TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment

TIP 2024

Dongzi Wang

August 24th, 2024

Abstract

Image Quality Assessment (IQA) is a fundamental task in computer vision that has witnessed remarkable progress with deep neural networks. Inspired by the characteristics of the human visual system, existing methods typically use a combination of global and local representations (i.e., multi-scale features) to achieve superior performance. However, most of them adopt simple linear fusion of multi-scale features, and neglect their possibly complex relationship and interaction. In contrast, humans typically first form a global impression to locate important regions and then focus on local details in those regions. We therefore propose a top-down approach that uses high-level semantics to guide the IQA network to focus on semantically important local distortion regions, named as TOPIQ. Our approach to IQA involves the design of a heuristic coarse-to-fine network (CFANet) that leverages multi-scale features and progressively propagates multi-level semantic information to low-level representations in a top-down manner. **A key component of our approach is the proposed cross-scale attention mechanism, which calculates attention maps for lower level features guided by higher level features.** This mechanism emphasizes active semantic regions for low-level distortions, thereby improving performance.

Observations and Motivation

To illustrate our motivation, we conducted a detailed analysis of two seminal **multi-scale approaches**: the MS-SSIM and LPIPS. We used example images from Fig. 1 and the TID2013 dataset for our analysis.

Figure 3 shows the spatial quality maps of MS-SSIM and LPIPS before pooling for example images from Fig. 1. We have the following observations:

• Both MS-SSIM and LPIPS appear to be distracted by the large background region in Image B, leading them to assign higher final scores to Image B. However, humans tend to focus more on the birds region and tend to prefer Image A.

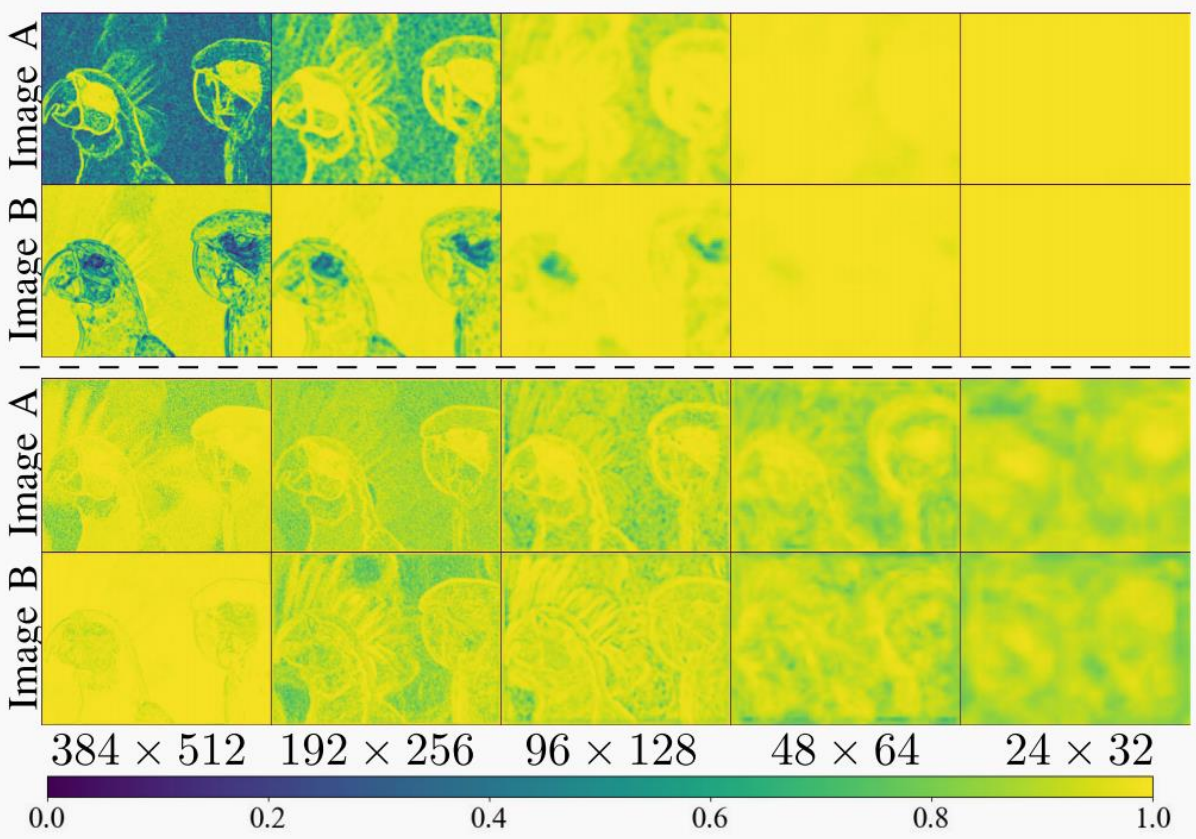


Fig. 3. Multi-scale spatial quality maps ($H \times W$) of MS-SSIM (top two rows) and LPIPS (bottom two rows) with example images (Image A and Image B) from Fig. 1. Please zoom in for best view. *Note: since LPIPS is lower better, we use (1 - LPIPS) here.*

Observations and Motivation

• For these two cases, the high-level differences between Image A and Image B are small. MS-SSIM appears to have difficulties in extracting semantic features, and the pixel-level differences after downsampling are also small. On the other hand, the backbone network of LPIPS is capable of extracting high-level semantics, but it tends to lose distortion differences. Therefore, it can be challenging to determine which image is better based on high-level feature differences alone.

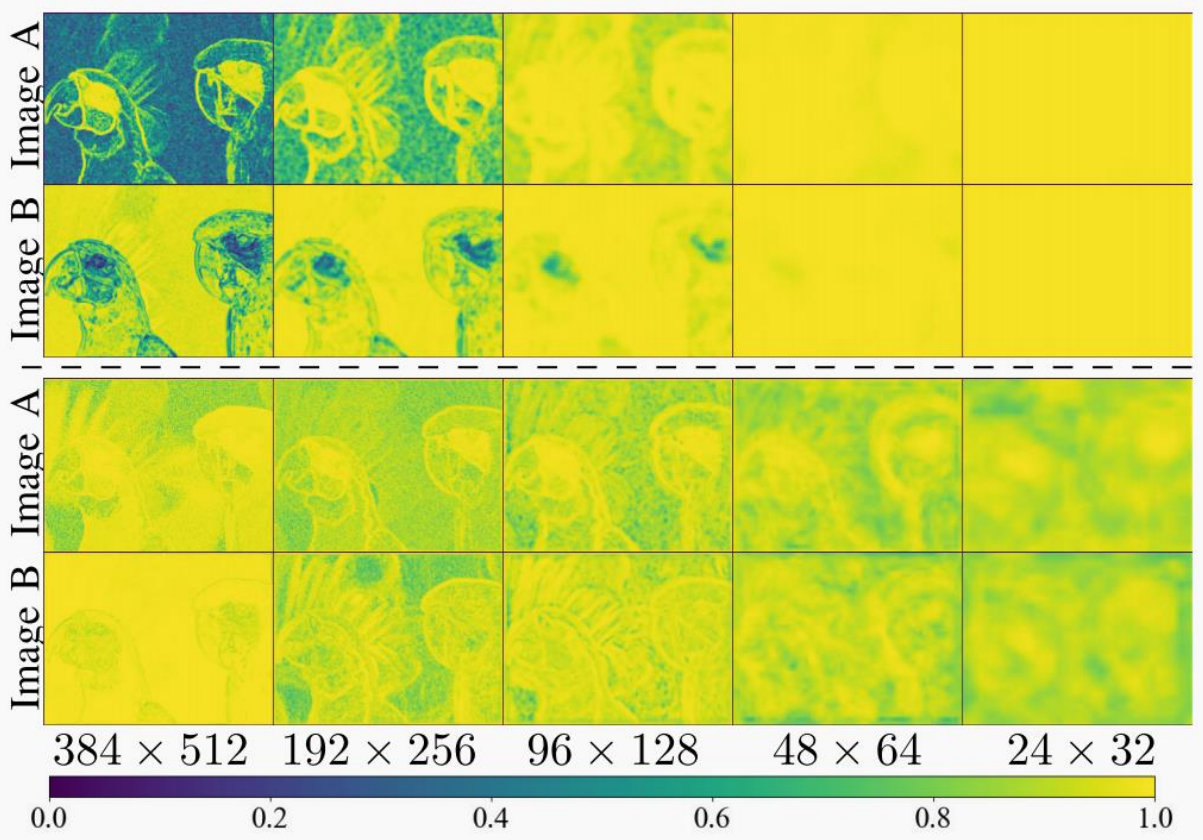


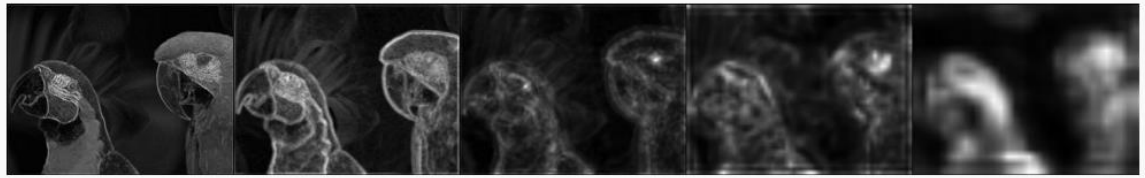
Fig. 3. Multi-scale spatial quality maps ($H \times W$) of MS-SSIM (top two rows) and LPIPS (bottom two rows) with example images (Image A and Image B) from Fig. 1. Please zoom in for best view. *Note: since LPIPS is lower better, we use (1 - LPIPS) here.*

Observations and Motivation

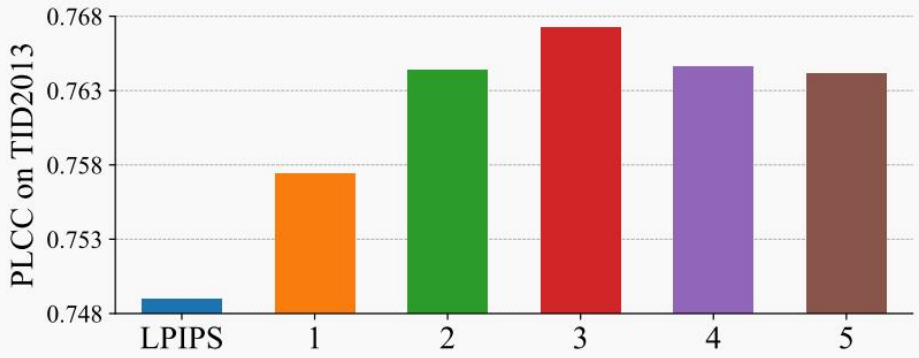
Based on these observations, we hypothesize that neither parallel nor bottom-up approaches can fully utilize multi-scale features. The parallel methods, such as MS-SSIM, have difficulties in extracting semantic representations. Conversely, for bottom-up approaches like LPIPS, although they can extract better semantic representations, they typically regress scores with different scale features independently, and therefore, are unable to focus on semantic regions as humans do.

The LPIPS+ Metric: To verify our hypothesis, we explore a simple extension of LPIPS by replacing the average pooling with weighted average pooling, denoted as **LPIPS+**. We take the feature maps of reference images as rough estimations of semantic weights. As is known, features with higher activation values in neural networks usually correspond to semantic regions, as shown in Fig. 4a for an example. Take reference features from i-th layer as $\mathbf{F}_i^r \in \mathbb{R}^{C_i \times H_i \times W_i}$, and the spatial quality map of m-th layer as $\mathbf{S}_m^r \in \mathbb{R}^{1 \times H_m \times W_m}$, LPIPS+ can be briefly formulated as follow:

$$\text{LPIPS+} = \sum_m \frac{\sum \text{Resize}(\mathbf{F}_i^r) \odot \mathbf{S}_m^r}{\sum \text{Resize}(\mathbf{F}_i^r)}, \tag{1}$$



(a) Example of multi-scale semantic activation maps in LPIPS from low-level to high-level. Please zoom in for best view.



(b) LPIPS+ using different layers as semantic weights.

Fig. 4. Empirical study of the LPIPS+ metric. (a) The feature activation maps can be roughly taken as semantic weight maps; (b) The third layer semantic features bring the most improvement compared with original LPIPS.

Coarse-to-Fine Attention Network

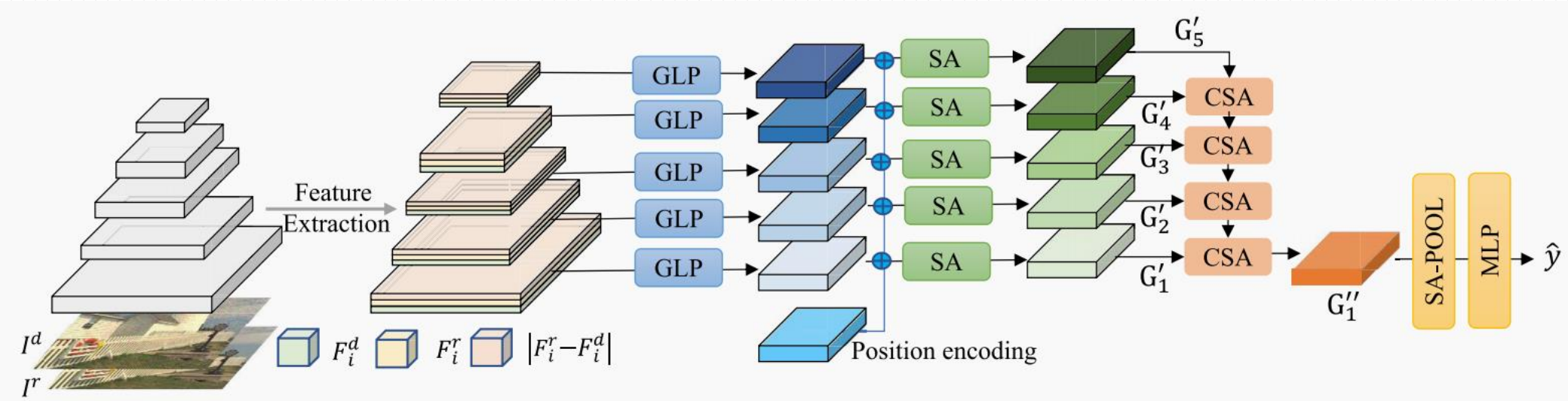


Fig. 5. Architecture overview of the proposed CFANet-FR. We use 5-scale features here same as previous works such as MS-SSIM and LPIPS.

The pipeline of CFANet-FR is presented in Fig. 5. Given distortion-reference image pairs as input, we first extract their multi-scale features using a backbone network. Next, we employ gated local pooling (GLP) to reduce the multi-scale features to the same spatial size, which are then enhanced using self-attention (SA) blocks. Subsequently, we progressively apply cross-scale attention (CSA) blocks from high-level to low-level features. Finally, we pool the semantic-aware distortion features and regress them to the quality score through a multilayer perceptron (MLP).

Gated Local Pooling

1) Gated Local Pooling: Denote input image pairs as $(I^d, I^r) \in \mathbb{R}^{3\times H\times W}$, the backbone features from block i as $(\mathbf{F}_i^d, \mathbf{F}_i^r) \in \mathbb{R}^{C_i\times H_i\times W_i}$, where H_i, W_i are height and width, C_i is the channel dimension, $i \in \{1, 2, \ldots, n\}$ and $n = 5$ for ResNet50. In general, low-level features are twice larger than their adjacent high-level features, and we have $H_i = H/2^i$. Therefore, directly compute correlation between large matrix like $\mathbf{F}_1$ and $\mathbf{F}_2$ is too expensive. For simplicity and efficiency, we reduce $\mathbf{F}_i$ to the same shape as the highest level features $\mathbf{F}_n$. A naïve solution is simple window average pooling. However, this would fuse features inside local window and make the distortion feature less distinguishable. Instead, we propose to select the distortion related features before pooling through a gated convolution [73], which has been proven to be useful in image inpainting. The problem here is how to calculate the gating mask. Notice that for FR task, the difference between $(\mathbf{F}_i^d, \mathbf{F}_i^r)$ is a strong clue for feature selection, we therefore formulate the gated convolution as

$$\mathbf{F}_i^{mask} = \sigma\left(\phi_i(|\mathbf{F}_i^d - \mathbf{F}_i^r|)\right) \cdot (\mathbf{F}_i^d \oplus \mathbf{F}_i^r \oplus |\mathbf{F}_i^d - \mathbf{F}_i^r|), \quad (2)$$

where σ is the sigmoid activation function that constrains the mask value to the range of $[0, 1]$, ϕ_i represents a bottleneck convolution block, and $\oplus$ denotes the concatenation operation. Please refer to Fig. 6 for further details. For efficiency, we use a single-channel mask, *i.e.*, $\phi_i(\cdot) \in \mathbb{R}^{1\times H_i\times W_i}$.

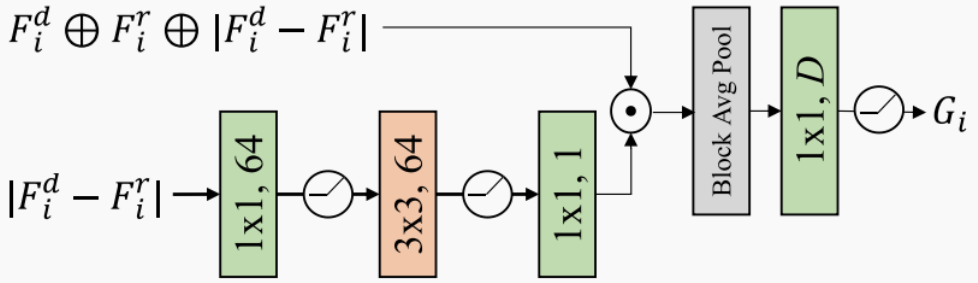


Fig. 6. The GLP block comprises a mask branch and a feature branch. The mask branch is a bottleneck convolution block with an internal channel dimension of 64. For FR datasets, we set the output dimension D to 256, and for NR datasets, we set it to 512. All convolution layers are followed by the GELU activation function.

Attention Modules

a) Self-attention: After GLP, we obtain a set of features from different scales, denoted by $\{\mathbf{G}_1, \ldots, \mathbf{G}_n\} \in \mathbb{R}^{(H_n \times W_n) \times D}$. As the receptive field of low-level features is limited, we first enhance $\mathbf{G}_i$ with a self-attention block as follows:

$$\mathbf{G}'_i = \mathrm{SA}(\mathbf{G}_i) = \mathrm{Attn}(\mathbf{G}_i W_q, \mathbf{G}_i W_k, \mathbf{G}_i W_v) + \mathbf{G}_i, \quad (5)$$

where $\mathbf{G}_i$ is projected onto $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ through simple linear projection. Through the SA block, $\mathbf{G}'_i$ aggregates features from other positions to enhance $\mathbf{G}_i$. In [45], they concatenate the multi-scale features and use several transformer layers to regress the score, without considering the fact that different semantic regions hold different importance to humans. This approach does not allow for interaction between high-level semantic features and low-level distortion features, and thus cannot model such relationships. Our proposed cross-scale attention method addresses this issue in a straightforward manner.

b) Cross-scale attention: Since the query feature $\mathbf{Q}$ in Eq. (4) naturally serves as a guide when computing the output, our cross-attention is designed by simply generating the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ with features from different scales, and we have:

$$\mathbf{Q} = \hat{\mathbf{G}}_{i+1} W_q, \ \mathbf{K} = \mathbf{G}'_i W_k, \ \mathbf{V} = \mathbf{G}'_i W_v, \quad (6)$$

$$\mathbf{A} = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k}), \quad (7)$$

$$\hat{\mathbf{G}}_i = \mathrm{CSA}(\mathbf{G}'_i, \hat{\mathbf{G}}_{i+1}) = \mathbf{A}\mathbf{V} + \hat{\mathbf{G}}_{i+1} \quad (8)$$

where $i \in \{1, \ldots, n-1\}$, and $\hat{\mathbf{G}}_n = \mathbf{G}'_n$. $\mathbf{G} \in \mathbb{R}^{(H_n \times W_n) \times D}$, $W \in \mathbb{R}^{D \times d_k}$, and $\mathbf{A} \in \mathbb{R}^{(H_n \times W_n) \times (H_n \times W_n)}$. D, d_k are the feature dimensions and $H_n \times W_n$ is the spatial size of the feature maps after gated local pooling (GLP). We can note that the features in $\hat{\mathbf{G}}_i$ are calculated using the attention maps $\mathbf{A}$ from $\hat{\mathbf{G}}_{i+1}$ to $\mathbf{G}'_i$. The attention weights in $\mathbf{A}$ represent the similarities between $\hat{\mathbf{G}}_{i+1}$ and $\mathbf{G}'_i$, highlighting features in $\mathbf{G}'_i$ that are more semantically active. Each feature vector in $\hat{\mathbf{G}}_{i+1}$ generates an attention map of size $H_n \times W_n$, which aggregates the "value" features from $\mathbf{V}$ through a weighted sum. Therefore, we can describe the high-level semantic features as clustering centers that can aggregate low-level features with more semantic significance. We provide visualization

Thanks for your attention