DustNet: Attention to Dust

Andreas Michel[1,2][0009−0003−1789−9104], Martin Weinmann[2][0000−0002−8654−7546], Fabian Schenkel[1,2], Tomas Gomez[3], Mark Falvey[3], Rainer Schmitz[3], Wolfgang Middelmann[1], and Stefan Hinz[2][0000−0002−7323−9800]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany
andreas.michel@iosb.fraunhofer.de
[2] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany
[3] Meteodata, Chile

Abstract. Detecting airborne dust in common RGB images is hard. Nevertheless, monitoring airborne dust can greatly contribute to climate protection, environmentally friendly construction, research, and numerous other domains. In order to develop an efficient and robust airborne dust monitoring algorithm, various challenges have to be overcome. Airborne dust may be opaque as well translucent, can vary heavily in density, and its boundaries are fuzzy. Also, dust may be hard to distinguish from other atmospheric phenomena such as fog or clouds. To cover the demand for a performant and reliable approach for monitoring airborne dust, we propose DustNet, a dust density estimation neural network. DustNet exploits attention and convolutional-based feature pyramid structures to combine features from multiple resolution and semantic levels. Furthermore, DustNet utilizes highly aggregated global information features as an adaptive kernel to enrich high-resolution features. In addition to the fusion of local and global features, we also present multiple approaches for the fusion of temporal features from consecutive images. In order to validate our approach, we compare results achieved by our DustNet with those results achieved by methods originating from the crowd-counting and the monocular depth estimation domains on an airborne dust density dataset. Our DustNet outperforms the other approaches and achieves a 2.5% higher accuracy in localizing dust and a 14.4% lower mean absolute error than the second-best approach.

Keywords: Dust Monitoring · Visual Regression · Deep Learning.

1 Introduction

Monitoring airborne dust emissions is a valuable and important task since airborne dust significantly affects climate, human health, infrastructure, buildings, and various socio-economic sectors. The emergence of airborne dust particles

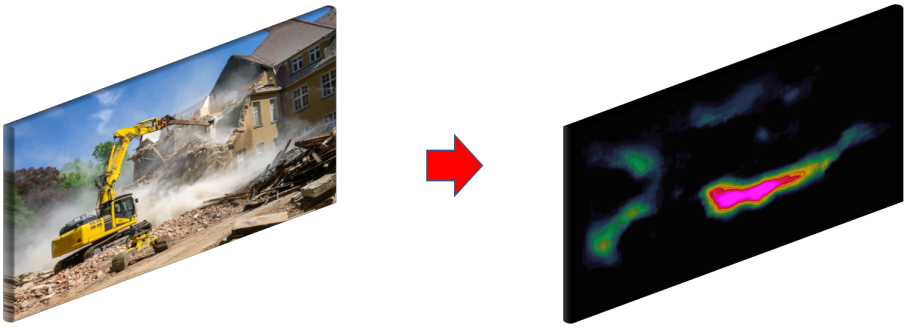


Fig. 1: **DustNet objective.** The objective of our method consists in estimating the dust level in a given RGB image or sequence. The left part shows a scene of a construction site [12], while the right part shows a dust density map of the scene.

can occur due to natural phenomena such as strong winds, wildfires, and seismic activities but may also be caused by human activity. Typical anthropogenic pollution sources are construction, traffic, or mining sites. Although completely mitigating dust emissions is not feasible, suppressing emissions with focused measures is possible. This e.g. includes watering untreated roads, speeding down vehicles, or reducing mining activities. However, optimizing dust mitigation strategies would require practical and economic dust emission monitoring. Most conventional instrument-based *in-situ* monitoring equipment focuses on identifying the impact of dust emissions but is limited in attributing responsibilities. On the other hand, remote sensing 3D dust scanning technologies, such as lidar, are not economically feasible on a large scale and may produce noisy and hard-to-interpret data in complex terrains. Hence, visual monitoring via camera-based systems would be preferable for identifying airborne dust emissions. However, visual dust density estimation is still underexplored. One of the possible reasons for the scarcity of research in this domain is that detecting dust in an image is a highly ill-posed problem. There is a multitude of issues why detecting dust is so hard: dust may vary significantly in terms of density, and it can be opaque as well as translucent. Furthermore, dust density variation can be very imbalanced. For example, in dry regions, the dense dust of dust storms will emerge less frequently than transparent dust due to low winds and only sporadically occur during particular meteorological configurations. Also, dust can be emitted at a wide range of locations and for various reasons. The transparency of dust implies that the appearance of dust is easily affected by environments, and its boundaries are usually fuzzy. As a result, images with dust appear partially blurry and usually have low spatial contrast. Classical algorithms usually cannot exploit these partial blur effects because other atmospheric effects like fog or clouds can cause similar effects. Also, for humans, detecting dust in an image sequence is much easier than in a single image, but exploiting temporal data by an algorithm is challenging. For example, moving clouds, vehicles, or shadows can easily distort results of optical flow-based methods. In addition, there is no

clear color scheme for dust. Opaque dust can show a brownish color like in a dust storm, but it can also have a black shade in a mining explosion. Overall, the aforementioned properties of dust indicate the need for a more sophisticated approach.

In the last decade, deep learning has had huge success in various tasks like classification [15], object detection [30], neural linguistic processing [35], and remote sensing [41]. However, airborne dust monitoring, obstructed by the aforementioned challenges, is not well researched, and most scientific papers focus on satellite images [16], or related tasks like smoke binary segmentation [37]. Recently, De Silva et al. published a binary dust segmentation dataset called URDE [7]. While this can be seen as a first important step towards dust monitoring, we believe that a regression approach could be more beneficial. In contrast to semantic segmentation, which predicts a label on a per-pixel basis, the continuous range of dust densities rather suits a regression strategy. Furthermore, the vague boundaries of dust make it challenging to create discrete hard labels.

Accordingly, this work focuses on density estimation (see Fig. 1) and thus is most related to DeepDust [27]. DeepDust estimates density maps on single images. In order to detect dust, it exploits multiscale feature maps by utilizing feature pyramid network [20] (FPN) structures. In contrast to that work, we also address the fusion of temporal information. Furthermore, we also exploit attention strategies and rely not only on a strictly convolutional method.

In order to validate the effectiveness of our approach, we compare achieved results to those of visual density estimation techniques originating from other domains, including monocular depth estimation (MDE) and crowd counting. MDE is the task of estimating the scene depth on a per-pixel basis, whereas crowd counting is the task of approximating the number of people in a given image. Though both tasks differ strongly from dust density estimation, our method is heavily influenced by ideas of both domains. In summary, the main contributions of this work are the following: (1) We research the underexplored field of airborne dust density estimation and propose various neural network architectures. (2) Our proposed neural networks combine attention-based and convolutional-based FPN structures to merge local and global features. (3) Our work addresses the fusion of temporal features in the field of dust density estimation. (4) In order to demonstrate the effectiveness of the proposed neural network architectures, we compare the achievements by our novel techniques with those of methods originating from the crowd counting and MDE domains on the Meteodata dust dataset.

2 Related Work

In this section, we briefly summarize related work with a focus on vision transformers, crowd counting and MDE.

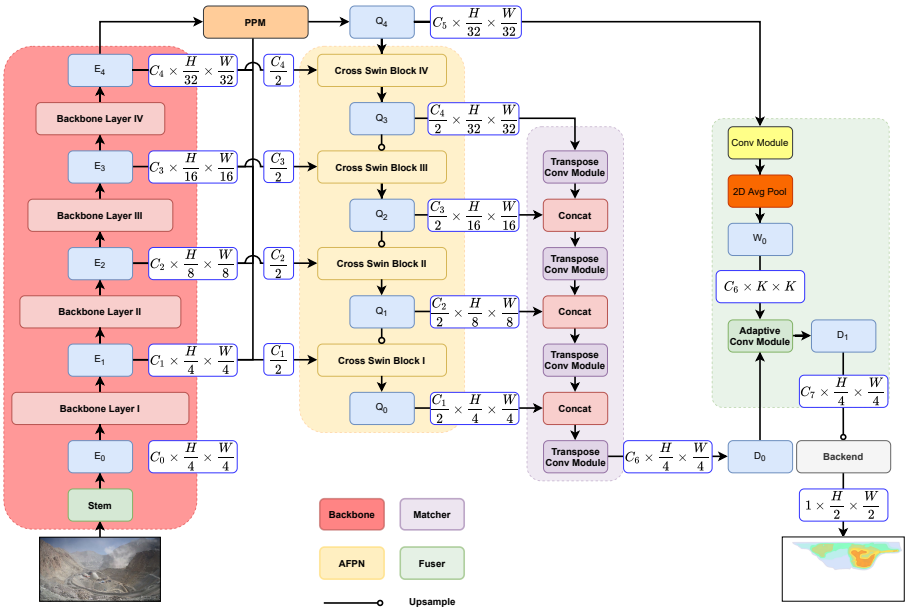


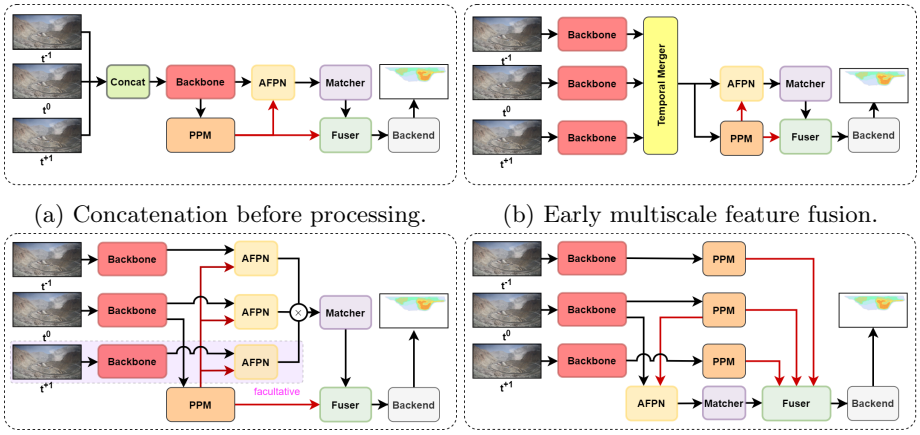
Fig. 2: **Overview of our proposed DustNet.** The basic blocks are a backbone, the AFPN, the matcher, the PPM, the fuser, and the backend.

2.1 Vision Transformer

Vaswani et al. [35] introduced the transformer architecture in 2017 in the field of natural language processing (NLP). The centerpiece of the transformer architecture is the multi-head attention module. Inspired by the success of the transformer architecture in NLP, the vision transformer (VIT) [8] introduced the transformer encoder successfully in the vision domain. In order to improve the performance of vision transformer especially in high-resolution settings, Liu et al. [23] introduced a hierarchical architecture, the swin transformer (Swin), which utilizes shifted windows to achieve a linear computational complexity. The attention is calculated only between patches, which are part of a specific window. The windows are shifted to create connections between features of the previous windows. The second version of the swin transformer [22] improves this approach further by an alternative positional encoding scheme and replacing the scaled dot attention with cosine attention, which performs better in higher resolutions.

2.2 Crowd Counting

Density estimation methods [39,31,19,21,26] have been used successfully in crowd counting. The objective of crowd counting is to predict a coarse density map of the relevant target objects, e.g. usually people. The ground truth is generated by smoothing center points with a multi-dimensional Gaussian distribution. Recent approaches are focused on increasing the spatial invariance [26] or dealing



(a) Concatenation before processing.

(b) Early multiscale feature fusion.

(c) Late multiscale feature fusion.

(d) Adaptive global feature fusion.

Fig. 3: **Different fusion approaches of DustNet.** The fusion strategy spans from early (a) to late (d) feature merging of each consecutive image.

with noise in the density maps [6]. Most works are designed for individual images, but Avvenuti et al. [2] take advantage of the temporal correlation between consecutive frames in order to lower localization and count error.

2.3 Monocular Depth Estimation

The first CNN-based method for monocular depth estimation was presented by Eigen et al. [9]. They utilized global and local information in order to predict a depth image from a single image. Further improvements of the pure CNN approaches focus on Laplacian pyramids [32], multi-scale convolutional fusion [36], structural information [17], the exploitation of coplanar pixels [28] to improve the predicted depth, reformulation of the depth prediction task as a classification-regression problem [11] and hybrids between CNN and vision transformer-based architectures [8]. Recently, the PixelFormer architecture [1] combines transformer architectures with an adaptive bin center approach inspired by [3] and adds skip connection modules to improve the feature flow between different encoder feature levels.

3 Method

In the following, we introduce our proposed DustNet architecture illustrated in Fig. 2. After presenting the submodules, we focus on the different temporal fusion approaches illustrated in Fig. 3.

3.1 Network Structure

Overview. DustNet processes input image sequences X of the dimensions $T \times H \times W \times 3$ to a continuous dust density map $\frac{H}{2} \times \frac{W}{2} \times 1$. T may consist of a

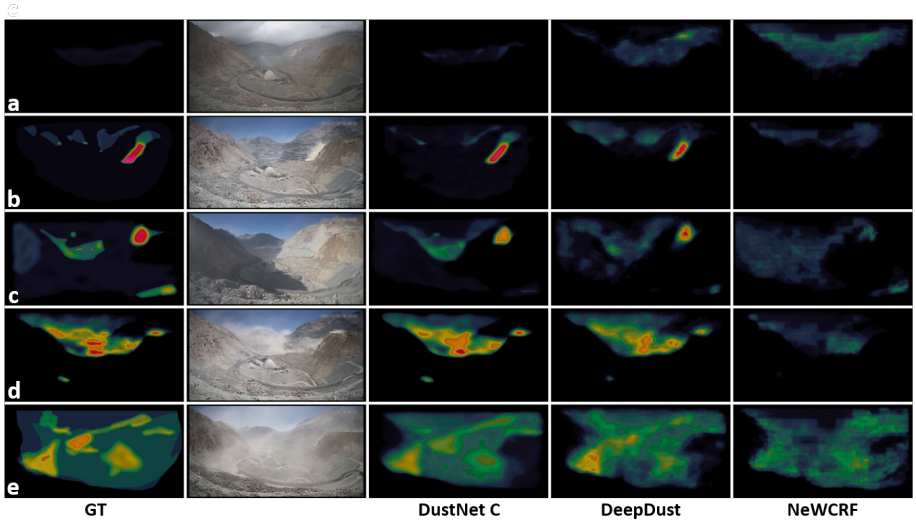


Fig. 4: **An opencast mine scene with varying dust densities from the Meteodata dust dataset.** Overall, the dust densities increase from **a** to **e**. Our proposed model DustNet C can well distinguish between clouds and dust, regress different dust levels, and has a low false positive rate. Overall, DustNet outperforms the other methods.

maximum of three consecutive images, where the target y is assigned to the image x_{t^0}. The images are fed into the backbone, which produces multiple feature maps with decreasing resolution and ascending information aggregation. The backbone features are passed to a pyramid pooling module (PPM) [40] and the attention feature pyramid network (AFPN). Hereby, in order to reduce the computational complexity, only half of the channels of feature maps are transferred to the AFPN. The PPM head aggregates global information fed into the AFPN and the Fuser module. The AFPN mixes the feature maps of different resolutions and information aggregation levels. The processed feature maps are transferred to the matcher module, accumulating the features maps into one high-resolution map. Then, the high-resolution features are merged with the global information features aggregated from the PPM head in the fuser module. Eventually, the combined features are processed by the backend, which consists of multiple sequences of CNNs, into a dust map.

Backbone. The backbone consists of a stem module and four blocks. We use this common backbone scheme in order to leverage pre-trained neural networks. We prefer a convolutional backbone like ResNet [13] instead of a transformer backbone due to the requirement to process high-resolution images. The backbone produces multiple feature maps with the resolution scales $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of the original images with the number of channels C of $\{256, 512, 1024, 2048\}$.

Pyramid pooling module. We utilized a PPM head [40] like in [38,1] to aggregate global information of the whole image. We use global average pooling

of scales $\{1, 2, 3, 6\}$ to extract the information. After extracting the features, we concatenate them and process them by a convolutional layer to the feature map Q_4 with the dimension of $512 \times \frac{H}{32} \times \frac{W}{32}$.

Attention feature pyramid network. The AFPN mixes high-resolution features with low semantics with low-resolution high semantic features. But instead of a traditional FPN like [20] utilizing CNNs, we are inspired by [1] and use four Swin blocks with cross window attention to improve the feature flow between the feature map layers. However, instead of applying scaled dot attention, we utilize cosine attention similar to [22]. This lead to increased performance in higher resolutions. The key and value matrix inputs are derived from the backbone feature maps, but to reduce computational complexity, we transfer only half of the channels. The query matrix is filled by the output of the upsampled stage before. The query matrix with the coarsest resolution originates from the global aggregated features of the PPM head.

Matcher. The matcher module also has an FPN-like architecture [20]. We upsample feature maps from the AFPN via a transpose convolution module (TCM). It consists of a 2D TCM with a stride and kernel size of two, followed by batch normalization [14] and a SiLU [10] activation function. Like [18] suggests, we apply only batch normalization without dropout [33]. The coarsest resolution feature map derived from the AFPN is fed to the first TCM block. The following AFPN feature maps are respectively concatenated to the output of the TCM block and processed via the next TCM block. The output of the matcher module has the dimension of $C_6 \times \frac{H}{4} \times \frac{W}{4}$. We choose a channel number C_6 of 256.

Fuser. The fuser module processes the high-resolution features D_0 by leveraging the aggregated features Q_4 of the PPM head. Q_4 is fed into a 2D pointwise convolutional kernel, followed by a SiLU activation function, and pooled by global average pooling to a feature map W_0 of the dimension $C_6 \times K \times K$. W_0 serves as an adaptive kernel for the adaptive convolutional layer [34], which enriches the feature map D_0 from the matcher with global information.

Backend. The backend consists of N blocks of a sequence of a 2D convolution layer, batch normalization, and SiLU activation functions that predict the dust map. We branch the features into two parallel blocks for each stage and accumulate the outputs. Hereby, we choose a dilation of three for one branch to increase the receptive field. After four stages, a pointwise convolutional layer predicts the dust maps.

3.2 Temporal Fusion

In order to leverage the temporal information between consecutive images, we developed and studied different approaches. Fig. 3 illustrates the different fusion strategies. Hereby the fusion strategy spans from early to late fusion.

Concatenate images. An obvious way to concatenate the images to $D \times H \times W$ is where the product of the number of images T and the number of channels C is the new channel dimension D. This approach is illustrated in Fig. 3a. Examples of this approach can be found in [5] or [24]. Hereby the backbones are usually specially adapted to 3D input.

Fig. 5: **Display of the generalization ability of DustNet.** The shown results of DustNet are produced on the URDE validation RandomDataset897 [7]. Hereby, DustNet S is trained on the Meteodata dust dataset and applied on the URDE dataset without finetuning.

Early multiscale feature fusion. Fig. 3b shows the methodology behind this fusion aspect. Features from three backbones, which share weights, are fed into the temporal merger (TM) neural network. TM subtracts the feature maps of image $x_{t^{-1}}$ and $x_{t^{+1}}$ respectively from x_{t^0} and multiplies the difference. We pass the new feature maps through a 2D pointwise convolutional layer and add skip connections from the feature maps of the image x_{t^0} to the output.

Late multiscale feature fusion. This approach represents a simple fusion of AFPN features (see Fig. 3c). Backbone and AFPN weights are shared between the instances. The PPM head is only fed with the backbone features from image x_{t^0}. In order to reduce the computational complexity and avoid convergence problems, only two consecutive images may be used.

Adaptive global information feature fusion. The goal hereby is to calculate the global aggregated features from a PPM head for each image. Backbone and PPM weights are shared. The local feature branch is only fed with the multiscale backbone features from image x_{t^0}. The fusion of the temporal information occurs in the fuser module. For each PPM head, an adaptive convolutional layer is added.

Table 1: Comparison of the best-performing density estimation methods on the Meteodata dust dataset.

	Params	MeM	Time	MAE	MSE	Acc	Pre	Rec
CanNet	**18 M**	**4.8 GB**	**7.1 ms**	20.21	855.40	0.79	0.80	0.79
DeepDust	101 M	32.8 GB	20.3 ms	19.60	749.36	0.78	0.80	0.79
PixelFormer	140 M	15.2 GB	43.3 ms	21.53	825.50	0.78	0.81	0.79
NeWCRF	140 M	16.8 GB	36.7 ms	20.00	822.68	0.78	0.80	0.78
DustNet S	64 M	8.8 GB	21.6 ms	19.27	705.47	0.80	0.81	0.80
DustNet A	65 M	8.8 GB	23.5 ms	18.77	701.73	0.79	0.81	0.80
DustNet B	67 M	9.6 GB	46.4 ms	26.60	1528.08	0.67	0.75	0.68
DustNet C	68 M	12.8 GB	37.8 ms	**16.77**	**601.49**	**0.81**	**0.83**	**0.82**
DustNet D	86 M	10.4 GB	45.0 ms	17.44	639.10	**0.81**	**0.83**	0.81

4 Experimental Results

In this section, we explain our experiment's implementation details, present and discuss our achieved results, and finally detail our ablation study and the limitations of our approach.

4.1 Dataset

We conduct experiments on the extended Meteodata dust dataset [27]. The extended dataset includes a variety of scenes from opencast mines with a wide range of dust levels, lighting conditions, and cloud conditions. The ground truth is aimed at mimicking the human perception of dust in a given image regarding opaqueness and estimated dust density levels. The dataset comprises 2298 consecutive RGB image triplets with a 1000×1920 pixels resolution. An image triplet consists of three consecutive images with a ground truth dust density map for the enclosed image. The average time between two consecutive images amounts to ten seconds. The pixels of the ground truth span are mapped to an 8-bit unsigned integer datatype, where the pixel values are proportional to the dust density. We split the dataset into a training dataset with 1906, a validation dataset with 144, and a test dataset with 248 image triplets. The challenges of this dataset are manifold. The high resolution of the images is a high computational burden.

Furthermore, the strong variance in dust levels, in combination with the highly imbalanced frequency of the different dust levels, complicates the estimation of the different dust levels. In order to display the generalization ability of our approach and the dust dataset, we also conduct a qualitative analysis on the URDE dataset [7]. It consists of images with a size of 1024×1024 pixels and contains scenes on dusty roads.

Table 2: Binned regression results of density estimation methods on the Meteodata dust dataset: The values of the pixels are binned into zero dust (ZB), low dust (LB), medium dust (MB), and high dust (HB) density. Overall, our proposed DustNet C outperforms the other methods.

	MAE					MSE				
	ØB	**ZB**	**LB**	**MB**	**HB**	**ØB**	**ZB**	**LB**	**MB**	**HB**
CanNet	38.08	12.04	23.61	42.65	74.01	2917.5	376.6	856.4	2540.3	7896.8
DeepDust	30.11	12.18	23.93	37.04	47.28	1640.2	367.5	873.3	1979.9	3340.2
PixelFormer	40.68	14.19	25.00	36.05	87.48	2864.6	377.1	892.5	1863.6	8325.3
NeWCRF	34.19	10.88	25.24	40.82	59.81	2255.0	324.9	954.5	2371.2	5369.3
DustNet S	31.35	12.41	22.35	39.45	51.20	1756.6	340.8	754.7	2152.6	3778.1
DustNet A	32.21	10.36	23.17	40.61	54.68	1837.3	281.8	795.0	2195.5	4076.9
Dustnet B	64.05	**7.95**	34.85	77.80	135.61	6948.3	**151.1**	1553.7	6698.8	19389.4
DustNet C	**27.29**	8.70	**21.97**	**33.29**	**45.19**	**1361.7**	257.4	**738.9**	**1579.2**	**2871.3**
DustNet D	31.13	8.24	22.36	41.59	52.33	1767.1	189.4	742.9	2295.1	3841.0

Table 3: Results of the ablation study on the Meteodata dust dataset: The base model is DustNet C with two consecutive images as inputs. The modules following a ✗ are replaced.

	Params	**MeM**	**Time**	**MAE**	**MSE**	**Acc**	**Pre**	**Rec**
1x Img Input	64 M	8.8 GB	**21.6 ms**	19.27	705.47	0.80	0.81	0.80
2x Img Input	68 M	12.8 GB	37.8 ms	**16.77**	**601.49**	0.81	**0.83**	0.82
3x Img Input	68 M	16.0 GB	64.6 ms	17.83	675.34	0.81	0.82	0.82
✗ AFPN	**32 M**	**7.2 GB**	27.3 ms	19.85	805.74	0.79	0.80	0.79
✗ Fuser	68 M	12.8 GB	38.0 ms	19.63	754.25	0.80	0.81	0.80
✗ Matcher	64 M	12.0 GB	29.5 ms	17.08	651.39	**0.83**	**0.83**	**0.83**

4.2 Benchmark Selection

Temporal dust density estimation is not well-researched. To our best knowledge, only DeepDust [27] focuses on a directly related task. Methods like temporal density estimation methods from other domains, like crowd counting [2], are not designed for a large time lag between two consecutive images and high-resolution scenes and therefore have convergence problems. From the methods available, we selected, in addition to DeepDust, CanNet [21], a lightweight fully convolutional crowd-counting approach and two state-of-the-art MDE models represented by NeWCRF [38] and PixelFormer [1]. For both MDE methods, we select as a backbone the base swin transformer [23] model with a window size of twelve.

4.3 Implementation Details

All experiments are conducted on four Nvidia A100 GPUs with 80 GB memory. We use l2 loss and the AdamW [25] optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a

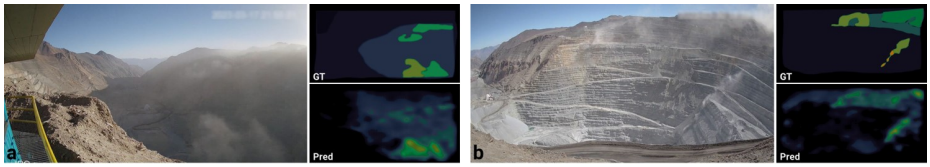


Fig. 6: **Opencast mines.** Results of DustNet C on mine sites that are not included in the training or validation dataset.

weight decay of 10^{-5}. During training, we start with a learning rate of $\alpha = 3{\cdot}10^{-4}$ and train for 50 epochs with a frozen backbone. After 50 epochs, we unfreeze the backbone and train for further 20 epochs.

Our proposed networks are trained with a ResNet101 [13] backbone. We deploy the following proposed architectures: **DustNet S** is the version of DustNet as shown in Fig. 2. It processes only a single image. **DustNet A** is equivalent to Fig. 3a. Compared to DustNet S, the capacity for input channels is increased to nine channels. **DustNet B** consists of three backbones with shared weights and a temporal merger module (Fig. 3b). **DustNet C** takes only two consecutive images as input (Fig. 3c). It merges the multiscale AFPN features. **DustNet D** consists of multiple backbones and PPM heads with shared weights (Fig. 3d). Each aggregated global information feature map from the PPM head is an adaptive weight.

4.4 Evaluation Metrics

We are interested in the localization and regression ability of our proposed models. We map all pixel values under 30 to zero and the remaining pixels to one to validate the localization aspect. This threshold was derived from the labeling process of the Meteodata dust dataset, which maps all values under 30 to zero. As a result, we can now apply standard classification metrics like accuracy (Acc), precision (Pre), and recall (Rec). In order to validate the quality of the predicted regression, we use standard metrics represented by mean absolute error (MAE) and mean squared error (MSE). Furthermore, to fairly assess the performance on tail values of imbalanced datasets, we consider the idea of balanced metrics [4]. Therefore, we bin our data into four bins: zero dust density bin (ZB), low dust density bin (LB), medium dust density bin (MB), and high dust density bin (HB). For each bin, we calculate the MAE and MSE. Following [4,29], we compute the mean across all bins and obtain the average binned mean absolute error ØB-MAE and the average binned mean squared error ØB-MSE.

4.5 Quantitative Results

Table 1 compares the best-performing density estimation methods on the Meteodata dust dataset. The best-performing model on a single image is our proposed

DustNet S, and the best-performing model for consecutive images is our DustNet C. In particular, in the case of the binned regression metrics (see Table 2), DustNet C outperforms the other methods vastly. Due to the imbalance of the dataset, a small difference in the metrics can lead to a big qualitative difference. CanNet is the most computationally efficient method tested on the Meteodata dust dataset and can classify dust decently but cannot regress well in the higher dust levels. PixelFormer performs worse regarding regression. Possible reasons include the inability of the Swin transformer backbone to process high-resolution images well and the adaptive bin center prediction module. The Swin transformer backbone and the conditional random field modules could also influence NeWCRF's performance. DustNet's performance gain over DeepDust could be caused by the improved feature flow, particularly between high- and low-level features.

4.6 Qualitative Results

Fig. 4 illustrates an opencast mining scene with varying dust densities from the Meteodata dust dataset. The overall airborne dust in the scene increases from **a** to **e**. Fig. 4**a** demonstrates the ability of DustNet to distinguish between dust, shadows, and clouds. The other methods lack the ability to differentiate in comparison to DustNet. From Fig. 4**c**, our method shows its performance on more specific singular dust plumes. Finally, in Fig. 4**d** and **e**, our method performs well in more dense dust scenes. Nearly in all cases, our model surpasses the other approaches.

Fig. 6 shows the result of DustNet C on further mining sites displayed. Here again, DustNet is able to produce good results. In order to demonstrate the generalization ability of our approach, we applied a DustNet S without retraining on the URDE dataset. Also, as seen in the images, our DustNet S displays a good performance. Due to the hard segmentation boundaries and the neglect of low dust density on the URDE dataset, a quantitative evaluation would, in our opinion, not be appropriate.

4.7 Ablation Study

In order to show the efficacy of our proposed method, we choose the best-performing proposed architecture DustNet C, change the number of input images and replace several modules (see Table 3). We replace the AFPN, the matcher, and the fuser module, respectively.

Inputs. We compare the performance change of our proposed method in varying the number of consecutive images. Backbone and AFPN weights are shared in our experiment. The use of two consecutive images outperforms the other options.

AFPN. In the AFPN ablation experiment, the AFPN is removed, and the features from the backbone are passed directly to the matcher. Removing the AFPN halves nearly the number of parameters, but it causes a big increase in MAE and MSE.

Matcher. In the matcher ablation experiment, we replace the matcher with a simple convolutional layer followed by an activation function for channel adaption. The feature map with the highest resolution from the AFPN is processed by the added convolutional layer and fed into the fuser module. The accuracy increases slightly for the tradeoff of a decreased regression ability. But the main reason for keeping the matcher is the increased differentiation ability between dust and similar visual effects like clouds. Removing the matcher leads to a significant drop.

Fuser. In the last ablation experiment, a traditional convolutional layer replaces the adaptive convolutional layer. Hence, the aggregated global information features cannot enrich the matcher's features. This leads to a significant increase in MAE and MSE.

4.8 Limitations and Future Work

Fig. 4**c** illustrates one of the shortcomings of our model. Our model can easily ignore small plumes of dust. We assume that neglecting small dust areas for the tradeoff of fewer false positives leads to a lower loss and therefore is a negative side effect of the training process. Also, we suppose that the l2 loss function leads to a worse MSE in bins with scarcer values (see Table 2). Furthermore, the ground truth of the Meteodata dust dataset has a high uncertainty compared to a classification or object detection dataset. For example, in Fig. 4**g**, we assume that our model represents the real dust conditions than the ground truth. The ground truth does not cover the small dust plume behind the truck, and overall undervalues the dust density in the dense dust plumes. Therefore a better metrical result causes not automatically a higher performance in a real-world scenario. An extensive comparison between density estimation and semantic segmentation of dust could be a valuable extension of this work. Furthermore, future work should address handling long-tailed visual regression data and increase the sensitivity for smaller dust plumes.

5 Conclusion

In this paper, we have presented DustNet, a dust density estimation neural network. DustNet computes for every pixel in a given image a dust density. Hereby DustNet exploits and fuses local, global, and temporal information. DustNet cannot only regress different dust levels but also distinguish between dust and similar visual effects like clouds. Our proposed approach outperforms a range of other approaches on the Meteodata dust dataset.

6 Acknowledgment

The images in the presented figures and those used for creating the Meteodata dust dataset are from the pit of Minera Los Pelambres, which collaborates with Meteodata in the advanced use of cameras for emission control strategies. The permission to use the images in this publication is kindly appreciated.

References

1. Agarwal, A., Arora, C.: Attention attention everywhere: Monocular depth prediction with skip attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5861–5870 (2023)
2. Avvenuti, M., Bongiovanni, M., Ciampi, L., Falchi, F., Gennaro, C., Messina, N.: A spatio-temporal attentive network for video-based crowd counting. In: Proceedings of the 2022 IEEE Symposium on Computers and Communications. pp. 1–6. IEEE (2022)
3. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proceedings of the 2010 20th International Conference on Pattern Recognition. pp. 3121–3124. IEEE (2010)
5. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
6. Cheng, Z.Q., Dai, Q., Li, H., Song, J., Wu, X., Hauptmann, A.G.: Rethinking spatial invariance of convolutional networks for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19638–19648 (2022)
7. De Silva, A., Ranasinghe, R., Sounthararajah, A., Haghighi, H., Kodikara, J.: A benchmark dataset for binary segmentation and quantification of dust emissions from unsealed roads. Scientific Data **10**(1), 14 (2023)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transattentions for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information Processing Systems **27** (2014)
10. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks **107**, 3–11 (2018)
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
12. gabort@AdobeStock: (2023), https://www.stock.adobe.com
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning. pp. 448–456 (2015)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
16. Lee, J., Shi, Y.R., Cai, C., Ciren, P., Wang, J., Gangopadhyay, A., Zhang, Z.: Machine learning based algorithms for global dust aerosol detection from satellite images: Inter-comparisons and evaluation. Remote Sensing **13**(3) (2021)

17. Lee, M., Hwang, S., Park, C., Lee, S.: Edgeconv with attention module for monocular depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2858–2867 (2022)
18. Li, X., Chen, S., Hu, X., Yang, J.: Understanding the disharmony between dropout and batch normalization by variance shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2682–2690 (2019)
19. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1091–1100 (2018)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
21. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5099–5108 (2019)
22. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transattention attention v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019 (2022)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transattention attention: Hierarchical vision transattention attention using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
24. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transattention attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211 (2022)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H.: Hybrid graph neural networks for crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11693–11700 (2020)
27. Michel, A., Weinmann, M., Schenkel, F., Gomez, T., Falvey, M., Schmitz, R., Middelmann, W., Hinz, S.: Terrestrial visual dust density estimation based on deep learning. In: Proceedings of the 2023 IEEE International Geoscience and Remote Sensing Symposium (2023)
28. Patil, V., Sakaridis, C., Liniger, A., Van Gool, L.: P3depth: Monocular depth estimation with a piecewise planarity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1610–1621 (2022)
29. Ren, J., Zhang, M., Yu, C., Liu, Z.: Balanced mse for imbalanced visual regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7926–7935 (2022)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
31. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 4031–4039. IEEE (2017)
32. Song, M., Lim, S., Kim, W.: Monocular depth estimation using laplacian pyramid-based depth residuals. IEEE Transactions on Circuits and Systems for Video Technology **31**(11), 4381–4393 (2021)

33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
34. Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., Kautz, J.: Pixel-adaptive convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11166–11175 (2019)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
36. Wang, L., Zhang, J., Wang, Y., Lu, H., Ruan, X.: Cliffnet for monocular depth estimation with hierarchical embedding loss. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V. pp. 316–331. Springer (2020)
37. Yuan, F., Zhang, L., Xia, X., Huang, Q., Li, X.: A wave-shaped deep neural network for smoke density estimation. IEEE Transactions on Image Processing **29**, 2301–2313 (2020)
38. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3916–3925 (2022)
39. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 589–597 (2016)
40. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)
41. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine **5**(4), 8–36 (2017)