Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 255–263
DOI: 10.15439/2025F8791
ISSN 2300-5963 ACSIS, Vol. 43

Keypoint-based metric for evaluating image super-resolution quality

Jakub Sadel*†, Tomasz Tarasiewicz*†, Pawel Kowaleczko†‡, Maciej Ziaja*†,
Daniel Kostrzewa*†, Pawel Benecki*, Michal Kawulok*†
0009-0006-9227-7470, 0000-0002-7706-1317, 0000-0003-4689-6343, 0009-0009-8285-5881,
0000-0003-2781-3709, 0000-0003-4674-5393, 0000-0002-3669-5110
*Silesian University of Technology, Gliwice, Poland
Email: {jakub.sadel | tomasz.tarasiewicz | maciej.ziaja | daniel.kostrzewa | pawel.benecki | michal.kawulok}@polsl.pl
†KP Labs, Gliwice, Poland
‡Warsaw University of Technology, Warsaw, Poland
Email: pawel.kowaleczko@pw.edu.pl

***Abstract*—Recent advances in single- and multi-image super-resolution have revealed the limitations of classical image similarity metrics (like peak signal-to-noise ratio), as they often fail to align with human perception when evaluating the visual quality of super-resolved outputs. In this paper, we explore how to exploit keypoint-based metrics to evaluate super-resolution image quality. Specifically, we explore two correlated metrics: (*i*) a *multi-scale index proposal* measure capturing salience of keypoints, and (*ii*) a *repeatability* metric quantifying how consistently the corresponding keypoints are identified in super-resolved and ground-truth images. Experiments on several simulated and real-world datasets show that the repeatability correlates with subjective judgments, and multi-scale index proposal can be helpful for difficult datasets when other metrics are insufficient.**

I. Introduction

LOW-RESOLUTION (LR) images pose a significant challenge in diverse imaging fields, from satellite surveillance to medical diagnostics. Physical constraints or adverse capture conditions often lead to low-quality data with insufficient detail. To address this, image *super-resolution* (SR) techniques reconstruct *high-resolution* (HR) images from LR inputs, effectively enhancing visual clarity and texture.

In recent years, deep learning-based SR methods have achieved impressive results, producing images with sharper edges and more natural details than classic interpolation schemes. However, these advances introduce a pressing need for robust *image quality assessment* (IQA) metrics tailored to super-resolved images. Traditional *full-reference* (FR) metrics such as *peak signal-to-noise ratio* (PSNR) and *structural similarity index measure* (SSIM) compare the SR output against a ground-truth reference at the pixel or local structural level. While useful for measuring basic fidelity, these methods often do not capture the subjective quality that humans perceive—especially in SR outputs that introduce plausible, but non-identical, high-frequency details.

This work was supported by the National Science Centre, Poland, under Research Grant 2022/47/B/ST6/03009. JS was supported by the SUT Statutory Research funds under Grant 02/080/BKM25/0058. MK was supported by the SUT funds through the Rector's Research and Development Grant 02/080/RGJ25/0053.

Ultimately, human viewers prioritize perceptual realism, focusing on sharp textures, coherent structures, and absence of distracting artifacts. An SR image that looks over-smoothed or contains noticeable distortions in critical regions may be subjectively worse, even if it has a high value of PSNR or SSIM metric. Conversely, an image that retains the essential scene layout while adding visually consistent fine details can be perceived as high quality despite larger pixel-wise differences from the ground truth. This observation drives the development of alternative IQA methods that better reflect human visual preferences and the complex trade-offs between fidelity and realism in SR.

The primary objective of this study is to explore keypoint-based metrics as a means of evaluating the quality of super-resolved images. Unlike traditional IQA metrics that focus on pixel-wise fidelity or perceptual similarity, our approach investigates the geometric and structural consistency of SR output through the lens of local feature detection. Keypoint-based metrics were also exploited for that purpose in [1]; however, that approach relies on conventional techniques that are non-differentiable, which limits its applicability for training deep neural networks. In this work, on the other hand, we aim at adapting methods based on deep learning, in order to allow for applying them to task-driven training in the future. To achieve this, we leverage the capabilities of Key.Net [2], a neural network designed for keypoint detection tasks that combines hand-crafted and learned convolutional filters within a shallow multiscale architecture. In particular, we examine the utility of the repeatability, a widely adopted metric in local feature evaluation [3], which measures the consistency of keypoint detection under various transformations, and the *multi-scale index proposal* (MSIP), introduced as a loss function for training Key.Net [2]. We extend this idea to the SR domain, under the assumption that a high-quality super-resolved image should exhibit similar keypoint number and distributions to the ground-truth reference image.

The paper is organized as follows. Section II reviews related work on SR methods and quality assessment metrics. Section III describes the datasets used for the evaluation.

Section IV details the proposed keypoint-based approach and metric configurations. Section V presents the experimental setup and results. Section VI summarizes the paper and provides directions for future research.

II. Related Work

A. SR Techniques: Major Categories and State of the Art

SR is highly relevant in domains where capturing higher-resolution imagery directly is impractical due to physical or cost-related constraints [4]. Applications span security surveillance, medical imaging, and remote sensing [5]. For instance, in satellite imaging, SR algorithms may be the only way to obtain finer details when sensor resolution is limited or hiqh-quality image acquisition is too expensive [4]. Over the past two decades, a vast array of SR methods have been proposed, generally categorized into *single-image SR* (SISR) approaches (relying on a single LR input) and *multi-image SR* (MISR) approaches (fusing multiple observations of the same scene).

Early SISR techniques were based on interpolation and example-based methods, long before the deep learning era. Interpolation-based methods such as bilinear or bicubic interpolation provide simple baselines by smoothing and enlarging pixel grids, but tend to produce blurry results due to their limited modeling of image structure. More sophisticated classical methods leveraged natural image priors. For example, patch-based dictionary learning and sparse coding became influential around 2010. Yang et al. [6] pioneered learning pairs of small patches from LR–HR images, known as dictionary atoms, so that a sparse linear combination of LR atoms can reconstruct high-frequency details via the corresponding HR atoms. This sparse representation approach achieved sharper results than interpolation by reusing high-frequency patterns from training data.

Convolutional neural network (CNN) models brought a breakthrough in SISR by learning the LR-to-HR mapping from large datasets. The seminal SRCNN [7] of Dong et al. was the first CNN for SR, with a three-layer network directly learning to improve the quality of interpolated LR inputs. Despite its simplicity, SRCNN significantly outperformed sparse-coding methods, demonstrating the power of data-driven feature extraction. Dong et al. later proposed FSRCNN [8], an accelerated variant that operates in the LR space for efficiency. Kim et al. introduced a *very deep SR* network (VDSR) [9] with ca. 20 layers and residual learning to ease training. Subsequently, the EDSR model removed batch normalization to allow for even deeper residual networks, achieving state-of-the-art performance in the 2017 NTIRE challenge [10].

By 2017, *generative adversarial networks* (GANs) emerged to tackle a core limitation of purely *mean squared error* (MSE) optimized CNNs: overly smooth outputs lacking high-frequency texture. Ledig et al. proposed SRGAN [11], introducing adversarial training to favor perceptually sharp results. SRGAN's generator (a deep ResNet architecture, also known as SRResNet, due to its use of residual learning blocks) is trained with a discriminator to produce outputs that are difficult to distinguish from real HR images. This GAN-based approach produced much crisper and more detailed images, at the cost of lower PSNR. Wang et al. later improved this with ESRGAN [12], which used residual-in-residual dense blocks and a more perceptually aligned loss, resulting in more natural textures. Beyond perceptual losses, task-driven objectives have emerged to tailor super-resolution to downstream applications. For instance, Zyrek and Kawulok [13] proposed a method that optimizes SISR for improved text recognition from scanned documents by combining image similarity with text detection objectives.

While not directly focused on SR, Bairi et al. [14] addressed the related problem of image reconstruction from compressive sensing data. Their dual-path framework integrates Vision Transformers and perceptual optimization, combining global contextual modeling with perceptual loss to enhance reconstruction quality beyond traditional fidelity-driven methods.

Most recently, diffusion probabilistic models have been applied to SISR, marking a new frontier in fidelity and diversity. Saharia et al. introduced SR3 [15], which adapts denoising diffusion models for SR. Starting from pure noise, SR3 iteratively denoises the input while being conditioned on the LR image, generating highly photo-realistic outputs.

Whereas SISR relies on learned priors to hallucinate missing details, MISR methods exploit actual complementary information from multiple observations of the same scene. By combining a sequence of LR images (often with subpixel shifts between them), MISR can surpass the limits of single-image fidelity. Traditional MISR techniques were rooted in multi-frame reconstruction theory and often cast as Bayesian or regularized optimization problems. A classical formulation assumes each LR image is a warped, blurred, downsampled, and noisy version of an unknown HR image. The SR task is then to invert this imaging model. Early solutions like the work [16] of Schultz and Stevenson employed a *maximum a posteriori* (MAP) estimator with a smoothness prior, solved via iterative back-projection or gradient descent. Farsiu et al. later proposed a fast and robust SR (FRSR) [17] MISR algorithm that combined an L1-norm fidelity term (robust to outliers) with a simplified regularization, and crucially computed registration error in the HR space to avoid repeated interpolations. Such approaches (e.g., Bayesian inference, maximum likelihood or MAP estimation) dominated MISR for years, incorporating robust alignment and prior terms to handle noise and misregistrations. However, they often required careful tuning and were computationally intensive.

With the rise of deep learning, MISR has seen new data-driven approaches addressing the key challenges of alignment and fusion. The first application of deep learning in MISR was EvoNet, presented in [4], where the method combines the advantages of multi-image fusion with learning the LR-to-HR mapping using deep networks, achieving superior reconstruction accuracy compared to state-of-the-art SR methods. A milestone in deep MISR was the PROBA-V Super Resolution Challenge [18], which spurred the development of several methods, including the HighRes-net by Deudon et al. [19] and the winning approach, DeepSUM [20]. HighRes-

net introduced an end-to-end trainable framework to jointly learn co-registration of multiple LR frames and their fusion into a single HR output. It uses a recursive fusion strategy with an implicit reference frame, eliminating the need for explicit motion compensation between inputs. DeepSUM, on the other hand, leverages deep neural networks to efficiently integrate spatial and temporal information from unregistered multi-temporal images, achieving superior reconstruction quality. Another notable model is the *residual attention MISR* (RAMS) network [21] by Salvetti et al. RAMS integrates attention mechanisms into a deep CNN to adaptively weight the contribution of each pixel from each image during fusion.

Current state-of-the-art MISR research often borrows architectures from SISR and video SR, enhanced to handle multiple unordered inputs. Transformer-based MISR models [22][23] have recently emerged, leveraging self-attention to capture long-range dependencies across frames. Another approach is to represent a stack of input images as a graph which is super-resolved with a graph neural network [24] or graph attention network [25].

B. Image Quality Assessment

Despite the progress in SR algorithms, evaluating the quality of super-resolved images remains challenging. Generally, IQA methods are classified into FR, *reduced-reference* (RR), and *no-reference* (NR) approaches, depending on how much ground truth information is available [26]. FR-IQA metrics (e.g., PSNR, SSIM) compare the output directly against a reference image, RR-IQA methods use partial information about the reference (such as features or the input LR image), and NR-IQA (also called blind IQA) relies only on the output itself. In conventional restoration tasks, FR metrics like PSNR and SSIM have been widely used since a ground-truth HR image is usually available for synthetic test cases. However, these classical metrics focus on signal fidelity and often correlate poorly with human visual perception when applied to SR results. A super-resolved image that looks sharp and realistic to a human observer, might score worse in PSNR/SSIM than a blurry image that stays pixel-close to the reference [27]. This is especially true for GAN-based SR methods which introduce high-frequency content: they achieve higher subjective quality but lower PSNR (the perception-distortion dilemma [28]). As a result, learning-based IQA metrics have gained traction. These include deep feature-based distances like *learned perceptual image patch similarity* (LPIPS) [29] (which compares images in a CNN feature space) and *deep image structure and texture similarity* (DISTS) [30], as well as neural networks trained to predict human opinion scores.

For SR evaluation, both fidelity and perceptual quality are important, and researchers have devised specialized metrics to handle the conflicts between them [31]. Several recent works propose SR-specific IQA methods that account for characteristics of SR outputs.

FR-IQA compares the super-resolved image directly with its HR ground truth. Metrics like PSNR and SSIM are most commonly used, offering objective and repeatable evaluation. PSNR measures pixel-wise fidelity based on mean squared error, while SSIM captures local luminance, contrast, and structure similarities. Despite their popularity, these metrics often fail to reflect perceptual quality, especially for SR results generated by GAN-based models, which prioritize sharpness and realism over strict pixel accuracy [26][27]. Due to structural differences from the ground truth, more advanced FR methods have been proposed that analyze images in gradient, wavelet, or perceptual feature domains. For example, phase congruency and gradient magnitude have been employed to better model human sensitivity to edge structures, as illustrated by the FSIM metric [32], which leverages both features within a full-reference evaluation framework. While FR metrics are ideal for synthetic SR benchmarks with available HR references, they are unsuitable in real-world settings where such references are missing.

RR-IQA methods evaluate quality using partial information from the reference, often in the form of extracted features or the LR input image itself. This makes RR approaches more practical than FR in real-world SR scenarios, where full ground truth is unavailable. Early RR methods focused on edge preservation, texture statistics, or wavelet-based similarities between LR and SR images [27]. More recent approaches, like PFIQA [5], use deep neural networks to combine perceptual and fidelity-aware branches, leveraging both the SR image and the LR input alongside auxiliary information (e.g., scale factor). These models better reflect human perception by assessing whether reconstructed details are consistent with the available input and visually plausible. RR-IQA is especially valuable for detecting artifacts that contradict known image structures or suggest overfitting (e.g., excessive sharpening). While RR approaches strike a balance between fidelity and perception, they remain limited in evaluating purely hallucinated textures, as they cannot fully determine whether added details are semantically correct.

The NR-IQA approach is the most flexible and challenging setting, as it operates solely on the super-resolved image without any access to ground truth or input. NR metrics are indispensable in real-world applications, where only the output is available. Early NR-IQA approaches relied on natural scene statistics, assuming that distortions in SR images would cause statistical deviations. Methods like *natural image quality evaluator* (NIQE) [33] are based on statistical features derived from natural images and estimate deviations associated with sharpness, noise, and unnatural textures. These were later combined into composite scores like the *perceptual index* (PI) [27], which became widely used in perceptual SR evaluation. Recent NR-IQA methods leverage deep learning to directly predict subjective quality from image content. For example, KLTSRQA [34] uses Karhunen-Loève transform-based features, while others employ CNNs or transformers trained with human opinion scores [27]. Knowledge distillation techniques have also been introduced to train NR-IQA models using pseudo-labels obtained from strong full-reference models [27]. Although modern NR-IQA metrics outperform traditional ones in terms of perceptual alignment, they still face challenges

related to generalization and distinguishing realistic textures from artifacts without reference anchors.

III. Datasets

In the field of SR, selecting appropriate datasets is crucial for evaluating the performance of quality assessment metrics. This research leverages three distinct datasets: CVIU-17 [35], SISAR [36], and MuS2 [37]. The first two are widely employed [27], [31], [26] in the field of IQA, whereas the third one presents an intriguing complement owing to its distinctive characteristics. Below, we provide detailed descriptions of these datasets, highlighting their composition, purposes, and contributions to advancing SR research.

A. CVIU-17

The CVIU-17 dataset is a publicly available resource tailored for evaluating IQA metrics in the context of SISR. It comprises 180 LR images and 1,620 HR images, generated using nine distinct SR algorithms across six integer scaling factors (see Fig. 1). LR images were generated by applying a combination of downsampling and blurring to the original images from the BSD200 dataset [38].

Fig. 1. An example from the CVIU dataset showing super-resolved images obtained using various SR methods. The visual comparison highlights the differences in reconstruction quality across methods.

B. SISAR

The SISAR dataset (SR image quality database with semi-automatic ratings) stands as the largest-of-its-kind resource for IQA. It contains 12,600 labeled HR images derived from 100 natural LR images, each with a resolution of 1024×768 (see Fig. 2). These HR images were generated using ten types of SR algorithms. SISAR's scale and diversity make it an invaluable asset for developing robust and generalizable IQA metrics, paving the way for more accurate quality predictions in SR applications.

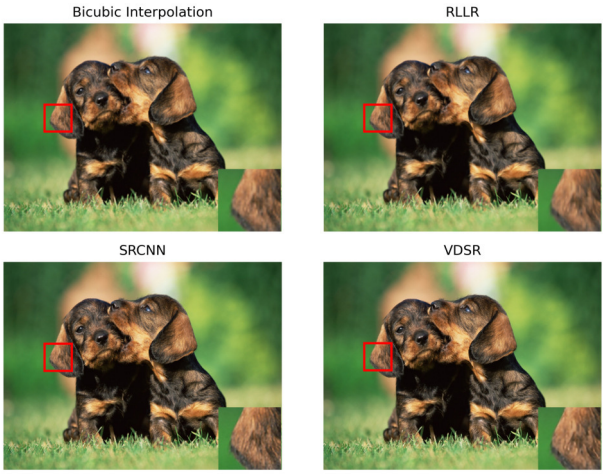

Fig. 2. An example scene from the SISAR dataset showing images super-resolved using various methods. The figure includes Bicubic Interpolation, RLLR, SRCNN, and VDSR, illustrating the visual differences in image quality achieved by each method.

C. MUS 2

The MuS2 dataset is a specialized benchmark for MISR of Sentinel-2 satellite [37]. It addresses the challenge of improving the spatial resolution of Sentinel-2 imagery, which has a ground sampling distance of 10 meters, by using HR WorldView-2 imagery as a reference. The dataset includes 91 scenes (see Fig. 3), each containing 14 or 15 Sentinel-2 LR images across 12 spectral bands, paired with corresponding WorldView-2 HR images. All LR and HR image pairs in MuS2 are pre-aligned, and no geometric misalignments are present.

MuS2 is particularly significant for remote sensing applications, where HR imagery is critical for tasks such as environmental monitoring, urban planning, and disaster management. Unlike many SR datasets that rely on simulated data, MuS2 provides real-world data, capturing the complexities of satellite imaging, such as atmospheric variations and temporal changes. This realism makes it an ideal testbed for developing and validating MISR algorithms that fuse information from multiple images to achieve superior reconstruction accuracy.

IV. Proposed Approach

The main goal of this study is to investigate the use of keypoint-based metrics for assessing the quality of super-resolved images. These metrics, derived from Key.Net, include MSIP and the repeatability.

Key.Net [2] is a hybrid keypoint detection model that combines handcrafted and learned CNN filters within a multi-scale architecture. The handcrafted filters are inspired by traditional methods like Harris [39] detectors, capturing first and second-order image derivatives. These filters act as soft anchors, reducing the number of learnable parameters and improving stability. Key.Net also incorporates a multi-scale pyramid representation, which enhances robustness to scale

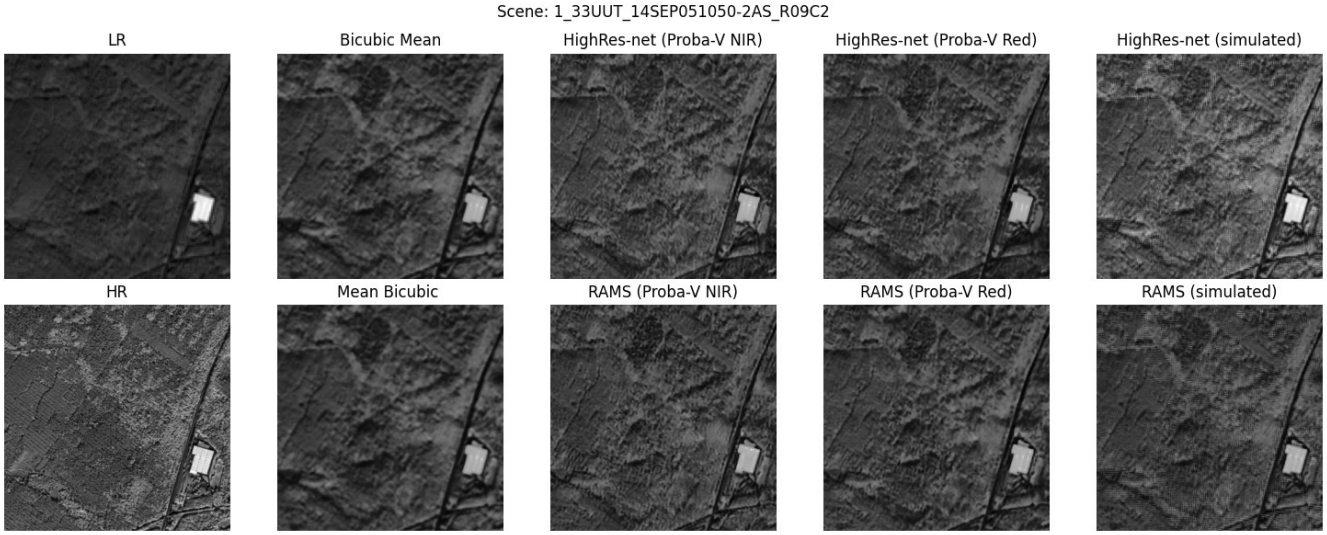

Fig. 3. An example scene from the MuS2 dataset, showing both LR and HR images. The figure includes bicubic interpolation of a mean LR, averaged image of bicubic interpolation outcomes from each LR image, as well as super-resolved images obtained from HighRes-net and RAMS networks trained using real-world PROBA-V images and from a simulated dataset.

changes by processing the input image at multiple scales. The learned filters are then applied to these multi-scale features to localize, score, and rank keypoints. The model uses a novel MSIP layer as loss function, which optimizes keypoint detection by leveraging both local and global information.

The MSIP module extends the concept of the *index proposal* (IP) layer by aggregating multiple IP layers applied at different spatial scales. Each IP layer operates on a shared response map using a local window of size $N_s \times N_s$, where N_s denotes the scale-specific window size. Within each window, a spatial softmax operator is used to convert raw activation values into a probability distribution, highlighting the most prominent local responses. This results in differentiable keypoint localization, computed as the expected coordinate (weighted average) of positions within the window.

By stacking several IP layers with increasing window sizes, MSIP is able to extract keypoints that are not only locally salient but also stable across spatial scales. This multi-scale strategy improves robustness and suppresses unstable detections, as only keypoints that persist across larger contexts are retained. This design encourages the detector to favor geometrically consistent keypoints that are resilient to scale variations, improving repeatability under transformation.

The repeatability quantifies how consistently keypoints are detected between two images of the same scene (e.g., the reference HR and its super-resolved counterpart). A keypoint is considered repeatable if it has a corresponding match within a defined spatial tolerance, often under geometric transformations such as scaling or homography. Formally, it is defined as:

$$\text{repeatability} = \frac{|\mathcal{K}_a \cap \mathcal{K}_b|}{\min(|\mathcal{K}_a|, |\mathcal{K}_b|)}, \tag{1}$$

where:

- $\mathcal{K}_a$ and $\mathcal{K}_b$ are the sets of keypoints detected in images I_a and I_b, respectively.
- $|\mathcal{K}_a \cap \mathcal{K}_b|$ is the number of matching keypoints between the two images.
- $\min(|\mathcal{K}_a|, |\mathcal{K}_b|)$ is the minimum number of keypoints detected in either image.

In this implementation, keypoints are defined by both their spatial coordinates and the scale at which they were detected. To determine whether two keypoints correspond, the *intersection-over-union* (IoU) is computed between the circular regions centered at each keypoint, scaled according to their associated detection scale. A match is established, if the IoU exceeds a fixed threshold. The repeatability can be measured at a single scale, reflecting keypoint consistency under fixed resolution, or across multiple scales, which accounts for robustness to scale variations. Multi-scale repeatability provides a more comprehensive assessment of detector stability under varying image resolutions.

The process of calculating keypoint-based metrics is presented in Fig. 4. To compute MSIP and Repeatability, the SR and HR images are converted to grayscale. Then, each of them is passed through the Key.Net network. As the output of Key.Net, a response map is obtained, and further calculations are performed depending on the metric:

- MSIP: Both response maps (SR and HR) are processed through the MSIP layer, which extracts sets of stable keypoints at different scales.
- Repeatability From the response maps, Non-Maximum Suppression (NMS) is applied to obtain sets of keypoints. IoU between corresponding keypoints from SR and HR images is calculated, resulting in the repeatability metric.

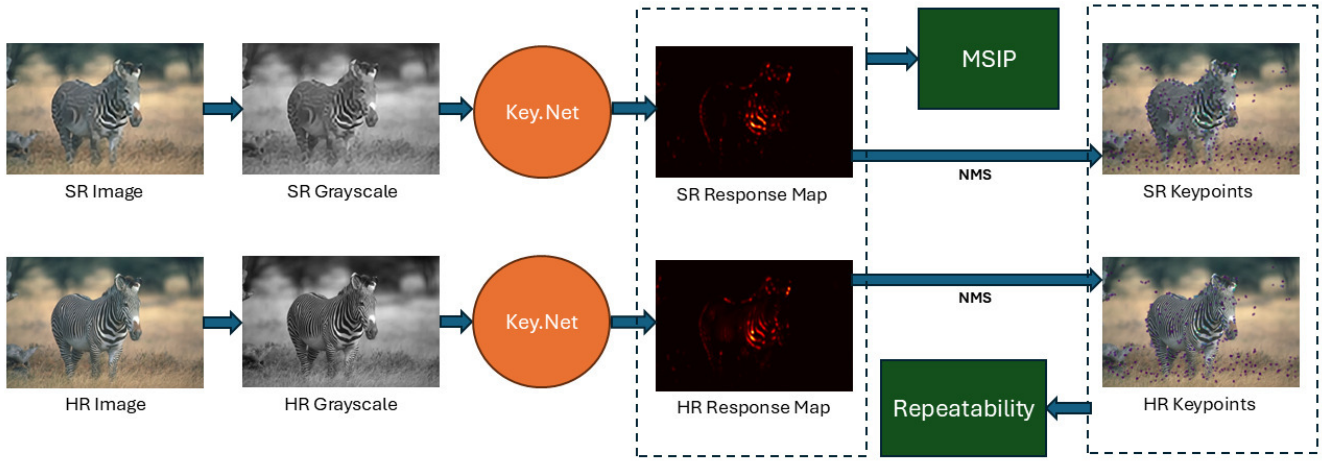

Fig. 4. A pipeline for keypoint-based metrics calculation in IQA for SR. Super-resolved and HR images are converted to grayscale, processed through Key.Net to obtain response maps, and further analyzed using MSIP and repeatability metrics.

We hypothesize that incorporating keypoint repeatability into IQA offers a promising alternative to perceptual or distortion-based measures. To verify this, we compute both conventional metrics—PSNR, SSIM, and LPIPS—as well as keypoint-based metrics, including single-scale repeatability (REP_S), multi-scale repeatability (REP_M), and the MSIP for all three datasets. Furthermore, for the CVIU17 and SISAR datasets, where *mean opinion scores* (MOS) are available, we measure correlation between metric predictions and human perceptual judgments.

V. Experiments

A. Experimental Setup

We carried out experiments on three standard IQA datasets for SR: CVIU-17, SISAR, and MuS2. Both CVIU-17 and SISAR provide HR images and MOS, enabling correlation analysis of objective metrics with human perception. In contrast, MuS2 lacks MOS labels and focuses on Sentinel-2 satellite images for MISR.

Since our approach does not require additional training, we directly used the super-resolved images and the corresponding HR images from each dataset. We applied the Key.Net model (PyTorch implementation) with the original pretrained weights; only minor modifications were introduced to handle grayscale conversion (for CVIU-17 and SISAR) and kernel sizes in the MSIP function. For the purposes of keypoint-based analysis, we assumed an identity homography transformation, implying no geometric discrepancies between the HR images and corresponding SR outcomes.

In our experiments, we evaluated several configurations of the MSIP function to understand how varying kernel sizes and weighting strategies affect the keypoint-based assessment of super-resolved images. We decided to use this division to systematically analyze the impact of different scale levels on the quality of keypoint detection.

- MSIP_default: The default Key.Net configuration, employing four scales with decreasing weights.
- MSIP_5: Single-scale variant using only the smallest kernel size for window.
- MSIP_17: Single-scale variant using only the largest kernel size for window.

Table I summarizes the kernel sizes and weights for each MSIP variant used in both scenarios.

TABLE I
MSIP CONFIGURATIONS USED FOR CVIU-17, SISAR, AND MUS2.

MSIP Variant	Kernel Sizes	Weights
MSIP_default	5×5, 9×9, 13×13, 17×17	1.0, 0.5, 0.25, 0.12
MSIP_5	5×5	1.0
MSIP_17	17×17	1.0

We computed the following metrics for each super-resolved and HR image pair: PSNR, SSIM, LPIPS, and the proposed set of keypoint-based metrics (MSIP, REP_S, REP_M). Where MOS annotations existed (CVIU-17, SISAR), we calculated the *Pearson linear correlation coefficient* (PLCC), *Spearman rank-order correlation coefficient* (SRCC), and *root mean squared error* (RMSE) between each metric's prediction and the MOS. To facilitate fair comparison across metrics, all correlation scores were normalized.

B. Results and Analysis on SISR IQA Benchmark Datasets

Table II shows that Dong11 yield the highest MOS (0.5491), surpassing simpler methods like bicubic interpolation (0.4603). Although these top MOS approaches do not always achieve the highest PSNR or SSIM, they align more closely with the keypoint-based REP_M scores. This finding supports the well-known perception-distortion tradeoff, wherein an image may visually please observers but deviate from the ground truth at the pixel level.

TABLE II
EVALUATION OF SR METHODS ON THE CVIU-17 DATASET.

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17	MOS
Bicubic	24.5020	0.6653	0.4435	87.7187	84.2950	0.1134	0.0420	0.1235	0.4603
BP	24.5766	0.6751	0.4143	87.8826	84.5593	0.1136	0.0412	0.1243	0.4867
Shan08	22.8392	0.5890	0.4627	83.9553	79.1901	0.1926	0.0534	0.2822	0.2834
Glasner09	24.3997	0.6786	0.3393	87.3181	83.6684	0.1188	0.0409	0.1587	0.4467
Yang10	24.5002	0.6752	0.3663	87.0764	83.5761	0.1173	0.0428	0.1295	0.5138
Dong11	24.1746	0.6661	0.3838	**88.4883**	**85.2192**	0.1308	0.0449	0.1637	**0.5491**
Yang13	24.6486	**0.6822**	0.3466	87.4525	84.1223	0.1110	0.0393	0.1392	0.5426
Timofte13	24.6546	0.6767	**0.3379**	86.9872	83.5923	**0.1090**	**0.0385**	0.1241	0.5248
SRCNN	**24.6925**	0.6802	0.3403	88.1164	84.6920	0.1141	0.0407	**0.1177**	0.5444
PLCC	0.5835	0.6317	0.7529	0.5899	0.6769	0.4816	0.4865	0.3364	-
SRCC	0.5716	0.6274	0.7401	0.6280	0.7164	0.5318	0.4929	0.3953	-
RMSE	4.9681	4.8174	4.8293	4.6891	4.7130	4.6094	4.7082	4.5634	-

TABLE III
EVALUATION OF SR METHODS ON THE SISAR DATASET

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17	MOS
BICUBIC	21.8163	0.6551	0.5001	56.3998	48.9003	0.0813	0.0141	0.1543	0.4494
RLLR	18.2244	0.5803	0.5227	46.9003	39.6196	0.1051	0.0163	0.2193	0.4409
SRCNN	18.1481	0.5748	0.5474	51.0858	42.6164	0.0960	0.0158	0.1893	0.4727
VDSR	**22.7156**	**0.6738**	**0.4664**	**64.3389**	**56.1876**	**0.0639**	**0.0130**	**0.1105**	**0.4896**
PLCC	0.6174	0.4830	0.6185	0.6957	0.7133	0.5679	0.4386	0.4956	-
SRCC	0.6129	0.4882	0.6059	0.6861	0.7021	0.5931	0.4263	0.5771	-
RMSE	0.2296	0.2611	0.2147	0.2098	0.2131	0.3241	0.2616	0.4238	-

Comparing the MSIP configurations in Table II, we observe that single-scale variants (MSIP_5, MSIP_17) can diverge significantly for certain SR methods, suggesting that analyzing only a single kernel size may overlook important local structures. Correlation part of this table, reinforces this observation: the correlation of MSIP_5 and MSIP_17 with MOS is generally lower than that of MSIP_default, indicating that decreasing weights over multiple kernels are more effective than a single-scale approach.

A similar trend appears in Table III, where VDSR achieves the highest MOS (0.49) but not the highest PSNR. The local structure measures REP_S and REP_M exhibit a relatively high correlation with MOS, implying that preserving essential edges and textures is key to subjective quality. As in the CVIU-17 dataset, the single-scale MSIP_5 and MSIP_17 lag behind MSIP_default in terms of correlation metrics, though MSIP_default still does not reach the correlation levels of REP_M. This suggests that while MSIP has potential for capturing perceptual differences in super-resolved images, it may require additional tuning or refined weighting strategies to better reflect subjective quality.

Both tables also presents a direct comparison of PLCC, SRCC, and RMSE against MOS for both datasets. As expected, PSNR and SSIM show modest correlations, consistent with their known limitations in modeling perceptual sharpness or high-frequency details. LPIPS reports a higher correlation, which is unsurprising, as it is designed to capture perceptual similarities. The keypoint-based metrics REP_S and REP_M outperform classic metrics in most cases, highlighting that matching local structure and geometry is highly relevant to subjective quality. In comparison, the MSIP_default variants of MSIP reveal a promising correlation but trail behind the best REP_M scores, indicating that further refinement of multi-scale configurations and weighting choices could improve MSIP-based evaluations.

C. Results and Analysis on the Real-World MISR Dataset

Below, we summarize the results of various SR approaches on the MuS2 dataset, which includes Sentinel-2 satellite imagery. Table IV covers both baseline methods (Bicubic) and SR models trained on real-world PROBA-V (NIR or RED bands) or simulated Sentinel-2 data. Specifically:

- Bicubic Mean: Single bicubic upsampling of an averaged LR frame.
- Mean Bicubic: Bicubic upsampling of each LR frame followed by averaging.
- HighRes-net (PROBA-V NIR) / HighRes-net (PROBA-V RED): HighRes-net variants trained on PROBA-V data (NIR or RED bands).
- HighRes-net (simulated): A HighRes-net variant trained on simulated Sentinel-2 data.
- RAMS (PROBA-V NIR) / RAMS (PROBA-V RED): RAMS variants trained on PROBA-V data (NIR or RED bands).
- RAMS (simulated): A RAMS variant trained on synthetic Sentinel-2 data.

A closer look at the metrics reveals several key issues:

1) Bicubic Mean and Mean Bicubic achieve the highest PSNR and relatively high SSIM values, yet their LPIPS and MSIP scores are very poor. Based on the article [37], they produce visually the smoothest (most blurred) outputs, confirming that high PSNR and SSIM alone may not indicate perceptually sharp images. MOS study was

TABLE IV
EVALUATION RESULTS ON THE MUS2 DATASET. IN THE BRACKETS, WE INDICATE THE DATASET USED FOR TRAINING (REAL-WORLD PROBA-V OR SIMULATED IMAGES).

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17
Bicubic Mean	24.1514	0.6012	0.5667	42.3759	24.0449	0.5470	0.0815	1.1352
Mean Bicubic	**24.1515**	0.6012	0.5667	42.3708	24.0607	0.5457	0.0815	1.1403
HighRes-net (PROBA-V NIR)	23.9385	0.5970	**0.4357**	42.7648	28.4317	0.5163	0.0806	1.0476
HighRes-net (PROBA-V RED)	23.9802	0.5995	0.4476	42.0976	26.5394	0.5244	**0.0798**	1.0775
HighRes-net (simulated)	23.6873	0.5415	0.4905	43.1661	28.4554	0.5349	0.0857	1.0651
RAMS (PROBA-V NIR)	23.9743	0.6004	0.4420	42.6352	**28.5710**	**0.5109**	0.0815	**1.0271**
RAMS (PROBA-V RED)	24.0145	**0.6015**	0.4593	41.0973	25.8652	0.5233	0.0808	1.0686
RAMS (simulated)	23.8417	0.5324	0.5621	**43.3115**	28.5120	0.5198	0.0831	1.0436

also done, but the methodology does not allow for such analysis as with CVIU/SISAR.

2) RAMS (PROBA-V NIR) delivers the best visual quality overall. It achieves consistently strong metrics across the board, showing the highest MSIP and REP_M scores and ranking second-best in LPIPS. This suggests an excellent balance between fidelity and perceptual quality.
3) The second-best visual performance is offered by HighRes-net (PROBA-V NIR), which maintains strong metrics and a low LPIPS, indicative of better perceptual alignment.
4) Models trained on synthetic Sentinel-2 data have lower PSNR and weaker SSIM values compared to their PROBA-V counterparts. Despite this, they achieve relatively high REP_S and REP_M scores, indicating their ability to reconstruct sharper local details, even in the absence of precise pixel-level alignment with the reference image.

In summary, Bicubic Mean and Mean Bicubic yield the highest PSNR and SSIM, but the worst perceptual outcomes and poor LPIPS and MSIP scores. By contrast, RAMS (PROBA-V NIR) stands out visually, backed by strong metrics, with HighRes-net (PROBA-V NIR) close behind. Meanwhile, the simulated Sentinel-2 models confirm that sharper details may be underappreciated by fidelity-based metrics (PSNR, SSIM) but can be captured by the repeatability and perceptual assessments.

VI. CONCLUSION

In this study, we explored the utility of keypoint-based metrics for assessing the quality of super-resolved images across single- and multi-image scenarios. By capturing local feature consistency through multi-scale repeatability, our approach complements conventional fidelity metrics (PSNR, SSIM) and perceptual measures (LPIPS), providing a stronger alignment with human subjective ratings. Experimental results on CVIU-17, SISAR, and MuS2 confirm that multi-scale keypoint detection highlights structurally salient regions that significantly influence observers' quality judgments. Moreover, the proposed MSIP variants offer an effective means to combine different spatial kernels and weights, focusing on various frequency components within super-resolved images.

Although the proposed metrics show strong potential, further work is needed to address current limitations and explore open challenges. First, the comparison with advanced full- and no-reference IQA metrics is limited; extending this analysis is essential to fully position keypoint-based evaluation within the broader IQA landscape. Ongoing research focuses on adapting the MSIP metric specifically for SR tasks, both as a standalone quality measure and as a trainable loss function in SR models. Preliminary results indicate its potential to be comparable with state-of-the-art perceptual metrics. Additionally, future work should address geometric misalignments by evaluating repeatability under non-identity homographies and by integrating registration-aware keypoint detectors. Finally, we plan to extend the MuS2 dataset with human opinion scores, enabling more robust evaluation of MISR methods in real-world settings.

REFERENCES

[1] P. Benecki, M. Kawulok, D. Kostrzewa, and L. Skonieczny, "Evaluating super-resolution reconstruction of satellite images," *Acta Astronautica*, vol. 153, pp. 15–25, 2018. doi: 10.1016/j.actaastro.2018.07.035

[2] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.Net: Keypoint detection by handcrafted and learned CNN filters," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00593 pp. 5835–5843.

[3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005. doi: 10.1109/TPAMI.2005.188

[4] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1062–1066, 2020. doi: 10.1109/LGRS.2019.2940483

[5] X. Lin, X. Liu, H. Yang, X. He, and H. Chen, "Perception- and fidelity-aware reduced-reference super-resolution image quality assessment," *IEEE Transactions on Broadcasting*, vol. 71, no. 1, pp. 323–333, 2025. doi: 10.1109/TBC.2024.3475820

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, vol. 19, no. 11, pp. 2861–2873, 2010. doi: 10.1109/TIP.2010.2050625

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016. doi: 10.1109/TPAMI.2015.2439281

[8] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46475-6_25 pp. 391–407.

[9] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.182 pp. 1646–1654.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. doi: 10.1109/CVPRW.2017.151 pp. 1132–1140.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.19 pp. 105–114.

[12] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-11021-5_5 pp. 63–79.

[13] M. Zyrek and M. Kawulok, "Task-driven single-image super-resolution reconstruction of document scan," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024. doi: 10.15439/2024F7855 p. 259–264.

[14] Z. Bairi, K. B. Bey, O. Ben-Ahmed, A. Amamra, and A. Bradai, "Dual-path image reconstruction: Bridging vision transformer and perceptual compressive sensing networks," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. PTI, 2023. doi: 10.15439/2023F978 pp. 347–354.

[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023. doi: 10.1109/TPAMI.2022.3204461

[16] R. Schultz and R. Stevenson, "Improved definition video frame enhancement," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 1995. doi: 10.1109/ICASSP.1995.479905 pp. 2169–2172 vol.4.

[17] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004. doi: 10.1109/TIP.2004.834669

[18] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of proba-v images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019. doi: 10.1007/s42064-019-0059-8

[19] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "HighRes-net: Recursive fusion for multi-frame super-resolution of satellite imagery," *arXiv preprint arXiv:2002.06460*, 2020. doi: 10.48550/arXiv.2002.06460

[20] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2020. doi: 10.1109/TGRS.2019.2959248

[21] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, 2020. doi: 10.3390/rs12142207

[22] J. Li, Q. Lv, W. Zhang, B. Zhu, G. Zhang, and Z. Tan, "Multi-attention multi-image super-resolution transformer (mast) for remote sensing," *Remote Sensing*, vol. 15, no. 17, 2023. doi: 10.3390/rs15174183

[23] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, "TR-MISR: Multiimage super-resolution based on feature fusion with transformers," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1373–1388, 2022. doi: 10.1109/JSTARS.2022.3143532

[24] T. Tarasiewicz and M. Kawulok, "A graph neural network for heterogeneous multi-image super-resolution," *Pattern Recognition Letters*, vol. 189, pp. 214–220, 2025. doi: 10.1016/j.patrec.2025.01.028

[25] ——, "A graph attention network for real-world multi-image super-resolution," *Information Fusion*, p. 103325, 2025. doi: 10.1016/j.inffus.2025.103325

[26] L. Shu, Q. Zhu, Y. He, W. Chen, and J. Yan, "A survey of super-resolution image quality assessment," *Neurocomputing*, vol. 621, p. 129279, 2025. doi: 10.1016/j.neucom.2024.129279

[27] H. Zhang, S. Su, Y. Zhu, J. Sun, and Y. Zhang, "Boosting no-reference super-resolution image quality assessment with knowledge distillation and extension," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10095465 pp. 1–5.

[28] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00652 pp. 6228–6237.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00068 pp. 586–595.

[30] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810

[31] K. Zhang, T. Zhao, W. Chen, Y. Niu, J. Hu, and W. Lin, "Perception-driven similarity-clarity tradeoff for image super-resolution quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5897–5907, 2024. doi: 10.1109/TCSVT.2023.3341626

[32] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. doi: 10.1109/TIP.2011.2109730

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. doi: 10.1109/LSP.2012.2227726

[34] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: A real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022. doi: 10.1109/TIP.2022.3154588

[35] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017. doi: 10.1016/j.cviu.2016.12.009

[36] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, "Learning-based quality assessment for image super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3570–3581, 2022. doi: 10.1109/TMM.2021.3102401

[37] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, "A real-world benchmark for sentinel-2 multi-image super-resolution," *Scientific Data*, vol. 10, no. 1, p. 644, 2023. doi: 10.1038/s41597-023-02538-9

[38] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001. doi: 10.1109/ICCV.2001.937655 pp. 416–423 vol.2.

[39] C. G. Harris and M. J. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988.