In this supplementary material, we first provide a brief description of the datasets used in our experiment Section (Section A). Next, the proof of Theorem 1 is provided in Section B. In Section C, we conduct ablation studies about the detectors' robustness towards unseen corruptions. Besides, we discuss the limitations of our proposed method in Section D. Finally, in Section E, we graphically illustrate the benefits of applying our QAD for deepfake detection problems.

A. Datasets

We describe here *seven* popular benchmark deepfake datasets used to verify our proposed QAD:

- **NeuralTextures**. Facial reenactment is a video re-rendering approach that uses the *Neural Textures* [58] technique. This method employs neural textures, which are learned feature maps placed on top of 3D model proxies, as well as a deferred neural renderer. The NeuralTextures dataset used in our study provide facial alterations to the mouth region, while the rest of the face remains unchanged.

- **Deepfakes**. Each autoencoder in the DeepFakes dataset is trained on the source face and the target face separately before encoding the data. An artificial face is created by using the decoder trained on the target face to decode the embedding representation of the source face. Note that though DeepFakes originally referred to a particular face swapping technique, the term has now come to apply to AI-generated facial modification approaches in general.

- **Face2Face**. Face2Face [59] is a method of real-time facial recreation in which the identity of the target individual is maintained while their expression mimics that of the source. More specifically, a series of manually chosen key-frames is used in conjunction with a flexible model-based bundling strategy to recover the identification associated with the target face. Target backdrop and illumination are preserved while expression coefficients from the source face are transmitted.

- **FaceSwap**. FaceSwap [6] is an easy-to-use program based on the visual architectures of the faces being swapped. To create a 3D representation of a person's face, 68 individual landmarks on the face are taken into account. Finally, it does color correction after projecting the facial areas back to the target face by reducing the pair-wise landmark errors.

- **FaceShifter**. FaceShifter [34] is a two-step face-swapping system. The first step employs a generator with Adaptive Attentional Denormalization layers and an encoder-based multi-level feature extractor for a target face. In particular, the Adaptive Attentional Denormalization creates a synthetic face by fusing a person's identity with their physical characteristics. To improve face occlusions, they then created a unique Heuristic Error Acknowledging Refinement Network in the second phase.

- **CelebDFv2**. CelebDFv2 [67] is a large, difficult dataset for deepfake forensics. It contains 590 original YouTube videos with subjects of various ages, ethnic backgrounds, and genders, as well as 5,639 DeepFake videos. The synthesized videos are generated by the Deepfake synthesis algorithm, followed by several refining steps targeting specific visual artifacts such as low resolution, color mismatch, and temporal flickering.

- **Face Forensics In the Wild (FFIW10K)**. FFIW10K consists of 10,000 forgeries videos of high quality, with an average of three human faces every frame. Each video is created using one of three face-swapping techniques: DeepFaceLab [44], FS-GAN [42], and FaceSwap [6], in order to increase the variety of manipulated videos.

For FaceForensics++ datasets, we follow the same preprocessing step as in ADD [32] for each modality. And, with a 3-tuple quality modality (raw, c23, and c40) in our experiment, we have 276,480, 53,760, and 53,750 for training, validation, and testing, respectively. Similarly, for both CelebDF-v2 and FFIW10K datasets, we used 360,000, 49,200, and 49,200 images for training, validating, and testing, respectively. To ensure the fair comparison, faces cropped from videos are resized to 128 × 128, then prepossessed to the desired input size of each benchmark detector, *e.g.*, 299 × 299 for XceptionNet.

B. Proof of Theorem 1

Let the optimization function be $\mathcal{L}(f(x), y) = 1 - \sigma_T(f(x, y))$, where σ_T is the softmax function with temperature $T > 0$:

$$\sigma_T(f(x), y) = \frac{\exp(f(x,y)/T)}{\sum_{k=1}^{2}\exp(f(x,k)/T)}. \tag{11}$$

We introduce a function class $\Phi_{\mathcal{W}} \subseteq [0,1]^{\mathcal{X}\times\mathcal{Y}}$ from the distribution of raw images: $\Phi_{\mathcal{W}} = \{(x_r, y) \mapsto \mathcal{L}(f(x_r), y) : f \in \mathcal{F}\}$.

Let $\nu \in \{1, 2, ..., \log_2(n)\}$ and $\tau_\nu = 2^{2-\nu}$, and we define function classes as follows:

$$\begin{aligned}\Phi_\nu = \{(x_c, y) \mapsto & \mathcal{L}(f(x_c), y) : \\ & \mathbb{E}_{\mathcal{D}}\left[|\sigma_T(f(x_r)) - \sigma_T(f(x_c))|\right] \le \tau_\nu\}.\end{aligned} \tag{12}$$

For any $\mathcal{L} \in \Phi_\nu$, and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the classical generalisation bound with the Rademacher complexity [2, 40] is defined as follows:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(f(x_c), y)] \leq \mathbb{E}_\mathcal{D}[\mathcal{L}(f(x_c), y)] + 2\mathfrak{R}_\mathcal{D}(\Phi_\nu) \\ + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right),\end{aligned} \tag{13}$$

where

$$\mathfrak{R}_\mathcal{D}(\Phi_\nu) = \frac{1}{n}\mathbb{E}_\pi\left[\sup_{\mathcal{L}\in\Phi_\nu}\sum_{i=1}^{n}\pi_i\mathcal{L}(f(x_c), y)\right], \tag{14}$$

and $\pi_1, ..., \pi_n$ are i.i.d. Rademacher random variables with $P(\pi_i = 1) = P(\pi_i = -1) = \frac{1}{2}$. The Rademacher complexity $\mathfrak{R}_\mathcal{D}$ measures the rate that the empirical risk converges to the population risk.

Moreover, we also have:

$$\begin{aligned}\mathfrak{R}_\mathcal{D}(\Phi_\nu) &= \frac{1}{n}\mathbb{E}_\pi\left[\sup_{\mathcal{L}\in\Phi_\nu}\sum_{i=1}^{n}\pi_i\mathcal{L}(f(x_c), y)\right] \\ &= \frac{1}{n}\mathbb{E}_\pi\Bigg[\sup_{\mathcal{L}\in\Phi_\nu}\sum_{i=1}^{n}\pi_i(\mathcal{L}(f(x_c), y) - \mathcal{L}(f(x_r), y) \\ &\qquad + \mathcal{L}(f(x_r), y))\Bigg] \\ &\leq \frac{1}{n}\mathbb{E}_\pi\left[\sup_{\mathcal{L}\in\Phi_\nu}\sum_{i=1}^{n}|\pi_i||\mathcal{L}(f(x_c), y) - \mathcal{L}(f(x_r), y)|\right] \\ &+ \frac{1}{n}\mathbb{E}_\pi\left[\sup_{\mathcal{L}\in\Phi_\mathcal{W}}\sum_{i=1}^{n}\pi_i\mathcal{L}(f(x_r), y)\right] \\ &\leq \tau_\nu + \mathfrak{R}_\mathcal{D}(\Phi_\mathcal{W})\end{aligned} \tag{15}$$

Replacing Eq. 15 into Eq. 13, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(f(x_c), y)] \leq &\mathbb{E}_\mathcal{D}[\mathcal{L}(f(x_c), y)] + 2\tau_\nu \\ &+ 2\mathfrak{R}_\mathcal{D}(\Phi_\mathcal{W}) + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right).\end{aligned} \tag{16}$$

In addition, for every $\mathcal{L}(f(x_c), y))$, there always exists τ_ν, such that:

$$\tau_\nu \geq \mathbb{E}_\mathcal{D}\left[\| \; \sigma_T(f(x_r)) - \sigma_T(f(x_c)) \; \|\right] \geq \frac{1}{2}\tau_\nu - 2^{1-\log_2(n)}. \tag{17}$$

Then

$$\tau_\nu \leq \frac{4}{n} + 2\mathbb{E}_\mathcal{D}\left[\| \; \sigma_T(f(x_r)) - \sigma_T(f(x_c)) \; \|\right] \tag{18}$$

Now, Eq. 16 can be rewritten as:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(f(x_c), y)] \leq &\mathbb{E}_\mathcal{D}[\mathcal{L}(f(x_c), y)] + \frac{8}{n} \\ &4\mathbb{E}_\mathcal{D}\left[|\sigma_T(f(x_r)) - \sigma_T(f(x_c))|\right] \\ &+ 2\mathfrak{R}_\mathcal{D}(\Phi_\mathcal{W}) + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right).\end{aligned} \tag{19}$$

Next, we rewrite the loss function for a compressed image, x_c, as follows:

$$\begin{aligned}\mathcal{L}(f(x_c), y) &= \frac{\sum_{i\neq y}\exp(f(x_c, i)/T)}{\sum_{i=1}^{2}\exp(f(x_c, i)/T)} \\ &= \frac{1}{1 + \frac{\exp(f(x_c, y)/T)}{\sum_{i\neq y}\exp(f(x_c, i)/T)}} \\ &= \frac{1}{1 + \exp(f(x_c, y)/T - \ln\big(\sum_{i\neq y}\exp(f(x_c, i)/T)\big))} \\ &= s\Big(-f(x_c, y)/T + \ln\big(\sum_{i\neq y}\exp\left(f(x_c, i)/T\right)\big)\Big),\end{aligned} \tag{20}$$

where $s(\cdot)$ is the sigmoid function. However, note that the second term in the sigmoid function of Eq. 20 belongs to the family of Log-Sum-Exp (LSE) function, and $s(\cdot)$ is a monotonically increasing function. Then, we have:

$$\mathcal{L}(f(x_c), y) \geq s\Big(-f(x_c, y)/T + f(x_c, \tilde{y})/T\Big), \tag{21}$$

where $\tilde{y} = \arg\max_{i\neq y} f(x_c, i)$. Since we always have $s(t) \geq \frac{1}{2}\mathbb{I}(t \geq 0), \forall t \in \mathbb{R}$, then

$$\begin{aligned}s\Big(-f(x_c, y)/T + f(x_c, \tilde{y})/T\Big) \geq \\ \frac{1}{2}\mathbb{I}\Big(-f(x_c, y)/T + f(x_c, \tilde{y})/T \geq 0\Big),\end{aligned} \tag{22}$$

or

$$\begin{aligned}2\mathcal{L}(f(x_c), y) &\geq \mathbb{I}\Big(f(x_c, y)/T \leq f(x_c, \tilde{y})/T\Big) \\ &= \mathbb{I}\Big(\hat{y}(x_c) \neq y\Big).\end{aligned} \tag{23}$$

Combining Eq. 19 and Eq. 23,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{I}\{\hat{y}(x_c) \neq y\;\}\right] \leq 2\mathbb{E}[\mathcal{L}(f(x_c), y)] \\ &\leq 2\mathbb{E}_\mathcal{D}[\mathcal{L}(f(x_c), y)] + \frac{16}{n} \\ &8\mathbb{E}_\mathcal{D}\left[\| \; \sigma_T(f(x_r)) - \sigma_T(f(x_c)) \; \|\right] \\ &+ 4\mathfrak{R}_\mathcal{D}(\Phi_\mathcal{W}) + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right).\end{aligned} \tag{24}$$

As softmax is a $L-Lipschitz$ function [14] with $L = 1/T$, we obtain:

$$\mathbb{E}\left[\mathbb{I}\{\hat{y}(x_c) \neq y\}\right] \leq 2\mathbb{E}_{\mathcal{D}}[\mathcal{L}(f(x_c), y)] + \frac{16}{n} \frac{8}{T}\mathbb{E}_{\mathcal{D}}\left[\| f(x_r) - f(x_c) \|\right] + 4\mathfrak{R}_{\mathcal{D}}(\Phi_{\mathcal{W}}) + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right). \tag{25}$$

C. Additional Experiments

Robustness against unseen corruptions. Although our QAD does not intend to defend against all image corruption types, we investigate the robustness of our model and other detectors under unseen perturbations. All defenders in our experiment are trained on FaceForensics++ [50] including *five* typical deepfakes: NeuralTextures, DeepFakes, Face2Face, FaceSwap, and FaceShifter, with their quality modalities: raw, c23, and c40. In the inference phase, we apply *five* operations with *five* severity levels, as given in [28]: saturation, contrast, block-wise distortion, white Gaussian noise, and blurring. The results are indicated in Fig. 6. Although none of the perturbations are included in the training phase, generally, our proposed QAD-E achieves the best robustness compared to previous SoTA approaches. One may notice that MAT [69] is a competitive defender; it obtains more robustness in the worst case by using a large input size of an image, *i.e.*, 380 × 380, which can alleviate the perturbations' effects. Finally, we believe that our method can be generalized to different corruptions when including them in the training phase, making detectors more robust. However, this is out of our study's scope, which mainly targets deepfake compression.

D. Limitations

We can point out two limitations of QAD. First, our proposed method relies on the existence of a M-tuple of quality modalities in the training dataset. While this requirement in the research environment is usually satisfied, a few deepfake datasets contain only videos in different qualities and conditions, such as DFDC [9]. Therefore, possible future research could focus on utilizing unpaired images of various qualities to robust the deepfake detector.

Secondly, as we target detecting deepfake in multi-quality and we did not mining fine-grained deepfake artifacts as in [69], our QAD can be less generalized when validating across datasets. As shown in Table 6, we trained all detectors on *five* FaceForensics++ [50] datasets with their *three* versions: raw, c23, and c40. In the inference phase, we test the pre-trained models on DFDC [9] and WildDeepfake [71] datasets. While our QAD, especially when integrating with EfficientNet-B1, outperforms previous works on the WildDeepfake, it still needs to improve on the DFDC dataset. Our future work will focus on designing a good metric learning framework that can be generalized towards both input quality and cross-domain deepfakes.

Method	Training set		Test set			
	FF++		DFDC		WildDeepfake	
	ACC	AUC	ACC	AUC	ACC	AUC
MesoNet [1]	61.6	65.6	60.3	71.4	54.4	55.8
Rössler *et al.* [50]	79.4	86.4	57.6	66.0	61.1	66.4
F^3Net [45]	75.4	84.2	54.1	66.4	58.9	64.5
MAT [69]	77.3	86.8	**65.1**	**71.9**	63.6	71.0
Fang & Lin [11]	80.7	89.0	63.2	70.2	63.1	67.7
SBIs [53]	68.9	86.0	60.5	70.9	59.2	65.5
QAD-R	85.3	93.4	61.4	67.0	62.5	68.9
QAD-E	**87.8**	**95.6**	56.5	65.3	**65.5**	**74.7**

Table 6. Cross-validation performance of models that trained on FF++ and validated on DFDC and WildDeepfake datasets.

E. Grad-CAM Results

Gradient-weighted Class Activation Mapping (Grad-CAM) [52] applies gradients from a high-probability prediction class to a lower level convolutions layer, resulting in a coarse localization map that highlights critical locations in the image. Positive layer outputs and higher gradient values generate more activation areas, which are illustrated in red in Fig. 7 and Fig. 8. On the other hand, negative pixels or low gradients produce fewer activation regions, which are represented in blue. In this experiment, we visualize ResNet50 baseline and our QAD-R on the *five* FaceForensics++ datasets: NeuralTextures, Deepfakes, Face2Face, FaceSwap, FaceShifter, and explain the benefits of our training frameworks as follows:

- **Consistently activating important regions across quality modalities.** Since ResNet-50 is trained without any regularization, it considers images of different quality as different images. Therefore, the activation regions can vary across quality modalities, which is indicated by the red arrows in Fig. 7, resulting in a large proportion of wrong predictions in low-quality images. In contrast, our QAD utilizes the HSIC to maximize the dependence between quality modalities, it regulates activation regions to be similar, ensuring its generalizability for detecting different quality deepfakes, as show in Fig. 7.

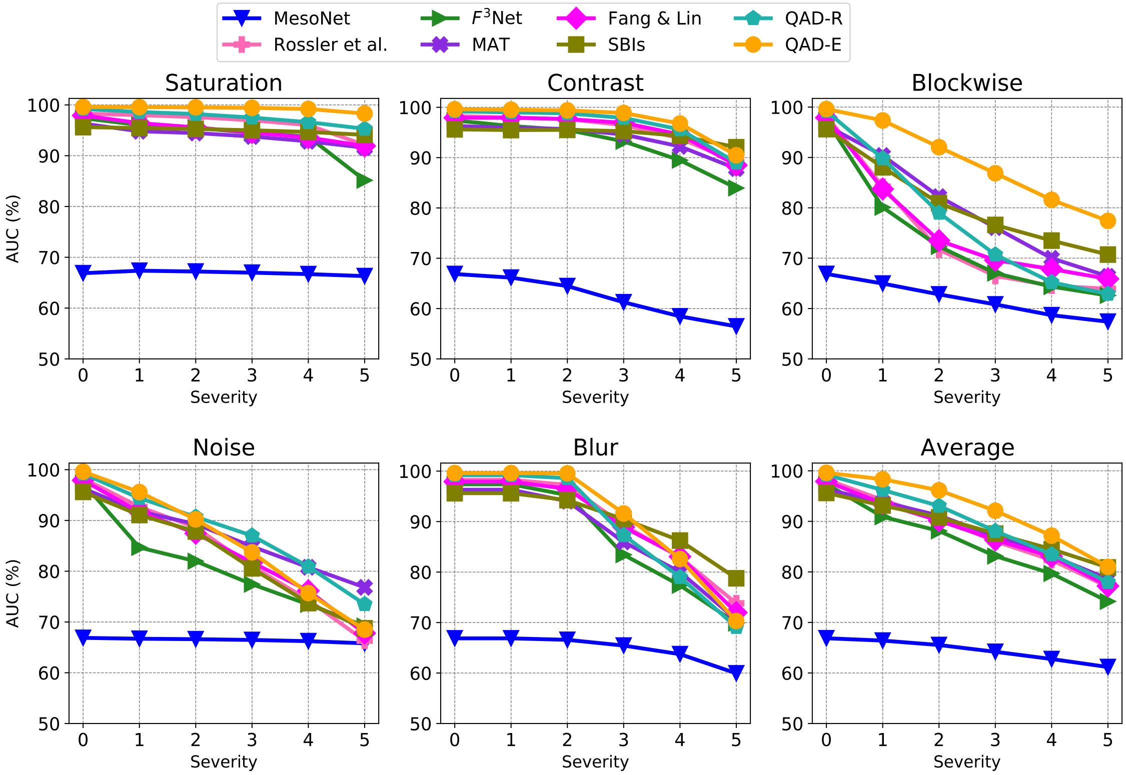

Figure 6. Classification performance (AUC) of deepfake detectors under various corruptions with different severity levels.

- **Expanding attention regions in low-quality images**. Under heavy compression, subtle differences and artifacts for distinguishing deepfakes can be diminished. As we can observe in Fig. 8, RESNET50 activates different regions, resulting in inconsistent prediction among quality modalities. Meanwhile, our QAD by utilizing the AWP assists in enlarging activation regions on the low-quality images, which are indicated by the green circles in Fig. 8. Therefore, it accumulates information from several regions inside the low-quality image to make an overall prediction, making the detector more accurate.

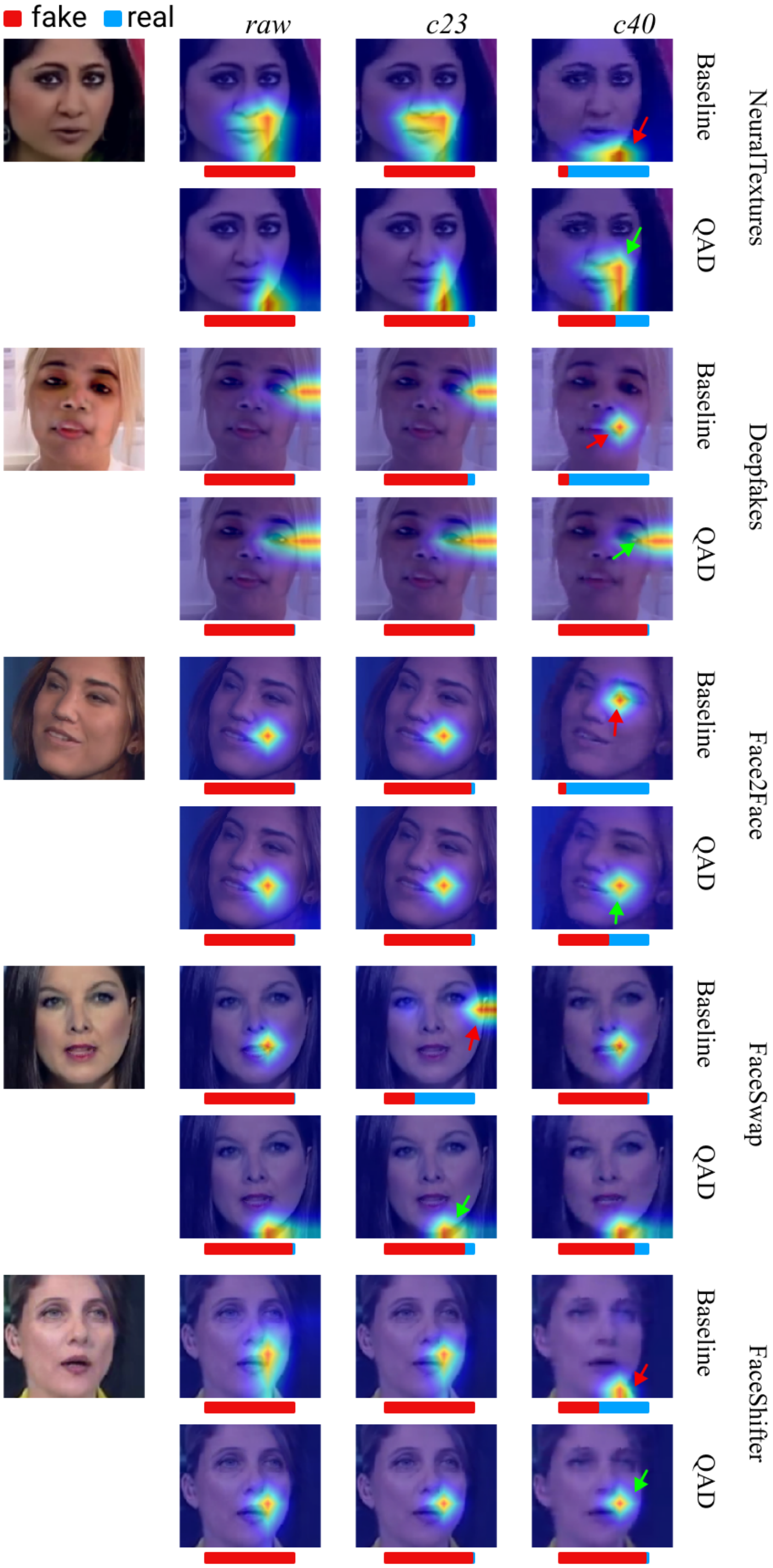

Figure 7. Grad-CAM activation maps of **deepfake images** from NeuralTextures, DeepFakes, Face2Face, FaceSwap and FaceShifter dataset. The red arrows indicate the inconsistent activation regions created by the RESNET50 baseline. The green arrows indicate the re-corrected activation regions created by our QAD framework.

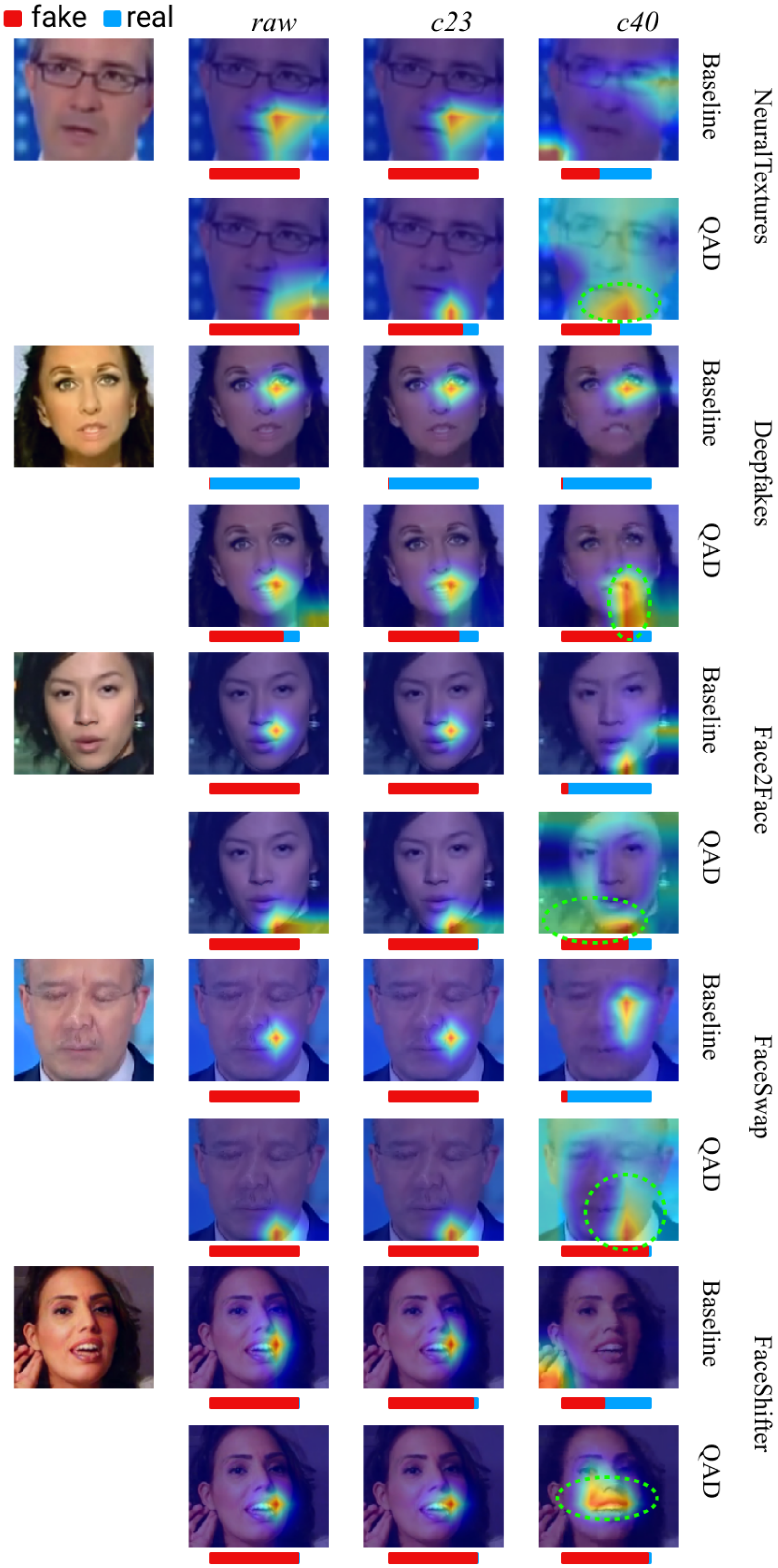

Figure 8. Grad-CAM activation maps of **deepfake images** from NeuralTextures, DeepFakes, Face2Face, FaceSwap and FaceShifter dataset. In contrast with inconsistent or wrong activation regions from the baseline, our QAD can enlarge the activation regions in the low-quality images and reconcile them with other quality modalities.