GEOMETRIC TRANSFORMATION INVARIANT IMAGE QUALITY ASSESSMENT USING CONVOLUTIONAL NEURAL NETWORKS

Kede Ma, Zhengfang Duanmu, and Zhou Wang

Department of ECE, University of Waterloo, Waterloo, ON, Canada
Email:{k29ma, zduanmu, zhou.wang}@uwaterloo.ca

ABSTRACT

Most existing full-reference (FR) image quality assessment (IQA) models assume that the reference and distorted images are perfectly aligned, and fail dramatically when the assumption does not hold. In this study, we first show that pre-registration, especially feature-based (as opposed to area-based) registration, is effective at reducing the performance drop of FR-IQA models. However, registration is an expensive process that often slows down the speed of the IQA algorithms by several orders of magnitude. This motivates us to construct an end-to-end convolutional neural network (CNN) for direct image quality prediction, which contains built-in invariance to geometric distortions. Our results show that when the training images are augmented by their geometrically transformed versions, the learned network performs at a high level without image registration, resulting in a fast and effective approach for geometric transformation invariant IQA.

***Index Terms*—** Image quality assessment, image registration, convolutional neural networks, geometric transformations, data augmentation

1. INTRODUCTION

Full reference (FR) image quality assessment (IQA) aims to quantify the perceptual quality of a possibly distorted image using its pristine-quality counterpart as reference [1]. When comparing two images in either pixel [2] or transform domain [3], most FR-IQA models assume they are perfectly aligned, *i.e.*, the geometrical relationship between the two images is an identity mapping. As a result, a tiny geometric transformation (*e.g.*, translation, rotation, and scaling) that may be imperceptible to humans could cause existing models to fail. Limited work has been dedicated to IQA invariant to geometric transformations. By incorporating geometric transformations as a special case of adaptive linear system decompositions of image error signals, an IQA method was developed in [4] that is capable of handling small geometric distortions. A substantially different approach is to extend the structural similarity (SSIM) index [2] into the complex wavelet transform domain. The resulting CW-SSIM [5] index was shown to be insensitive to consistent relative phase distortions and therefore can handle up to 4 degrees of rotations and 7 pixels of translations [5].

In this work, we first probe the sensitivity of FR-IQA models under gentle geometric transformations. Specifically, we equip existing knowledge-driven FR-IQA models with mature image registration techniques [6, 7, 8, 9, 10] to combat the misalignment problem. Our results show that feature-based alignment [9] is more robust than area-based (direct) alignment [6, 7, 8] and results in less performance drop. However, a major drawback of incorporating image registration as a preprocessing step is the significantly increased computational complexity, especially for those methods that involve iterative and multi-scale optimizations.

The drawback of pre-registration-based approaches motivates us to look for novel solutions that can directly predict image quality without registration. As a powerful class of models, convolutional neural networks (CNNs) have reshaped the fields of image processing and computer vision, achieving state-of-the-art results in image classification [11], semantic segmentation [12], and many low-level vision tasks [13, 14, 15]. Recently, CNN-based data-driven IQA models [16, 17] have been shown to surpass the performance of knowledge-driven models, which rely heavily on domain expertise.

In this work, we focus on studying the hierarchical representations of CNNs in their ability to handle geometric transformations. Specifically, we construct a fully convolutional network, which consists of four stages of convolution, subsampling, batch normalization [18], and ReLU nonlinearity [19]. The predicted distortion measure is computed as the mean squared error (MSE) between the last-stage model responses of the original and distorted images. We train our network by maximizing the Pearson correlation between the predicted distortion scores and the human mean opinion scores (MOSs). We show that the network is able to learn invariance to translation, rotation, and scaling on the training data augmented by their geometrically transformed versions, resulting in state-of-the-art performance without image registration.

2. EXPERIMENTAL SETUPS AND PERFORMANCE OF KNOWLEDGE-DRIVEN MODELS

We use the LIVE IQA database [20] as the starting point, which contains 29 original and 779 distorted images, and augment it by considering four types of geometric transformations—translation, rotation, scaling, and their mixtures. To mimic real-world scenarios, an image should first be geometrically transformed (*e.g.*, camera movement) and then distorted (*e.g.*, compressed by JPEG). An equivalent but much simpler implementation that we adopt is to directly apply the transform to the original image. Specifically, 5 translated, 5 rotated, and 5 scaled images are generated by randomly shifting a LIVE reference image in the horizontal and/or vertical directions with two offsets sampled uniformly between -15 and 15 pixels, randomly rotating the image between -5° and 5°, and randomly scaling the image by a factor between 0.85 and 1.15. 15 images of mixed transforms are generated by applying translation, rotation, and scaling simultaneously. All parameters are carefully chosen so that the transformations do not hurt the perceptual quality. Therefore, it is reasonable to assume that the difference of MOS (DMOS) of each distorted image remains unchanged when compared with its geometrically transformed reference versions. In total, the augmented LIVE database contains $25,048$ images.

Table 1. SRCC results of knowledge-driven FR-IQA models on the augmented LIVE database

Images	SSIM	MS-SSIM	CW-SSIM	VIF
Not aligned	0.154	0.133	0.141	0.230
Perfectly aligned	0.932	0.954	0.806	0.968
LK aligned	0.585	0.572	0.515	0.622
ECC aligned	0.726	0.724	0.638	0.764
DIC aligned	0.889	0.904	0.780	0.926
SURF aligned	0.921	0.942	0.812	0.956

Table 2. Average execution time in seconds on $10,000$ images in the augmented LIVE database

Algorithm	SSIM [2]	SURF+RANSAC [9, 21]
Environment	MATLAB	MATLAB+MEX
Time (s)	0.015 ± 0.002	5.663 ± 12.567

We select four image registration algorithms to help FR-IQA models combat the misalignment between the reference and distorted images. These are area-based alignment 1) the Lucas-Kanade algorithm (LK) [6], 2) ECC [7], 3) DIC [8], and feature-based alignment 4) SURF+RANSAC [9, 21]. The LK algorithm [6] is a pioneering work in image alignment, which minimizes the MSE between the warped image and the template. As a variant of the LK algorithm, ECC [7] obtains the optimum transformation by maximizing the enhanced correlation coefficient between the warped image and the template. DIC [8] expands upon the inverse compositional scheme to jointly estimate a group of geometric and photometric transformations, making it robust to intensity variations. In SURF+RANSAC [9, 21], the feature correspondences between images are provided by SURF feature matching [9] and the transformation that best explains these correspondences is obtained by RANSAC [21]. All algorithms are implemented in the image alignment toolbox [10] and tested with the default settings.

We test four knowledge-driven FR-IQA models without and with image registration on the augmented LIVE database. These include SSIM [2], MS-SSIM [22], CW-SSIM [5], and VIF [3]. The Spearman's ranking correlation coefficient (SRCC) [23] results are listed in Table 1, from which we have several useful observations. First, without registration, all competing models fail dramatically as expected. Second, compared with the three area-based alignment algorithms, the feature-based SURF+RANSAC [9, 21] achieves the least performance drop. Specifically, under mild and moderate distortion levels, SURF still performs at a reasonably high level, finding sufficient keypoint matches for robust transform estimation. By contrast, area-based alignment begins to fail, especially when the underlying transformation is complex (*e.g.*, when translation, rotation, and scaling are mixed). Under severe distortions, no algorithm is able to successfully align two images because of destructive structure loss. However, this does not appear to be a big problem because FR-IQA models are likely to give a low quality score to a severely distorted image regardless of registration accuracy. Third, since CW-SSIM [5] is originally designed to compare pattern similarity (*e.g.*, two binary edge maps), it is less competitive on photographic images. However, the capability to tolerate small geometric transformations (resulting from errors in image registration) allows CW-SSIM to reduce more performance gaps than the other models.

Although adding image alignment as a preprocessing step empowers existing knowledge-driven FR-IQA models to handle geometric transformations, it also brings substantial computational complexity. Specifically, state-of-the-art image registration algorithms require solving nonlinear and nonconvex optimization problems in an iterative and multi-scale fashion, and the number of iterations often increases with the complexity of the underlying transformation and with the level of distortion. We compare the execution time between SSIM [2] and SURF+RANSAC [9, 21] on $10,000$ images in the augmented LIVE database on a computer with 3.4GHz CPU and 16G RAM. From Table 2, we see that SURF+RANSAC runs 379 times slower than SSIM with a much larger standard deviation. In summary, existing image alignment methods are effective at resolving the misalignment problem in FR-IQA, but may be impractical for real-world applications due to substantially increased computational burden.

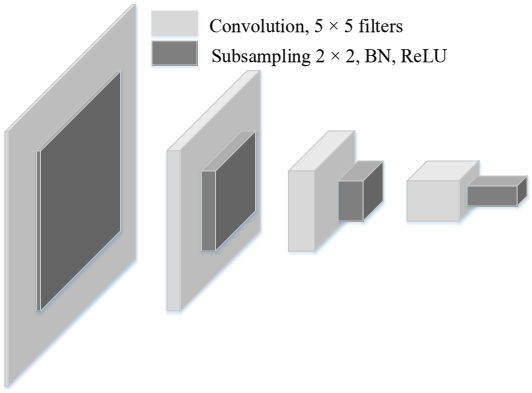

Fig. 1. Architecture of the proposed convolutional network for FR-IQA. BN stands for batch normalization.

3. CNN FOR GEOMETRIC TRANSFORMATION INVARIANT IQA

The results presented in Section 2 and the success of CNNs inspire us to develop a CNN-based solution for geometric transformation invariant IQA.

3.1. Network Architecture and Training

The mini-batch training data are denoted by $\{(\mathbf{X}_r^{(l)}, \mathbf{X}_d^{(l)}, d^{(l)})\}_{l=1}^{L}$, where $\mathbf{X}_r^{(l)}$ and $\mathbf{X}_d^{(l)}$ are the l-th reference and distorted images, respectively, $d^{(l)}$ is the DMOS, and L is the mini-batch size. The network architecture is shown in Fig. 1, where the reference and distorted images are mapped into the same perceptual space for comparison. It consists of four stages of convolution, subsampling, batch normalization, and ReLU nonlinearity, whose parameters are collectively denoted by $\mathbf{W}$. The size of convolution filters is fixed to 5×5 for all stages, while the number of filters at the first stage is set to 6 and is increased by a factor of 2 for each subsequent stage. The convolution responses are subsampled by a factor of 2 along both horizontal and vertical directions, which can be efficiently implemented with stride convolution for computational efficiency. Before applying ReLU nonlinearity, we employ batch normalization to accelerate training, where responses are jointly normalized across the mini-batch and over all spatial locations [18]. Unlike standard CNN architectures, we avoid using fully connected layers to make the net-

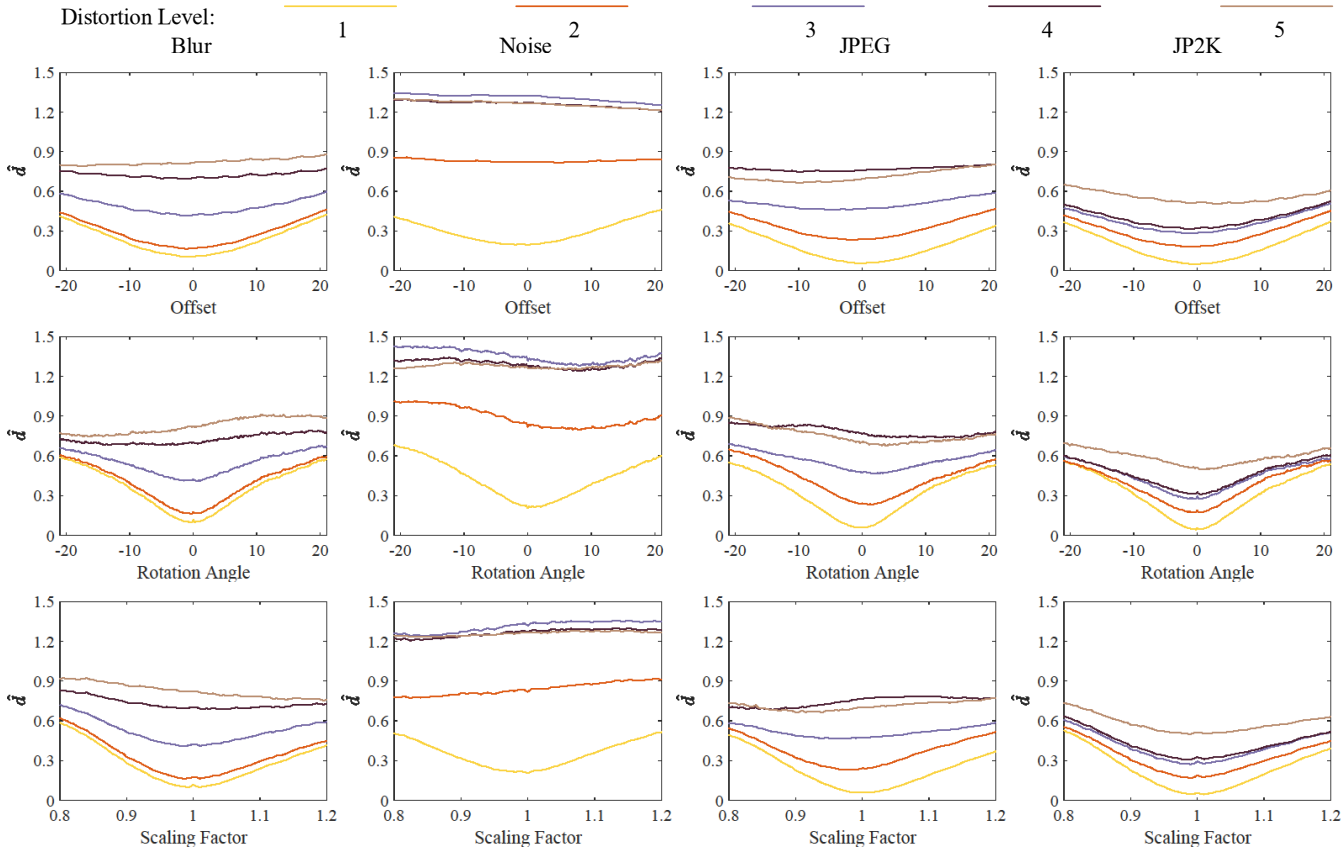

Fig. 2. Sensitivity of the proposed CNN-based model to Gaussian blur (Column 1), white Gaussian noise (Column 2), JPEG compression (Column 3), and JPEG2000 compression (Column 4) under translation (Row 1), rotation (Row 2), and scaling (Row 3) for the "Chemist" image from [24]. $\hat{d}$ is the predicted distortion score in Eq. (1).

work fully convolutional. As a result, our network accepts inputs of arbitrary size and produces an overall distortion score.

We compute the distortion measure as the MSE between the last-stage model responses of the original image $f(\mathbf{X}_r^{(l)})$ and the distorted image $f(\mathbf{X}_d^{(l)})$ [25]

$$\hat{d}^{(l)}(\mathbf{W}) = \frac{16}{3MN} \sum_{i=1}^{\frac{M}{16}} \sum_{j=1}^{\frac{N}{16}} \sum_{k=1}^{48} \left(f(\mathbf{X}_r^{(l)})_{ijk} - f(\mathbf{X}_d^{(l)})_{ijk} \right)^2, \tag{1}$$

where M and N denote the height and the width of input images. The factor of 16 is introduced due to subsampling in four stages.

During training, the optimal parameters $\mathbf{W}^\star$ are obtained by maximizing the Pearson correlation between the predicted distortion scores and the DMOSs

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\operatorname{argmax}} \ \operatorname{corr}(\hat{\mathbf{d}}(\mathbf{W}), \mathbf{d}), \tag{2}$$

where both $\hat{\mathbf{d}} = [\hat{d}^{(1)}, \hat{d}^{(2)}, \cdots, \hat{d}^{(L)}]^T$ and $\mathbf{d} = [d^{(1)}, d^{(2)}, \cdots, d^{(L)}]^T$ are length-L vectors in the current mini-batch. For end- to-end blind IQA [17, 26], the ℓ_p-norm induced metric

$$\ell_p(\hat{\mathbf{d}}(\mathbf{W}), \mathbf{d}) = \|\hat{\mathbf{d}}(\mathbf{W}) - \mathbf{d}\|_p = \sum_{l=1}^{L} |\hat{d}^{(l)} - d^{(l)}|^p \tag{3}$$

Table 3. SRCC results of CNN-based FR-IQA models on the augmented LIVE test set

Images	CNN trained w/o data augmentation	CNN trained with data augmentation
Not aligned	0.140	0.902
Perfectly aligned	0.967	0.939
LK aligned	0.577	0.917
ECC aligned	0.769	0.929
DIC aligned	0.924	0.934
SURF aligned	0.957	0.933

is often used as the empirical loss, where p is set to 1 or 2. For FR-IQA, however, Eq. (3) is less preferable than a correlation loss for optimization because a network with a random initialization tends to perform at a reasonable level in terms of rank correlation. For example, our network with He's initialization [27] achieves an SRCC of 0.876 on the original LIVE database. Directly maximizing SRCC is difficult due to its non- differentiability. Therefore, we adopt Pearson correlation in Eq. (2). This optimization framework also reminds us of listwise learning-to-rank approaches [28], which have been exploited in blind IQA [24].

During testing, we estimate the population mean and variance of batch normalization with exponential moving average, and perform

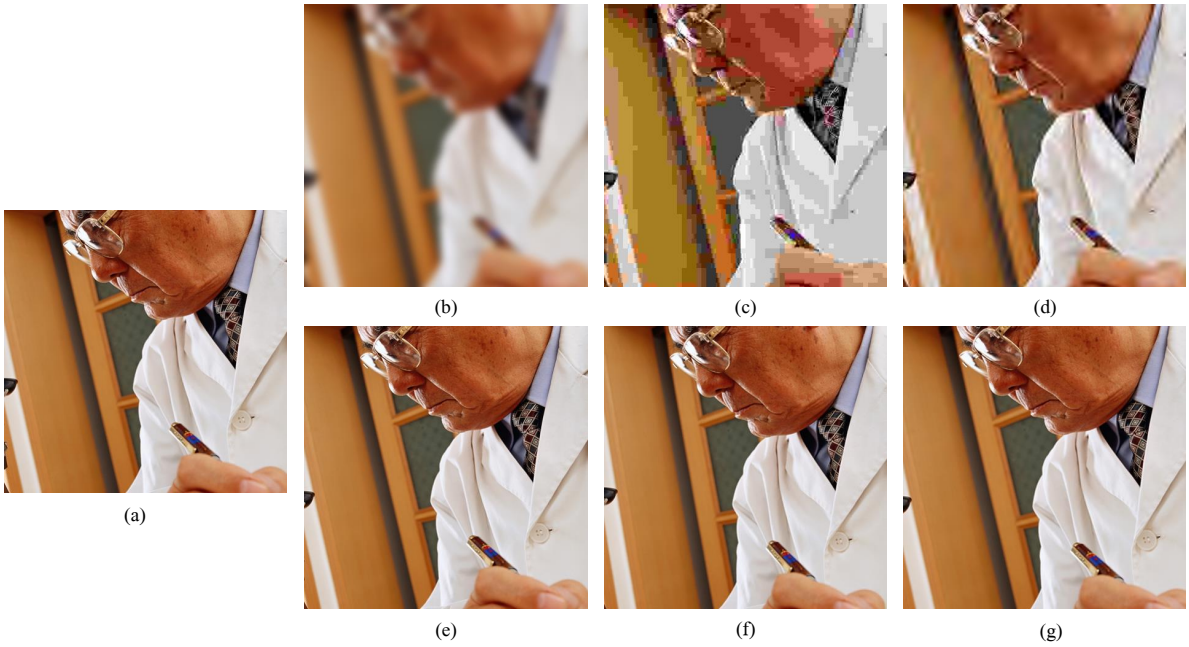

Fig. 3. The proposed CNN-based model is robust to geometric transformations as compared to other distortions of strong visual impairment. This is in stark contrast to traditional FR-IQA methods such as SSIM [2]. (a) Original image. (b) Gaussian blur. SSIM = 0.726, $\hat{d}$ = 0.423. (c) JPEG compression. SSIM = 0.740, $\hat{d}$ = 0.760. (d) JPEG2000 compression. SSIM = 0.710, $\hat{d}$ = 0.515. (e) Translation (by 5 pixels). SSIM = 0.612, $\hat{d}$ = 0.037. (f) Rotation (by 3 degrees). SSIM = 0.585, $\hat{d}$ = 0.061. (g) Scaling (by a factor of 1.05). SSIM = 0.599, $\hat{d}$ = 0.069.

a standard forward propagation to output the distortion score of $\mathbf{X}_d$ using $\mathbf{X}_r$ as reference.

We first train a CNN without data augmentation. Specifically, we randomly select 476 and 166 images from the original LIVE database [20] for training and validation, respectively. We test the model on 5,146 images augmented from the remaining 166 images as described in Section 2. No content overlap occurs among training, validation, and test sets. We then train another CNN with data augmentation, where the network is exposed to geometric transformations during training and validation, trying to learn robust feature representations against them. The two models are trained using the Adam optimization algorithm [29] with a mini-batch size of 64. The learning rate α is initialized to 10^{-2} and is subsequently lowered by a factor of 10 when the loss plateaus, until $\alpha = 10^{-4}$. Other parameters in Adam are set by default. The learning stops when the maximum epoch number 500 is reached and the weights that achieve the highest correlation in the validation set are used for testing.

3.2. Experimental Results

Table 3 lists the SRCC results of the two CNNs on the augmented LIVE test set. Without data augmentation, the data-driven network fails in a similar way as knowledge-driven models. When trained with the augmented data, the plain network learns invariance to translation, rotation, scaling, and mixed transform, without adding advanced modules that are dedicated to spatial manipulation of data such as spatial transformer [30]. As a result, we achieve comparable performance to the perfectly aligned case without using image registration techniques.

To take a close look at the sensitivity of the CNN-based model to geometric transformations, we test our network with a wider range of translation offsets (in pixels), rotation angles (in degrees), and scaling factors on the "Chemist" image from [24]. The results are drawn in Fig. 2, from which we can see that under mild distortions, our network is robust to translation (up to 10 pixels), rotation (up to 5 degrees), and scaling (up to a factor of 0.9 and 1.1). A visual illustration is also shown in Fig. 3. Under severe distortions, the predicted distortion score is approximately constant, which is expected because most perceptually meaningful structures would have been damaged and alignment is of little importance.

The network runs at 9 ms and 36 ms per image on an Nvidia GTX Titan X GPU and on a 3.4 GHz CPU, which are 630 and 157 times, respectively, faster than SSIM [2] with SURF+RANSAC [9, 21].

4. CONCLUSION

In this study, we focus on geometric transformation invariant IQA. Our study suggests that while image pre-registration is effective at improving the quality prediction performance of traditional knowledge-driven FR-IQA models, its wide usage in practice may be severely impeded by the high computational cost. Therefore, we construct an end-to-end IQA model using a CNN-based approach for direct image quality prediction without registration. We show that such a model, when trained with images augmented by their geometrically transformed versions, leads to a simple yet efficient solution to the geometric transformation invariant IQA problem.

5. REFERENCES

[1] Zhou Wang and Alan C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[2] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Hamid R. Sheikh and Alan C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[4] Zhou Wang and Eero P. Simoncelli, "An adaptive linear system framework for image distortion analysis," in *IEEE International Conference on Image Processing*, 2005, pp. 1160–1163.

[5] Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan C. Bovik, and Mia K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009.

[6] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *The 7th International Joint Conference on Artificial Intelligence*, 1981, pp. 121–130.

[7] Georgios D. Evangelidis and Emmanouil Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, Oct. 2008.

[8] Adrien Bartoli, "Groupwise geometric and photometric direct image registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2098–2108, Dec. 2008.

[9] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[10] G. Evangelidis, "IAT: A Matlab toolbox for image alignment," `http://www.iatool.net`, 2013.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[13] Viren Jain and Sebastian Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776.

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[15] Kede Ma, Huan Fu, Tongliang Liu, Zhou Wang, and Dacheng Tao, "Local blur mapping: Exploiting high-level semantics by deep neural networks," *CoRR*, vol. abs/1612.01227, 2016.

[16] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[17] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[18] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[19] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *IEEE International Conference on Machine Learning*, 2010, pp. 807–814.

[20] Hamid R. Sheikh, Zhou Wang, Alan C. Bovik, and Lawrence Cormack, "Image and video quality assessment research at LIVE," 2006, [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[21] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[22] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh IEEE Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[23] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000, [Online]. Available: http://www.vqeg.org.

[24] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[25] Alexander Berardino, Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "Eigen-distortions of hierarchical representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 3533–3542.

[26] Jongyoo Kim and Sanghoon Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[28] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: From pairwise approach to listwise approach," in *International Conference on Machine Learning*, 2007, pp. 129–136.

[29] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[30] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.