Helping Visually Challenged People Take Better Quality Pictures

Anonymous CVPR 2022 submission

Paper ID 11433

Abstract

Computer vision technologies can help visually challenged users take better quality pictures by automated guidance, empowering them to interact more confidently on social media. Our goal is to help them avoid the most common distortions, such as blur, exposure, and noise, complementing work on aesthetic aspects such as framing. To advance progress on the problem of assessing visually challenged user-generated content (VC-UGC), we built the largest subjective image quality and distortion dataset. It contains 40K real-world distorted VC-UGC images, 40K patches, and 2.7M human perceptual quality and distortion labels. Using this resource, we created a blind picture quality and distortion predictor that learns local-to-global spatial quality relationships and achieves state-of-the-art performance on VC-UGC pictures, significantly outperforming standard models. Using a multi-task model, we also created a prototype feedback system that guides users to mitigate quality issues and take better pictures. The new dataset and prediction models will be made public following the review process.

1. Introduction

Computer vision breakthroughs have the power to build community and make technologies more accessible at the largest scales. One example is making social media more accessible to visually challenged people. Being able to automatically understand picture and video content by AI-driven assistance could benefit low-vision/blind users when selecting pictures to upload on social media. While there has been progress on building visual tools to assist visually challenged users on other tasks [8, 5, 52, 1], studies [51, 4, 28, 45] have shown that such users often still rely on friends for several of these tasks, making them feel vulnerable and disempowered. These studies have shown that the visually impaired often ask for information about, and assistance with picture quality.

Current No-Reference Image Quality Assessment (NR-IQA) models can predict both perceptual quality and distortion types, based on perceptually relevant 'quality-aware' statistics [30], yet, obtaining high accuracy remains a challenging problem [49, 40, 20]. Moreover, pictures captured by visually challenged users, without any guiding feedback, usually suffer from higher levels of distortions [8]. Since state-of-the-art NR-IQA predictors are trained on datasets containing pictures captured by users having normal vision, they do not transfer well [8] to visually challenged user-generated content (VC-UGC). Creating accurate IQA models for VC-UGC content needs large and appropriately labeled datasets, which currently do not exist.

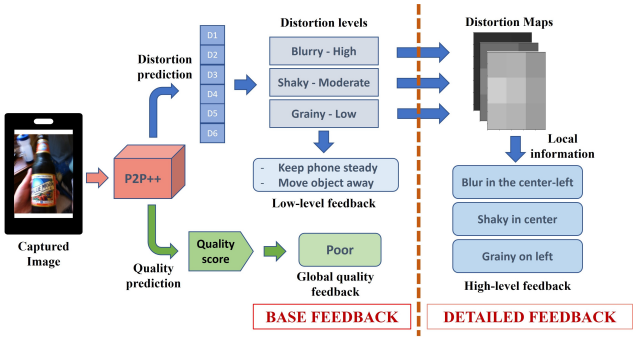

Fig. 1: **Quality feedback to assist visually challenged users:** The captured image is passed through P2P++ (Sec. 4), which generates global quality and distortion scores. The predicted scores are used to provide suitable feedback on distortions present and ways to mitigate them. (Sec. 5)

To better assist visually impaired people to take better pictures, a feedback system should describe the distortions that occur as pictures are being captured. Distortions arise because of imperfect capture devices, focusing issues, stabilization problems, and sub-optimal lighting conditions, and all are often amplified in VC-UGC. Moreover, multiple distortions often intermix, making them harder to rate and classify. The absolute degree of distortion is not the same as perceptual quality, since the latter is also deeply affected by the content and by perceptual processes such as masking [44, 32].

In this work, we show that automatic picture quality prediction models can supply guidance and feedback to address this problem. Recent work on perceptual QA [49, 48] has shown that modeling the relationship between local and global distortions can lead to better visual quality predictions. Inspired by this, our proposed dataset comprises of

images taken from VizWiz, along with both randomly selected and salient patches extracted from them. We also conducted a large-scale visual psychometric study on both the images and extracted patches, whereby we collected human subjective quality scores and distortion labels. This large dataset enabled us to design new IQA models that can accurately predict perceptual quality scores and also categorize distortions.

Further, we used these learned models to build a feedback system to help visually challenged users take better pictures. While pictures captured by visually challenged users often suffer from aesthetic flaws such as incorrect framing or orientation [42, 19, 5], our work focuses on helping users *improve the perceptual quality* of their captured pictures, rather than photographic aesthetics. Ultimately, a feedback system should be able to assist users to improve both aspects.

The contributions we make on these impactful but challenging problems are summarized:

- **We built the largest subjective image quality and distortion database targeting pictures captured by visually challenged users.** This new resource contains about 40K images collected from VizWiz [8] and 40K patches (half randomly selected and half salient). We conducted a large-scale subjective picture quality study on them collecting 2.7M each of quality labels and distortion labels. This is also the largest publicly available distortion classification dataset. We also collected about 75K ratings on 2.2K ORBIT images (frames extracted from ORBIT [29] videos) (Sec. 3).
- **We created a state-of-the-art blind VC-UGC picture quality and distortion predictor.** Using a deep neural architecture based on the recent PaQ-2-PiQ model [49], we created a multi-task system able to predict both the perceptual quality of pictures captured by visually impaired users, and the possible presence of five common picture distortions. Since this model is trained on patches, we can use it to predict spatial maps of both quality and distortion types. Our proposed model referred to as P2P++ achieves top performance on the new dataset and an independent mini-dataset (ORBIT images) when compared to other NR-IQA models. (Sec. 4 and 5.2)
- **Using the multi-task model, we also created a unique prototype feedback system to assist visually challenged users to take better quality pictures.** We provide feedback on overall quality, along with suggestions on how to mitigate quality issues.

2. Related Work

Image Quality and Distortion Datasets: The most heavily-used image quality datasets are older corpora of synthetically generated distortions of natural pictures [37, 25, 34, 35]. Since synthetic distortions are quite different from authentic, real-world distortions, NR-IQA models trained on them perform poorly on real-world content [12, 26, 49]. These resources however do not include labels on distortion types [12]. The Flickr-Distortion dataset [3] contains synthetic distortion labels on 804 Flickr UGC images, but it does not contain any quality labels, and is not public. The VizWiz-QualityIssues [8] dataset contains images taken by visually challenged users, along with distortion labels for a few common impairments and aesthetic flaws. The images in the dataset were generated under real use conditions via the VizWiz mobile app [5]. However, since it lacks picture quality scores and supplies only 5 subjective distortion scores per image, it cannot be used in its current form for our purposes.

Image Quality Models: In this application, we require NR-IQA (blind) models, which have recently advanced significantly. Popular NR-IQA models [30, 31, 47, 13, 11, 24, 6, 23] work well on legacy single synthetic distortion datasets [37, 34, 35] but perform poorly on real-world UGC data [26, 49]. PaQ-2-PiQ [49] leverages relationships between local and global quality to achieve current high performance on all IQA datasets. A few multi-task models like IQA-CNN++ [21] and QualNet [14] are available that use relationships between quality and distortion features to predict both picture quality scores and distortion categories. These perform well on synthetic datasets, but struggle on real-world UGC pictures and distortions. The authors of [8] used an Xception backbone trained on ImageNet [10] to predict distortions on the VizWiz pictures and achieved promising results.

Assisting the Visually Challenged: Several applications now exist to help visually challenged users capture better images through audio feedback, mostly built for visual recognition tasks [1, 19, 5]. *TapTapSee* [1] helps users take focus-adjusted images, while *VizWiz Ver2* [5] and *EasySnap* [19] use simple darkness and blur detection algorithms. The authors of [42] developed an assisted photography framework to help users better frame their photos, using an image composition model to assess aesthetic quality. The *Scan Search* [52] application uses the Lucas-Kanade [27] optical flow method to track feature points and determine camera stability. The authors of [8] developed algorithms to detect the recognizability and answerability of images captured by

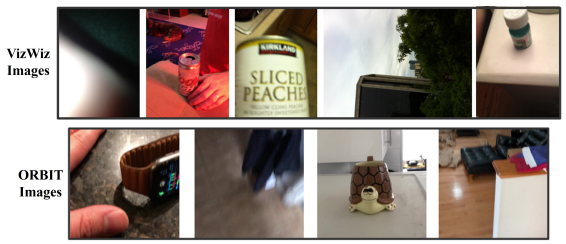

Fig. 2: **Sample images from the two datasets - VizWiz (top row) and ORBIT (bottom row)**, each resized to fit. The actual images are of highly diverse sizes and resolutions.

blind users. None of these tasks is predicated on perceptual models, trained on human data, or directed towards VC-UGC quality prediction.

3. Dataset and Human Study

Here, we outline the details of our new VC-UGC dataset and the online human study used to collect subjective quality labels. The proposed dataset contains 39,660 images (from VizWiz [8], sample images in Fig. 2) and 39,660 patches extracted from them, half of which are salient patches, the other half cropped at random. We also collected 2.7M quality ratings and distortion labels on them. This is the first quality-focused dataset dedicated to developing assistive technology for visually challenged photographers.

3.1. Dataset construction

Categorizing Distortions: The focus of our work is to study technical distortions such as blur, over/under exposure, and noise, but not aesthetic flaws such as framing, mood, and content selection. While the latter are important, they involve different capture problems and perceptual processes and should be treated differently. As mentioned previously, natural distortions are extremely diverse and commingle with each other, making it hard to exhaustively categorize them [12, 26]. Since the images in our dataset were captured by the visually impaired, they are more heavily distorted. We focused on obtaining the labels on five major distortion categories: focus blur ('blurry'), motion blur ('shaky'), overexposure ('bright'), underexposure ('dark'), and noise ('grainy'), and two other categories: 'none' (the absence of distortion) and 'other' (non-identifiable distortions)

Cropping Patches: Relationships between local and global spatial quality have been shown to be important and, when modeled, lead to improved quality predictions [49, 48]. We carried these ideas further by studying the impact of the choice of patches on quality prediction. To do this, we divided the entire dataset into two random halves. On half the images, a random patch was cropped to 40% of each its linear spatial dimensions. On the other half, using the pyramid feature attention network [50], a most visually salient patch of the same (40%) dimensions was cropped. All of the patches have the same aspect ratios as the original image they were cropped from (refer Fig. 3).

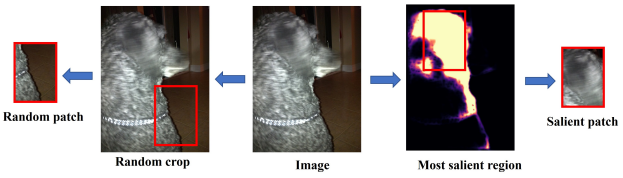

Fig. 3: **Two kinds of spatial patches** were cropped from images, all to 40% of the original image dimensions: randomly selected and salient patches cropped from disjoint halves of the overall image corpus.

Fig. 4: **Study workflow** for both image and patch sessions.

ORBIT Data: To study the cross-dataset performance of our model, we also created a separate, smaller dataset of images from the ORBIT database [29]. We captured 2,235 frame images and gathered global quality ratings on them (suppl. material for details).

3.2. Subjective Quality and Distortion Study

The human study was carried out on AMT in 3 stages - image, patch, and ORBIT sessions. Overall, 3,945 subjects participated in the study and, after rejection and cleaning, we collected an average of 34 ratings on each image and each patch. Our study was accessible to all platforms and geographical locations.

3.2.1 AMT Study Design

The study workflow is shown in Fig. 4. The subjects were asked to participate in two tasks - image quality rating and distortion type identification. Each subject was asked to read separate per task instructions, followed by a quiz to assess their understanding. After passing the quiz, there was a training phase containing five sample distorted images, to familiarize with the task (details in suppl. material). Following training is the testing phase where each subject rated 110 images, followed by answering a questionnaire.

3.2.2 Subject Rejection

As indicated in previous studies [49, 48, 8, 12, 39, 26, 43], online crowdsourcing carries the risk of distracted, inadequately equipped, disengaged, or even frankly dishonest subjects, so there is often a high percentage of unreliable labels. We used various strategies to screen the subjects using criteria applied both during and after the study.

During Task: During the instruction phase, we checked whether the subject's browser window, browser and OS version, and zoom (non-magnified) condition satisfied the requirements stated in the instructions. If they did not, their participation was ended. To detect dishonest workers, at the halfway point of each testing phase, we processed the scores already given to determine whether they had been giving unchanging ratings (only nudging the slider / supplying haphazard scores) on either tasks and where rejected accordingly.

Post Task: Of the 110 images viewed in a session, 5 were randomly repeated. A subject was rejected if their "repeat" scores were not consistently similar to the scores given the first time. We also included 5 images from the LIVE-FB

dataset [49] as "gold" set and screened the subjects if their ratings did not match the golden ones. Overall, we rejected the scores given by 814 subjects.

3.2.3 Data Cleaning

The remaining scores after subject rejection were processed by a series of data cleaning steps: **(1)** removed 43 images (1.3K ratings) of a constant value. **(2)** removed the ratings provided by subjects who did not wear their prescribed lenses (0.9% of total ratings removed). **(3)** applied the ITU-R BT.500-14 [18] (Annex 1, Sec 2.3) subject rejection protocol to screen 56 more outlier subjects **(4)** For each image and patch, we also rejected outliers from the individual score distributions, as follows. We first calculated the kurtosis [2] to determine the normality of the scores. If they were determined to be normal, the Z-score outlier rejection method [17] was applied. Otherwise, the Tukey IQR method [41] was applied. Overall, including all subject and score outlier rejections, around 1.7% of the ratings were tossed. We were left with about 2.7M subject scores (1.36M on images, 1.33M on patches) on VizWiz images, and 76K ratings on the ORBIT images.

3.2.4 Data Analysis

Inter-subject quality consistency: An inter-subject consistency test [49, 48] was carried out by randomly dividing the subject pool into two disjoint sets of equal size, then calculating the Spearman Rank Correlation Coefficient (**SRCC**) [22] between the two corresponding sets of MOS (mean opinion score). The average SRCC over 50 such random splits yields a useful measure of inter-subject consistency. The average SRCC on VizWiz images was **0.93**, on patches was **0.90** (**0.87** on random and **0.92** on salient patches), and on ORBIT was **0.93**. These results substantially validates the efficacy of our data collection and subject rejection processes.

Intra-subject quality consistency: We computed the Linear Correlation Coefficient (LCC) [36] between the mean of the ratings on the 5 "golden" images with the original scores. The median PCC value over all subjects was **0.90** for the VizWiz image study, **0.90** for the patch study, and **0.87** for the ORBIT study. Again, these high correlations help validate our overall subjective study protocol.

Patch vs Image quality: Fig. 5 shows scatter plots of image MOS against patch MOS, for both kinds of patches. The SRCC between image and patch MOS was **0.84** indicating a strong relationship between local and global image quality. The SRCC between image MOS and random and salient patch MOS was **0.82** and **0.86**, respectively, suggesting that salient patches could play a stronger role when representing global picture quality. This may be because some

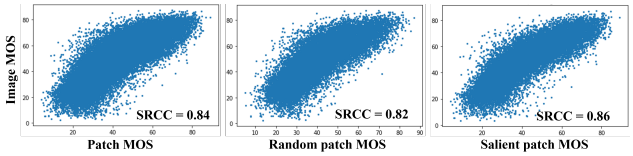

Fig. 5: **Scatter plots of patch vs image MOS correlations.** Image MOS vs all patches (left), random patch (middle) and salient patch (right) MOS cropped from the same image.

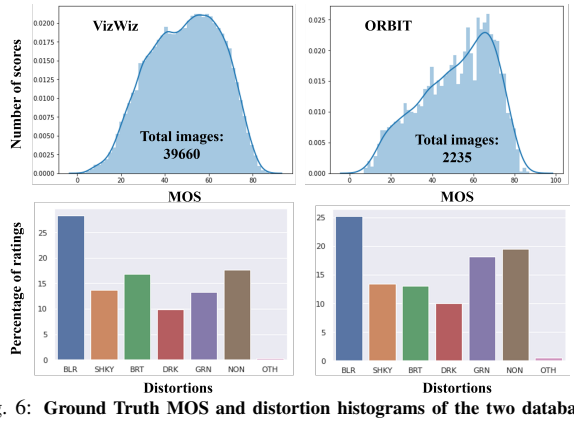

Fig. 6: **Ground Truth MOS and distortion histograms of the two databases**. Left column is the data collected on VizWiz [8] images, and right column is the data collected on ORBIT [29] images. The plots below show the distribution of the distortions in each dataset.

distortions are salient, and/or that distortions on salient regions are more annoying.

Distortion Score Analysis: To conduct a consistency analysis, we converted the binary distortion labels into probabilistic values by dividing the number of positive labels of each distortion by the total number of labels collected and then computed the correlations between the resulting vectors. The average SRCC (inter-subject consistency) values for the distortion categories were: blurry (**0.75**), shaky (**0.62**), bright (**0.68**), dark (**0.60**), grainy (**0.35**), and none (**0.85**). Some distortion categories were harder to consistently identify than others. The high agreement on 'none' shows that it is easier to determine the absence of distortions. Similarly, the SRCC values computed between image and patch distortions were: blurry (**0.73**), shaky (**0.68**), bright (**0.60**), dark (**0.62**), grainy (**0.46**), and none (**0.73**). The lower correlations for some distortions (like 'dark') suggests that the perception of distortions that are globally apparent may be more weakly impacted by local quality.

MOS and Distortion Distributions: Fig. 6 (top) plots the MOS distribution of the images in the VizWiz and ORBIT datasets. Fig. 6 (bottom) shows the proportional distribution of distortion ratings in the two datasets. As expected, 'blurry' was the most prominent distortion in both datasets. Overall, ORBIT contains more images of higher quality and with fewer distortions. Also, 'grainy' and 'dark' images, which occurred less often, are associated with less consistent ratings and are thus harder to predict. As we show in Sec. 4.2.2, this non-uniform distribution of distortion types makes it harder to train models that can perform equally well on all classes.

4. Modeling and Experiments

Our goal is to create an application that accurately predicts and provides feedback on the perceptual quality and distortions present in a picture. We designed efficient model architectures able to predict quality and distortions, while accounting for local-to-global percepts. Our model analyzes quality and distortion types both separately and together using a multi-task framework.

4.1. Data pre-processing

Unlike the quality scores, which could be used in their raw form, the distortion labels had to be transformed into suitable output labels for training. Since distortion type prediction is a classification task, binarizing the labels was our first choice. We considered multiple options to decide a threshold on the per-distortion proportions of ratings. However, binarization led to poor consistencies in the data (suppl. material), leading to worse predictions since hard labels reduce robustness on out-of-distribution samples [33]. Hence, we finally decided to train and test all our models using probabilistic labels.

4.2. Modeling

Given the two prediction tasks at hand, we studied both no-reference image quality models and distortion classification models, with a goal of building a single multi-task models capable of both tasks.

4.2.1 Image Quality Models

Architecture: The model structure consists of a deep CNN backbone, followed by two-dimensional global average pooling, then two fully connected layers of size 512 and 32, respectively, and a single output neuron with sigmoid activation. The model was trained for 10 epochs using Adam optimizer with MSE loss. The learning rate was set to 5×10^{-4} for the first 5 epochs, then with a decay rate of 0.1 per epoch. We experimented with ResNet-50V2 [16], Xception [9], and ResNeXt-50 [46] backbones pre-trained on ImageNet [10] and fine-tuned on VizWiz images.

Dataset Splits: We used the same train-validation-test split as provided by the authors of VizWiz-QualityIssues [8]. The training, validation, and testing set consists of 23.9K (60.3%), 7.7K (19.6%), and 8K (20.1%) images respectively. Similar split was applied to the patch dataset.

Baselines and evaluation metrics: The trained models were compared against several baselines, including shallow and deep learners (whose code was publicly available). We included the popular image quality prediction models, BRISQUE [30], NIQE [31], and FRIQUEE [13], which extract perceptually relevant statistical image features to train an SVR. We also compared against deep picture quality models such as CNNIQA [20] and NIMA [40] (with a VGG-16 [38] backbone and a single regressed quality score as output), PaQ-2-PiQ baseline, and PaQ-2-PiQ RoIPool with backbones pre-trained on LIVE-FB [49], then fine-tuned on our dataset. Similar to all standard work in the field of image quality assessment, we evaluated the model performances using SRCC and LCC.

Table 1: **Performance of image quality models evaluated on the new dataset.** Higher values indicate better performance.

Model	SRCC	LCC
BRISQUE [30]	0.71	0.72
NIQE [31]	0.68	0.70
FRIQUEE [13]	0.72	0.69
CNNIQA [20]	0.78	0.79
NIMA [40]	0.83	0.83
P2P-Baseline (ResNet-18) [49]	0.87	0.88
P2P-RoIPool (ResNet-18) [49]	**0.90**	**0.90**
Xception	0.86	0.88
ResNeXt-50	**0.90**	0.89
ResNet-50V2	**0.90**	**0.90**

Results: From Table 1, we note that models trained with shallow learners on extracted features yielded lower prediction accuracy than the deep models, reflecting the limited abilities of traditional features to capture complex distortions of natural images. CNNIQA [20], which is a shallow CNN model, outperformed the traditional algorithms, but fell short of the performances of deeper models. We observed that performance generally was higher for deeper models (ResNet-50V2, ResNeXt-50, and Xception) outperformed NIMA [40] implemented with a VGG-16 backbone. The PaQ-2-PiQ RoIPool model achieved the best performance, demonstrating the efficacy of exploiting the relationship between local (patch) and global quality prediction. The performances of the deeper backbones were all similar (and close to human performance – SRCC 0.93 as stated in Sec. 3.2.4), suggesting that more heavier models would not produce better performances.

4.2.2 Distortion Prediction Models

Architecture and Implementation: Because our models generates continuous probabilistic outputs, as described in Sec. 4.1, we treat distortion prediction as a regression problem. Similar to the quality model architecture, our proposed model consists of a deep CNN backbone, followed by global pooling, and two fully connected layers. Instead of producing a single output, it has seven output neurons, each expressing a score for a separate distortion class. As before, we experimented with three backbones - ResNet-50V2 [16], Xception [9], and ResNeXt-50 [46]. The hyperparameters were kept the same except the initial learning rate was set to 10^{-3}. The same train-validation-test split was used.

Baselines and evaluation: The trained models were compared against other deep models. We include two models from [3], where the authors used pre-trained Atrous VGG-

Table 2: **Performances of distortion prediction models** on the new dataset. All values are SRCC, where higher values indicate better performance.

Model	BLR	SHK	BRT	DRK	GRN	NON
AtrousVGG-16 [3]	0.75	0.73	0.60	0.69	0.45	0.77
ResNet-101 [3]	0.81	0.77	**0.69**	0.75	0.45	0.81
Xception [8]	0.79	0.75	0.65	0.80	**0.56**	0.78
ResNeXt-50	**0.82**	0.75	0.68	0.79	**0.56**	**0.82**
ResNet-50V2	0.81	**0.78**	0.67	**0.81**	0.50	**0.82**

16 [7] and ResNet-101 [15] backbones with a single fully-connected head layer to predict synthetic distortions. We also tested the model [8] composed of an Xception [9] backbone (pre-trained on ImageNet [10]), fine-tuning the head layers only. The distortion labels and outputs lie within [0,1], and we again used SRCC to evaluate the performance.

Results: As may be observed from Table 2, the fine-tuned deep models outperformed the baselines. Atrous VGG-16 [3] performed the worst, whereas the ResNet-50V2 [16] and ResNeXt-50 [46] models consistently performed best on most distortion classes. All the fine-tuned models yielded similar performances across all classes. However, the distortion distribution in the dataset is quite skewed (Fig. 6), hence the prediction performances varied across classes. The low performance on the 'bright' and 'grainy' classes is consistent with the low agreement among the subjects on these distortions (Sec. 3.2.4).

4.2.3 Multi-task Models

Architecture and Implementation: Combining both tasks – quality and distortion type predictions – into a single model bears two advantages: a) fewer computations and faster inferencing, crucial for supplying real-time feedback to users (Sec. 5.2); b) shared distortion and quality features can lead to better predictions [14]. Starting with PaQ-2-PiQ (P2P) RoIPool [49] as a base, we modified it by attaching a multi-task head to conduct both quality and distortion predictions. This multi-task model, which we call P2P++, produces quality and distortion predictions for each class, on both entire images and local patches simultaneously (Fig. 7). The head has a shared layer of size 512, followed by two separate layers of size 32, dedicated to separate tasks. In addition to training P2P++ (which has a ResNet-18 [15] backbone pre-trained on LIVE-FB [49]), we also experimented with ResNet-50V2 [16] and Xception [9] baselines trained on images only. The hyperparameters for training were the same as for the distortion model setup, using the same train-validation-test split.

Baselines and evaluation: We compared the performance of our models against two multi-task deep models - IQACNN++ [21] and QualNet [14]. IQACNN++ consists of a shallow CNN backbone, whereas QualNet contains a VGG-16 [38] backbone and predicts global quality using fused distortion and quality features.

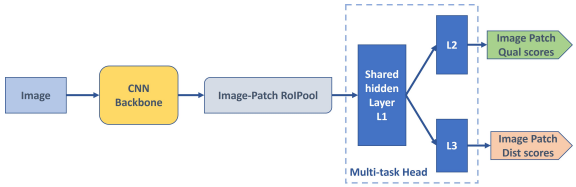

Fig. 7: **The proposed P2P++ model** extends the PaQ-2-PiQ [49] RoIPool model by including a multi-task head that simultaneously produces both quality and distortion scores, at both patch and whole image scales.

Table 3: **Performance of the multi-task models** on the new subjective test dataset. All values are SRCC, and higher values indicate better performance.

Model	BLR	SHK	BRT	DRK	GRN	NON	Qual
IQACNN++ [21]	0.65	0.52	0.27	0.57	0.40	0.62	0.78
QualNet [14]	0.70	0.60	0.46	0.70	0.29	0.71	0.81
Xception	0.78	0.73	**0.64**	0.75	0.51	**0.78**	0.88
ResNet-50V2	0.80	0.76	0.62	0.77	0.51	0.76	**0.90**
P2P++	**0.82**	**0.77**	0.60	**0.78**	**0.53**	**0.78**	**0.90**

Table 4: **Quality prediction results on the patches** in the new subjective dataset. Higher values indicate better performance.

	All Patches		Salient		Random	
Model	SRCC	LCC	SRCC	LCC	SRCC	LCC
IQACNN++ [21]	0.71	0.71	0.71	0.70	0.72	0.72
QualNet [14]	0.77	0.77	0.78	0.77	0.77	0.76
Xception	0.84	0.84	0.85	0.84	0.84	0.83
ResNet-50V2	0.87	**0.87**	0.87	0.87	0.86	0.86
P2P++	**0.88**	**0.87**	**0.88**	**0.88**	**0.87**	**0.87**

Results: From Table 3, it may be observed that the shallow IQACNN++ [21] model yielded the worst results. QualNet [14] was able to outperform IQACNN++, but struggled on multiple distortion categories. The larger models equipped with ResNet-50V2 and Xception backbones performed very well, but the much lighter P2P++ model was able to achieve the best performance on almost all categories. Again, by inferencing on learned local-to-global quality and distortion features, better results were obtained at lower cost. As before, all of the models had more difficulty predicting the 'bright' and 'grainy' distortion types.

4.3. Ablations

Performance on patches: Table 4 summarizes the quality performance of the multi-task models on patches (distortion results in Suppl. material). This is important, since giving feedback on local distortion occurences may further assist visually impaired users. P2P++ performs the best on both the random and salient patches, closely followed by ResNet-50V2. This validates the localization capabilities of the patch model. The performance on salient patches was slightly better than on random patches, since, perhaps they often capture visibly obvious and annoying distortions that draw attention and are easier to predict.

Failure Cases: Fig. 8 (a) was rated high (MOS = 76.2) by

(a) Predicted: 50.2 Ground Truth MOS: 76.2 (b) Predicted: 62.7 Ground Truth MOS: 42.5

Fig. 8: **Failure cases:** VizWiz images where predictions differed the most from human quality scores. Reduce fig size

the humans, but a low predicted score (MOS = 50.2) from P2P++. Perhaps the blurry "bokeh" effect regions of the hand and background were less noticeable to raters than the high quality (salient) foreground. The image Fig. 8 (b) was rated as worse (MOS = 42.5) by the subjects than by P2P++ (MOS = 62.7). The non-uniformity of the blur across the image could've caused this discrepancy. While the sharp and distorted regions of the image are of roughly equal areas, the distortions were likely more salient to the human subjects, causing them to rate it severely. The same regions were likely predicted as less salient by P2P++. These results suggest more work needs to be done on understanding the interplay between saliency, distortion annoyance, and bokeh.

5. Applications of the proposed model

The models described on Sec. 4 can be extended to provide visualization and feedback to directly assist visually impaired users, which we describe next.

5.1. Predicting quality and distortion maps

The P2P++ model can be used to compute both spatial quality maps and distortion classification maps. Since it is trained on both global and local patch labels, it is flexible enough to compute the quality predictions and distortion type predictions on any number and sizes of image patches. Inspired by [49], we utilized these outputs to create perceptual quality and distortion classification maps that span the entire image space. To generate spatial quality maps, the image is divided into non-overlapping patches of size $N \times N$, on which predicted quality scores are obtained from the model output on every patch. Similarly, on each patch, a predicted distortion vector is obtained, with multiple values corresponding to each distortion type. Each distortion output can be used to generate a corresponding distortion-specific map. The patch size (N) is easily varied, allowing the generation of finer or coarser maps.

Fig 9 shows the predicted quality and distortion maps (for the two most prominent predicted distortions) computed on a sample test image. The quality map accurately predicted the bottom-right part of the image to be of highest quality, while the distortion maps predicted the bottom-left area to be blurry, and the top most region of the image to be underexposed. This example shows how perceived quality and distortion localization in an image affect each other.

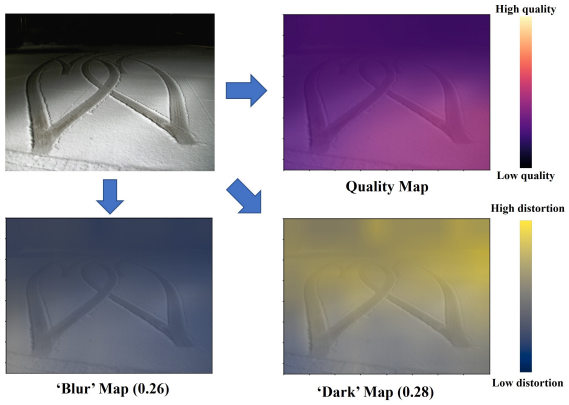

Fig. 9: **Spatial quality and distortion maps:** Predicted perceptual quality and distortion maps were generated on a sample image from our new database 5.1. The top-right image shows the predicted perceptual quality map (blended with the original image using a magma colormap). The bottom two images show the 'blur' and 'dark' distortion maps (and their global scores) blended with a cividis colormap. Best viewed in color.

5.2. Feedback to assist visually challenged

Guided Photography: Our overarching goal is to build a system able to provide feedback to visually challenged users so that they can take better pictures. This is a very challenging, multi-dimensional, and human-oriented problem, which requires extensive ergonomic and validation studies with visually impaired volunteer subjects. At this point, we built a prototype early-stage, guided feedback system as a demonstration of how our work can be used to assist visually challenged users to take better photos. The currently implemented framework is illustrated in Fig. 10. The assistive model has two parts, a quality feedback loop and a distortion feedback loop. The high-level, immediate model outputs are an approximate English translation of the global picture quality prediction and expressions of the predicted distortion levels. Specifically, the user is provided an image rating from among 'Bad' (0-20), 'Poor' (20-40), 'Fair' (40-60), 'Good' (60-80), and 'Excellent' (80-100). If the user is satisfied with the quality, he/she can choose to save it, or ask for distortion feedback otherwise. In our current prototype which is implemented on a workstation (but see below for parallel work), the feedback is given by output text; naturally, transcribed audio expressions will be used in practice. If the quality is substandard, then further feedback is required to make the application useful. If feedback on the distortion is requested, the user is informed of the three major distortions determined to be present in the image, along with the severity of each: High (> 0.50), Moderate ($0.20 - 0.50$), and Low (< 0.20).

Based on the nature and severities of the distortions detected, our system also suggests simple ways (**base feedback**) to mitigate them. A description of the feedback that

is currently given is available in Suppl. material. As the user becomes more adept at the P2P++ system, they will be able to request and take advantage of additional, more **detailed feedback** on the picture distortions. To facilitate this, P2P++ also generates 3×3 distortion maps for the three most dominant impairments, and informs the user of their relative location in the image (top-left, bottom-right, center, etc.), as also depicted in Fig. 10.

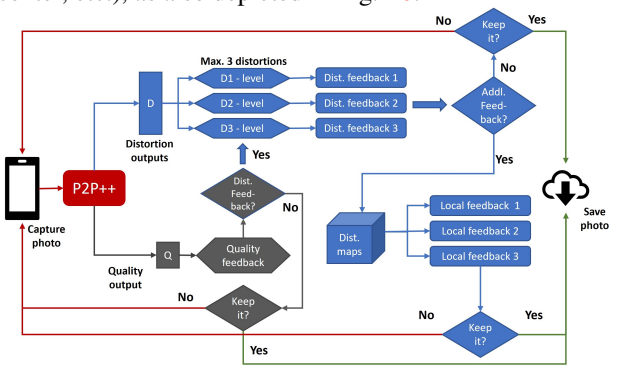

Fig. 10: **Guided Photography Framework:** Flowchart of the proposed assistive photography framework (Sec. 5.2), showing the series of prompts and advice given to guide visually challenged users, from capturing through saving a satisfactory photo. Feedback examples shown in Fig. 1.

Automated Photography: Although guided photography promises to be a transformative technology, we acknowledge that much work remains in terms of developing an ideal feedback language and interface, which in term will require working with visually challenged subjects to test and advance the system. In the interim, there are more immediate ways to assist visually challenged users to take better pictures via much simpler, albeit less comprehensive application, which can automatically help them take a better quality photo. This can be accomplished by capturing a short video clip of the scene the subject is trying to photograph, that includes and is approximately centered at the moment the 'shutter button' is depressed. By using a broad sampling of a single frame per second, a fairly wide range of qualities may be presented. Given the sampled frames, P2P++ then computes the global quality of each to determine the frame having the highest perceptual quality. The user is provided a feedback on this quality ('Poor' to 'Excellent') and given the option to save the image. A simple demonstration on an ORBIT [29] video is shown in Fig. 11.

Performance on ORBIT: To study the generalizability of our model, and the representation capabilities of our dataset, we also sought to test P2P++ on other, independent VC-UGC data than our new dataset. Since we could not find any such datasets, we evaluated and compared the multi-task models on a special-purpose excerpt we created from the ORBIT dataset, consisting of frames sampled from the ORBIT [29] videos. As may be observed from Table 5, P2P++ performs very well, and generally better than the much

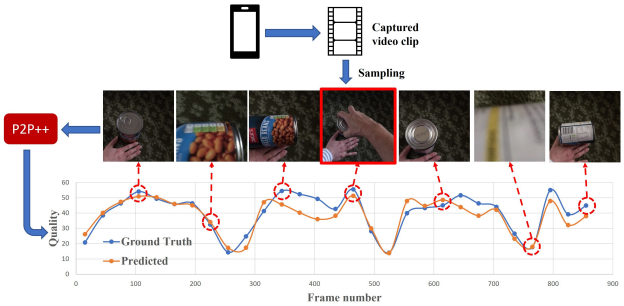

Fig. 11: **Automated photography:** P2P++'s quality outputs on a sequence are used to determine the highest quality frame among a temporally sampled ORBIT [29] video (highlighted in red). Best viewed in color.

Table 5: **Performance of the multi-task models** when trained on our new dataset and tested on the excerpted ORBIT dataset. All values are SRCC, and higher values indicate better performance.

Model	BLR	SHK	BRT	DRK	GRN	NON	Qual
IQACNN++ [21]	0.56	0.37	0.06	0.83	0.05	0.38	0.78
QualNet [14]	0.69	0.59	0.11	0.79	0.17	0.70	0.85
Xception	0.70	0.64	0.27	0.81	**0.37**	**0.72**	0.83
ResNet-50V2	0.68	0.69	0.17	**0.84**	0.18	0.65	**0.86**
P2P++	**0.72**	**0.71**	**0.30**	0.83	**0.37**	**0.72**	**0.86**

heavier ResNet-50V2 and Xception models. All of our models outperform other multi-task models when trained on ournew database and tested on the excerpted ORBIT dataset. Attaining such high performance on most distortion classes on ORBIT validates the generalizability of P2P++ to other VC-UGC media. The lower performance (of all models) on the 'bright' and 'grainy' categories is again due to subject ambiguity on these classes. Fig. 11 illustrates the actual performance and outputs produced by P2P++ when compared to the ground truth quality scores obtained on an ORBIT video.

6. Concluding Remarks

The success of computer vision algorithms can be largely measured by the benefits granted to ordinary people to enhance their quality of life. To that end, assisting visually impaired people to take better quality pictures can give them more prominent voices on social media platforms, and can also assist them with other visual tasks such as recognition and captioning. Assessing perceptual quality and distortions on VC-UGC is a difficult, but important and little-addressed problem. Our work makes substantive progress towards that goal by the proposed VC-UGC targeted dataset, a VC-UGC quality and distortion prediction model, and a prototype system that supplies specialized feedback to help guide, assist, automate, and improve their photographic efforts. Of course, while this work is a step in the right direction, this field is still nascent with very significant challenges remaining.

References

[1] TapTapSee. [Online] Available: https://taptapseeapp.com/. 1, 2

[2] *Measure of Kurtosis*, pages 343–343. Springer New York, New York, NY, 2008. 4

[3] N Ahn, B Kang, and K Sohn. Image distortion detection using convolutional neural network. *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 220–225, 2017. 2, 5, 6

[4] C. L. Bennett, E. Jane, M. E. Mott, E. Cutrell, and M. R. Morris. *How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2018. 1

[5] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: Nearly real-time answers to visual questions. UIST '10, page 333–342, New York, NY, USA, 2010. Association for Computing Machinery. 1, 2

[6] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *2016 IEEE Int'l Conf. Image Process. (ICIP)*, pages 3773–3777, Sep. 2016. 2

[7] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2016. 6

[8] T Chiu, Y Zhao, and D Gurari. Assessing image quality issues for real-world problems. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653, 2020. 1, 2, 3, 4, 5, 6

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2016. 5, 6

[10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vision and Pattern Recogn.*, pages 248–255, June 2009. 2, 5, 6

[11] D. Ghadiyaram and A. C. Bovik. Blind image quality assessment on real distorted images using deep belief nets. In *IEEE Global Conference on Signal and Information processing*, volume pp. 946–950, pages 946–950, Atlanta, GA, 2014. 2

[12] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372-387, Jan 2016. 2, 3

[13] D. Ghadiyaram and A. C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, vol. 17, no. 1, art. 32, pp. 1-25, January 2017. 2, 5

[14] S. A. Golestaneh and K. Kitani. No-reference image quality assessment via feature fusion and multi-task learning, 2020. 2, 6, 8

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vision and Pattern Recogn.*, pages 770–778, 2016. 6

[16] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016. 5, 6

[17] B. Iglewicz and D. C. Hoaglin. Volume 16: How to Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques*, 1993. 4

[18] International Telecommunication Union. ITU-R BT.500-14, methodologies for the subjective assessment of the quality of television images. [Online] Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-14-201910-I!!PDF-E.pdf. 4

[19] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, page 203–210, New York, NY, USA, 2011. Association for Computing Machinery. 2

[20] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 1733–1740, June 2014. 1, 5

[21] L Kang, P Ye, Y Li, and D Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2791–2795, 2015. 2, 6, 8

[22] Maurice George Kendall. Rank correlation methods. 1948. 4

[23] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE J. of Selected Topics in Signal Process.*, vol. 11, no. 1, pp. 206-220, Feb 2017. 2

[24] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130-141, Nov 2017. 2

[25] E. C. Larson and D. M. Chandler. Categorical image quality (CSIQ) database, 2010. [Online] Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23. 2

[26] H. Lin, V. Hosu, and D. Saupe. Koniq-10K: Towards an ecologically valid and large-scale IQA database. *arXiv preprint arXiv:1803.08489*, March 2018. 2, 3

[27] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, page 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 2

[28] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5988–5999, New York, NY, USA, 2017. Association for Computing Machinery. 1

[29] D Massiceti, L Zintgraf, J Bronskill, L Theodorou, M T Harris, E Cutrell, C Morrison, K Hofmann, and S Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. 2021. 2, 3, 4, 8

[30] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. 1, 2, 5

[31] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely blind" image quality analyzer. *IEEE Signal Processing Letters*, vol. 20, pp. 209-212, 2013. 2, 5

[32] J. Park, S. Lee, and A.C. Bovik. VQpooling: Video quality

pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610-620, Feb. 2013. 1

[33] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, 2019. 5

[34] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. . J. Kuo. Color image database TID2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing*, volume vol. 30, pp. 106-111, pages 106–111, June 2013. 2

[35] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009. 2

[36] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988. 4

[37] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov 2006. 2

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 5, 6

[39] Z. Sinno and A.C. Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612-627, Feb. 2019. [Online] LIVE VQC Database: http://live.ece.utexas.edu/research/LIVEVQC/index.html. 3

[40] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998-4011, Aug 2018. 1, 5

[41] J. Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co, Reading, Mass, 1977. 4

[42] M Vázquez and A Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. 21(5), Nov. 2014. 2

[43] Y Wang, S Inguva, and B Adsumilli. Youtube UGC dataset for video compression research. 2019. 3

[44] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98-117, Jan 2009. 1

[45] S Wu, J Wieland, O Farivar, and J Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1180–1192, New York, NY, USA, 2017. Association for Computing Machinery. 1

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 5, 6

[47] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 1098–1105, June 2012. 2

[48] Z Ying, M Mandal, D Ghadiyaram, A Bovik University of Texas at Austin, and AI Facebook. Patch-vq: 'patching up' the video quality problem. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14014–14024, 2021. 1, 3, 4

[49] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. C. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3572–3582, 2020. 1, 2, 3, 4, 5, 6, 7

[50] T. Zhao and X Wu. Pyramid feature attention network for saliency detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3080–3089, 2019. 3

[51] Y Zhao, S Wu, L Reynolds, and S Azenkot. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), Dec. 2017. 1

[52] Y. Zhong, P. J. Garrigues, and J. P. Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, New York, NY, USA, 2013. Association for Computing Machinery. 1, 2