Content-Diverse Comparisons improve IQA

William Thong[1,2*], Jose Costa Pereira[3], Sarah Parisot[3]
Aleš Leonardis[3], Steven McDonagh[3]

University of Amsterdam[1] Sony AI[2] Huawei Noah's Ark Lab[3]

william.thong@sony.com, jose.c.pereira@huawei.com, steven.mcdonagh@huawei.com

BMVC 2022

Perceptual Image Quality Assessment (IQA)

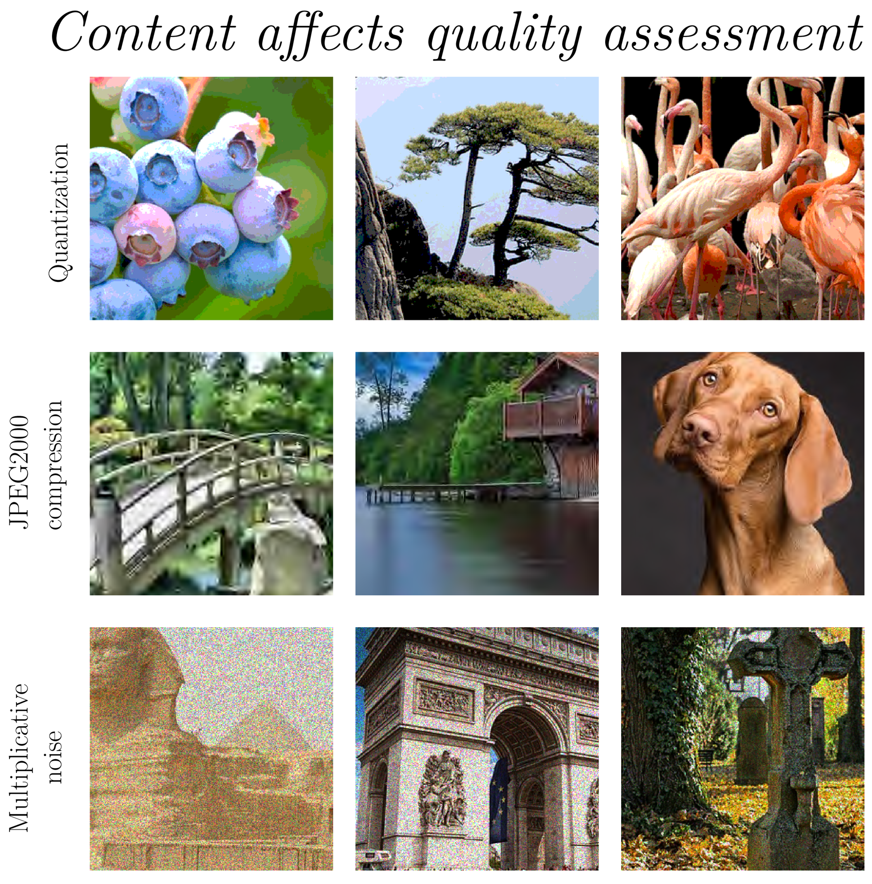
Task

Assign quantitative scores to rank images by their perceptual quality

- Straightforward task for humans; yet effective automation is challenging
- Traditionally done by ranking PSNR or SSIM scores
- Improvements arise by learning a deep network to compare *image pairs* of similar content

Challenge

Content affects quality assessment



- Fixing image-pair content restricts diversity and limits model training exposure, in terms of heterogeneity

Contributions

Content-Diverse IQA training

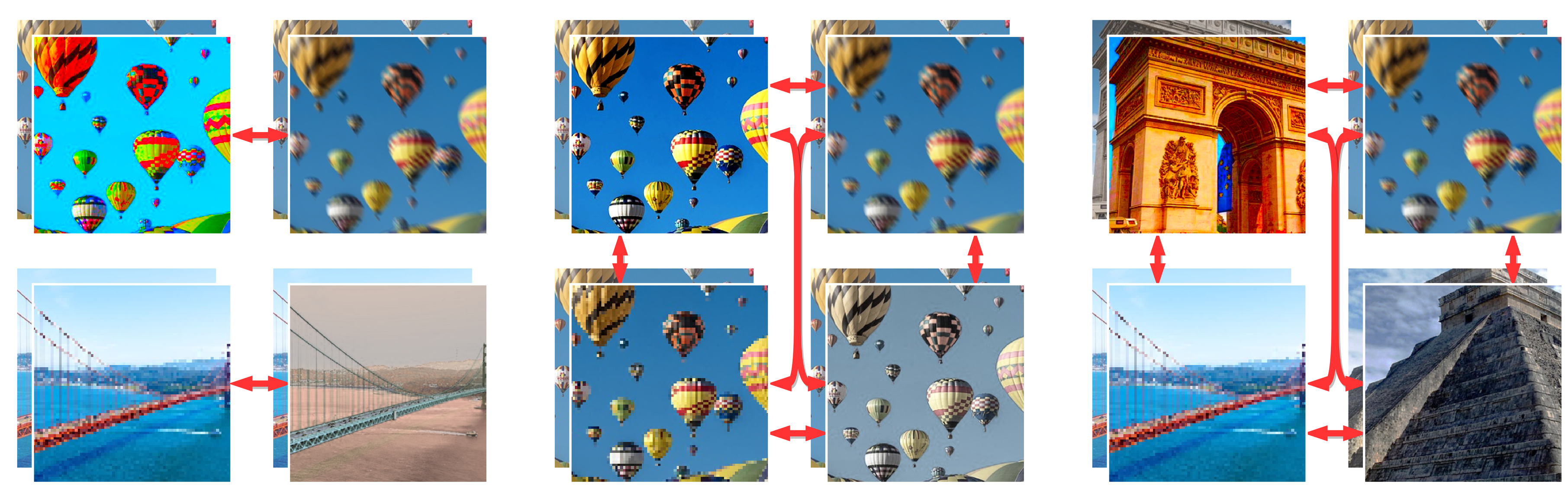
1. Relax pairwise constraints to enable comparisons with differing content
2. Derive three differentiable regularizers for listwise comparisons at the mini-batch level
3. Comprehensive evaluation across eight IQA datasets with state-of-the-art performance

Benefits

- Emulate wide latent factors under consideration during human IQA
- Applicable to any model architecture without structural changes
- Improvements to downstream reconstruction tasks (with IQA as a training objective)

* Work done during an internship at Huawei Noah's Ark Lab

Pairwise Formation



(a) *Fixed* pairs *similar* content (b) *All* pairs *similar* content (c) *All* pairs *differing* content

- Dataset: $\{x^i, x^i_{\text{ref}}, y^i\}_{i=1}^M$ with M distorted images x, scalar quality score labels y
- Learn function f to predict quality $\hat{y} = f(x, x_{\text{ref}})$; *i.e.* "Full-Reference" IQA

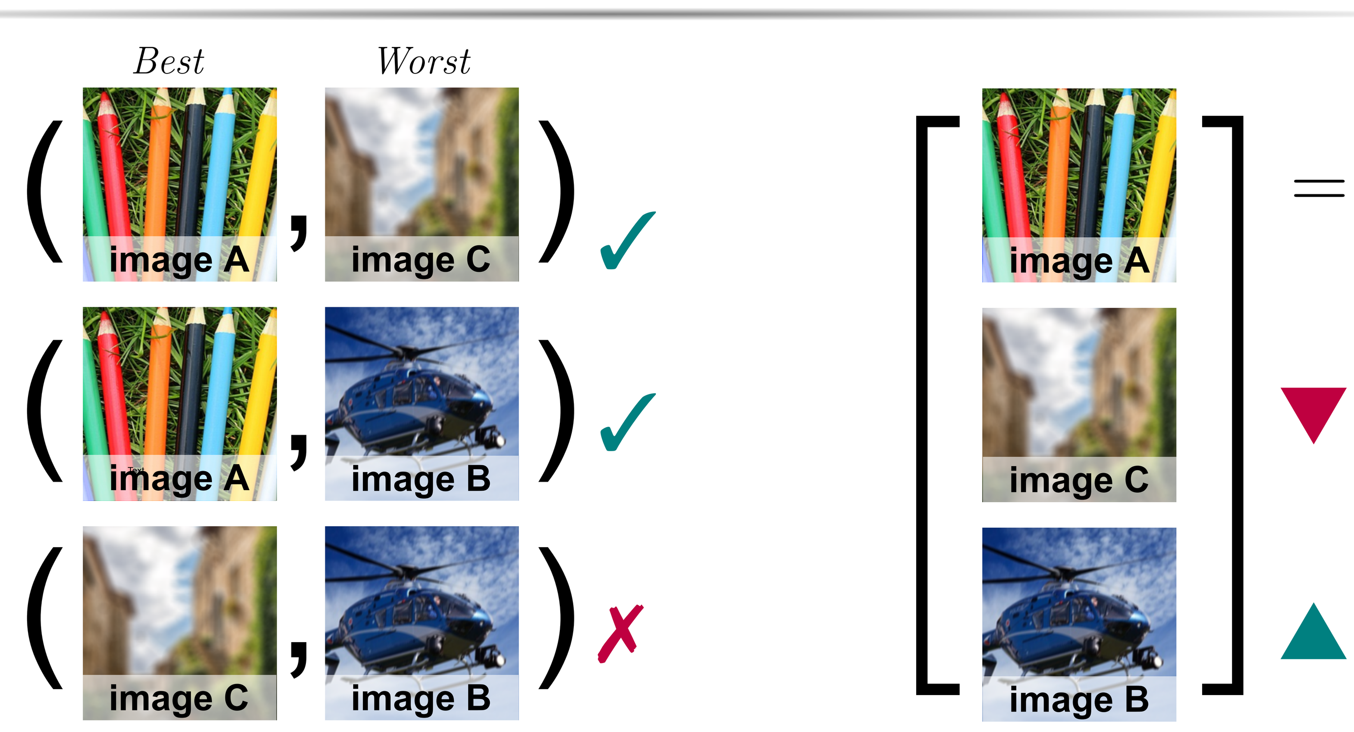
Pairwise Training:

- Learning of f typically relies on *pairwise* training: images x^i and x^j with $i \neq j$
- If label $y^i > y^j$, then we desire: $\hat{y}^i = f(x^i, x^i_{\text{ref}}) > \hat{y}^j = f(x^j, x^j_{\text{ref}})$
- Learn to produce faithful *image rankings* *c.f.* regressing directly to y

Proposal:

- Consider *all* possible image pairs; allow image content to *differ* within a pair
- Pairwise relaxation: **broader definition** of valid pairwise comparisons
- No longer imposes a constraint on the number of comparisons in a mini-batch
- Probabilistic model of $y^i > y^j$ via Bradley-Terry sigmoid and cross-entropy: $\mathcal{L}_c = \mathbb{1}[y^i > y^j] \cdot \log(p(y^i > y^j)) + \mathbb{1}[y^i < y^j] \cdot \log(1 - p(y^i > y^j))$

Listwise comparisons and Correlation Coefficients



(a) Pairwise scores (b) Listwise scores

Observation:

- Pairwise comparisons provide only two distorted images; limit training visibility

Proposal:

- $\hat{Y} = \{\hat{y}^1, \ldots, \hat{y}^L\}$ predicted scores with Y related GT scores for set of L images
- Listwise comparisons via differentiable correlation coefficients
  - Pearson coefficient R_r encourages linearity: $r(Y, \hat{Y}) = \text{cov}(Y, \hat{Y}) / \sigma_Y \sigma_{\hat{Y}}$
  - Spearman R_ρ, linearity of ranks: $\rho(Y, \hat{Y}) = \rho(\text{rank}_Y, \text{rank}_{\hat{Y}})$
  - Kendall R_τ, ordinal ranking: $\tau(Y, \hat{Y}) = \frac{2}{L(L-1)} \sum_{i<j} \text{sgn}(Y^i - Y^j) \, \text{sgn}(\hat{Y}^i - \hat{Y}^j)$
  - rank and sgn are approximated with temperature-based sigmoid and tanh
- Regularizers are derived, final loss function becomes: $\mathcal{L} = \mathcal{L}_c + \lambda(R_r + R_\rho + R_\tau)$

Ablative studies

Regularizer R_r	R_ρ	R_τ	LIVE [45] PLCC	SRCC	KRCC	CSIQ [21] PLCC	SRCC	KRCC	TID2013 [40] PLCC	SRCC	KRCC
			0.963	0.968	0.842	0.950	0.954	0.809	0.908	0.897	0.717
✓			0.962	0.967	0.839	0.952	0.956	0.812	0.906	0.896	0.715
	✓		0.960	0.966	0.835	0.953	0.957	0.815	0.910	0.901	0.723
		✓	0.962	0.968	0.840	0.950	0.955	0.811	0.910	0.900	0.722
✓	✓		0.960	0.966	0.837	0.954	0.959	0.819	0.908	0.899	0.718
	✓	✓	0.961	0.967	0.838	0.941	0.960	0.821	0.909	0.900	0.721
✓		✓	0.960	0.966	0.837	0.954	0.959	0.820	0.912	0.903	0.725
✓	✓	✓	**0.964**	**0.969**	**0.843**	**0.957**	**0.960**	**0.824**	**0.915**	**0.907**	**0.731**

- Proposed terms encourage linear properties and rank preservation

Quantitative results

Method	LIVE [45] PLCC	SRCC	KRCC	CSIQ [21] PLCC	SRCC	KRCC	TID2013 [40] PLCC	SRCC	KRCC
PSNR	0.865	0.873	0.680	0.819	0.810	0.601	0.677	0.687	0.496
SSIM [56]	0.937	0.948	0.796	0.852	0.865	0.680	0.777	0.727	0.545
MS-SSIM [55]	0.940	0.951	0.805	0.889	0.906	0.730	0.830	0.786	0.605
VSI [61]	0.948	0.952	0.806	0.928	0.942	0.786	**0.900**	0.897	**0.718**
MAD [21]	**0.968**	0.967	0.842	**0.950**	**0.947**	0.797	0.827	0.781	0.604
VIF [44]	0.960	0.964	0.828	0.913	0.911	0.743	0.771	0.677	0.518
FSIM [60]	0.961	0.965	0.836	0.919	0.931	0.769	0.877	0.851	0.667
NLPD [20]	0.932	0.937	0.778	0.923	0.932	0.769	0.839	0.800	0.625
GMSD [58]	0.957	0.960	0.827	0.945	0.950	**0.804**	0.855	0.804	0.634
WaDIQaM [6]	0.940	0.947	0.791	0.901	0.909	0.732	0.834	0.831	0.631
PieAPP [41]	0.908	0.919	0.750	0.877	0.892	0.715	0.859	0.876	0.683
LPIPS [62]	0.934	0.932	0.765	0.896	0.876	0.689	0.749	0.670	0.497
DISTS [11]	0.954	0.954	0.811	0.928	0.929	0.767	0.855	0.830	0.639
IQT [10]	–	**0.970**	**0.849**	–	0.943	0.799	–	**0.899**	0.717
Ours	**0.964**	**0.969**	**0.843**	**0.957**	**0.960**	**0.824**	**0.915**	**0.907**	**0.731**

*Please see our paper for corresponding references and additional benchmarks.

Downstream tasks

- Image quality metrics for ×4 super-resolution
- Training objectives for ESRGAN [53]
- We reduce subtle over-sharpening artifacts present in other methods

Takeaways

- Image content matters in image quality assessment
- Formulate through diverse pairwise and listwise comparisons

Links

Contact:
william.thong@sony.com
steven.mcdonagh@huawei.com

Paper: **Code:**