Swin-AQI: Depth-Aware Vision Transformer with Physics-Informed Learning for Particulate Matter Estimation from Ground Imagery

Van Truong Thinh Nguyen [1,2], Thanh-Hai Tran [1,2], and Xuan-Bach Le [1,2⋆]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
{thinh.nguyen0334,lexuanbach}@hcmut.edu.vn

Abstract. Street-level imagery is an appealing, low-cost complement to air-quality stations, but a camera is only useful when its pixels respond to the pollutant of interest. For particulate matter they do: aerosols scatter light along the line of sight, so distant buildings, sky boundaries, and horizon regions lose contrast as aerosol loading rises. *Swin-AQI* is a depth-aware Swin Transformer for monocular $PM_{2.5}$ and PM_{10} estimation from geotagged urban images: MiDaS relative depth biases shifted-window attention toward distant tokens, learnable Fourier features encode the capture hour, and Beer–Lambert and Koschmieder terms keep predictions consistent with visibility physics. We supervise with SILAM atmospheric fields. On 1,977 daylight images from Ho Chi Minh City, Swin-AQI reaches R^2=0.80 for $PM_{2.5}$ and R^2=0.76 for PM_{10}, outperforming every CNN and transformer baseline under identical inputs. Our ablations are candid about why: hour-of-day and uncertainty weighting contribute the most, while depth-aware attention and the physics terms add smaller but separable gains. We are also deliberate about scope. Swin-AQI is a single-city, moderate-pollution estimator whose gaseous outputs stay contextual.

Keywords: Particulate matter estimation · Depth-aware vision transformer · Physics-informed learning · Street-level imagery

1 Introduction

Urban particulate exposure varies over short distances [26,17]. Reference stations are sparse, and low-cost sensors require calibration and drift management [12,16,20]. Street-level imagery offers a complementary signal because particulate matter changes image formation: aerosols scatter and absorb light, reducing far-field contrast, obscuring distant structure, and weakening sky boundaries as optical path length increases [1,11]. A visual particulate estimator should therefore be geometric as well as semantic. When haze is the evidence, distant regions should carry more weight than nearby texture.

⋆ Corresponding author

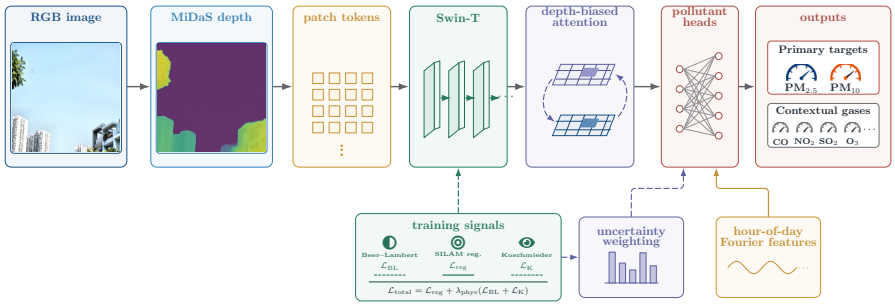

Fig. 1: Swin-AQI framework. MiDaS estimates relative scene depth, a Swin-T backbone receives depth-aware attention biases, the depth-attention features are concatenated with the encoded capture time, and the pollutant heads are trained with uncertainty-weighted regression and particulate-specific physics losses.

Existing image-based methods get part of the way there. Hand-crafted visibility features exploit contrast, edge strength, and sky appearance, but remain sensitive to illumination and scene layout [13,27]. Convolutional regressors improve robustness, yet typically learn a generic image-to-pollution mapping without explicitly privileging path-length-dependent haze evidence [28].

Figure 1 presents *Swin-AQI*, a depth-aware Swin Transformer for estimating $PM_{2.5}$ and PM_{10} from ground imagery that addresses both the modeling and the evaluation gaps identified above. The model preserves the hierarchical shifted-window design of Swin [14] but introduces a MiDaS-derived relative-depth bias that steers attention toward distant key tokens, where particulate haze is most strongly expressed. It further incorporates hour-of-day Fourier features and weak Beer–Lambert/Koschmieder consistency terms, and retains the gaseous outputs only as auxiliary contextual tasks rather than as primary RGB-sensing claims.

We make four contributions. First, we assemble a Ho Chi Minh City street-image dataset of 1,977 geotagged daylight images paired with hourly SILAM fields. Second, we introduce a monocular-depth attention bias for shifted-window transformers that follows the path-length dependence of particulate scattering. Third, we formulate a scoped training objective with hour-of-day encoding, uncertainty-weighted multi-task regression, and weak visibility-consistency losses. Fourth, we use modern CNN and transformer baselines under identical inputs.

The remainder of the paper is organized as follows. Section 2 details the Swin-AQI architecture, its depth-aware attention bias, and the time-encoding and physics-informed training objective. Section 3 describes the dataset, supervision, and baseline protocol, and Section 4 reports the comparison against CNN and transformer baselines together with ablations. Section 5 discusses diagnostics, scope, and limitations, Section 6 situates our approach within prior work, and Section 7 concludes.

2 Methodology

Algorithm 1 walks through Swin-AQI from input to prediction. We treat $PM_{2.5}$ and PM_{10} as the primary outputs, since distance-dependent haze gives them a physically grounded image cue. The four gases CO, NO_2, SO_2, and O_3 ride along as auxiliary context, and we interpret them only in that role. Figure 1 draws the same computation as a diagram: the solid path is inference, while the dashed paths are the weak training signals that nudge particulate predictions toward physics without ever replacing the SILAM supervision.

2.1 Depth-Aware Visual Encoding

Within each shifted window, standard Swin attention computes

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + B_{\mathrm{pos}}\right) V. \tag{1}$$

Standard attention treats near and far tokens alike once they share a window. Haze, however, accumulates along the line of sight, so a distant token usually carries more particulate information than a nearby one, and we want the model to use that fact. We estimate a normalized relative depth map $D \in [0,1]^{H\times W}$ with MiDaS [19]. Per-token depth values are batch-normalised to $[0,1]$, and for every token pair $(i,\, j)$ we define a pairwise depth-similarity bias:

$$\hat{D}_i = \frac{D_i - \min_k D_k}{\max_k D_k - \min_k D_k}, \qquad B_{\mathrm{depth}}(i,j) = \frac{0.1}{\tau}\,\exp\!\left(-\frac{|\hat{D}_i - \hat{D}_j|}{\tau}\right), \tag{2}$$

where $\hat{D}_i \in [0,1]$ is the batch-normalised depth of token i and τ is a single learnable scalar temperature that controls the sharpness of the depth-similarity decay. Then, a depth-aware attention module is applied to the post-Swin features $\mathbf{F} \in \mathbb{R}^{B\times N\times C}$. The attention update over the post-backbone features becomes

$$\mathrm{Attn}_{\mathrm{depth}}(\mathbf{F}) = \mathrm{softmax}\!\left(\frac{QK^\top}{\sqrt{d_h}} + B_{\mathrm{depth}}\right) V, \qquad [Q,\, K,\, V] = \mathbf{F}\, W_{QKV}, \tag{3}$$

where $B_{\mathrm{depth}} \in \mathbb{R}^{N\times N}$ is broadcast identically across all attention heads and d_h is the per-head dimension. The bias does not override the learned image interactions, but encodes a depth-coherence prior. Token pairs at similar estimated depths receive a stronger additive attention bonus, while pairs separated by a large depth difference are downweighted, with the sharpness of that transition governed by τ.

2.2 Time Encoding and Prediction Heads

Urban particulate levels typically follow diurnal traffic, activity, and boundary-layer cycles, so hour-of-day supplies a complementary temporal prior. Because

Algorithm 1 Swin-AQI training and inference

Require: RGB image $\mathbf{I}$, capture hour t, pollutant labels $\mathbf{y}$ during training
Ensure: Estimates $\hat{\mathbf{y}} = [\hat{y}_{\mathrm{PM}_{2.5}}, \hat{y}_{\mathrm{PM}_{10}}, \hat{y}_{\mathrm{CO}}, \hat{y}_{\mathrm{NO}_2}, \hat{y}_{\mathrm{SO}_2}, \hat{y}_{\mathrm{O}_3}]$
1: Estimate MiDaS relative depth D and downsample it to each Swin stage
2: Tokenize $\mathbf{I}$ and process tokens with shifted-window Swin-T blocks
3: Add depth bias B_{depth} to positional attention logits
4: Encode capture hour as Fourier features $\phi(t)$ and concatenate with pooled visual features
5: Predict six pollutant outputs with pollutant-specific regression heads
6: During training, minimize uncertainty-weighted smooth L1 loss plus PM-ordering and weak visibility-consistency terms
7: **return** $\hat{\mathbf{y}}$

such a prior can also act as a shortcut when the visual signal is weak, we treat capture time as an auxiliary input rather than a substitute for image evidence. We encode the capture hour $t \in \{0, \ldots, 23\}$ with learnable Fourier features:

$$\theta_i(t) = 2\pi \left(f_i |s_i|\right) \frac{t}{24}, \qquad \phi(t) = \bigoplus_{i=1}^{N} \left[\sin \theta_i(t), \cos \theta_i(t)\right]. \tag{2}$$

The encoding follows transformer positional encodings [25] and Fourier feature mappings [23]. We concatenate $\phi(t)$ with the pooled depth-aware visual representation and pass the fused vector to pollutant-specific MLP heads. The six outputs are trained with smooth L1 losses and homoscedastic task uncertainties σ_k^2 [10].

2.3 Training Objective

The training objective combines supervised regression, particulate ordering, and weak physics consistency:

$$\mathcal{L}_{\mathrm{PM}} = \frac{1}{N} \sum_{i=1}^{N} \max\!\big(0,\ \hat{c}^{(i)}_{PM_{2.5}} - \hat{c}^{(i)}_{PM_{10}}\big), \tag{3}$$

$$\mathcal{L}_{\mathrm{total}} = \sum_{k=1}^{6} \left(\frac{1}{2\sigma_k^2} \mathcal{L}_k + \log(1 + \sigma_k) \right) + \lambda_{PM} \mathcal{L}_{PM} + \lambda_{physics} \mathcal{L}_{\mathrm{physics}}, \tag{4}$$

where $\mathcal{L}_k$ is the smooth L1 loss [5]. The PM term penalizes only predictions that violate $\mathrm{PM}_{2.5} \leq \mathrm{PM}_{10}$, reflecting that fine particles are a subset of PM_{10} mass. In our work, λ_{PM} is fixed to 10.0, although it can also be set as a hyperparameter. The uncertainty term uses $\log(1+\sigma_k)$ to avoid negative contributions when both uncertainty and regression loss are small.

The physics term is a regularizer based on image proxies rather than calibrated radiometric supervision:

$$\mathcal{L}_{\mathrm{physics}} = \mathcal{L}_{\mathrm{BL}} + \mathcal{L}_{\mathrm{K}}. \tag{5}$$

We compute a spectral extinction coefficient from the particulate predictions. Writing the coarse mass as $m_{\text{coarse}} = \max(\hat{c}_{PM_{10}} - \hat{c}_{PM_{2.5}}, 0)$,

$$\beta = \alpha_{\text{fine}} \cdot \hat{c}_{PM_{2.5}} + \alpha_{\text{coarse}} \cdot m_{\text{coarse}} + \beta_{\text{Rayleigh}}, \tag{6}$$

following the particulate structure of the modified IMPROVE equation [6]. Beer–Lambert extinction gives

$$T = e^{-\beta \cdot d}, \tag{7}$$

with MiDaS inverse depth converted to scene-relative path length

$$d_{\text{km}} = \left\langle 1 - \frac{\text{disp}}{\max(\text{disp})} \right\rangle_{\text{spatial}} \times d_{\text{assumed}}. \tag{8}$$

After converting β from Mm^{-1} to km^{-1}, predicted transmittance is

$$\hat{T} = \exp\Big(-\beta \times 10^{-3} \times d_{\text{km}}\Big). \tag{9}$$

On the image side we use a normalized raw brightness value $\tilde{V} =\in [0, 1]$, and match it to the predicted transmittance:

$$\mathcal{L}_{\text{BL}} = \left\|\hat{T} - \tilde{V}\right\|_1. \tag{10}$$

The second term uses Koschmieder visibility $V = K/\beta$, normalized to

$$\hat{V} = \text{clip}\left(\frac{V \times 10^3 \text{ (km)}}{100 \text{ (km)}},\, 0,\, 1\right), \tag{11}$$

and compared with the same raw brightness $\tilde{V}$ as before :

$$\mathcal{L}_{\text{K}} = \left\|\hat{V} - \tilde{V}\right\|_1. \tag{12}$$

Together, these terms keep the particulate predictions compatible with the observed contrast decay, while the SILAM labels remain the supervision target.

3 Experiments

This section evaluates Swin-AQI on geotagged urban imagery with SILAM labels. We then report backbone comparisons and ablations under the same spatial protocol. The dataset and model related scripts are included here [3] [4].

[3] Dataset link.
[4] Model repository.

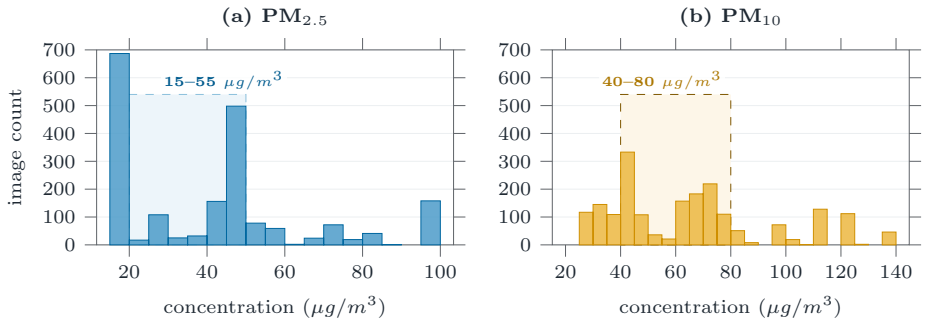

Fig. 2: Histograms of particulate matter observations. The distributions concentrate in a moderate-pollution range, which defines the empirical scope of the reported particulate results.

3.1 Dataset and Label Validation

The dataset contains 1,977 geotagged daylight images from Ho Chi Minh City, Vietnam, collected during a two-week period in 2026 within a 5 km-radius study area. Each image has resolution 4272×2848 and is paired with GPS coordinates, capture time, and SILAM atmospheric-composition fields [21]. SILAM provides hourly estimates of $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3 at 1 km resolution, and we interpolate these fields to each image location and timestamp.

The collection mainly covers a moderate particulate regime. Figure 2 shows that most observations fall between 15–55 $\mu g/m^3$ for $PM_{2.5}$ and 40–80 $\mu g/m^3$ for PM_{10}, with only light tails at higher concentrations. The dataset therefore tests common daylight urban conditions in the study area, not severe haze episodes or heavily polluted regions.

Figure 3 places the particulate targets beside the four auxiliary gases. The wider CO scale and the quieter gas panels illustrate why the paper reports operational claims for $PM_{2.5}$ and PM_{10}, while treating gaseous outputs as contextual multi-task signals rather than stand-alone sensing results.

Images are resized to 224×224 for the Swin-Tiny backbone and enhanced with CLAHE, which improves local contrast while preserving color structure. We deliberately avoid further augmentation, since aggressive transformations would distort the haze, contrast, and visibility cues on which the model relies.

3.2 Baselines and Training

We compare against ResNet-50 [8], EfficientNet-B4 [22], ConvNeXt-Tiny [15], ViT-Small [3], DeiT-Small [24], and standard Swin-T [14]. All models use ImageNet pretraining, AdamW, learning rate 10^{-4} with reduction on plateau, batch size 32, early stopping, and the same spatial protocol.

Input fairness. Every baseline receives the same information as Swin-AQI: resized image, encoded capture hour concatenated to pooled backbone features,

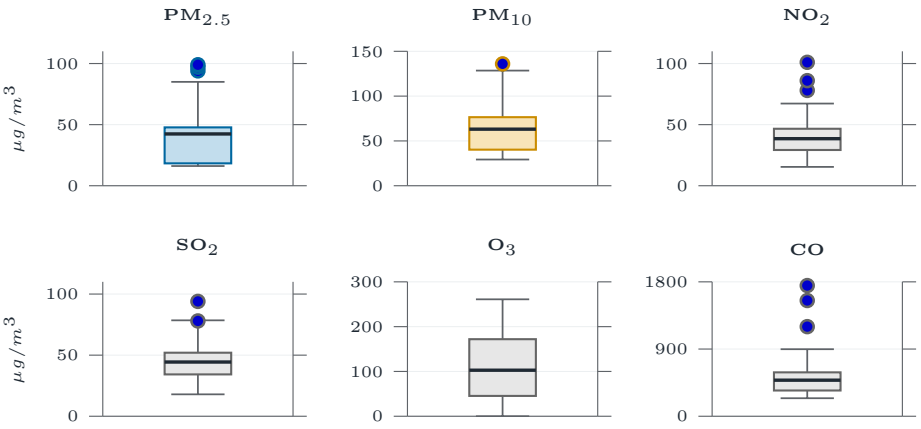

Fig. 3: Box plots of all six pollutants (concentrations in $\mu g/m^3$). The colored panels ($PM_{2.5}$, PM_{10}) are the optical targets, and the gray panels are contextual gases. Note the enlarged vertical scale for CO.

and the identical six-output uncertainty-weighted head. Depth-aware attention and physics losses are Swin-AQI-specific, so their effect is isolated through ablation rather than transplanted to unrelated backbones.

Baseline scope and availability. The baselines are modern general-purpose backbones, not direct copies of earlier image-to-AQI systems [13,27,28], which target different datasets and label sources. The image dataset cannot be publicly released under its collection agreements, but model code and spatial-protocol scripts will be made available on request.

4 Results

We ask a focused question here: once every model sees the same spatial split, do our depth-aware and physics-informed components actually help? We answer it with two tables. The first measures end-to-end accuracy against the baselines, and the second peels Swin-AQI back one component at a time. Throughout, we report mean absolute error (MAE) and R^2.

4.1 Baseline comparisons

Table 1 puts Swin-AQI next to the baselines, all of which we hold to the same blocked split, image input, capture-hour feature, six-output head, and training schedule. On the targets we care about most, Swin-AQI comes out ahead, reaching R^2=0.80 for $PM_{2.5}$ and R^2=0.76 for PM_{10}. We find the contrast with standard Swin-T the most telling: the two share a backbone family and the same metadata, and all we add is depth-aware attention and physics-informed training. We read this as support for our inductive biases under this protocol, though we are careful not to read it as evidence for visual-only sensing.

Table 1: Performance comparison under the same spatial split, metadata inputs, six-output head, and training schedule. Higher R^2 and lower MAE are preferred, and the best value in each column is in **bold**.

	$PM_{2.5}$		PM_{10}		Auxiliary gases	
Model	$R^2 \uparrow$	**MAE**$\downarrow$	$R^2 \uparrow$	**MAE**$\downarrow$	$R^2 \uparrow$	**MAE**$\downarrow$
ConvNeXt-Tiny	0.27	16.15	0.32	18.57	0.18	54.83
EfficientNet-B4	0.38	13.99	0.40	16.40	0.44	43.66
ResNet-50	0.56	12.35	0.54	14.31	0.49	38.12
ViT-Small	0.22	16.78	0.25	20.22	0.11	56.84
Swin-T	0.24	16.30	0.27	18.86	0.14	54.57
DeiT-Small	0.31	14.12	0.28	17.80	0.29	48.92
Swin-AQI	**0.80**	**4.95**	**0.76**	**7.11**	**0.71**	**24.32**

Table 2: Ablation comparison under a shared 50-epoch budget. Drops from the full Swin-AQI row show the contribution of each removed component. Higher R^2 and lower MAE are preferred, and the best value in each column is in **bold**.

	$PM_{2.5}$		PM_{10}		Auxiliary gases	
Removed component	$R^2 \uparrow$	**MAE**$\downarrow$	$R^2 \uparrow$	**MAE**$\downarrow$	$R^2 \uparrow$	**MAE**$\downarrow$
None (Full Swin-AQI)	**0.79**	**6.07**	**0.74**	**8.53**	**0.70**	**29.17**
Depth-aware attention	0.69	7.36	0.64	9.94	0.49	40.57
Time embedding	0.58	8.73	0.51	11.74	0.44	41.60
Uncertainty weighting	0.60	8.81	0.57	11.33	0.59	33.36
Physics loss	0.72	6.90	0.69	9.27	0.36	44.25

4.2 Ablation

In Table 2 we switch off one piece at a time, namely depth-aware attention, the time embedding, uncertainty weighting, and the physics loss. To keep the study affordable, we give every variant, including our full reference, the same fixed 50-epoch budget and otherwise identical hyperparameters. That budget is why our "None (Full Swin-AQI)" row sits just under the fully trained headline of Table 1 (0.79 vs. 0.80 for $PM_{2.5}$), so we read the ablation through relative drops rather than absolute numbers.

The biggest hit comes when we drop the time embedding: R^2 falls by 0.21 for $PM_{2.5}$ and 0.23 for PM_{10} against our full reference. Uncertainty weighting matters almost as much, especially for $PM_{2.5}$. Depth-aware attention adds a further 0.10/0.10, and the physics loss buys a smaller particulate gain while, tellingly, accounting for the largest decline on the contextual gases.

We take this as the honest reading of our method. Swin-AQI is not a visual-only estimator, and we do not claim it to be: hour-of-day carries real predictive signal through diurnal traffic and meteorological cycles. Yet the "Time embedding" row still keeps our visual depth and physics mechanisms, and it stays well above standard Swin-T, which sees the same time input but none of those mecha-

nisms. We therefore see time, uncertainty weighting, depth-aware attention, and weak physics regularization as complementary contributors, and we see a clear next step in time-shuffled and held-out-hour controls to pin the effect down.

5 Discussion

Swin-AQI is at its best on daylight scenes where the particulate cue is genuinely visible: clear to moderately hazy weather, some distant structure or a sky boundary, and concentrations below the level at which contrast loss saturates. In that regime the depth-aware attention is well matched to image formation, because far-field contrast decay is a meaningful stand-in for aerosol loading.

MiDaS gives relative, not metric, depth, so reflective glass, blank sky, dense vegetation, an unusual camera pitch, or rain on the lens can all distort the depth prior. Our labels are interpolated SILAM fields rather than per-image measurements, so a sharp local plume can be smoothed away. And many everyday conditions, from fog and mist to backlighting, hard shadows, and exposure shifts, lower visibility for reasons that have nothing to do with particulates.

For these reasons we keep the operational claim narrow: $PM_{2.5}$ and PM_{10} as a complementary sensing layer. The gaseous heads stay contextual, the data come from a single city at mostly Good-to-Moderate levels, and the model leans partly on the capture hour. The obvious next steps, which we are actively pursuing, are multi-city and seasonal tests, time-shuffled and held-out-hour controls, humidity-aware fog rejection, and selective fusion with calibrated sensors.

6 Related Work

Image-based air-quality estimation has used either hand-crafted visibility cues or learned image regression. Feature-based systems rely on contrast, edge strength, sky appearance, and related haze indicators [13,27], but these cues are sensitive to illumination, exposure, weather, and scene layout. Deep regressors reduce manual feature design [28,2], yet often treat the whole image as generic appearance evidence rather than modeling where particulate haze should be informative.

Atmospheric optics provide that missing structure. Beer–Lambert attenuation links extinction to path length, and Koschmieder visibility relates extinction to far-field contrast [1,11]. Dehazing methods exploit related assumptions to estimate transmission or recover a clearer image [4,7]. Swin-AQI repurposes this physical intuition for supervised particulate regression rather than image restoration.

Modern CNN and transformer backbones improve representation capacity, but backbone strength alone does not encode the path-length cue. CNNs learn hierarchical appearance patterns, and transformers learn token interactions [3,15]. Swin contributes efficient shifted-window attention [14], but standard attention has no built-in preference for distant keys. Swin-AQI instills this preference through a MiDaS relative-depth bias.

The remaining components follow established learning principles. Sinusoidal positional encodings and Fourier feature mappings help neural networks represent periodic inputs [25,23], which motivates hour-of-day encoding. Physics-informed learning adds process-level constraints to supervised objectives [18,9], and uncertainty-weighted multi-task learning balances tasks with different noise levels [10]. Swin-AQI combines these ideas while keeping its operational claim limited to particulate matter.

7 Conclusion

We set out to estimate particulate matter from ordinary street images by teaching a Swin-T backbone to read distance-dependent scattering, and Swin-AQI is the result: relative-depth attention, hour-of-day Fourier features, uncertainty-weighted multi-task regression, and weak Beer–Lambert/Koschmieder consistency terms in a single model. On a single-city Ho Chi Minh City study with spatially blocked evaluation it reaches R^2=0.80 for $PM_{2.5}$ and R^2=0.76 for PM_{10}. The ablations keep us honest: the temporal prior and uncertainty weighting do most of the work, while depth-aware attention and the physics loss add smaller, separable gains. We therefore offer camera imagery as a dense complement to reference instruments under moderate daylight conditions, valuable precisely where stations are sparse, rather than as a substitute for them.

References

1. Bohren, C.F., Huffman, D.R.: Absorption and scattering of light by small particles. John Wiley & Sons (1983).
2. Chakma, A., Vizena, B., Cao, T., Lin, J., Zhang, J.: Image-based air quality analysis using deep convolutional neural network. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3949–3952 (2017).
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021),
4. Fattal, R.: Single image dehazing. ACM Transactions on Graphics **27**(3), 1–9 (2008).
5. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015).
6. Hand, J.L., Copeland, S.A., Day, D.E., Dillner, A.M., Indresand, H., Malm, W.C., McDade, C.E., Moore Jr., C.T., Pitchford, M.L., Schichtel, B.A., Watson, J.G.: Spatial and seasonal patterns and temporal variability of haze and its constituents in the united states: Report v. Tech. rep., Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO (2011),
7. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1956–1963 (2009).

8. 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).
9. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L.: Physics-informed machine learning. Nature Reviews Physics **3**(6), 422–440 (2021).
10. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018).
11. Koschmieder, H.: Theorie der horizontalen sichtweite. Beiträge zur Physik der freien Atmosphäre **12**, 33–53 (1924)
12. Kumar, P., Patton, A.P., Durant, J.L., Frey, H.C.: A review of factors impacting exposure to pm2.5, ultrafine particles and black carbon in asian transport microenvironments. Atmospheric Environment **187**, 301–316 (2018).
13. Liu, C., Tsow, F., Zou, Y., Tao, N.: Particle pollution estimation based on image analysis. PLOS ONE **11**(2), e0145955 (2016).
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002 (2021).
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022).
16. Morawska, L., Thai, P.K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G.S.W., Jayaratne, R., Kumar, P., Lau, A.K.H., Louie, P.K.K., Mazaheri, M., Ning, Z., Motta, N., Mullins, B., Rahman, M.M., Ristovski, Z., Shafiei, M., Tjondronegoro, D., Westerdahl, D., Williams, R.: Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environment International **116**, 286–299 (2018).
17. Pope, C.A., Coleman, N., Pond, Z.A., Burnett, R.T.: Fine particulate air pollution and human mortality: 25+ years of cohort studies. Environmental Research **183**, 108924 (2020).
18. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics **378**, 686–707 (2019).
19. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1623–1637 (2022).
20. Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S.W., Shelow, D., Hindin, D.A., Kilaru, V.J., Preuss, P.W.: The changing paradigm of air pollution monitoring. Environmental Science & Technology **47**(20), 11369–11377 (2013).
21. Sofiev, M., Vira, J., Kouznetsov, R., Prank, M., Soares, J., Genikhovich, E.: Construction of the silam eulerian atmospheric dispersion model based on the advection algorithm of michael galperin. Geoscientific Model Development **8**(11), 3497–3522 (2015).
22. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114 (2019),

23. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020),
24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357 (2021),
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017),
26. World Health Organization: WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva (2021),
27. Zhang, C., Yan, J., Li, C., Wu, H., Bie, R.: End-to-end learning for image-based air quality level estimation. Machine Vision and Applications **29**(4), 601–615 (2018).
28. Zhang, Q., Fu, F., Tian, R.: A deep learning and image-based model for air quality estimation. Science of The Total Environment **724**, 138178 (2020).