dehazed image, d_{gt} (ground truth) relative depth values from Depth Anything and d_{pred} the predicted depth values from the model. For single image dehazing $L1$ loss was chosen and for depth estimation a combination of 2 losses was implemented Structural Similarity ($SSIM$) [30] and $L1$ loss.

The dehazing and depth loss were combined to get the final loss function.

$$\mathscr{L}_{\text{Haze}}(y_p, y_{gt}) = |y_p - y_{gt}| \tag{2}$$

$$\mathscr{L}_{\text{Depth}}(d_p, d_{gt}) = \alpha \cdot \mathscr{L}_{\text{SSIM}}(d_p, d_{gt}) + \beta \cdot |d_p - d_{gt}| \tag{3}$$

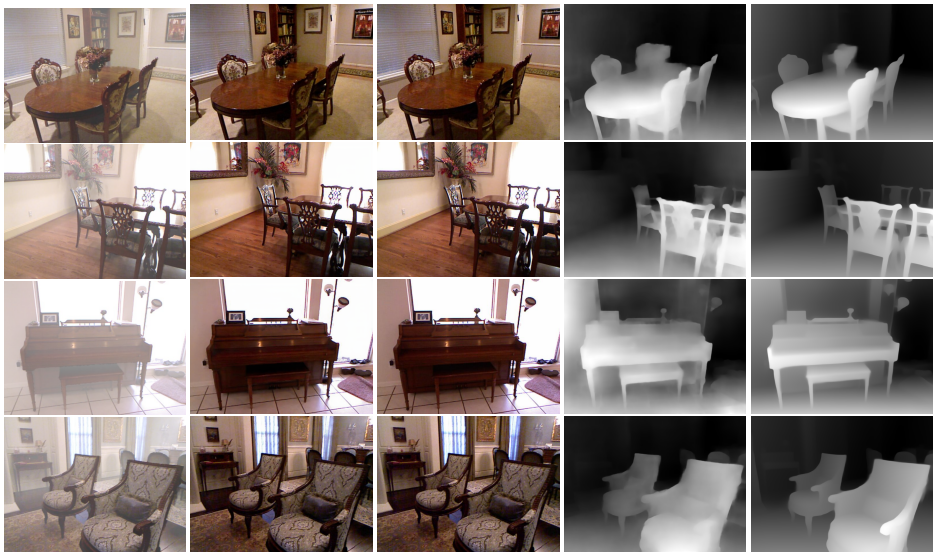
$$\begin{aligned} \mathscr{L}_{\text{Combined}}(y_p, d_p, y_{gt}, d_{gt}) = \gamma \cdot \mathscr{L}_{\text{Haze}}(y_p, y_{gt}) \\ + \delta \cdot \mathscr{L}_{\text{Depth}}(d_p, d_{gt}) \end{aligned} \tag{4}$$

4 EXPERIMENTS

In this section, the evaluation of the proposed model is conducted using standard benchmarks for image dehazing. The presented implementation is compared with existing lightweight and heavyweight models.

Method	RESIDE-IN		RESIDE-OUT		RESIDE(IN+OUT)		Inference (ms)
	PSNR(M)	SSIM	PSNR(M)	SSIM	PSNR(M)	SSIM	
DCP [9] (2010)	16.627	0.818	19.13	0.815	17.875	0.816	-
MSCNN [26] (2016)	-	-	-	-	17.57	0.8125	-
AOD-Net [18] (2017)	20.51	0.816	24.14	0.920	22.325	0.868	-
FFA-NET [24] (2021)	36.39	0.9886	33.57	0.9840	34.98	0.8963	310.15
Light-DehazeNet [29] [29] (2021)	-	-	-	-	28.39	0.9487	9.28
MixDehazeNet-S [20] (2023)	39.47	0.995	35.09	0.985	37.28	0.99	131.62
Proposed Model	30.339	0.965	-	-	-	-	43.01

Table 1: Performance evaluation of the models on RESIDE-SOTS dataset [19].



(a) Hazy Image (b) Dehazed (c) Clear (d) Depth (e) GT depth [33]

Figure 2: Qualitative analysis on RESIDE-SOTS indoors dataset for both dehazing and depth [19]. The proposed model takes four hazy images as input. Ground truth (GT) depth is determined using DepthAnything outputs.

4.1 Implementation details

PyTorch [23] was the framework used for the implementation. A training regimen comprising 40 epochs was adopted, with an initial learning rate of 3×10^{-4} for the first 20 epochs, followed by a reduction to 3×10^{-5} for the next 20 epochs. Adam optimizer with default settings [15] and a batch size of 4 was utilized for both models.

4.2 Datasets

RESIDE [19] provides a comprehensive collection of hazy images, encompassing both real-world and synthetic scenes. Three subsets of the dataset were utilized: RESIDE-IN(ITS), comprising 13,990 hazy images along with their corresponding clear counterparts from indoor environments, RESIDE-OUT(OTS), comprising 50,874 hazy images and their corresponding clear images captured in outdoor settings, and Synthetic Objective Testing task (SOTS), consisting of 1000 clear images from indoor and outdoor scenes, each paired with its hazy counterpart.

For every clear image Depth anything [33] was employed to extract the relative depth map, serving as the ground truth value.

Two separate models were trained, one using the images from the indoor scenes and the other with the images from the outdoor scenes. For each model there are two ground truth values, the clear image and the relative depth map and one input value which is the hazy image.

Evaluation metrics: The standard evaluation metrics employed for single image dehazing are utilized.

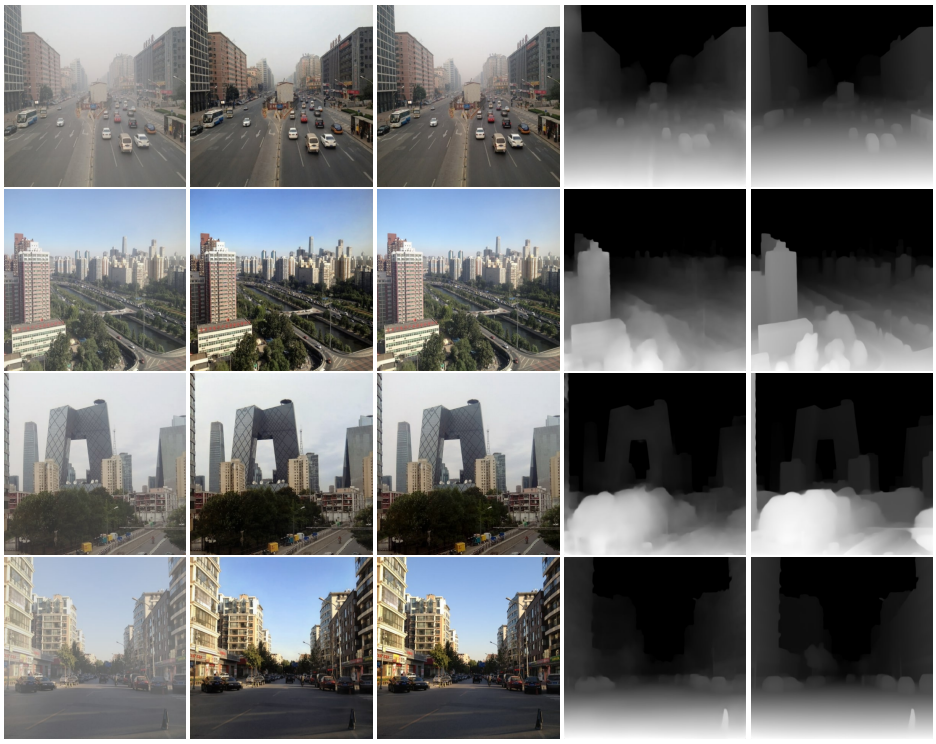
- Structural Similarity Index(SSIM):

$$\frac{(2\mu_{\hat{y}}\mu_y + C_1) + (2\sigma_{\hat{y}y} + C_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{y}^2} + \sigma_y^2 + C_2)} \tag{5}$$

- Peak Signal-to-Noise Ratio (PSNR):

$$20 \cdot \log_{10} \frac{1}{RMSE} \tag{6}$$

where $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$



(a) Hazy Image (b) Dehazed (c) Clear (d) Depth (e) GT depth [33]

Figure 3: Qualitative analysis on RESIDE-SOTS outdoors dataset for both dehazing and depth [19]. The proposed model takes four hazy images as input. Ground truth (GT) depth is determined using DepthAnything outputs.

4.3 Evaluation protocol

The model was tested on the indoor set which consists of 500 image pairs of RESIDE-SOTS (indoors) [19] at full resolution. The GPU used for the calculation of inference time is NVIDIA GeForce RTX 3060. The proposed model produces images at a fixed size, whereas the samples provided by RESIDE-SOTS (outdoors) exhibit varying resolutions. Consequently, direct comparison of metrics on the RESIDE-SOTS (outdoors) with other methods that output images at full resolution may lead to unfair assessments of performance.

4.4 Results

A selection of dehazed images and their corresponding relative depth values will be presented for qualitative analysis for both RESIDE-SOTS indoors Fig.2 and RESIDE-SOTS outdoors Fig.3. The implementations are compared to other models Table.1 and a quantitative comparison is provided Fig.4. While the proposed method is lightweight, the results exhibit remarkable clarity in both the dehazed and depth images, showing the efficiency and robustness of the approach. The qualitative evaluation with the heavyweight models reveals minimal disparities, whereas with lightweight models, the distinctions are prominently noticeable. In comparison to the implementations detailed in Table 1, the proposed model demonstrates better performance in terms of inference time, with the exception of Light-DehazeNet. Notably, the proposed architecture stands out as the sole model providing the relative depth map, a factor that influences inference time.

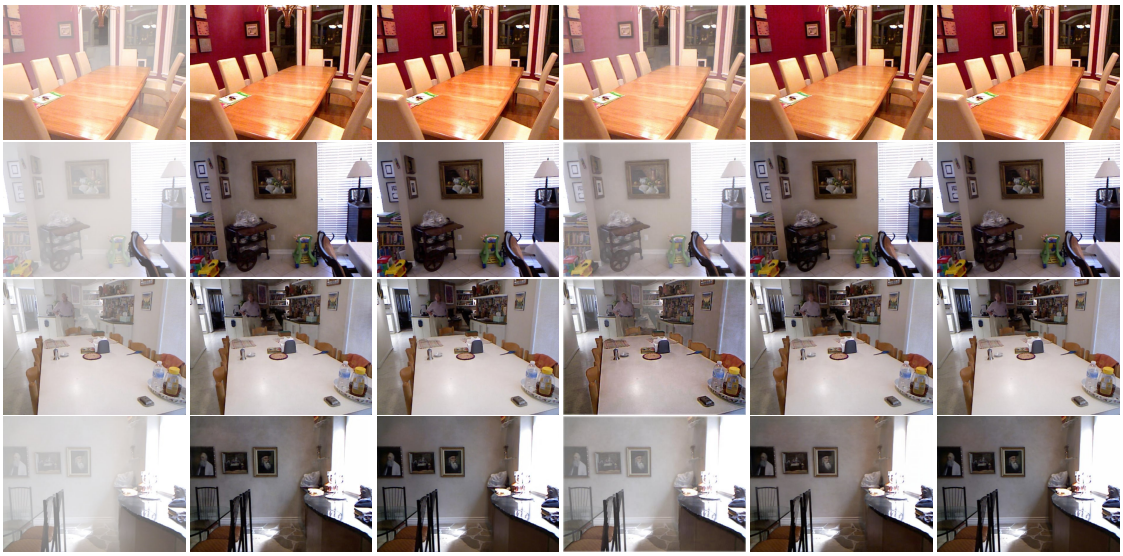
4.5 Ablation study

To showcase the efficacy of the proposed architecture, an ablation study was conducted to analyze the number of need blocks of MobileVit and the loss function.

The model underwent testing where each of the four intermediate blocks of the Encoder was systematically omitted, allowing for an assessment of their individual impact on performance. Additionally, an evaluation without any of the intermediate blocks is provided. As shown in Table 2, the removal of any one of these four intermediate blocks yields negative effects on performance. The first block exhibits the most pronounced impact, while the fourth block shows the least. When

Blocks	PSNR	SSIM
1,2,3	29.958	0.963
1,2,4	30.195	0.963
1,3,4	30.039	0.962
2,3,4	28.373	0.931
No Blocks	22.179	0.691

Table 2: Performance Impact of removing intermediate encoder blocks



(a) Hazy Image (b) FFA-NET (c) MixDehazeNet (d) LightDehazeNet (e) Proposed model (f) Clear

Figure 4: Qualitative analysis on RESIDE-SOTS indoors [19].

all four blocks are discarded, the performance drops significantly, underscoring the value they provide.

We also conducted tests on the constants within the loss function (described in Table 3), which combines the haze loss with the depth loss.

γ	δ	PSNR	SSIM
0.5	0.5	29.02	0.958
0.6	0.4	29.572	0.960
0.8	0.2	30.339	0.965
0.4	0.6	28.56	0.947
0.2	0.8	28.05	0.941

Table 3: Impact of constants in the loss function

5 CONCLUSION

In conclusion, this paper presents a novel lightweight architecture tailored for single image dehazing, leveraging a MobileVit encoder for efficiency and speed, alongside a fully convolutional lightweight decoder featuring skip connections for enhanced feature extraction. By integrating depth estimation into the dehazing task within a single network in a supervised manner, depth information aids in scene understanding. Moreover, the shortage of datasets providing both dehazing and relative depth ground truths is overcome by employing state-of-the-art networks like Depth Anything for relative depth extraction. The main contribution lies in proposing a lightweight solution for image dehazing, utilizing MobileVit in the encoder, incorporating relative depth values to empower dehazing, and employing a fully convolutional decoder with skip connections for efficient and accurate processing. The relative depth output can be leveraged in various other computer vision tasks to enhance their performance and robustness. By providing depth information alongside dehazed images, the proposed model not only improves visual clarity but also enriches the data available for downstream tasks, thus contributing to more accurate and comprehensive computer vision solutions. Overall, this framework offers a promising implementation for enhancing scene understanding in challenging environmental condition

6 ACKNOWLEDGMENTS

This research has been supported by the European Commission funded program RESCUER, under H2020 Grant Agreement 101021836.

7 REFERENCES

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[2] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing*, 25(11):5187–5198, 2016.

[3] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13001–13011, 2023.

[4] Raul de Queiroz Mendes, Eduardo Godinho Ribeiro, Nicolas dos Santos Rosa, and Valdir Grassi Jr. On deep learning techniques to boost monocular depth estimation for autonomous navigation. *Robotics and Autonomous Systems*, 136:103701, 2021.

[5] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale

boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[8] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5812–5820, 2022.

[9] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462–3471, 2020.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[13] Hyungjoo Jung, Youngjung Kim, Dongbo Min, Changjae Oh, and Kwanghoon Sohn. Depth prediction from a single image with conditional adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1717–1721. IEEE, 2017.

[14] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022.

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

[16] Dong-Jae Lee, Jae Young Lee, Hyunguk Shon, Eojindl Yi, Yeong-Hun Park, Sung-Sik Cho, and Junmo Kim. Lightweight monocular depth estimation via token-sharing transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4895–4901. IEEE, 2023.

[17] 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.

[18] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017.

[19] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018.

[20] LiPing Lu, Qian Xiong, DuanFeng Chu, and BingRong Xu. Mixdehazenet: Mix structure block for image dehazing network. *arXiv preprint arXiv:2305.17654*, 2023.

[21] Earl J McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York*, 1976.

[22] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[24] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11908–11915, 2020.

[25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[26] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 154–169. Springer, 2016.

[27] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018.

[28] Michael Rudolph, Youssef Dawoud, Ronja Güldenring, Lazaros Nalpantidis, and Vasileios Belagiannis. Lightweight monocular depth estimation through guided decoding. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2344–2350. IEEE, 2022.

[29] Hayat Ullah, Khan Muhammad, Muhammad Irfan, Saeed Anwar, Muhammad Sajjad, Ali Shariq Imran, and Victor Hugo C de Albuquerque. Light-dehazenet: a novel lightweight cnn architecture for single image dehazing. *IEEE transactions on image processing*, 30:8968–8982, 2021.

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

[31] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019.

[32] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021.

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

[34] Dong Zhao, Jia Li, Hongyu Li, and Long Xu. Complementary feature enhanced network with vision transformer for image dehazing. *arXiv preprint arXiv:2109.07100*, 2021.