Supplementary Material: Advances and Challenges in Infrared–Visible Image Fusion

1 Overview

This Supplement accompanies the main manuscript and is intended to be self–contained. It serves three purposes: (i) to provide compact, deployment–oriented matrices of representative IVIF methods (Tables 1–3); (ii) to collect formal definitions and equivalent forms of widely used objective metrics with unified notation (Sec. 3); and (iii) to state minimal reproducibility conventions for reporting efficiency and robustness. The new matrices are aligned with the method taxonomy in the article and use a consistent set of columns—*label supervision* (label–free or not), *registration awareness* (explicit/implicit), *robustness cues* (misregistration, low–light, noise), a qualitative *runtime class* (fast/medium/slow), and concise *notes* capturing the core mechanism. Symbols (✓/△/×) indicate supported, partial/indirect, and not–addressed aspects, respectively; runtime classes summarize author–reported behavior and widely reproduced complexity and are *not* cross–paper speed rankings.

2 Method Matrices for Infrared–Visible Image Fusion

Table S1 collates *generative* approaches (GAN and diffusion), highlighting the trade–off between training/sampling cost and stability/chroma/detail preservation; Table S2 summarizes *discriminative/hybrid* designs (CNN/AE baselines, CNN–Transformer hybrids, physics-/model–driven or self–supervised variants) with their characteristic fidelity–robustness–efficiency balances; Table S3 groups *multi–task joint fusion* into task–guided fusion (detection/segmentation), registration–fusion synergy, and multi–perspective fusion/stitching, clarifying when alignment or task heads improve target utility and when they introduce annotation or compute overhead.

2.1 Generative Methods

To preserve readability while maintaining full methodological coverage, Table S1 collates representative *generative* IVIF systems into a uniform, scan-friendly matrix. We group GAN- and diffusion-based designs and encode, via compact columns, whether training is label-free, whether registration awareness is explicit or implicit, qualitative robustness under adverse conditions (misregistration, low light, noise), and a coarse runtime proxy derived from author reports and widely reproduced settings. This table complements the architectural discussion in the main article.

Table S1: Generative IVIF methods: strengths, limitations, and suitable scenarios.

Method	Strengths	Limitations	Suitable Scenarios
FusionGAN [1]	End-to-end adversarial learning discovers fusion rules automatically; adaptively integrates IR saliency and VIS texture without manual design	Training instability and mode bias are common in GANs; robustness drops under strong nonlinear illumination changes	General-purpose fusion where handcrafted rules are hard to design; offline processing with moderate compute
GANMc [2]	Multi-class constraints alleviate modality/class imbalance; improves weakly-supervised representation learning	Dual-discriminator gradients make optimization unstable; risk of semantic drift during adversarial alignment	Datasets with class imbalance or rare targets; surveillance needing minority-class retention
Dense-FG [3]	Dense connections enhance feature propagation; better recovery of local details	Possible overfitting to loss design; computationally expensive, limits real-time use	Detail-sensitive analysis (e.g., edges/structures); scenarios where runtime is secondary
DCDR-GAN [4]	Dense-connected disentangled representations suppress cross-modality interference; lower effective information loss	Suboptimal color/detail enhancement; slow convergence during training	Scenes requiring clearer modality separation; thermal targets over textured backgrounds
DDcGAN [5]	Dual discriminators balance dual-modal learning; reduced risk of source-information loss	Loss coordination is tricky; potential geometric distortion in fused outputs	High-fidelity fusion where preserving both modalities is critical; offline/near-real-time with tuning budget
Attention-FGAN [6]	Multi-scale attention enhances global context and saliency; mitigates single-discriminator bias	Compute-intensive attention; visible-salient target characterization may still be insufficient	Wide-FOV scenes with large context range; analysis emphasizing global consistency

(Continued on next page)

(Continued from previous page)

Method	Strengths	Limitations	Suitable Scenarios
CSPA-GAN [7]	Cross-scale pyramid attention automates feature matching; label-free adversarial training reduces labeling demand	Feature-map losses less interpretable; cross-scale attention adds heavy compute	Data-scarce settings; exploratory fusion without dense labels
AT-GAN [8]	Intensity attention + semantic conversion improves robustness on low-quality inputs; ∼30% faster inference (reported)	Depends on predefined lighting priors; limited adaptability to fast-changing scenes	Low-light/low-quality surveillance with mild dynamics; quasi-real-time pipelines
DUGAN [9]	Dual fusion paths + U-shaped discriminator combine global-difference modeling and local denoising; improved SNR and edges	∼50% higher training cost (dual paths); degrades under extreme noise	Moderate-noise environments; applications valuing denoising and edge sharpness
DDFM [10]	Bayesian *conditional* diffusion with prior-guided generation; stable convergence, interpretable sampling; label-free feasible	Long training (likelihood correction increases cost); quality depends on prior statistics	Safety/mission-critical use favoring stability/interpretability; heterogeneous scenes with variable conditions
DIf-Fusion [11]	Color-aware diffusion loss preserves VIS chromatic fidelity; reduced IR overexposure sensitivity	Assumes near-linear color space; performance drops in very low-light	Color-critical tasks (daytime driving, inspection); scenes where chroma realism matters
Diff-IF [12]	Knowledge-guided diffusion injects domain priors; fine-grained controllability with fewer labels	Costly prior distillation; transferred knowledge may not generalize broadly	Task-driven fusion with available priors; detection/segmentation co-training

(Continued on next page)

(Continued from previous page)

Method	Strengths	Limitations	Suitable Scenarios
FusionDiff [13]	Dual-source–guided multi-stage denoising; strong noise resistance; preserves sharp edges and multi-focus cues	Requires accurate registration; lacks explicit dynamic-interaction mechanisms	Multi-focus/static scenes; noisy inputs with reliable alignment
LFDT-Fusion [14]	UNet skip connections + cross-modality Transformer attention inside diffusion; dynamic feature interaction; high-resolution fidelity; good generalization	>100M parameters hinder deployment on edge; extreme-lighting robustness under-validated	High-resolution, quality-first fusion on GPU servers; offline or near-real-time with ample compute

Across entries, GAN variants are typically label-free with *medium* inference cost; their stability and illumination robustness hinge on loss composition and discriminator balance. Diffusion variants trade higher training/sampling cost for stable, interpretable sampling and strong detail/chroma preservation, suiting server-side or offline deployment. For transparency and reproducibility, we recommend standardizing input size, precision and timer scope, disclosing seeds and repeat counts, and clearly separating author-reported from reproduced runtimes.

2.2 Discriminative and Hybrid Methods

Table S2 summarizes *discriminative* IVIF methods across three pragmatic families: (i) CNN/auto-encoder baselines, (ii) hybrid CNN–Transformer designs, and (iii) physics-/model-driven or self-supervised variants. Each entry reports label efficiency, registration awareness, qualitative low-light behavior, a runtime class (fast/medium/slow), and a concise note capturing the core mechanism (e.g., dense/nested connections, correlation- or cross-attention-based interaction, algorithm unrolling). The matrix exposes fidelity–robustness–efficiency trade-offs at a glance.

Table S2: Analysis of IVIF methods based on discriminative fusion, outlining core innovations, technical advantages, and key limitations.

Method	Core Innovation	Technical Advantages	Key Limitations
IFCNN [15]	Perception loss-based framework for cross-modality feature alignment	Generalizable fusion capability across diverse imaging tasks, Adaptive feature selection through perception loss regularization	Lacks IVIF-specific fusion rule optimization, Sub-optimal performance in low-light/high-contrast scenarios
SCGAFusion [16]	Group convolution with residual non-local attention mechanism	Enhances local feature correlation through group convolution, Captures pixel-wise long-range dependencies using residual attention	Insufficient local detail preservation, Training sensitivity to kernel parameters
DCTNet [17]	CNN-Transformer hybrid architecture with frequency-domain decomposition	Breaks CNN's input size limitation through DCT-based feature extraction, Achieves parallel processing efficiency with dual branches	High computational cost of DCT when not block-processing, Risk of high-frequency information loss during reconstruction
MACCNet [18]	Multi-scale attention with cross-convolution modules	Local significance-guided feature refinement, Cross-modality feature complementarity through convolutional interactions	Computationally intensive hyperparameter tuning, High memory consumption due to multi-scale attention layers
PFCFuse [19]	PoolFormer-CNN-based dual branching architecture	Reduces parameter redundancy in multi-branch networks, Balances feature extraction and fusion precision through average pooling	Spatial information loss due to pooling operations, Hardware acceleration required for real-time deployment
AUIF [20]	Physics-driven dual-channel network design	Simplified architecture with interpretable feature extraction, Effective retention of thermal radiation and visual details	Higher training complexity compared to CNNs, Deployment challenges in resource-constrained environments
DenseFuse [21]	Dense block-based feature preservation architecture	Maximizes information retention through intermediate layer extraction, Simplified training process with feature reuse	Insufficient infrared target emphasis, Limited adaptability to fusion strategy variations
NestFuse [22]	Nested connection with spatial/channel attention modules	Multi-scale feature preservation through hierarchical connections, Minimizes semantic gap between modalities	Attention weight dependency on pre-trained networks, Information loss due to global pooling operations

Continued on next page

Table S2 – Continued from previous page

Method	Core Innovation	Technical Advantages	Key Limitations
SEDRFuse [23]	Symmetric residual network with contrast enhancement capability	Beyond-pixel-level fusion through contrast optimization, Efficient computation through shared channel weights	Modal-specific information loss due to shared features, Pre-training bias towards common features
Xie et al. [24]	Spectral analysis-driven attention framework	Frequency-wise complementary feature learning, Automated feature selection mechanism	Sensitivity to specific features leading to information imbalance, High computational cost of spectral transformations
SSL-WAEIE [25]	Self-supervised weighted autoencoder for feature exchange	Multi-stage feature extraction without labeled data, Automatic information alignment between modalities	Vulnerability to outliers/noise, Extended training and inference durations
DIVFusion [26]	Visual perception-enhanced fusion framework for low-light conditions	Coupled image enhancement and fusion, Robustness to low-light environments through joint optimization	Overexposure handling deficiency in bright regions, Need for additional IR integration strategies
GuideFuse [27]	Gradient-guided semi-automated fusion strategy	Visualizable feature selection through gradient analysis, Dynamic weight adjustment for flexible fusion	Manual intervention requirement, Memory consumption from intermediate results storage
DeDNet [28]	Decomposition-driven fusion and denoising network	Joint optimization of fusion and denoising processes, Improved low-quality image recovery	Performance degradation under non-uniform noise, Computational burden from decomposition modules
DeFusion [29]	Lightweight self-supervised feature decomposition framework	Unsupervised extraction of shared/private features, Simplified training without complex loss functions	Feature loss during non-paired data training, Limited detail preservation due to simplistic fusion rules
SwinFuse [30]	Swin-Transformer-based lightweight architecture	Enhanced long-range dependency modeling, Parameter efficiency superior to conventional Transformers	Simple fusion strategy lacking innovation, Real-time performance constrained by Transformer compute load
TFIV [31]	Pure Transformer-based fusion model	Alleviates inter-modality differences at pixel positions, Global context awareness through self-attention	High computational complexity, Memory issues with long sequence modeling

Continued on next page

Table S2 – Continued from previous page

Method	Core Innovation	Technical Advantages	Key Limitations
CrossFuse [32]	CNN-Transformer hybrid with two-stage training	Complementary feature extraction through cross-stage optimization, Combines CNN's locality with Transformer's globality	Risk of information loss during fusion, Increased deployment complexity with two-stage training
YDTR [33]	Y-shaped dynamic Transformer framework	Local-global feature synergy through dynamic branching, Adaptability to source image similarity variations	Incomplete resolution of modality differences in Y-structure, Training instability from dynamic weights
CDDFuse [34]	Feature-decoupled dual-branch Transformer-CNN architecture	Correlation-driven feature decomposition, Explicit separation of shared/private features	High-frequency feature loss during propagation, Parameter redundancy in dual-branch design
TGFuse [35]	Transformer-GAN lightweight framework	Balanced image-level perception optimization, GAN-induced detail preservation	Loss function focus on structural similarity, Detail loss during GAN training phase
M2Fnet [36]	Transformer-based multimodal fusion network	Adaptive feature alignment under varying illumination conditions, Automated feature space mapping	Manual analysis cost for modality-specific distributions, Deployment complexity with multimodal inputs
DATFuse [37]	Dual-attention Transformer with unsupervised feature extraction	Local-global feature joint modeling, Unsupervised complementary information capture	High registration accuracy requirement, Computational burden from dual attention mechanisms
TCCFusion [38]	Transformer-correlation hybrid architecture	Preserves complementary information beyond receptive field limitations, Long-distance dependency modeling through correlations	Limited fusion processing capacity, Additional computational cost from correlation calculations

CNN/AE baselines remain attractive for lightweight, general-purpose fusion but only partially address misregistration and strong illumination changes. Hybrids introduce long-range context and cross-modal interaction, frequently improving the accuracy–efficiency balance at *medium* runtime; implicit registration cues help but do not eliminate alignment errors. Physics-/model-driven and self-supervised variants enhance interpretability, denoising, or low-light resilience with modest overhead. Reporting should follow the same reproducibility protocol as in Table S1.

2.3 Multi-task Joint Fusion

Table S3 organizes *multi-task joint fusion* into three streams that are most relevant to deployment: (i) fusion guided by detection/segmentation heads to bias features toward task utility;

(ii) registration–fusion synergy that iteratively reduces alignment error via coarse-to-fine warps or recurrent correction; and (iii) multi-perspective fusion and stitching, where parallax and occlusion require disparity/pose handling beyond single-view pipelines. The unified matrix format aligns columns with Tables S1–S2 to facilitate cross-family comparison.

Table S3: Analysis of IVIF methods based on multi-task joint fusion, highlighting core innovations, technical advantages, and key limitations.

Method	Core Innovation	Technical Advantages	Key Limitations
MURF [39]	End-to-end framework integrating registration and fusion with iterative error compensation	Real-time registration error correction through feature similarity refinement, Cross-modal alignment enhancement via constraint propagation	Limited robustness to extreme deformations, Dependency on large-scale training data
ReCoNet [40]	Dual-branch residual network with cross-layer feature interaction	Separation of feature extraction and fusion pathways, Enhancement of modality-specific complementary features through hierarchical interactions	Computational redundancy due to parameter duplication, Limited performance in non-rigid registration scenarios
Jiang et al. [41]	Joint optimization of fusion and stitching through multi-view feature alignment	Revelation of intrinsic correspondences through perspective-aware feature learning, Cross-spectral feature complementarity for wide-baseline alignment	High computational complexity of feature alignment, Sensitivity to low-texture regions
Wang et al. [42]	Cross-modal generative registration paradigm (CGRP) with style transfer	Unsupervised misalignment fusion of IR and visible light images, Geometric structure preservation through style migration, Reduction of ghosting artifacts	Implicit assumptions in generative modeling, Requirement for intermediate result storage
Nie et al. [43]	Heterogeneous feature interaction via CNN-Transformer hybrid architecture	Enriched feature diversity through modality-specific feature extraction, Improved target detection accuracy, Adaptive multimodal input handling	Limited robustness to varying illumination conditions, Generalization bottleneck in cross-domain scenarios
RFNet [44]	Feedback-driven fusion-registration synergy with coarse-to-fine refinement	Fusion feedback for registration error correction, Coarse-global disparity estimation followed by fine-local adjustment, Local misalignment rectification capability	Error amplification in global disparity estimation, Computational overhead in fine-grained refinement stages

Continued on next page

Table S3 – Continued from previous page

Method	Core Innovation	Technical Advantages	Key Limitations
TarDAL [45]	Target-driven attention learning framework	Efficient extraction of target region features, Low computational complexity, Real-time inference capability	Limited adaptability to complex backgrounds, Requires high-quality target annotations
SegFormer [46]	Fusion-segmentation cascade network with dual-scale feature extraction	Dual-scale feature preservation through dilated residual dense blocks (DRDB), Semantic guidance for fusion decisions, Real-time processing capability	Annotation dependency for segmentation networks, Potential high-frequency detail loss in fusion
DetFusion [47]	Detection-driven fusion network with shared attention mechanism	Target-oriented feature selection through detection auxiliary branch, Information utilization enhancement via shared attention modules, Pixel intensity distribution preservation	Detection inaccuracies leading to fusion biases, Performance degradation in complex scenes
SeAFusion [48]	Semantic-aware fusion framework with segmentation guidance	Semantic loss-driven fusion optimization through segmentation predictions, Information reflux mechanism for feature consistency, Deployable real-time processing capability	Pre-training data dependency for segmentation networks, Limited adaptability to extreme lighting conditions

Task-driven designs raise AP/mAP and mIoU while keeping competitive fusion scores, but can inherit biases from imperfect detectors/segmenters. Registration–fusion co-optimization reduces ghosting and seam artifacts at additional compute cost. Multi-view/stitching broadens spatial coverage and geometric consistency yet demands robust pre-alignment (e.g., SLAM/VO priors) and often heavier memory. We recommend coupling fusion metrics with task accuracy and device-level efficiency, and plotting condition–performance curves over misregistration and illumination strata.

3 Evaluation metrics

Because no single fusion algorithm uniformly dominates across modalities and imaging conditions, the quality of fused images must be evaluated with care. In practice, two complementary paradigms are used: *subjective* evaluation, which relies on human visual judgement, and *objective* evaluation, which relies on mathematically defined criteria.

Subjective assessment is intuitive but suffers from variability across observers (e.g., cognition, attention, fatigue) and is costly to conduct at scale. Consequently, for reproducible benchmarking and method development, the community primarily depends on objective metrics.

Objective evaluation aims to approximate aspects of human perception while remaining quantifiable and repeatable. Commonly used metrics can be grouped into four families: (i)

Table S4: Image fusion quality assessment metrics

Category	Metric	Description	Best values
Information theory-based	Entropy (EN)	Measures the information content in fused image	Higher
	Mutual information (MI)	Quantifies shared information between source and fused images	Higher
	Qabf	Edge-based similarity metric	Higher
Visual Quality-based	Standard deviation (SD)	Measures contrast in fused image	Higher
	PSNR	Peak signal-to-noise ratio	Higher
	VIF	Visual information fidelity index	Higher
Image Features-based	Edge intensity (EI)	Measures sharpness of edges	Higher
	Spatial Frequency (SF)	Combines row and column frequency	Higher
	Average gradient (AG)	Measures local contrast	Higher
Structural Similarity-based	Correlation coefficient (CC)	Linear correlation between images	Higher
	SSIM	Structural similarity index	Higher
	RMSE	Root mean squared error	Lower

information-theoretic criteria, (ii) feature/descriptor–based criteria, (iii) structural-similarity measures, and (iv) perception-inspired models [49]. Table S4 summarizes representative metrics in these families, indicating their intuition, category, and the preferred direction (higher/lower) when assessing fusion quality.

The subsections that follow provide concise, formal definitions (and equivalent forms where useful), specify symbol conventions at first use, and note practical considerations such as admissible ranges and typical sensitivities (e.g., to noise, misregistration, window size, histogram estimation). This appendix focuses exclusively on objective metrics to support transparent and reproducible evaluation in the main text.

3.1 Information-theoretic Metrics

These metrics quantify how much information from source images is preserved.

3.1.1 Entropy (EN)

Entropy (EN))[50] , a key concept in information theory, quantifies the uncertainty or randomness of information within a system. In image processing, entropy is commonly used to evaluate the amount of information present in an image. Typically, a higher entropy value reflects richer information in the fused image, implying superior fusion quality. The mathematical formulation is defined as follows:

$$\mathrm{EN} = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

Here, p_i represents the probability of the $i-th$ gray level occurring, and n is the total number of gray levels in the image. The formula calculates the entropy by summing the product of the probability of each possible gray level and its logarithm (taken as a negative value), thereby quantifying the richness of information in the image.

3.1.2 Mutual Information (MI)

Mutual Information (MI) is a key concept in information theory, used to measure the interdependence between two random variables and evaluate the amount of information they share. In the field of image fusion, MI reflects how much information from the original images is retained in the fused image. A higher MI value indicates a stronger correlation between the random variables, meaning the fused image preserves more information from the source images. The calculation formula is as follows:

$$\mathrm{MI} = \mathrm{MI}_{A,F} + \mathrm{MI}_{B,F} \tag{2}$$

In $MI_(A,F)$,$MI_(B,F)$,they represent the mutual information between the fused image and the source images, respectively. The mutual information between two random variables is defined as follows:

$$\mathrm{MI}_{A,F} = \sum_{a\in A}\sum_{f\in F} p_{A,F}(a,f)\,\log\frac{p_{A,F}(a,f)}{p_A(a)\,p_F(f)} \tag{3}$$

In this formula,$p_A(a)$and $p_F(f)$represent the edge probability histogram of source image A and fusion image F, respectively. $p_{A,F}(a,f)$represents the joint histogram of source image A and fusion image F.

3.1.3 Qabf

Qabf[51] is a measure that evaluates how well significant information from the source image is preserved in the fused image without introducing distortion. The Qabf value directly reflects the quality of the fused image. Defined as follows:

$$Q(a,b,f) = \frac{1}{|W|}\sum_{w\in W}\left(\lambda(w)Q_0(a,f|w) + (1-\lambda(w))\,Q_0(b,f|w)\right) \tag{4}$$

Where:

$$\lambda(w) = \frac{s(a|w)}{s(a|w)+s(b|w)}, Q_0(x,f) = \frac{\sigma_{xf}}{\sigma_x\sigma_f} \times \frac{2\bar{x}\bar{f}}{(\bar{x})^2+(f)^2} \times \frac{2\sigma_x\sigma_f}{\sigma_x^2+\sigma_f^2} \tag{5}$$

W represents a set of local windows in the image, where a and b are the source images,f is the fused image.$Q_0(a,f|w)$ and $Q_0(b,f|w)$ denote the similarity measures between the source images a and b , respectively, and the fused image f within the window w, These measures are related to contrast, sharpness, and entropy. Q_0 is a metric for measuring image correlation loss, brightness distortion, and contrast distortion [52], where σ is the standard deviation of the image.

3.1.4 Sum of the Correlations of Differences (SCD)

Sum of correlation of Differences(SCD)[53] stands out by not directly evaluating the correlation between the source images and the fused image, but rather by assessing the correlation

between the source images and the difference images derived from the fused image. This approach more accurately reflects the performance of the fused image in preserving information from the source images. Let S_1 and S_2 be defined as the two source images, and the difference images D_1 and D_2 be defined as:

$$D_1 = F - S_2, D_2 = F - S_1 \tag{6}$$

The formula for calculating SCD is:

$$\text{SCD} = r(D_1, S_1) + r(D_2, S_2) \tag{7}$$

wherer$(,)$ is a function that computes the correlation between two images. Assuming S_k and D_k represent the average pixel values of image S_k and D_k, respectively, the correlation is defined as:

$$r(D_k, S_k) = \frac{\sum_i \sum_j (D_k(i,j) - \bar{D}_k)(S_k(i,j) - \bar{S}_k)}{\sqrt{(\sum_i \sum_j (D_k(i,j) - \bar{D}_k)^2)(\sum_i \sum_j (S_k(i,j) - \bar{S}_k)^2)}} \tag{8}$$

In general, a higher SCD value indicates better image fusion performance.

3.2 Perceptual Quality Metrics

These metrics reflect visual fidelity and perceptual quality[54].

3.2.1 Standard Deviation (SD)

Standard Deviation (SD) [55] is a metric used to evaluate the distribution of grayscale and contrast in fused images. Since the human eye is more sensitive to high-contrast regions, a larger standard deviation generally indicates better visual quality. The formula for calculating SD is as follows:

$$\text{SD} = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (F(i,j) - \mu)^2} \tag{9}$$

Here M and N represent the height and width of the fused image, respectively, $F(i,j)$ denotes the pixel value at position(i,j), and μ is the mean pixel value of the image. A higher SD value indicates greater contrast in the fused image, resulting in better visual quality.

3.2.2 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) evaluates the degree of distortion in an image by comparing the ratio between the maximum possible power (peak power) of the image and the noise power affecting its quality. A higher PSNR value indicates a higher signal-to-noise ratio and better image quality. The formula is defined as:

$$\text{PSNR} = 10 \log_{10}(\frac{MAX_I^2}{MSE}) \tag{10}$$

Here,MAX_I represents the maximum possible pixel value of the image, typically 255 for 8-bit images. MSE stands for Mean Squared Error, which measures the average squared error between the original image and the distorted image. PSNR provides a quantitative way to assess distortion during image compression or processing, although it does not always align well with human perception of image quality.

3.2.3 Visual Information Fidelity (VIF)

Visual Information Fidelity (VIF) [56, 57] combines natural image statistics, a distortion model, and a human visual system (HVS) model to assess image quality. Compared with PSNR and SSIM, VIF often correlates better with human perception. The workflow includes image decomposition into subbands, block-wise modeling, and information (mutual information) aggregation.

$$\mathrm{VIF} = \frac{\sum_{k}\sum_{b} \log_2\left(1 + \dfrac{g_{k,b}^2\,\sigma_{x,k,b}^2}{\sigma_{v,k,b}^2 + \sigma_n^2}\right)}{\sum_{k}\sum_{b} \log_2\left(1 + \dfrac{\sigma_{x,k,b}^2}{\sigma_n^2}\right)}. \tag{11}$$

Where k and b index the subband and block, respectively; $g_{k,b}$ is the scalar gain relating the reference to the distorted coefficients in block b of subband k; $\sigma_{x,k,b}^2$ is the variance (signal power) of the reference coefficients; $\sigma_{v,k,b}^2$ is the variance of the distortion (signal-independent) in that block; and σ_n^2 is the HVS internal noise variance. This standard parametrization avoids the non-standard symbol C_u (previously described as a "channel capacity constant") and removes stray I terms.

VIF is typically in $[0, 1]$ for natural images; higher values indicate better preservation of visual information.

3.2.4 Universal Image Quality Index (UIQI)

The Universal Image Quality Index (UIQI) [52], also referred to as the Q index, assesses fusion performance by comparing the correlation, luminance similarity, and contrast similarity between the fused image and the reference image. The formula for UIQI is as follows:

$$Q(F, R) = \frac{\sigma_{FR}}{\sigma_F \sigma_R} \times \frac{2\mu_F \mu_R}{\mu_F^2 + \mu_R^2} \times \frac{2\sigma_F \sigma_R}{\sigma_F^2 + \sigma_R^2} \tag{12}$$

In the formula, σ_F and σ_R represent the standard deviations of the fused image F and the reference image R,respectively, while$\sigma_F R$ denotes the covariance between the fused image F and the reference image R. μ_F and μ_R represent the mean values of the fused image F and the reference image R, respectively. The terms in the formula correspond to correlation, average luminance similarity, and contrast similarity. A higher UIQI value indicates greater similarity between the fused image and the reference image.

3.3 Feature-based Metrics

These metrics focus on structural detail and edge sharpness.

3.3.1 Edge Intensity (EI)

Edge strength is a metric used to quantify the sharpness or prominence of edges in an image, defined by measuring rapid changes in pixel intensity. Currently, the Sobel operator is a widely used method for calculating edge strength. It convolves the image with two distinct filters (designed to detect horizontal and vertical edges, respectively) and then computes the gradient magnitude. The formula is as follows:

$$EI(F) = \frac{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} s_x(i,j)^2 + s_y(i,j)^2}}{M * N} \tag{13}$$

Where s_x and s_y represent the results of the Sobel operator convolution in the horizontal and vertical directions, respectively. A higher edge strength value indicates sharper or more prominent edges, thereby enabling more accurate image analysis and processing.

3.3.2 Spatial Frequency (SF)

Spatial Frequency (SF) is a gradient-based metric used to measure the rate of gray-level changes in an image, where the gradient can be divided into horizontal and vertical gradients, corresponding to row frequency (RF) and column frequency (CF), respectively. The definition of SF is as follows:

$$\text{SF} = \sqrt{RF^2 + CF^2} \tag{14}$$

The formulas for calculating RF and CF are as follows:

$$\text{RF} = \sqrt{\frac{1}{mn}\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}(F(i,j) - F(i,j+1))^2} \tag{15}$$

$$\text{CF} = \sqrt{\frac{1}{mn}\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}(F(i,j) - F(i+1,j))^2} \tag{16}$$

m and n are the number of rows and columns in the image, and $F(i,j)$ is the gray value of the image at position (i,j).

3.3.3 Average Gradient (AG)

Average Gradient (AG) measures the sharpness and texture details of an image by calculating the average gradient value within the image. A higher AG value indicates richer texture information in the fused image and better image quality. The calculation method for AG is as follows:

$$\text{AG} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|\nabla I(i,j)| \tag{17}$$

Where $|\nabla I(i,j)|$ is the gradient magnitude of the pixel at position(i,j).

3.4 Structure-based Metrics

These metrics evaluate structural similarity to source or reference images.

3.4.1 Correlation coefficient (CC)

The correlation coefficient (CC)[58] is a measurement tool for evaluating the strength and direction of the linear relationship between the fused image and the source images. It quantifies the similarity between two images to measure their linear correlation by calculating the sum of the products of the deviations of the intensity values of the two images from their respective mean values, and then dividing it by the product of the square roots of the sums of the squares of their respective deviations. The calculation formula is as follows:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N} \left(I_i - \bar{I}\right)\left(R_i - \bar{R}\right)}{\sqrt{\sum_{i=1}^{N} \left(I_i - \bar{I}\right)^2} \sqrt{\sum_{i=1}^{N} \left(R_i - \bar{R}\right)^2}} \tag{18}$$

Where $I_i$and R_i re the intensity values of the fused image and the source image at the i-th pixel point, respectively. $\bar{I}$ And $\bar{R}$ are the average intensity values of the fused image and the source image, respectively. N is the total number of pixels in the image. The value of CC ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear correlation.

3.4.2 Structural similarity index measure (SSIM)

The Structural Similarity Index Measure (SSIM) evaluates the loss and distortion of a fused image by comparing it with a reference image. It combines luminance, contrast, and structural similarities into a single quality score. The formula is

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}. \tag{19}$$

Where μ_x and μ_y are the mean luminances of the two images; σ_x^2 and σ_y^2 are the variances (image contrast); and σ_{xy} is the covariance (structural similarity). The stabilizing constants follow the standard setting $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where K_1 and K_2 are small positive numbers (typically $K_1 = 0.01$, $K_2 = 0.03$) and L is the dynamic range of pixel values (e.g., $L = 255$ for 8-bit images). These constants prevent the denominators from approaching zero and *should not be set to 0.*

$$\mathrm{SSIM} \;=\; \omega_x\,\mathrm{SSIM}_{xf} \;+\; \omega_y\,\mathrm{SSIM}_{yf}. \tag{20}$$

Where ω_x and ω_y are the weights for the contributions of images x and y to the overall structural similarity (commonly $\omega_x, \omega_y > 0$ and $\omega_x + \omega_y = 1$).

3.4.3 Root-Mean-Square Error (RMSE)

The root mean squared error (RMSE) is commonly used in the evaluation of multispectral image fusion. It reflects the average error magnitude between the fusion result and the reference image from the pixel level. It is calculated by taking the average of the sum of the squared errors of all pixel points and then taking the square root, thus avoiding the problem of positive and negative values in the sum of squared errors canceling each other out. The specific formula is as follows:

$$\mathrm{RMSE} = \left[\frac{1}{M \times N \times B} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{B} [F_{1(i,j,k)} - R_{2(i,j,k)}]\right]^2 \tag{21}$$

M and N are the number of rows and columns of the image, respectively. B represents the number of bands of the image. $F_{1(i,j,k)}$ is the pixel value of the fused image at position (i, j), and $R_{2(i,j,k)}$ is the pixel value of the reference image at position (i, j). The smaller the RMSE, the closer the pixels of the fusion result are to those of the reference image, that is, the better the fusion effect.

References

[1] Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: FusionGAN: A Generative Adversarial Network for Infrared and Visible Image Fusion. Information Fusion **48**, 11–26 (2019) https://doi.org/10.1016/j.inffus.2018.09.004

[2] Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H.: GANMcC: A Generative Adversarial Network with Multiclassification Constraints for Infrared and Visible Image Fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–14 (2021) https://doi.org/10.1109/TIM.2020.3038013

[3] Xu, X., Shen, Y., Han, S.: Dense-fg: A fusion gan model by using densely connected blocks to fuse infrared and visible images. Applied Sciences **13**(8), 4684 (2023) https://doi.org/10.3390/app13084684

[4] Gao, Y., Ma, S., Liu, J.: DCDR-GAN: A Densely Connected Disentangled Representation Generative Adversarial Network for Infrared and Visible Image Fusion. IEEE Transactions on Circuits and Systems for Video Technology **33**(2), 549–561 (2023) https://doi.org/10.1109/TCSVT.2022.3206807

[5] Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.: DDcGAN: A Dual-Discriminator Conditional GAN for Multi-Resolution Image Fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020) https://doi.org/10.1109/TIP.2020.2977573

[6] Li, J., Huo, H., Li, C., Wang, R., Feng, Q.: AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. IEEE Transactions on Multimedia **23**, 1383–1396 (2021) https://doi.org/10.1109/TMM.2020.2997127

[7] Yin, H., Xiao, J., Chen, H.: Cspa-gan: A cross-scale pyramid attention gan for infrared and visible image fusion. IEEE Transactions on Instrumentation and Measurement **72**, 5014116 (2023) https://doi.org/10.1109/TIM.2023.3317932

[8] Rao, Y., Wu, D., Han, M., Wang, T., Yang, Y., Lei, T., Zhou, C., Bai, H., Xing, L.: AT-GAN: A Generative Adversarial Network with Attention and Transition for Infrared and Visible Image Fusion. Information Fusion **92**, 336–349 (2023) https://doi.org/10.1016/j.inffus.2022.12.007

[9] Chang, L., Huang, Y., Li, Q., Zhang, Y., Liu, L., Zhou, Q.: DUGAN: Infrared and Visible Image Fusion Based on Dual Fusion Paths and a U-Type Discriminator. Neurocomputing **578**, 127391 (2024) https://doi.org/10.1016/j.neucom.2024.127391

[10] Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: Ddfm: denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on

Computer Vision, pp. 8082–8093 (2023)

[11] Yue, J., Fang, L., Xia, S., Deng, Y., Ma, J.: Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. IEEE Transactions on Image Processing **32**, 5705–5720 (2023) https://doi.org/10.1109/TIP.2023.3322046

[12] Yi, X., Tang, L., Zhang, H., Xu, H., Ma, J.: Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. Information Fusion **110**, 102450 (2024) https://doi.org/10.1016/j.inffus.2024.102450

[13] Li, M., Pei, R., Zheng, T., Zhang, Y., Fu, W.: FusionDiff: Multi-Focus Image Fusion Using Denoising Diffusion Probabilistic Models. Expert Systems with Applications **238**, 121664 (2024) https://doi.org/10.1016/j.eswa.2023.121664

[14] Yang, B., Jiang, Z., Pan, D., Yu, H., Gui, G., Gui, W.: Lfdt-fusion: A latent feature-guided diffusion transformer model for general image fusion. Information Fusion **113**, 102639 (2025) https://doi.org/10.1016/j.inffus.2024.102639

[15] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: A general image fusion framework based on convolutional neural network. Information Fusion **54**, 99–118 (2020) https://doi.org/10.1016/j.inffus.2019.07.011

[16] Zhu, D., Ma, J., Li, D., Wang, X.: Scgafusion: A skip-connecting group convolutional attention network for infrared and visible image fusion. Applied Soft Computing **163**, 111902 (2024)

[17] Li, J., Liu, L., Song, H., Huang, Y., Jiang, J., Yang, J.: DCTNet: A Heterogeneous Dual-Branch Multi-Cascade Network for Infrared and Visible Image Fusion. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2023) https://doi.org/10.1109/TIM.2023.3325520

[18] Yang, Y., Zhou, N., Wan, W., Huang, S.: Maccnet: Multiscale attention and cross-convolutional network for infrared and visible image fusion. IEEE Sensors Journal **24**(10), 16587–16600 (2024) https://doi.org/10.1109/JSEN.2024.3385638

[19] Hu, X., Liu, Y., Yang, F.: PFCFuse: A PoolFormer and CNN Fusion Network for Infrared-Visible Image Fusion. IEEE Transactions on Instrumentation and Measurement **73**, 1–14 (2024) https://doi.org/10.1109/TIM.2024.3450061

[20] Zhao, Z., Xu, S., Zhang, J., Liang, C., Zhang, C., Liu, J.: Efficient and model-based infrared and visible image fusion via algorithm unrolling. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1186–1196 (2022) https://doi.org/10.1109/TCSVT.2021.3075745

[21] Li, H., Wu, X.: DenseFuse: A Fusion Approach to Infrared and Visible Images. IEEE Transactions on Image Processing **28**, 2614–2623 (2019) https://doi.org/

10.1109/TIP.2018.2887342

[22] Li, H., Wu, X., Durrani, T.: NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. IEEE Transactions on Instrumentation and Measurement **69**(12), 9645–9656 (2020) https://doi.org/10.1109/TIM.2020.3005230

[23] Jian, L., Yang, X., Liu, Z., Jeon, G., Gao, M., Chisholm, D.: SEDRFuse: A Symmetric Encoder-Decoder with Residual Block Network for Infrared and Visible Image Fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–14 (2021) https://doi.org/10.1109/TIM.2020.3022438

[24] Xie, Z., Zong, S., Li, Q., Cai, P., Zhan, Y., Liu, G.: Interactive residual coordinate attention and contrastive learning for infrared and visible image fusion in triple frequency bands. Scientific Reports **14**(1) (2024) https://doi.org/10.1038/s41598-023-51045-9

[25] Zhang, G., Nie, R., Cao, J.: Ssl-waeie: Self-supervised learning with weighted auto-encoding and information exchange for infrared and visible image fusion. IEEE/CAA Journal of Automatica Sinica **9**(9), 1694–1697 (2022) https://doi.org/10.1109/JAS.2022.105815

[26] Tang, L., Xiang, X., Zhang, H., Gong, M., Ma, J.: DIVFusion: Darkness-Free Infrared and Visible Image Fusion. Information Fusion **91**, 477–493 (2023) https://doi.org/10.1016/j.inffus.2022.10.034

[27] Zhang, Z., Li, H., Xu, T., Wu, X.-J., Fu, Y.: Guidefuse: A novel guided auto-encoder fusion network for infrared and visible images. IEEE Transactions on Instrumentation and Measurement **73** (2024) https://doi.org/10.1109/TIM.2023.3306537

[28] Huang, J., Li, X., Tan, H., Yang, L., Wang, G., Yi, P.: DeDNet: Infrared and Visible Image Fusion with Noise Removal by Decomposition-Driven Network. Measurement **237**, 115092 (2024) https://doi.org/10.1016/j.measurement.2024.115092

[29] Liang, P., Jiang, J., Liu, X., Ma, J.: Fusion from decomposition: A self-supervised decomposition approach for image fusion. In: European Conference on Computer Vision, pp. 719–735 (2022). Springer

[30] Wang, Z., Chen, Y., Shao, W., Li, H., Zhang, L.: Swinfuse: A residual swin transformer fusion network for infrared and visible images. IEEE Transactions on Instrumentation and Measurement **71** (2022) https://doi.org/10.1109/TIM.2022.3191664

[31] Li, J., Yang, B., Bai, L., Dou, H., Li, C., Ma, L.: TFIV: Multigrained Token Fusion

for Infrared and Visible Image via Transformer. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2023) https://doi.org/10.1109/TIM.2023.3312755

[32] Li, H., Wu, X.: CrossFuse: A Novel Cross Attention Mechanism-Based Infrared and Visible Image Fusion Approach. Information Fusion **103**, 102147 (2024) https://doi.org/10.1016/j.inffus.2023.102147

[33] Tang, W., He, F., Liu, Y.: YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer. IEEE Transactions on Multimedia **25**, 5413–5428 (2023) https://doi.org/10.1109/TMM.2022.3192661

[34] Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) **null**, 5906–5916 (2022) https://doi.org/10.1109/CVPR52729.2023.00572

[35] Rao, D., Xu, T., Wu, X.: TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. IEEE Transactions on Image Processing **32**, 5705–5720 (2023) https://doi.org/10.1109/TIP.2023.3273451

[36] Jiang, C., Ren, H., Yang, H., Huo, H., Zhu, P., Yao, Z., Li, J., Sun, M., Yang, S.: M2FNet: Multi-Modal Fusion Network for Object Detection from Visible and Thermal Infrared Images. International Journal of Applied Earth Observation and Geoinformation **130**, 103918 (2024) https://doi.org/10.1016/j.jag.2024.103918

[37] Tang, W., He, F., Liu, Y., Duan, Y., Si, T.: DATFuse: Infrared and Visible Image Fusion via Dual Attention Transformer. IEEE Transactions on Circuits and Systems for Video Technology **33**(7), 3159–3172 (2023) https://doi.org/10.1109/TCSVT.2023.3234340

[38] Tang, W., He, F., Liu, Y.: TCCFusion: An Infrared and Visible Image Fusion Method Based on Transformer and Cross Correlation. Pattern Recognition **137**, 109295 (2023) https://doi.org/10.1016/j.patcog.2022.109295

[39] Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12148–12166 (2023) https://doi.org/10.1109/TPAMI.2023.3283682

[40] Huang, Z., Liu, J., Fan, X., Liu, R., Zhong, W., Luo, Z.: ReCoNet: Recurrent Correction Network for Fast and Efficient Multi-Modality Image Fusion. In: European Conference on Computer Vision (ECCV 2022), pp. 524–541 (2022). https://doi.org/10.1007/978-3-031-19797-0_31

[41] Jiang, Z., Zhang, Z., Fan, X., Liu, R.: Towards All Weather and Unobstructed Multi-Spectral Image Stitching: Algorithm and Benchmark. In: Proceedings of the 30th ACM International Conference on Multimedia (2022). https://doi.org/10.1145/3503161.3547966

[42] Wang, D., Liu, J., Fan, X., Liu, R.: Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), pp. 3508–3515 (2022). https://doi.org/10.24963/ijcai.2022/487

[43] Nie, J., Sun, H., Sun, X., Ni, L., Gao, L.: Cross-Modal Feature Fusion and Interaction Strategy for CNN-Transformer-Based Object Detection in Visual and Infrared Remote Sensing Imagery. IEEE Geoscience and Remote Sensing Letters **21** (2024) https://doi.org/10.1109/LGRS.2023.3339214

[44] Xu, H., Ma, J., Yuan, J., Le, Z., Liu, W.: Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19679–19688 (2022)

[45] Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), pp. 5792–5801 (2022). https://doi.org/10.1109/CVPR52688.2022.00571

[46] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)

[47] Sun, Y., Cao, B., Zhu, P., Hu, Q.: DetFusion: A Detection-Driven Infrared and Visible Image Fusion Network. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4003–4011 (2022)

[48] Tang, L., Yuan, J., Ma, J.: Image Fusion in the Loop of High-Level Vision Tasks: A Semantic-Aware Real-Time Infrared and Visible Image Fusion Network. Information Fusion **82**, 28–42 (2022) https://doi.org/10.1016/j.inffus.2021.12.004

[49] Haghighat, M.B.A., Aghagolzadeh, A., Seyedarabi, H.: A Non-Reference Image Fusion Metric Based on Mutual Information of Image Features. Computers & Electrical Engineering **37**(5), 744–756 (2011) https://doi.org/10.1016/j.compeleceng.2011.07.012

[50] Roberts, J.W., Van Aardt, J.A., Ahmed, F.B.: Assessment of Image Fusion Procedures Using Entropy, Image Quality, and Multispectral Classification. Journal of Applied Remote Sensing **2**(1), 023522 (2008) https://doi.org/10.1117/1.2945910

[51] Piella, G., Heijmans, H.: A New Quality Metric for Image Fusion. In: Proceedings of the 2003 IEEE International Conference on Image Processing, vol. 3, p. 173 (2003). https://doi.org/10.1109/ICIP.2003.1247209

[52] Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Letters **9**(3), 81–84 (2002) https://doi.org/10.1109/97.995823

[53] Aslantas, V., Bendes, E.: A New Image Quality Metric for Image Fusion: The Sum of the Correlations of Differences. AEU – International Journal of Electronics and Communications **69**(12), 1890–1896 (2015) https://doi.org/10.1016/j.aeue.2015.09.004

[54] Chen, Y., Blum, R.S.: A New Automated Quality Assessment Algorithm for Image Fusion. Image and Vision Computing **27**(10), 1421–1432 (2009) https://doi.org/10.1016/j.imavis.2007.12.002

[55] Eskicioglu, A.M., Fisher, P.S.: Image Quality Measures and Their Performance. IEEE Transactions on Communications **43**(12), 2959–2965 (1995) https://doi.org/10.1109/26.477498

[56] Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006) https://doi.org/10.1109/TIP.2005.859378

[57] Han, Y., Cai, Y., Cao, Y., Xu, X.: A New Image Fusion Performance Metric Based on Visual Information Fidelity. Information Fusion **14**(2), 127–135 (2013) https://doi.org/10.1016/j.inffus.2011.08.002

[58] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)