VQualA 2025 Challenge on Face Image Quality Assessment: Methods and Results — Supplementary Material

Sizhuo Ma* Wei-Ting Chen* Qiang Gao* Jian Wang* Chris Wei Zhou*
Wei Sun Weixia Zhang Linhan Cao Jun Jia Xiangyang Zhu Dandan Zhu
Xiongkuo Min Guangtao Zhai Baoying Chen Xiongwei Xiao Jishen Zeng
Wei Wu Tiexuan Lou Yuchen Tan Chunyi Song Zhiwei Xu
MohammadAli Hamidi Hadi Amirpour Mingyin Bai Jiawang Du Zhenyu Jiang
Zilong Lu Ziguan Cui Zongliang Gan Xinpeng Li Shiqi Jiang Chenhui Li
Changbo Wang Weijun Yuan Zhan Li Yihang Chen Yifan Deng Ruting Deng
Zhanglu Chen Boyang Yao Shuling Zheng Feng Zhang Zhiheng Fu
Abhishek Joshi Aman Agarwal Rakhil Immidisetti Ajay Narasimha Mopidevi
Vishwajeet Shukla Hao Yang Ruikun Zhang Liyuan Pan Kaixin Deng
Hang Ouyang Fan Yang Zhizun Luo Zhuohang Shi Songning Lai Weilin Ruan
Yutao Yue

1. Proposed Methods

In this section, we present the proposed methods from teams that are not included in the main report due to space constraints.

1.1. MobileNetV3-Based FIQA (by Conquerit)

The *Conquerit* Team used a MobileNetV3-Large [5] architecture. To evaluate and guide the training of their quality regression model, they adopt a correlation-based loss that encourages both value accuracy and ranking consistency between predicted scores $\hat{y}$ and ground-truth quality scores $y \in [0, 1]$.

$$\mathcal{L}_{\text{corr}} = \alpha \cdot \mathcal{L}_{\text{PLCC}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Rank}}, \tag{1}$$

where $\alpha \in [0, 1]$ controls the balance between the two losses. The Pairwise Rank Loss approximates SRCC:

$$\mathcal{L}_{\text{Rank}} = E_{ij}[\log(1 + \exp(-(\hat{y_i} - \hat{y_j}) \cdot \text{sign}(y_i - y_j)))]. \tag{2}$$

To enhance robustness against label noise and focus learning on harder examples, they propose a modified regression loss that combines label smoothing with a focal weighting scheme. This formulation, referred to as Focal Label Smoothing Loss, builds on the intuition of focal loss while adapting it to continuous regression targets. Given predicted score $\hat{y}$ and ground truth score y, they first apply label smoothing:

$$\tilde{y} = y + \epsilon \cdot \mathcal{N}(0, 1), \tag{3}$$

where ϵ is the smoothing strength. This helps regularize overconfident predictions and improves generalization. Next, they compute the squared error between predictions and smoothed targets:

$$\text{MSE} = (\hat{y} - \tilde{y})^2. \tag{4}$$

They then apply a *focal weighting* to emphasize difficult samples:

$$w_{\text{focal}} = (1 - e^{-\text{MSE}})^{\gamma}, \tag{5}$$

where γ is the focusing parameter that increases the loss contribution of harder examples (*i.e.* larger errors). The loss is computed as

$$\mathcal{L}_{\text{focal-smooth}} = s \cdot w_{\text{focal}} \cdot \text{MSE}, \tag{6}$$

where s is an optional scaling factor for numerical stability or emphasis tuning.

The final training objective balances the correlation loss and the focal label smoothing loss using another weighting parameter λ:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{corr}} + (1 - \lambda) \cdot \mathcal{L}_{\text{focal-smooth}}. \tag{7}$$

*Sizhuo Ma (sma@snap.com), Wei-Ting Chen (weitingchen@microsoft.com), Qiang Gao (qgao@snap.com), Jian Wang (jwang4@snap.com) and Chris Wei Zhou (zhouw26@cardiff.ac.uk) are the challenge organizers. The other authors are participants of the VQualA 2025 Challenge on Face Image Quality Assessment.

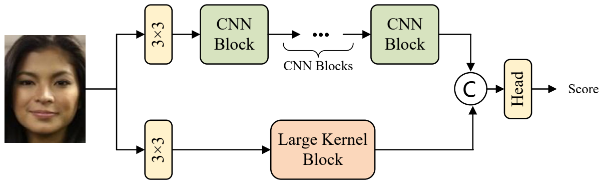

Figure 1. BIT_ssvgg.

Training details. They use ReduceLROnPlateau learning rate annealing during training. They use input size of (224,224,3) with normalization to (-1,1).

Testing details. Input images are resized to (240,240,3), center-cropped to (224,224,3) and normalized to (-1,1).

1.2. Dual-Branch Network with Local and Global Perception for Face Image Quality Assessment (by BIT_ssvgg)

The *BIT_ssvgg* Team observed that, to effectively assess the perceptual quality of facial images, it was crucial to capture both fine-grained local distortions and holistic structural degradations. Motivated by this, they adopted a dual-branch network design that leveraged complementary inductive biases, as shown in Fig. 1. Specifically, one branch was based on MobileNetV2 [9], which captured localized texture variations and high-frequency artifacts through its hierarchical depthwise convolutional design and small receptive fields. However, while efficient, MobileNetV2 tended to lack global context aggregation, which is essential for assessing structure-aware degradations such as defocus blur, over-smoothing, or global compression artifacts.

To compensate for this limitation, they introduced a second branch composed of large-kernel convolutions [3], which are known to emulate long-range dependencies without the computational overhead of attention modules. This branch emphasized global spatial interactions, enabling the model to better perceive large-scale structural degradation patterns across the entire face region. By aggregating the outputs from both local-detail-oriented and global-context-aware branches, their network effectively balanced sensitivity to local quality distortions with awareness of holistic structural integrity, which is especially critical in face IQA tasks where both local fidelity and global facial symmetry are important.

Training details. They implemented their model using the PyTorch framework and trained it on a single NVIDIA RTX 4090 GPU. The network was optimized using the Adam optimizer with a learning rate of 2e-5 and a weight decay of 5e-4. The training process took approximately 2 hours. They adopted standard data preprocessing strategies including resizing, normalization, and random horizontal flipping. They trained for a total of 50 epochs.

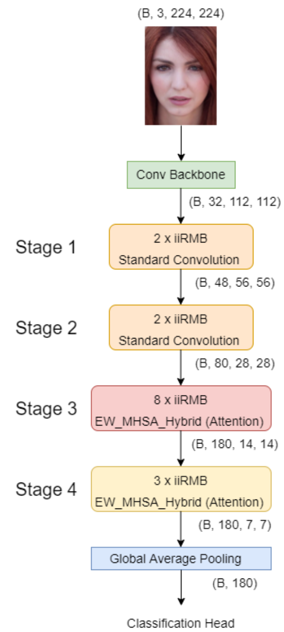

Figure 2. 2077Agent.

Testing details. The input images are resized to 224×336 for testing.

1.3. Knowledge Distillation for Improved Efficient MOdel(EMO) Image Quality Assessment (by 2077Agent)

The *2077Agent* Team first trained a high-capacity Efficient Model (EMO) [11] on the Generic Face IQA (GFIQA) dataset to establish the teacher model. To meet the competition's constraints on parameter count and computational complexity, they then designed and deployed a lightweight student model (shown in Fig. 2), enhancing its performance through a knowledge distillation strategy. Furthermore, they introduced optimized loss functions tailored to the evaluation metrics SRCC and PLCC to further improve the model's predictive accuracy.

Training details. Teacher and student networks share the same overall design—an initial convolutional layer that downsamples the input, followed by four hierarchical stages of improved inverted residual mobile blocks (iiRMB), and ending in a small regression head—but differ in depth, width, and resolution [11].

The teacher network processes 512×512 inputs and contains roughly 5 million parameters. After a three-layer stem (two 3×3 strided convolutions and a 1×1 projection), it has four stages with depths $\{3, 3, 9, 3\}$. The embedding dimensions at each stage are $\{48, 72, 160, 288\}$ with expansion ratios $\{2.0, 3.0, 4.0, 4.0\}$. Early stages use BatchNorm+SiLU; later stages use LayerNorm+GELU. Regularization is minimal (`attn_drop`=0, `drop`=0), with a drop-path rate of 0.05 and layer-scale initialized to 10^{-6}. Finally, the feature map is global-average-pooled and passed through two BatchNorm+ReLU bottleneck layers before a sigmoid output [9–12].

The student network scaled down to 1 million parameters and 224×224 inputs, the student mirrors the teacher's layout but with depths $\{2, 2, 8, 3\}$ and embedding dimensions $\{32, 48, 80, 180\}$. Expansion ratios are $\{2.0, 2.5, 3.0, 3.5\}$, attention head dims $\{16, 16, 20, 20\}$, and a 7×7 window. It uses the same BN+SiLU / LN+GELU scheme and a drop-path rate of 0.04036. Its regression head is identical to the teacher's, ensuring comparable output capacity at a fraction of the size.

This training was conducted in two stages. First, the teacher model with approximately 5×10^6 parameters was trained for 100 epochs on a general face image quality assessment dataset. During training, they applied the following augmentations: `Normalize`; `RandCropOrResize` to randomly crop or resize images to 512×512; `RandHorizontalFlip`; and `ToTensor`. Optimization was performed with Adam (initial learning rate 1×10^{-4}, weight decay 1×10^{-5}) alongside a `StepLR` schedule that decayed the learning rate by a factor of 0.1 every 5 epochs. Upon completion, the teacher's weights were frozen.

Next, they instantiated a student model under the same data loading and augmentation pipeline, and they performed widely used knowledge distillation method [2, 8] using a `HybridLoss` ($\alpha = 0.75$) together with an MSE distillation loss (temperature = 4.0, weight = 0.5) to boost both SRCC and PLCC performance. The entire process ran on a single RTX 3090 GPU and took about 20 hours of training time.

Testing details. To avoid excessive computational load, they directly resize all test data to (224, 224) before feeding them into the network.

1.4. LightHSPA: An Efficient and Lightweight Face Image Quality Assessment Network (by DERS)

The *DERS* Team proposed LightHSPA, a lightweight and efficient end-to-end convolutional neural network designed specifically for face image quality assessment (FIQA) under strict computational constraints. The architecture is composed of three main modules:

- Efficient Backbone: They employed a highly efficient feature extraction backbone inspired by MobileNetV2 [9]. It utilizes depthwise separable convolutions and inverted residual blocks to minimize parameter count and computational complexity (FLOPs). To further enhance feature representation with minimal overhead, they integrated a lightweight channel attention and an efficient spatial attention mechanism at the end of the backbone.
- Lightweight HSPA Attention: To capture non-local dependencies and global facial context, they developed LightHSPA, a lightweight version of the High-order Self-attention with Projection Awareness (HSPA) mechanism. This module uses depthwise separable convolutions and adaptive pooling to significantly reduce the computational cost of the self-attention operation, making it suitable for an efficient model. This module was used during experimentation to explore performance trade-offs.
- Compact Quality Head: A compact regression head predicts the final quality score. It uses multi-scale global pooling (average and max) to create a robust feature vector from the final feature map. This vector is then processed by a small multi-layer perceptron (MLP) with dropout and batch normalization to produce a single scalar quality score.

The entire model is designed to be configurable (e.g., via width multiplier) to balance the trade-off between performance (SRCC/PLCC) and efficiency (parameters/FLOPs).

Training details. Their model was implemented in PyTorch and trained on a single NVIDIA 4090 GPU for 200 epochs, taking approximately 9 hours. They used the Adam optimizer with an initial learning rate of 2×10^{-3} and a weight decay of 1×10^{-4}, employing a step learning rate scheduler that decayed the rate by a factor of 0.1 every 5 epochs. Only the official 30,000 in-the-wild face images from the competition were used for training. Their training strategy involved minimizing Mean Squared Error (MSE) loss and applying data augmentation, including random cropping to 224×224 patches and random horizontal flipping.

Testing details. Their model takes a single face image as input and outputs a single perceptual quality score. No test-time augmentation (TTA) or other post-processing steps were used. They ran the model on the original image resolutions as provided in the test set.

2. Details about RankCORE

This appendix provides details about the RankCORE model developed by the ISeeCV Team.

Self-Supervised Adaptive Ranking (SSAR). Traditional Perceptual Quality Assessment (PQA) methods [6, 7] require large datasets of images annotated with scalar quality scores—a process that is *expensive*, *subjective*, and *time-consuming*. *Self-supervised learning* (SSL) [1] sidesteps this requirement by exploiting *relative relationships* between samples rather than absolute labels.

They observed that common image degradations—such as Gaussian blur, additive noise, and interpolation artifacts—*monotonically degrade* perceptual quality: if I is an original face, then its transformed version $I' = T_s(I)$ always satisfies

$$q(I') < q(I),$$

where T_s is a transformation at severity $s \in [0, 1]$. They leveraged this inherent ordering to train a PQA network using *margin-based ranking loss*, enabling the model to *learn absolute prediction* of quality without ever seeing ground-truth scores.

For each pair (I, I'), with predicted scores $\hat{q}(I)$, $\hat{q}(I')$, and severity s, they defined margin

$$m(s) = \lambda\, s, \qquad \lambda > 0$$

and optimized:

$$\mathcal{L}_{\text{adaptive}}(I, I'; s) = \max\big(0,\, -(\hat{q}(I) - \hat{q}(I')) + m(s)\big).$$

Here, larger severities enforce larger margins, compelling the network to respect greater quality gaps when distortions are harsher.

Score-Stratified Uniform Sampler (SSUS). They introduced SSUS, which uniformly selects score-strata and then uniformly samples within each stratum to guarantee balanced coverage of the full score distribution. This ensured that rare or underrepresented score regions were seen as often as dense ones, reducing bias and improving model robustness. It also promoted gradient diversity, stabilized training, and could be extended by weighting specific bins.

CoReFace: Correlation-Robust Face Quality Estimation. To penalize both local prediction errors and global ranking mistakes, they adopted a composite Wing-PLCC loss. It is well known that MSE is not sensitive to outliers, while L1 loss doesn't penalize mid-range errors in regression. WingLoss [4] brings the best of both worlds. To the best of their knowledge, WingLoss has not been fundamentally explored for face perceptual quality assessment. WingLoss [4] ($\omega = 0.03$, $\epsilon = 2$) gives logarithmic sensitivity around small absolute errors while retaining linear behavior for larger discrepancies, helping the network focus on hard-to-distinguish quality levels, as shown in Eq. (8). Additionally, they leveraged the Pearson Linear Correlation Coefficient (PLCC) metric as a differentiable loss function. They augmented this with a global PLCC term to align predictions with the mean-opinion-score distribution, as defined in Eq. (9).

$$WingLoss = \begin{cases} w \ln\left(1 + \frac{|x|}{\epsilon}\right), & \text{if } |x| < w \\ |x| - C, & \text{otherwise} \end{cases} \tag{8}$$

$$PLCCLoss(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{9}$$

The final CoReFaceLoss is defined as the weighted sum of Eq. (8) and Eq. (9):

$$CoReFaceLoss = \alpha \cdot WingLoss + \beta \cdot PLCCLoss \tag{10}$$

3. Teams and Affiliations

VQualA 2025 FIQA Track Organizers

Members: Sizhuo Ma[1] (sma@snap.com), Wei-Ting Chen[2] (weitingchen@microsoft.com), Qiang Gao[1] (qgao@snap.com), Jian Wang[1](jwang4@snap.com) and Chris Wei Zhou[3] (zhouw26@cardiff.ac.uk)
Affiliations:
[1]Snap Inc.
[2]Microsoft
[3]Cardiff University

ECNU-SJTU VQA Team

Proposed Method: Efficient Face Image Quality Assessment via Self-training and Knowledge Distillation
Members: Wei Sun[1] (sunguwei@gmail.com), Weixia Zhang[2] (zwx8981@sjtu.edu.cn), Linhan Cao[2] (caolinhan@sjtu.edu.cn), Jun Jia[2](jiajun0302@sjtu.edu.cn), Xiangyang Zhu[3] (zhuxiangyang@pjlab.org.cn), Dandan Zhu[1] (ddzhu@mail.ecnu.edu.cn), Xiongkuo Min[2] (minxiongkuo@sjtu.edu.cn) and Guangtao Zhai[2] (zhaiguangtao@sjtu.edu.cn)
Affiliations:
[1]East China Normal University
[2]Shanghai Jiao Tong University
[3]Shanghai Artificial Intelligence Laboratory

MediaForensics

Proposed Method: MSPT: A Lightweight Face Image Quality Assessment Method with Multi-Stage Progressive Training
Members: Baoying Chen[1] (1900271059@email.szu.edu.cn), Xiongwei Xiao[2] (xiongweixiaoxxw@gmail.com) and Jishen Zeng[1] (jishen.zjs@alibaba-inc.com)
Affiliations:
[1]Alibaba Group
[2]The Hong Kong Polytechnic University

Next

Proposed Method: Towards Robust No-Reference Image Quality Assessment via Prompt-Aware Alignment and Multi-Level Distillation
Members: Wei Wu[1] (wuwei.lorenzo@gmail.com), Tiexuan Lou[2] (xuanxuanqaq@gmail.com), Yuchen Tan[2] (12334068@zju.edu.cn), Chunyi Song[2] (cysong@zju.edu.cn) and Zhiwei Xu[2] (xuzw@zju.edu.cn)
Affiliations:
[1]Donghai Laboratory
[2]Ocean College, Zhejiang University

ATHENAFace

Proposed Method: Face Image Quality Assessment via Lightweight Ensemble Learning and Correlation-Driven Optimization
Members: MohammadAli Hamidi[1] (mohammadali.hamidi@unica.it) and Hadi Amirpour[2] (hadi.amirpour@aau.at)
Affiliations:
[1]University of Cagliari
[2]University of Klagenfurt

NJUPT-IQA-Group

Proposed Method: Lightweight Spatial-Frequency Fusion Network for Blind Face Image Quality Assessment
Members: Mingyin Bai (1223014043@njupt.edu.cn), Jiawang Du (1223013838@njupt.edu.cn), Zhenyu Jiang (1224014031@njupt.edu.cn), Zilong Lu (1224014206@njupt.edu.cn), Ziguan Cui (cuizg@njupt.edu.cn) and Zongliang Gan (ganzl@njupt.edu.cn)
Affiliations:
[1]School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications

ECNU VIS Lab

Proposed Method: MobileNetV4 in FIQA
Members: Xinpeng Li[1] (51275901118@stu.ecnu.edu.cn), Shiqi Jiang[1] (52265901032@stu.ecnu.edu.cn), Chenhui Li[1] (chli@cs.ecnu.edu.cn) and Changbo Wang[1] (cbwang@dase.ecnu.edu.cn)
Affiliations:
[1]East China Normal University

JNU620

Proposed Method: Progressive Learning Strategy for Face Image Quality Assessment
Members: Weijun Yuan[1] (yweijun@stu2022.jnu.edu.cn), Zhan Li[1] (lizhan@jnu.edu.cn), Yihang Chen[1] (ehang@stu.jnu.edu.cn), Yifan Deng[1] (dyf010408@stu.jnu.edu.cn), Ruting Deng[1] (routine@stu2022.jnu.edu.cn), Zhanglu Chen[1] (czhanglu@stu2024.jnu.edu.cn), Boyang Yao[1] (yaoboy@stu.jnu.edu.cn), Shuling Zheng[1] (3440989938@qq.com), Feng Zhang[1] (1569259893@qq.com) and Zhiheng Fu[1] (2557502986@qq.com)
Affiliations:
[1]Jinan University

ISeeCV

Proposed Method: RankCORE: Ranking-Aware Correlation Optimized Regression Estimator
Members: Abhishek Joshi[1] (abhishek.j@aftershoot.com) and Aman Agarwal[1] (aman.a@aftershoot.com)
Affiliations:
[1]Aftershoot

RegNet

Proposed Method: Face Image Quality Assessment in the Wild with RegNet
Members: Rakhil Immidisetti[1] (samrakhil@gmail.com) and Ajay Narasimha Mopidevi[2] (ajaymopidevi@gmail.com)
Affiliations:
[1]Blue River Technology
[2]Lendbuzz

Conquerit

Proposed Method: MobileNetV3-Based FIQA
Members: Vishwajeet Shukla[1] (vishwajeet993511@gmail.com)
Affiliations:
[1]Adobe Systems

BIT_ssvgg

Proposed Method: Dual-Branch Network with Local and Global Perception for Face Image Quality Assessment
Members: Hao Yang[1] (3120235187@bit.edu.cn), Ruikun Zhang[1] (ruikun.zhang@bit.edu.cn) and Liyuan Pan[1] (liyuan.pan@bit.edu.cn)
Affiliations:
[1]School of Computer Science & Technology, Beijing Institute of Technology

2077Agent

Proposed Method: Knowledge Distillation for Improved Efficient MOdel(EMO) Image Quality Assessment
Members: Kaixin Deng[1] (kaixin.deng.t0@elms.hokudai.ac.jp), Hang Ouyang[2] (1246276321@qq.com), Fan Yang[2] (173001344@qq.com), Zhizun Luo[2] (1151507490@qq.com) and Zhuohang Shi[3] (shi_zhuo_hang@yeah.net)
Affiliations:

[1]Hokkaido University
[2]Chengdu University of Technology
[3]Hebei University of Technology

DERS

Proposed Method: LightHSPA: An Efficient and Lightweight Face Image Quality Assessment Network

Members: Songning Lai[1] (songninglai@hkust-gz.edu.cn), Weilin Ruan[1] (wruan792@connect.hkust-gz.edu.cn) and Yutao Yue[1] (yutaoyue@hkust-gz.edu.cn)
Affiliations:
[1]The Hong Kong University of Science and Technology (Guangzhou)

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. 4

[2] Kaixin Deng, Jianuo Lei, Xinrui Li, Shiyu Shuai, Mingyue Lin, and Siyuan Li. An improved lightweight segmentation neural network for dermoscopic lesion images based on knowledge distillation. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pages 268–271. IEEE, 2023. 3

[3] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 2

[4] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018. 4

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 1

[6] Lin Kang, Pei Ye, Yubin Li, and David Doermann. Blind image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 23(1):310–325, 2014. 4

[7] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. No-reference image quality assessment in spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 4

[8] Hang Ouyang, Tailin Li, Chongru Wang, Yuanting Gu, Fan Yang, and Kaixin Deng. Improving road defect detection precision and efficiency with structural pruning techniques. In *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–6. IEEE, 2024. 3

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2, 3

[10] Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhengkai Jiang, Tianxin Huang, Yabiao Wang, and Chengjie Wang. Rethinking mobile block for efficient attention-based models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1389–1400. IEEE Computer Society, 2023.

[11] Jiangning Zhang, Teng Hu, Haoyang He, Zhucun Xue, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao. Emov2: Pushing 5m vision model frontier. *arXiv preprint arXiv:2412.06674*, 2024. 2

[12] Shi Zhuohang. Dynamic convolution-based image dehazing network. *Multimedia Tools and Applications*, 83(16):49039–49056, 2024. 3