NTIRE 2025 Challenge on Video Quality Enhancement for Video Conferencing: Datasets, Methods and Results

Varun Jain[1], Zongwei Wu[2], Quan Zou[1], Louis Florentin[1], Henrik Turbell[1], Sandeep Siddhartha[1], Radu Timofte[2], Qifan Gao*, Linyan Jiang*, Qing Luo*, Jack Song*, Yaqing Li*, Summer Luo*, Mae Chen*, Stefan Liu*, Danie Song*, Huimin Zeng*, Qi Chen*, Ajeet Verma*, Shweta Tripathi*, Vinit Jakhetiya*, Badri N Subhdhi*, Sunil Jaiswal*

[1]Microsoft and [2]University of Würzburg

jain.varun@microsoft.com, zongwei.wu@uni-wuerzburg.de, {quan.zou, lflorentin, heturbel, ssiddhartha}@microsoft.com, radu.timofte@uni-wuerzburg.de

Abstract

This paper presents a comprehensive review of the 1st Challenge on Video Quality Enhancement for Video Conferencing held at the NTIRE workshop at CVPR 2025, and highlights the problem statement, datasets, proposed solutions, and results. The aim of this challenge was to design a Video Quality Enhancement (VQE) model to enhance video quality in video conferencing scenarios by (a) improving lighting, (b) enhancing colors, (c) reducing noise, and (d) enhancing sharpness — giving a professional studio-like effect. Participants were given a differentiable Video Quality Assessment (VQA) model, training, and test videos. A total of 91 participants registered for the challenge. We received 10 valid submissions that were evaluated in a crowdsourced framework. Additional materials can be found on the project website [1,2].

1. Introduction

Light is a crucial component of visual expression and is the key to controlling texture, appearance, and composition. Professional photographers often have sophisticated studio lights and reflectors to illuminate their subjects such that the true visual cues are expressed and captured. Similarly, tech-savvy users with modern desk setups employ a sophisticated combination of key and fill lights to give themselves control over their illumination and shadow characteristics. However, many users are constrained by their physical environment, which may lead to poor positioning of ambient lighting or lack thereof. It is also commonplace to encounter flares, scattering, and specular reflections that may come from windows or mirror-like surfaces. Problems can be compounded by poor-quality cameras that may introduce sensor noise. This leads to poor visual experience during video calls and can have a negative impact on downstream tasks such as denoising, super-resolution, segmentation, and face detection.

The current light correction solution in Microsoft Teams, called AutoAdjust, finds a global mapping of input to output colors which is updated sporadically. Since this mapping is global, it gives more importance to foreground colors, which may lead to improper exposure of or color shifts in the background. On the other hand, popular single-image portrait relighting methods [48] estimate local correction in only the foreground and preserve the background by an implicit in-network matte layer. A possible side effect of local correction can be the reduction of local contrast, which often serves as a proxy to convey depth in 2D images, making people appear dull in some cases.

We conducted P.910 [26] studies totaling 350,000 pairwise comparisons that measured people's preference for AutoAdjust and portrait relighting over no effect and images manually edited by experts in Adobe Lightroom. We used the Bradley–Terry model [1] to estimate the scores for each method and observed that people preferred AutoAdjust more than any other method.

To take the next step towards achieving studio-grade video quality, one would need to (a) understand what people

* These members were participants who co-authored this report detailing their methodologies, not the challenge organizers. Please refer to Appendix A for their correspondence details.

[1] https://www.microsoft.com/en-us/research/academic-program/ntire-2025-vqe/

[2] https://github.com/varunj/cvpr-vqe/

Figure 1. Ground truth from (top) our synthetics framework, (bottom) the AutoAdjust solution. The top row shows the input with suboptimal foreground illumination which is fixed by adding a studio light setup in front of the subject which is simulated in synthetics and predicted via global changes in the real data.

prefer and construct a differentiable Video Quality Assessment (VQA) metric, and (b) be able to train a Video Quality Enhancement (VQE) model that optimizes this metric. To solve the first problem, we have trained a VQA model that, given a pair of videos x_1 and x_2, gives the probability that x_1 is better than x_2 as described in Sec. 2.5. Given a standard test set, this information can be used to construct a ranking order of a given set of methods.

We invited researchers to participate in a challenge aimed at developing Neural Processing Unit (NPU) friendly VQE models that leverage our trained VQA model to improve video quality.

This challenge was one of the NTIRE 2025 [3] Workshop associated challenges on: ambient lighting normalization [33], reflection removal in the wild [40], shadow removal [32], event-based image deblurring [30], image denoising [31], XGC quality assessment [23], UGC video enhancement [29], night photography rendering [10], image super-resolution (x4) [5], real-world face restoration [6], efficient super-resolution [28], HR depth estimation [43], efficient burst HDR and restoration [16], cross-domain few-shot object detection [11], short-form UGC video quality assessment and enhancement [18, 19], text to image generation model quality assessment [12], day and night raindrop removal for dual-focused images [17], video quality enhancement for video conferencing, low light image enhancement [24], light field super-resolution [37], restore any image model (RAIM) in the wild [20], raw restoration and super-resolution [7] and raw reconstruction from RGB on smartphones [8].

[3] https://www.cvlai.net/ntire/2025/

2. Challenge

2.1. Problem Statement

The task was to enhance video quality in video conferencing scenarios. We only looked at the following properties of a video to judge its studio-grade quality:

1. Foreground illumination – the person (all body parts and clothing) should be optimally lit.
2. Natural colors – correction may make local or global color changes to make videos pleasing.
3. Temporal noise – correct for image and video encoding artefacts and sensor noise.
4. Sharpness - to ensure that correction algorithms do not introduce softness, the final image should at least be as sharp as the input.

We understand that there may be many other aspects to a good video. For simplicity, we discounted all except the ones mentioned above. Specifically, we did not measure the following:

1. Egocentric motion – unstable camera may introduce sweeping motion or small vibrations that we did not aim to correct.

2. Makeup and beautification – it is commonplace for users to apply beautification filters that alter their skin tone and facial features such as those found on Instagram and Snapchat. We did not aim for that aesthetic.
3. Removal of reflection on glasses and lens flare - although it is a common occurrence in video teleconference scenarios, we did not aim to remove reflections that may come from screens and other light sources onto users' glasses due to the risk associated with altering eye appearance and gaze direction.
4. Avatars – A solution that synthesizes a photorealistic avatar of the subject and drives it based on the input video would score the highest in terms of noise, illumination, and color. If it indeed minimizes the total cost function that takes into account all these factors, it would be acceptable.

2.2. Baseline Solution

Since the AutoAdjust model was ranked higher than expert-edited images and portrait relighting methods, we provided participants a baseline solution so that they could reproduce the AutoAdjust feature as currently shipped in Microsoft Teams. It was provided as a Python script that calls the AutoAdjust executable, and includes code for post-processing.

2.3. Compute Constraints

The goal was to have a computationally efficient solution that can be offloaded to NPU for CoreML inference. We established a qualifying criterion of CoreML uint8 or fp16 models with at most $20.0\text{x}10^9$ MACs per frame for an input resolution of 1280×720. We estimate such a model to have a per frame processing time of 9 ms on an M1 Ultra powered Mac Studio and 5 ms on an M4 Pro powered Mac Mini for the given input resolution. Submissions that did not meet this criterion were not considered for the P.910 evaluation.

2.4. Dataset

2.4.1. Unpaired Real Data

We host a web service that reaches users all over the world and prompts them to sit in front of a laptop or a PC. We then record minute-long videos while users perform hand gestures and body movements. We sampled $13,000$ videos from this dataset for training, validation, and testing of VQE methods. The videos are 10 s long, encoded at 19 FPS on average, and amount to a total of $3,900,000$ frames. We kept $3,000$ (23%) videos for testing and ranking submissions and make $10,000$ (77%) available to the teams. They could choose to split it between the training and validation sets as they desire. The teams were also free to use other publicly available datasets, while being mindful about data drift.

Of the $13,000$ videos, we selected 300 high quality videos where P.910 raters voted strongly in favor of the AutoAdjust result, as shown in the bottom half of Figure 1. We assumed these to be the ground truth. P.910 done on these videos shows a Mean Opinion Score (MOS) [26] of 3.58 in favor of the target.

Figure 2. Comparison of lighting setup in the Synthetic Portrait Relighting dataset. (left) Lighting from the HDRI, (center) key light with HDRI lighting turned off, and (right) key and fill lights with HDRI lighting turned off. Note that the HDRI is only used as a background when using the studio lighting and does not contribute to the illumination of the subject.

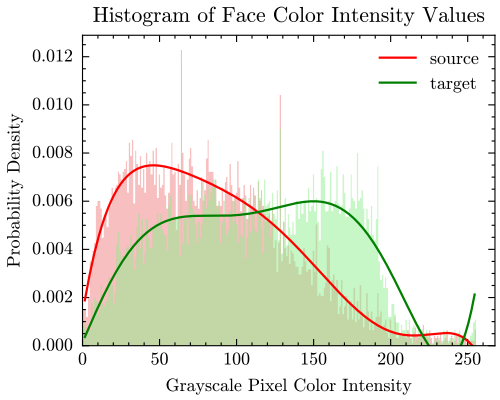

Figure 3. Color intensity in source and target images of our Synthetic Portrait Relighting dataset. The source images are dark with intensity centered around 50. The target fixes this by boosting illumination – making it more uniform and span a larger range.

2.4.2. Paired Synthetic Portrait Relighting Data

In addition to these data, we also provided paired data for fully supervised learning as shown in the top half of Figure 1. Note that it is possible to learn a correction which is different from these ground truth labels and achieve a higher MOS. Hence, these labels had to be treated as suggestive improvements, and not as global optima.

We use a physically-based path tracer and photorealistic assets in Blender to render $1,500$ videos for training and 500 videos for testing. Each video is 5 s long encoded at 30 FPS. The source image has lighting only from the High Dynamic Range Image (HDRI) environment. For the target,

Teamname	Input Resolution	Inference Resolution	Training Time	Epochs	Ensemble	LUT	Attention	#MACs/frame	Latency/frame	GPU/NPU
TMobileRestore	(720, 1280, 3)	(720, 1280, 3)	1 day	100	Yes	Yes	No	16.8×10^9	200ms	V100
Summer	(720, 1280, 3)	(720, 1280, 3)	2 days	120	Yes	Yes	No	15.4×10^9	180ms	V100
XTeam	(720, 1280, 3)	(720, 1280, 3)	6 hrs	25	Yes	Yes	No	16.8×10^9	200ms	V100
Velta	(720, 1280, 3)	(720, 1280, 3)	12 hrs	30	Yes	Yes	No	15.4×10^9	180ms	V100
DeepView	(720, 1280, 3)	(720, 1280, 3)	7 days	80	Yes	No	Yes	106.4×10^9	238ms	V100
Auv	(720, 1280, 3)	(720, 1280, 3)	1 day	5	Yes	No	Yes	106.4×10^9	238ms	V100
Meeting	(720, 1280, 3)	(720, 1280, 3)	3 hrs	10	Yes	Yes	No	16.8×10^9	200ms	V100
Maqic	(720, 1280, 3)	(720, 1280, 3)	7 days	157	No	Yes	No	13.0×10^3	28 ms	V100
LUT	(720, 1280, 3)	(720, 1280, 3)	4 days	80	No	Yes	No	13.0×10^3	27 ms	V100
Wizard	(720, 1280, 3)	(720, 1280, 3)	15 days	50	Yes	No	Yes	114.2×10^9	170ms	V100

Table 1. Final results of the NTIRE 2025 Challenge on Video Quality Enhancement for Video Conferencing held at CVPR 2025.

we added 2 diffuse light sources to simulate a studio lighting setup. Refer Figure 2 to visualize the effect of these light sources and Figure 3 for statistics on the color intensity values on the face. These are the same images that were used to finetune the portrait relighting method.

To ensure that these data generalize well in the wild, we refer to the image-level degradations used in Real-ESRGAN [36] and applied them to the source image. To simulate out-of-focus blur, we applied generalized Gaussian blur kernels [25] that have ramp edges and flat top areas – better modeling the combined effects of lens defocusing and light diffraction. For color noise, we used channel-independent additive Gaussian noise, and gray noise was added by applying the same Gaussian noise to all 3 channels. Finally, sensor noise was modeled by sampling from a Poisson distribution. Lastly, we applied random resizing and JPEG compression.

P.910 done on these videos shows a MOS of 4.06 in favor of the target indicating that these make for a better target compared to the baseline AutoAdjust solution. Some examples of these pairs are shown in Figure 1 and more details about the rendering framework can be found in [13].

2.5. VQA Model

$$p_{x_1,x_2}, A_{x_1}, A_{x_2} = VQA_\theta(x_1, x_2) \tag{1}$$

We provided teams with a pre-trained Siamese [15] Video Quality Assessment model VQA_θ that was trained on $22,553$ videos and 11 enhancement models. Ground-truth was collected by prompting human raters with $315,636$ side-by-side video comparisons. For high-level semantic understanding, we used our own models that were pre-trained on a collection of real and synthetic images for the tasks of person segmentation, face quality and image aesthetics. For low-level features such as noise, flicker and video coding artifacts we used the DOVER [38] model. We took the penultimate feature maps of both models, performed average pooling across temporal and spatial dimensions and concatenated them. We then used a set of projections to predict the final logits.

Given a pair of images or videos x_1 and x_2, the model predicts the probability of x_1 being preferred over x_2 in a P.910 study. It also provides 11 auxiliary scores for each input $A = [a_1, a_2, ...a_{11}]$ that correspond to factors such as image aesthetic, color harmonization, color liveliness, key-lighting, noise, image composition, face capture quality etc. These are supervised with metrics obtained from publicly available Apple Vision APIs.

2.6. Metrics and Evaluating Submissions

The final goal was to rank the submissions according to the P.910 scores. We asked the teams to submit their predictions on the $3,000$ real-video test set. We then compared the submissions to the given input, the baseline, and against each other. As shown in Figure 4, comparison using the Bradley–Terry model gives us the score for each submission that maximizes the likelihood of the observed P.910 voting. Our P.910 framework has a throughput of $210,000$ votes per week. In case two methods had statistically insignificant difference in subjective scores, we used the objective score shown in Equation (4) to break ties.

$$S_{obj}^{real}(\hat{Y}, X, \theta) = \frac{1}{12n} \sum_{i=1}^{n} \{p_{\hat{y}_i,x_i}, A_{\hat{y}_i}\}|_{VQA_\theta(\hat{y}_i,x_i)} \tag{2}$$

$$S_{obj}^{synth}(\hat{Y}, Y) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \sqrt{E[(Y - \hat{Y})^2]}} \tag{3}$$

$$S_{obj}(\hat{Y}, Y, X, \theta) = S_{obj}^{real}(\hat{Y}, X, \theta) \times S_{obj}^{synth}(\hat{Y}, Y) \tag{4}$$

Due to the infeasibility of getting P.910 scores in real-time, teams could use the objective score S_{obj} for continuous and independent evaluation. For the $3,000$ unsupervised videos, teams were required to submit the per-video VQA score $p_{\hat{y}_i,x_i}$ along with the 11 auxiliary scores $A_{\hat{y}_i}$ predicted by the VQA model as shown in Equation (1). For the synthetic test set, the teams reported the Root Mean Squared Error (RMSE) per video. These scores were also published on the leaderboard so that participants could track their progress relative to other teams. However, we did not rank the teams based on these objective metrics since it was possible to learn a correction that is different from and subjectively better than the ground truth provided.

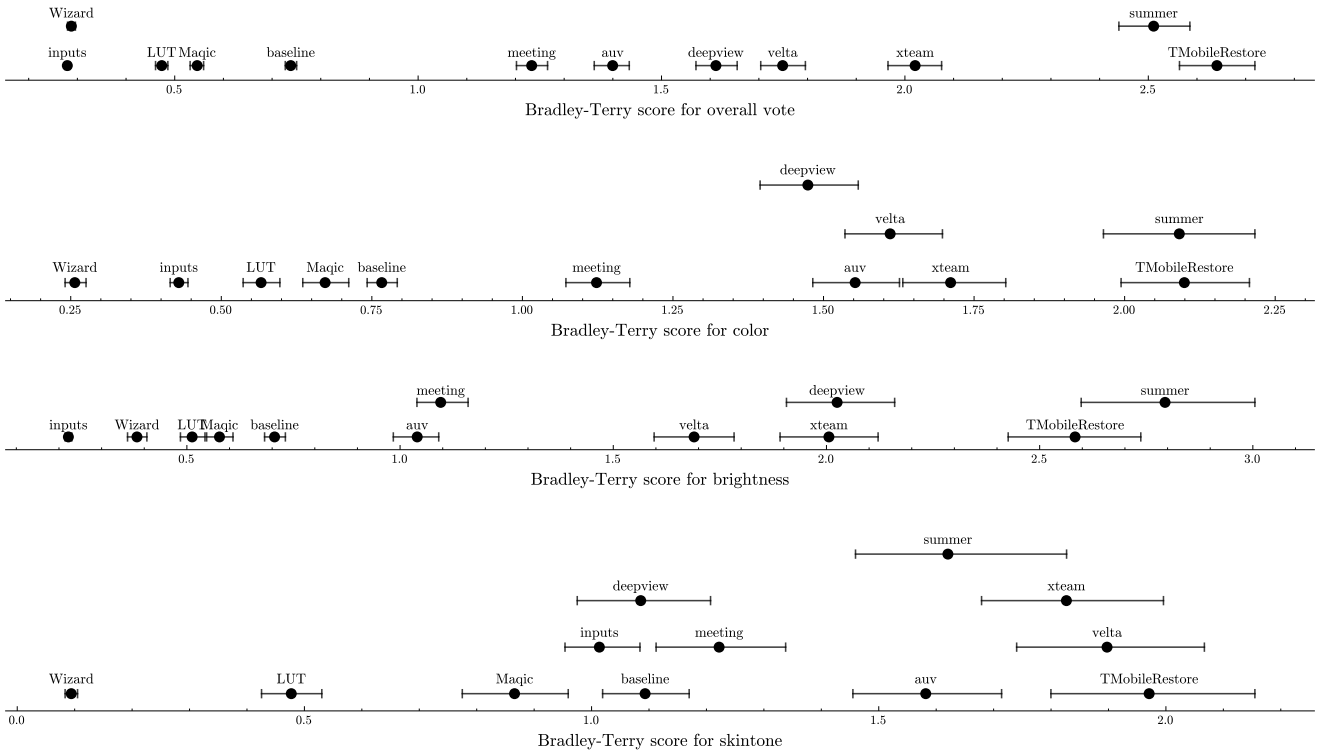

Figure 4. Interval plots illustrating the mean P.910 Bradley-Terry scores and their corresponding 95% confidence intervals for the 10 submissions, input videos, and the provided baseline. (Top) Overall preference, and (bottom) factors influencing preference.

3. Results

We received 5 complete submissions for both the mid-point and final evaluations. For each team's submission, we utilized our crowd-sourced framework to evaluate their 3,000-video test set. This involved presenting human raters with 270,000 side-by-side video comparisons. The raters were asked to provide their preference on a scale of 1 to 5, where 1 and 5 represent strong preference for the left and right video respectively, and 2 and 4 represent weak preference. A rating of 3 indicates no preference. Furthermore, raters were prompted to specify if their decision was primarily influenced by (a) image colors, (b) image brightness, or (c) skin tone. The Bradley–Terry scores for each team that maximize the likelihood of the observed P.910 voting are shown in Figure 4.

4. Challenge Methods

This section outlines the methodologies and datasets used by the highest-ranking submissions. We observe that Look-Up Table (LUT) based solutions TMobileRestore and DeepView scored the highest. This can be attributed to the efficient, yet temporally stable nature of the correction when compared to methods that predict dense pixel-to-pixel mapping between the input and output image pairs.

4.1. TMobileRestore

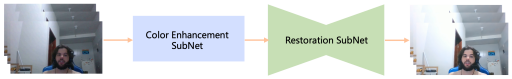

Figure 5. Two stage video conferencing enhancement framework proposed by team TMobileRestore.

4.1.1. Description

They propose a video enhancement algorithm designed to tackle common issues found in video conferencing videos, such as noise, compression artifacts, pathological illumination, and visual inconsistencies. Their algorithm employs a two-stage training process to achieve optimal results: the first stage uses a LUT for brightness and color correction, while the second stage focuses on removing compression noise, sensor noise, and enhancing the overall video quality, as shown in Figure 5.

The first stage of the video enhancement framework is designed to tackle color distortions that are typically found in video conferencing footage, such as inconsistent lighting and color shifts, which considerably lowers visual quality.

To effectively rectify these distortions, they use a combination of Clookup table (CLUT) based methods [46] and convolutional neural network structures. During this phase, the network processes input frames by extracting features at multiple levels, allowing it to concurrently extract image features. They implement a CLUT, where the neural network predicts content-dependent weights from downsampled input to merge basic CLUTs into an image-adaptive one, thereby enhancing the original input image.

The second stage is dedicated to rectifying low-level distortions such as noise and compression artifacts. To address these problems, they utilize a lightweight U-Net architecture with skip connections, specifically engineered for effective and robust restoration. The network extracts features at various scales, enabling it to concurrently address both local artifacts (like blocky compression noise) and global degradations. Skip connections between the encoder and decoder ensure the preservation of fine-grained details throughout the restoration process. The first phase includes 21 convolutional layers that have the ability to broaden receptive fields and carry out both global and local refinement for image distortions. This allows the network to restore the natural and visually pleasant context throughout the video.

4.1.2. Datasets

To train the two-stage network, they used a combination of public datasets, including LDV3 [41], REDS [22], and datasets provided in this competition. For realistic degradation simulation, they model mixed distortions to create training data that closely resembled real-world scenarios. The training data for the first stage incorporated color distortions, such as random saturation shifts and contrast adjustments. For the second stage, the training data included randomized degradations such as Poisson-Gaussian noise, motion blur, and H.265/H.264 compression. The degradation parameters were dynamically sampled per batch to improve robustness. For both stages, they applied spatial augmentations such as rotation, flipping, and chromatic aberration, as well as temporal jitter techniques such as frame dropping and shuffling to prevent overfitting.

4.1.3. Experiments & Results

In stage 1, the CLUT is trained with a hybrid loss function combining L1 loss and cosine color shift loss, over $200,000$ iterations with a batch size of 32 and a patch size of 512×512. The initial learning rate was set to 0.0002 and halved every $10,000$ iterations, using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_1 = 0.99$.

For the second stage, the sub-network was optimized using a combination of L2 loss, perceptual loss, LPIPS and GAN loss to enhance textures without over-smoothing, over $300,000$ iterations with a batch size of 16 and a patch size of 512×512. After pretraining both stages independently, they jointly fine-tuned the network for an additional $20,000$ iterations with a reduced learning rate of 0.00001. Details are listed in Table 1. XTeam and Meeting are similar to this method with early training termination at $50,000$ and $10,000$ iterations respectively.

4.2. Summer

This method also consists of the color enhancement subnetwork and the video restoration sub-network. The color enhancement network uses a set of five pre-trained 3D LUTs to dynamically adjust the color and tone of video frames in real-time [44]. These LUTs are trained using the provided supervised VQE dataset, which ensures that each LUT represents a distinct style of color and tone transformation tailored for video content. To predict the optimal combination of these LUTs for each frame, the method employs a convolutional neural network (CNN) with seven convolutional blocks. This CNN extracts global features from the downsampled video frames, capturing essential characteristics that influence the color-enhancement process. By analyzing these features, the network predicts the weights for blending the five 3D LUTs, resulting in a final LUT that is adapted to the specific content of each video frame.

The video restoration sub-network utilizes seven residual blocks that progressively refine video frames, reducing noise, correcting blurriness, and restoring details. This deep learning approach effectively learns to map from degraded images to high-quality images, ensuring clear and detailed video frames. Velta is similar to this method, with training ending early in $30,000$ iterations.

4.3. DeepView

They propose a video conference enhancement network that addresses both degradation distortion repair and color enhancement to ensure high-quality video communication. For distortion repair, it is the same as TMobileRestore.

For color enhancement, they adopt the HVI-CIDNet approach [39], which includes the HVI color space and the CIDNet architecture. The HVI color space minimizes noise and compresses low-light regions, while CIDNet's dual-branch network handles chromatic denoising and brightness enhancement. By processing images in the HVI color space and applying cross-attention mechanisms, this approach restores natural colors and details, providing vibrant and accurate color representation in video conferencing. Auv is similar to this method, with training ending early in $5,000$ iterations.

4.4. Maqic

4.4.1. Description

Online video streams suffer from the physical environment, including poor positioning of ambient lighting or lack thereof, leading to poor visual experience during video calls and may perturb the downstream tasks. Considering

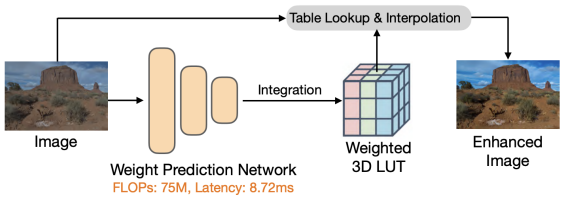

Figure 6. Typical 3DLUT-based retouching pipeline.

that human-region should be the focus of the video calls, they interpreted this challenge as the task of video portrait retouching, which aims to improve the aesthetic quality of input portrait photos and especially requires human-region priority [45]. While deep learning-based methods [4, 9, 14, 47] largely elevate the retouching efficiency and provide promising retouched results, most of them concentrate on the image tasks, which leads to the efficiency bottleneck when translating to the video task. Therefore, they consider an efficient solution, *i.e.*, a look-up table (LUT) based retouching, which performs fast inverse tone mapping according to the trained look-up table for each pixel value.

For the video portrait retouching task, to improve the temporal consistency, existing video-based enhancement methods [3, 21] typically include an additional optical flow estimator (*e.g.*, SpyNet [27]) to propagate information from adjacent frames. However, this is not suitable for a highly efficient LUT-based solution, where the online SpyNet inferencing inevitably slows down the whole retouching pipeline. Therefore, they choose the image-based ICELUT [42] (as shown in Figure 7) as the retouching backbone. Their contributions can be summarized as choosing an efficient backbone for video portrait retouching and the stage-wise training strategy to achieve perceptual satisfying retouched results.

The typical LUT solution (as shown in Figure 6) provides only the LUT-based pixel value transfer to accelerate the retouching process. For the portrait scenario, the retouching should be two-fold: (a) retouching for both the background and foreground, and (b) the focus on highlighting the human region instead of the background. Therefore adopting a solution with a region-wise adaptation (*e.g.*, ICELUT [42]) is necessary to filter out the background and focus on the portrait region. As shown in Figure 7, given the low-quality input frame, the adopted method adaptively performs the fusion of multiple LUTs and composes the 3D LUT, which then performs tone mapping to obtain the visually satisfying result for each frame.

4.4.2. Datasets

They notice that the given dataset contains limited image resolution, with degradations such as noise, motion blur, and flicker. Training with these data complexes the target goal, requiring the model to simultaneously perform retouching and video restoration. Therefore, they adopt a stage-wise training strategy to separate the aforementioned goals.

In the first stage, they train with MIT-Adobe FiveK [2], which is a high-quality tone mapping dataset without image degradations. This enables the LUTs to retouch input videos, which adaptively changes the tone of frames and adjusts the light condition. Then based on the pre-trained LUTs, they conduct fine-tuning on the supervised subset given in the challenge to equip the LUTs with restoration ability.

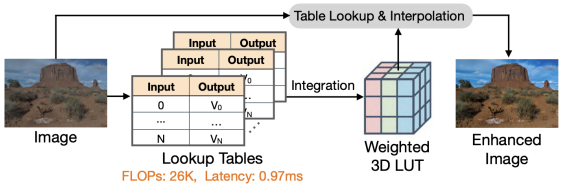

Figure 7. The adopted ICELUT used by Maqic constructs weighted 3D LUT with multiple lookup table candidates to perform real-time video retouching.

4.4.3. Experiments & Results

The model is implemented with the PyTorch framework, they conduct all the experiments on a single NVIDIA Tesla V100 GPU. They include additional details in Table 1. LUT is similar to this method with early training termination.

4.5. Wizard

4.5.1. Description

Their approach [34] integrates both technical and aesthetic quality assessment algorithms with the video enhancement task. Building upon the HVI-CIDNet [39] framework, they introduce a perceptual quality-aware color and intensity decoupling network that leverages the Lighten Cross-Attention (LCA) mechanism. In addition, they incorporate a quality loss function based on CLIP-IQA metrics to enhance the perceptual quality of the output video, ensuring alignment with human visual preferences.

To accelerate convergence and enhance performance, they initialize the model with pretrained HVI-CIDNet [39] weights, leveraging prior knowledge for effective spatial and chromatic feature handling in video frames. A major limitation of traditional image enhancement models is that they often prioritize technical fidelity over perceptual quality. However, in video conferencing applications, visual appeal is just as important as technical accuracy. To address this, they introduce a perceptual quality loss function that incorporates metrics from CLIP-IQA [35], in combination with the Video Quality Assessment metrics (Equation (1)).

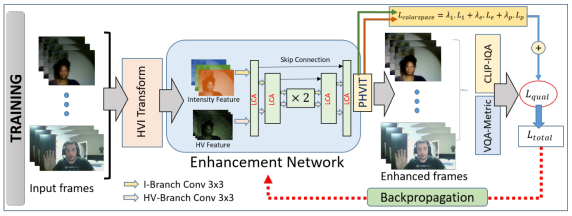

Figure 8. Overview of the quality-aware CIDNet proposed by Wizard. During training, they use extracted frames as input. These inputs are first passed through HVI transform network, the obtained HVI features are then processed in CIDNet, and lastly Perceptual-inverse HVI Transform (PHVIT) is applied to get sRGB-enhanced image. The outputs from the VQE model are evaluated using CLIP-IQA for perceptual quality assessment, and the resulting scores are utilized as a quality loss. The $\mathcal{L}_{colorspace}$ is constructed as combination of L1 Loss (L_1), Edge Loss (L_e), and Perceptual Loss (L_p).

The quality loss component encourages the model to prioritize aesthetic factors, ensuring that the output video frames align closely with human visual preferences.

As given in the block diagram Figure 8, during training, let x_i be the input frames and $\hat{x}_i$ be the enhanced frames generated by the enhancement network. Enhanced frames are fed into the CLIP-IQA model, which computes the quality scores $Q_{\hat{x}_i}$. Scores typically range between $[0-1]$, where a higher score indicates better perceptual quality, the objective is to maximize the score. To incorporate this into the loss function while ensuring optimization, the mean quality score $\bar{Q}(\hat{x}_i)$ across a batch is normalized as follows:

$$\mathcal{L}^{clip}_{quality} = 1 - \bar{Q}_c(\hat{x}_i), \tag{5}$$

and $\bar{Q}_c(\hat{x}_i)$ is given as,

$$\bar{Q}_c(\hat{x}_i) = \frac{1}{N}\sum_{i=1}^{N} Q_c(\hat{x}_i), \tag{6}$$

where, N represents the batch size. Here, a value of 1 corresponds to the highest quality, in the same manner, another term from VQA model (Equation (1)) is constructed as:

$$\mathcal{L}^{VQA}_{quality} = 1 - \bar{Q}_v(\hat{x}_i) \tag{7}$$

Finally, quality loss is given as:

$$\mathcal{L}_{quality} = \mathcal{L}^{clip}_{quality} + \mathcal{L}^{VQA}_{quality} \tag{8}$$

The term $\mathcal{L}_{quality}$ is then integrated into the overall loss function to guide the training process. The total loss function used during training is a weighted sum of the standard colorspace loss and the perceptual quality loss :

$$\mathcal{L}_{total} = \mathcal{L}_{colorspace} + \lambda \cdot \mathcal{L}_{quality} \tag{9}$$

where $\mathcal{L}_{colorspace}$ is the loss term used in base-model [39]. The hyperparameter λ controls the contribution of the perceptual quality term, ensuring a balance between technical accuracy and visual quality.

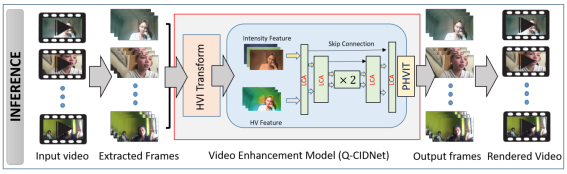

Figure 9. Inference pipeline of proposed quality-aware CIDNet. First the frames are extracted from input video, extracted frames are fed to Q-CIDNet for enhancement.

4.5.2. Datasets

To develop and evaluate the Video Quality Enhancement (VQE) model, they selectively utilized subsets of the real and synthetic datasets provided in this challenge. Out of the total $10,000$ real videos and $1,500$ synthetic videos, they opted for a focused approach to training by leveraging representative samples from each dataset:

- Real Dataset Utilization: From the $10,000$ real videos provided for training and validation, they selected 300 videos for training. This subset was chosen to capture diverse lighting conditions, variations in ambient reflections, and noise characteristics while maintaining a balance between complexity and model training efficiency.
- Synthetic Dataset Utilization: From the $1,500$ synthetic videos provided for training, they selected 300 videos for training. These synthetic samples were curated to include a variety of lighting configurations generated by adding diffuse light sources to simulate a studio setup. This data was instrumental in fine-tuning the model's ability to handle lighting corrections and improve visual appeal.

4.5.3. Experiments & Results

After initializing the model with pretrained weights from HVI-CIDNet, they fine-tuned it using the Adam optimizer with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 50 epochs. The parameter λ is set to 0.75 for quality loss inclusion. The learning rate was initially set to 1×10^{-4} and gradually decreased to 1×10^{-7} using a cosine annealing schedule during the training process. On an input of $3 \times 720 \times 1280$, the model requires 114.249 GMACs, equivalent to 228.498 GFLOPs, with 1.973M parameters. The measured inference latency is 170 ms per frame. Details are listed in Table 1 and inference detailed in Figure 9.

Acknowledgments

The authors thank the Human Understanding Toolkit team, led by Tadas Baltrusaitis, for their support in using their synthetic data generation framework and adapting it to our

needs. We especially thank Lohit Petikam for his critical help in designing the studio lighting setup in Blender and for collaborating on writing the rendering scripts. This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2025 sponsors: ByteDance, Meituan, Kuaishou, and University of Wurzburg (Computer Vision Lab).

A. Teams & Affiliations

NTIRE 2025 Team

Title: NTIRE 2025 Challenge on Video Quality Enhancement for Video Conferencing

Members:
Varun Jain[1] (jain.varun@microsoft.com),
Zongwei Wu[2] (zongwei.wu@uni-wuerzburg.de),
Quan Zou[1] (quan.zou@microsoft.com),
Louis Florentin[1] (lflorentin@microsoft.com),
Henrik Turbell[1] (heturbel@microsoft.com),
Sandeep Siddhartha[1] (ssiddhartha@microsoft.com),
Radu Timofte[2] (radu.timofte@uni-wuerzburg.de)
Affiliations:
[1] Microsoft, Redmond WA, USA
[2] Computer Vision Lab, University of Würzburg, Germany

TMobileRestore

Members:
Qifan Gao[1] (qf_gao@outlook.com),
Linyan Jiang[1] (jly724215288@gmail.com),
Qing Luo[1] (luoqing.94@qq.com),
Jie Song[2] (553252129sj@gmail.com),
Yaqing Li[1] (lyqstudy@qq.com)
Affiliations:
[1] ShannonLab, Tencent, China
[2] Xidian University, China

Summer

Members:
Summer Luo[1] (185471613@qq.com),
Mae Chen[2] (chenxm_m@nankai.edu.cn)
Affiliations:
[1] Independent researcher, China
[2] Nankai Uninversity, China

DeepView

Members:
Stefan Liu[1] (stefan1026@163.com),
Danie Song[2] (hypox128@gmail.com)
Affiliations:
[1] Shanghai Jiao Tong University, China
[2] Shenzhen University, China

Maqic

Members:
Huimin Zeng[1] (zeng.huim@northeastern.edu),
Qi Chen[2] (qchen76@jh.edu)
Affiliations:
[1] Northeastern University, USA
[2] Johns Hopkins University, USA

Wizard

Title: Q-CIDNet: Perceptual Quality Aware Color and Intensity Decoupling Network for Video Quality Enhancement

Members:
Ajeet Kumar Verma[1] (ajeet.verma@iitjammu.ac.in),
Shweta Tripathi[1] (2023pcs0041@iitjammu.ac.in),
Vinit Jakhetiya[1] (vinit.jakhetiya@iitjammu.ac.in),
Badri N Subhdhi[2] (subudhi.badri@iitjammu.ac.in),
Sunil Jaiswal[3] (sunil.jaiswal@k-lens.de)
Affiliations:
[1]Department of Computer Science and Engineering, IIT Jammu, India
[2]Department of Electrical Engineering, IIT Jammu, India
[3]K|Lens, GmbH, Germany

References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 1

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 7

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 7

[4] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *CVPR*, pages 6306–6314, 2018. 7

[5] Zheng Chen, Kai Liu, Jue Gong, Jingkai Wang, Lei Sun, Zongwei Wu, Radu Timofte, Yulun Zhang, et al. NTIRE 2025 challenge on image super-resolution (×4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[6] Zheng Chen, Jingkai Wang, Kai Liu, Jue Gong, Lei Sun, Zongwei Wu, Radu Timofte, Yulun Zhang, et al. NTIRE 2025 challenge on real-world face restoration: Methods and

results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[7] Marcos Conde, Radu Timofte, et al. NTIRE 2025 challenge on raw image restoration and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[8] Marcos Conde, Radu Timofte, et al. Raw image reconstruction from RGB on smartphones. NTIRE 2025 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[9] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACM MM*, pages 870–878, 2018. 7

[10] Egor Ershov, Sergey Korchagin, Alexei Khalin, Artyom Panshin, Arseniy Terekhin, Ekaterina Zaychenkova, Georgiy Lobarev, Vsevolod Plokhotnyuk, Denis Abramov, Elisey Zhdanov, Sofia Dorogova, Yasin Mamedov, Nikola Banic, Georgii Perevozchikov, Radu Timofte, et al. NTIRE 2025 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[11] Yuqian Fu, Xingyu Qiu, Bin Ren Yanwei Fu, Radu Timofte, Nicu Sebe, Ming-Hsuan Yang, Luc Van Gool, et al. NTIRE 2025 challenge on cross-domain few-shot object detection: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[12] Shuhao Han, Haotian Fan, Fangyuan Kong, Wenjie Liao, Chunle Guo, Chongyi Li, Radu Timofte, et al. NTIRE 2025 challenge on text to image generation model quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[13] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, et al. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 4

[14] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, pages 3277–3285, 2017. 7

[15] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, pages 1–30. Lille, 2015. 4

[16] Sangmin Lee, Eunpil Park, Angel Canelo, Hyunhee Park, Youngjo Kim, Hyungju Chun, Xin Jin, Chongyi Li, Chun-Le Guo, Radu Timofte, et al. NTIRE 2025 challenge on efficient burst hdr and restoration: Datasets, methods, and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[17] Xin Li, Yeying Jin, Xin Jin, Zongwei Wu, Bingchen Li, Yufei Wang, Wenhan Yang, Yu Li, Zhibo Chen, Bihan Wen, Robby Tan, Radu Timofte, et al. NTIRE 2025 challenge on day and night raindrop removal for dual-focused images: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[18] Xin Li, Xijun Wang, Bingchen Li, Kun Yuan, Yizhen Shao, Suhang Yao, Ming Sun, Chao Zhou, Radu Timofte, and Zhibo Chen. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Kwaisr dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[19] Xin Li, Kun Yuan, Bingchen Li, Fengbin Guan, Yizhen Shao, Zihao Yu, Xijun Wang, Yiting Lu, Wei Luo, Suhang Yao, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[20] Jie Liang, Radu Timofte, Qiaosi Yi, Zhengqiang Zhang, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Lei Zhang, et al. NTIRE 2025 the 2nd restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[21] Jing Lin, Xiaowan Hu, Yuanhao Cai, Haoqian Wang, Youliang Yan, Xueyi Zou, Yulun Zhang, and Luc Van Gool. Unsupervised flow-aligned sequence-to-sequence learning for video restoration. In *International Conference on Machine Learning*, pages 13394–13404. PMLR, 2022. 7

[22] Hongying Liu, Zhubo Ruan, Peng Zhao, Chao Dong, Fanhua Shang, Yuanyuan Liu, Linlin Yang, and Radu Timofte. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 55(8):5981–6035, 2022. 6

[23] Xiaohong Liu, Xiongkuo Min, Qiang Hu, Xiaoyun Zhang, Jie Guo, et al. NTIRE 2025 XGC quality assessment challenge: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[24] Xiaoning Liu, Zongwei Wu, Florin-Alexandru Vasluianu, Hailong Yan, Bin Ren, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2025 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[25] Yu-Qi Liu, Xin Du, Hui-Liang Shen, and Shu-Jie Chen. Estimating generalized gaussian blur kernels for out-of-focus image deblurring. *IEEE Transactions on circuits and systems for video technology*, 31(3):829–843, 2020. 4

[26] Babak Naderi and Ross Cutler. A crowdsourcing approach to video quality assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2810–2814. IEEE, 2024. 1, 3

[27] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 7

[28] Bin Ren, Hang Guo, Lei Sun, Zongwei Wu, Radu Timofte, Yawei Li, et al. The tenth NTIRE 2025 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[29] Nickolay Safonov, Alexey Bryntsev, Andrey Moskalenko, Dmitry Kulikov, Dmitriy Vatolin, Radu Timofte, et al. NTIRE 2025 challenge on UGC video enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[30] Lei Sun, Andrea Alfarano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, et al. NTIRE 2025 challenge on event-based image deblurring: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[31] Lei Sun, Hang Guo, Bin Ren, Luc Van Gool, Radu Timofte, Yawei Li, et al. The tenth ntire 2025 image denoising challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[32] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Cailian Chen, Zongwei Wu, Radu Timofte, et al. NTIRE 2025 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[33] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Radu Timofte, et al. NTIRE 2025 ambient lighting normalization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[34] Ajeet Kumar Verma, Shweta Tripathi, Vinit Jakhetiya, Badri N Subudhi, and Sunil Jaiswal. Q-cidnet: Perceptual quality aware color and intensity decoupling network for video quality enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 7

[35] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 7

[36] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 4

[37] Yingqian Wang, Zhengyu Liang, Fengyuan Zhang, Lvli Tian, Longguang Wang, Juncheng Li, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2025 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[38] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 4

[39] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. *arXiv preprint arXiv:2502.20272*, 2025. 6, 7, 8

[40] Kangning Yang, Jie Cai, Ling Ouyang, Florin-Alexandru Vasluianu, Radu Timofte, Jiaming Ding, Huiming Sun, Lan Fu, Jinlong Li, Chiu Man Ho, Zibo Meng, et al. NTIRE 2025 challenge on single image reflection removal in the wild: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[41] Ren Yang, Radu Timofte, et al. AIM 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *European Conference on Computer Vision Workshops*, 2022. 6

[42] Sidi Yang, Binxiao Huang, Mingdeng Cao, Yatai Ji, Hanzhong Guo, Ngai Wong, and Yujiu Yang. Taming lookup tables for efficient image retouching. In *European Conference on Computer Vision*, pages 144–159. Springer, 2024. 7

[43] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, Samuele Salti, Stefano Mattoccia, et al. NTIRE 2025 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2

[44] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2020. 6

[45] Huimin Zeng, Jie Huang, Jiacheng Li, and Zhiwei Xiong. Region-aware portrait retouching with sparse interactive guidance. *IEEE Transactions on Multimedia*, 26:127–140, 2023. 7

[46] Fengyi Zhang, Hui Zeng, Tianjun Zhang, and Lin Zhang. Clut-net: Learning adaptively compressed representations of 3dluts for lightweight image enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6493–6501, 2022. 6

[47] Yongbing Zhang, Siyuan Liu, Chao Dong, Xinfeng Zhang, and Yuan Yuan. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *TIP*, 29:1101–1112, 2020. 7

[48] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 1