Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 22, No. 2, 2026

CoNet++: Bilinear and Multi-Scale Feature Fusion for Real-Time Exposure Correction

Qihong Su

School of Electronic Information, Southwest Minzu University, Chengdu 610225, Sichuan, China

Abstract: Learnable three-dimensional lookup tables provide efficient global color transformation for exposure correction, but their pixel-value mapping is weak at spatially localized degradation. Collaborative transformation frameworks partly address this limitation by combining a lookup-table branch with a low-resolution pixel-wise branch, although simple cross-branch fusion and scale aggregation can still blur local details and misalign contextual features. We introduce CoNet++, an extension of CoTF that inserts a multi-scale feature fusion module at the encoder bottleneck and a bilinear fusion module before deep cross-branch interaction. The former aggregates block-wise local and global context with channel gating, whereas the latter jointly models spatial and channel responses for adaptive feature alignment. On the LCDP test set, CoNet++ achieves 24.1819 dB PSNR, 0.8666 SSIM, and 0.1430 LPIPS, compared with 23.9292 dB, 0.8559, and 0.1507 for the reproduced CoTF baseline. The full model contains 0.3177 million parameters and requires 0.9330 GFLOPs for the local 256-pixel profiling setting. These results show that the two modules are complementary, while the increased latency and the single-dataset evaluation define the present applicability boundary.

Keywords: Exposure correction; Image enhancement; Three-dimensional lookup table; Feature fusion; Attention.

1. Introduction

Exposure correction aims to recover visually plausible brightness, contrast, color, and detail from images captured under undesirable illumination. It is a basic component of computational photography and supports downstream perception in applications such as autonomous systems and video understanding. Classical histogram, curve-mapping, and Retinex formulations remain computationally attractive, but their hand-crafted priors can be brittle in spatially complex scenes. Learning-based methods improve representation capacity, yet dense pixel-wise transformations usually scale their computational burden with image resolution [1-4].

Learnable three-dimensional lookup tables (3D LUTs) offer a different efficiency profile. A compact network predicts image-dependent weights for basis LUTs, and trilinear interpolation then maps each RGB value with little spatial computation [5-7]. This design is well suited to high-resolution enhancement, but a conventional LUT performs a global color mapping: pixels with identical RGB values receive identical outputs even when their local contexts differ. Consequently, local contrast, texture, and exposure boundaries may remain sub-optimal. Uniformly downsampling an image to predict LUT coefficients can further discard color information from content-rich regions [5,8].

CoTF addresses these problems by collaborating an image-adaptive 3D-LUT transformation with a low-resolution pixel-wise transformation, using relation-aware modulation and adaptive sampling to combine global appearance adjustment with local context [8]. Its reported results establish a strong accuracy-efficiency balance. Our reproduced implementation nevertheless exposes two narrower engineering bottlenecks: bottleneck features are aggregated with limited cross-scale context, and the enhanced LUT output and convolutional features are fused without explicit joint spatial-channel refinement. These limitations motivate an extension rather than a replacement of the CoTF framework.

We propose CoNet++, which preserves CoTFs collaborative transformation and adaptive sampling components while introducing two targeted modules. First, multi-scale feature fusion (MSFF) aggregates block-wise local and global context at the encoder bottleneck and applies channel gating before residual reintegration. Second, a bilinear fusion module (BFM) constructs spatially reweighted features and channel-modulated features, then learns their residual contribution before subsequent cross-branch interaction. The method is intentionally lightweight and changes neither the 3D-LUT principle nor the original adaptive sampling claim.

The contributions are threefold. (1) We formulate MSFF to reduce scale-semantic mismatch in the low-resolution feature branch. (2) We formulate BFM to refine spatial and channel alignment before final transformation fusion. (3) We provide a controlled LCDP ablation using a reproduced CoTF baseline, report fidelity, perceptual quality, parameters, computation, and latency, and explicitly analyze the cases in which accuracy gains incur deployment cost.

2. Related Work

2.1. Exposure correction

Recent exposure-correction research increasingly separates the recovery of complementary image properties. CLIP-RestoreX aligns structural and perceptual features and uses frequency-domain enhancement to recover both types of information [1]. Conditional Laplacian pyramid networks process low- and high-frequency components in a guided coarse-to-fine hierarchy [2]. Decoupling-and-aggregating convolution explicitly separates contrast enhancement from detail restoration without adding inference cost after re-parameterization [3]. Luminance-aware color transforms encode exposure conditions and estimate multiple transformation functions for both underexposed and overexposed inputs [4]. These developments improve local

adaptation, but dense or multi-stage processing can remain costly for high-resolution images.

2.2. Learnable lookup tables

Image-adaptive 3D LUTs combine several basis tables according to coefficients predicted from an input image, enabling real-time global enhancement [5]. Spatial-aware LUTs and adaptive-interval LUTs improve spatial sensitivity or sample allocation [6,7], and separable formulations compress the LUT representation [9]. The remaining structural limitation is that lookup itself is primarily indexed by color rather than full local context. CoTF therefore combines LUT-based global transformation with a low-resolution pixel-wise branch [8]. CoNet++ retains that collaboration and concentrates on feature aggregation and alignment within it.

2.3. Attention and multi-scale restoration

Channel attention recalibrates feature responses using global statistics, and efficient restoration transformers model longer-range interactions with reduced computational overhead [10,11]. Our modules borrow the general principles of channel reweighting, residual learning, and efficient context aggregation, but they are specialized to the two failure points observed in the CoTF-derived implementation: bottleneck scale fusion and pre-interaction cross-branch refinement.

3. Proposed Method

3.1. Problem formulation and baseline

Let I ∈ [0,1]^{H×W×3} denote an exposed RGB image, Y its normally exposed target, and fθ the restoration network with learnable parameters θ. Paired exposure correction is posed as empirical risk minimization:

$$\hat{Y} = f_\theta(I)$$
$$\theta *= arg\ min\theta\ E_{(I,Y)\sim D}[L(f_\theta(I), Y)]. \quad (1)$$

The global branch predicts an image-adaptive 3D lookup table T. For pixel p with RGB coordinate I_p, trilinear interpolation evaluates the eight vertices surrounding its normalized color-grid cell (i,j,k). Let δr, δg, and δb be the fractional coordinates within that cell. The interpolation weights are products of the corresponding lower- or upper-vertex distances:

$$G(I_p) = \Sigma^1_{a,b,c=0}\ \omega_{abc}(I_p)\ T[i+a, j+b, k+c]$$
$$\Sigma_{a,b,c}\ \omega_{abc}(I_p) = 1. \quad (2)$$
$$\omega_{abc}(I_p) = \prod_{d\in RGB}[a_d\delta_d + (1-a_d)(1-\delta_d)]. \quad (3)$$

CoTF combines this global LUT result G(I) with a low-resolution contextual feature produced by an encoder E and decoder D. CoNet++ modifies the feature processing preceding the retained cross-branch interaction H and RGB mapping R. With Fm denoting the MSFF bottleneck output and Fb the BFM-refined decoder feature, the computational path is

$$F_m = MSFF\big(E(I)\big),\quad F_b = BFM\big(D(F_m)\big),$$
$$\hat{Y} = R(H(G(I), F_b)). \quad (4)$$

This decomposition retains the high-resolution LUT path while confining most contextual computation to compact feature maps. It also separates the roles of global color transformation, multi-scale context aggregation, and joint spatial-channel refinement.

3.2. Multi-scale feature fusion

MSFF operates at the encoder bottleneck, where spatial resolution is small and contextual aggregation is inexpensive. A 1 × 1 projection first matches the target channel dimension: X0 = P(X). For scale s ∈ {2,4}, Ps partitions X0 into non-overlapping s × s regions. Channel averaging yields a region descriptor zs, which is compressed, layer-normalized, and expanded. A softmax response and a non-negative cosine gate with learnable query qs then modulate the expanded descriptor hs:

$$z_s = Mean_c(P_s(X_0)),\quad h_s = E_s(LN(C_s(z_s))). \quad (5)$$
$$g_s = h_s \times softmax(h_s) \times max(0, cos(h_s, q_s))$$
$$C_s(X_0) = B_s(U_s(g_s W_s)). \quad (6)$$

Here Cs and Es are learned linear compression and expansion, Ws is a learned projection initialized as an identity map, Us denotes bilinear upsampling, and Bs is a 1 × 1 blending convolution. The local and global context outputs are added to both the original and projected shortcuts:

$$M = X + X_0 + C_2(X_0) + C_4(X_0). \quad (7)$$

The implementation subsequently forms a spatial gate from channel-wise average and maximum responses. Although named ChannelGate in the code, this operation produces a spatial mask shared across channels. The final residual scaling coefficient 0.1 limits the initial perturbation of the identity path:

$$A(M) = \sigma(Conv_{7\times7}([Avg_c(M); Max_c(M)]))$$
$$F_m = X + 0.1(M \times A(M)). \quad (8)$$

The two block sizes supply complementary context granularity without a full self-attention matrix. Because aggregation occurs at the bottleneck, the additional spatial cost is substantially lower than applying the same operations at the input resolution.

3.3. Bilinear fusion module

BFM refines the decoded feature F before the final cross-branch interaction. Its spatial branch applies 3 × 3 average pooling, pointwise projection, separable horizontal and vertical depthwise convolutions, and a sigmoid mask. The spatial response is

$$S(F) = F\ \times\ \sigma(PW2(DWv(DWh(PW1(AvgPool3\times 3(F)))))). \quad (9)$$

In parallel, a 1 × 1 projection produces Q, K, and V. Two lightweight operators are applied sequentially to Q and K. The first uses local depthwise filtering, normalization, activation, and a pointwise sigmoid mask; the second uses global average pooling and a channel-wise pointwise sigmoid mask:

$$[Q, K, V] = split(PW_{qkv}(F))$$
$$O_1(Z) = Z \times \sigma(PW(ReLU(BN(DW_{3\times3}(Z))))). \quad (10)$$
$$O_2(Z) = Z \times \sigma(PW(GAP(Z)))$$
$$Q' = O_2(O_1(Q))\quad K' = O_2(O_1(K)). \quad (11)$$

The enhanced query and key are first combined and depthwise filtered, and the result multiplicatively modulates V. A depthwise output projection and dropout produce the channel-refined response:

$$C(F) = Drop(DW_{3\times3}(DW_{3\times3}(Q' + K') \times V)). \quad (12)$$

The actual implementation uses fixed residual coefficients rather than normalized trainable fusion weights. Thus, the BFM output is exactly

$$F_b = F + 0.1S(F) + 0.1C(F). \quad (13)$$

The products in S(F) and C(F) make the refinement input-dependent: spatial evidence controls where features are preserved, whereas the Q/K-derived response controls which channel interactions modulate V. BFM conditions the input to

the subsequent CoTF-derived interaction block rather than replacing that block.

3.4. Training objective

All compared variants follow the same paired restoration objective and data protocol used by the local CoTF reproduction. The implemented objective sums pixel, structural-similarity, and perceptual terms:

$$L_{total} = 10L_1 + 5L_{SSIM} + 0.1L_{perc}. \quad (14)$$

For N scalar image samples, the reconstruction term and structural term are

$$L_1 = \frac{1}{N} \Sigma_n |\hat{Y}_n - Y_n|$$

$$L_{SSIM} = 1 - SSIM(\hat{Y}, Y). \quad (15)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (16)$$

The perceptual term is an L1 distance between fixed VGG-19 features. Following the configuration, the layer weights for conv1_2, conv2_2, conv3_4, conv4_4, and conv5_4 are 0.1, 0.1, 1, 1, and 1, respectively:

$$L_{perc} = \Sigma_l \alpha_l \|\varphi_l(\hat{Y}) - \varphi_l(Y)\|_1$$

$$\alpha = \{0.1, 0.1, 1, 1, 1\}. \quad (17)$$

No style loss or metric-specific test-time optimization is used. This controlled objective isolates architectural effects, although training budgets differ for the later MSFF experiments as reported in Table 3; the limitation is considered when interpreting the ablation.

3.5. Computational characteristics

For a fixed 3D-LUT grid, trilinear interpolation accesses eight vertices per pixel and therefore scales linearly with image area, O(HW). MSFF processes bottleneck features and uses block-wise descriptors instead of a dense pairwise attention matrix. For a feature tensor of size Hf × Wf × C and an effective depthwise kernel footprint K, the dominant BFM terms can be summarized as

$$C_{BFM} = O(H_f W_f C^2 + H_f W_f CK)$$

$$M_{BFM} = O(H_f W_f C). \quad (18)$$

The quadratic channel term arises from the QKV pointwise projection, while depthwise filtering is linear in C. These asymptotic expressions explain why the parameter increase remains small, but they do not imply unchanged wall-clock latency: operator scheduling, memory movement, and device-specific kernel efficiency still affect runtime, as quantified in Section 5.3.

4. Experimental Setup

4.1. Dataset and metrics

Experiments use the LCDP exposure-correction dataset under the local split of 1,415 training images and 218 test images. Performance is measured by peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS). Higher PSNR and SSIM indicate better signal fidelity and structural preservation, whereas lower LPIPS indicates better perceptual similarity [12,13]. Parameters and floating-point operations characterize model complexity; CPU latency is reported only for variants profiled under the same local 256-input setting.

4.2. Compared variants

Four controlled variants are evaluated: the reproduced CoTF baseline, baseline plus BFM, baseline plus MSFF, and the complete CoNet++ with both modules. External method values in Table 1 are quoted from the official CoTF paper under its reported LCDP settings [8]. They provide historical context but are not mixed with our locally profiled runtime because hardware and input resolution differ.

Table 1. LCDP performance context. Published values are quoted from Li et al. [8]; the final row is the local experiment and is not a hardware-matched reproduction of published runtime results.

Method	PSNR	SSIM	LPIPS	Source
RetinexNet	16.20	0.6304	0.2940	[8]
SID	21.89	0.8082	0.1781	[8]
ENC-DRBN	23.08	0.8302	0.1536	[8]
LCDPNet	23.24	0.8420	0.1368	[8]
CoTF official	23.89	0.8581	0.1035	[8]
CoNet++ local	24.1819	0.8666	0.1430	Ours

5. Results and Discussion

5.1. Quantitative ablation

Table 2 and Figure 1 show the controlled final-test ablation. BFM provides the stronger standalone improvement: relative to the baseline, it increases PSNR by 0.0917 dB and SSIM by 0.0093 while reducing LPIPS by 0.0069. MSFF alone increases PSNR by 0.0393 dB and reduces LPIPS by 0.0031, but SSIM decreases by 0.0014. The isolated MSFF result therefore supports a limited perceptual and signal-fidelity benefit rather than a universal improvement across all metrics.

Combining BFM and MSFF gives the best overall local result: 24.1819 dB PSNR, 0.8666 SSIM, and 0.1430 LPIPS. The gains over the reproduced baseline are 0.2527 dB, 0.0107, and 0.0077, respectively. The combined gain being larger than either standalone PSNR gain suggests complementarity between cross-scale context aggregation and spatial-channel feature refinement. Because only one run per setting is available, this observation should not be interpreted as statistical significance.

Table 2. Stepwise ablation on the local LCDP test set. Higher PSNR/SSIM and lower LPIPS are better.

Setting	BFM	MSFF	PSNR	SSIM	LPIPS
Baseline	-	-	23.9292	0.8559	0.1507
BFM	Yes	-	24.0209	0.8652	0.1438
MSFF	-	Yes	23.9685	0.8545	0.1476
CoNet++	Yes	Yes	24.1819	0.8666	0.1430

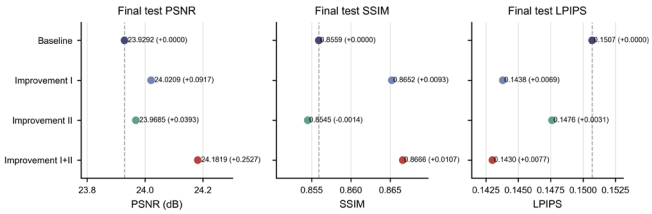

Figure 1. Final-test PSNR, SSIM, and LPIPS for the reproduced baseline and the three architectural variants. Parentheses show change relative to baseline.

5.2. Optimization behavior

The loss trajectories in Figure 2 decrease stably for all

variants, while the validation curves in Figure 3 show that the combined configuration reaches its strongest PSNR before the final training iteration. The best recorded validation point for CoNet++ is 24.1930 dB at 100,000 iterations with SSIM 0.8665. The reported final test result remains the independent comparison used for the main claim. Mini-batch loss is not an epoch-averaged uncertainty estimate, so convergence plots are treated as optimization diagnostics rather than evidence of statistical superiority.

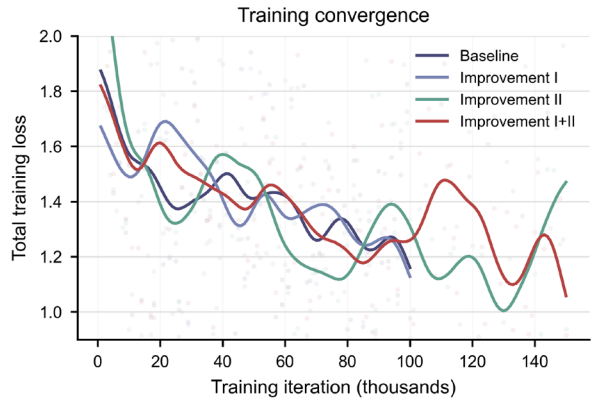

Figure 2. Smoothed mini-batch training loss curves for the compared configurations.

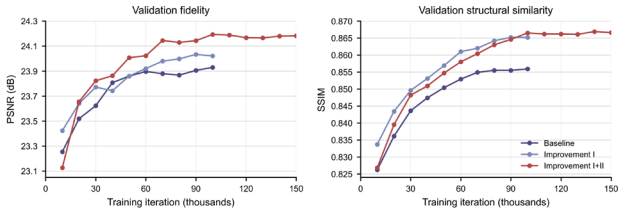

Figure 3. Validation PSNR and SSIM over training iterations. Curves are used to assess optimization behavior, not test-set significance.

5.3. Complexity and deployment trade-off

Table 3 quantifies the cost of the proposed modules. CoNet++ increases the parameter count from 0.3100 million to 0.3177 million, a relative increase of approximately 2.49%. Computation increases from 0.8004 to 0.9330 GFLOPs, and local CPU latency increases from 89.54 to 135.24 ms. BFM contributes more latency than MSFF in this profiling setting. Thus, the method preserves a compact parameter footprint but does not preserve baseline latency.

The local FLOPs and latency values cannot be directly compared with the official CoTF results reported for 1024 x 1024 images on an NVIDIA TITAN V [8]. They are valid only for relative comparison among the four locally profiled variants. For latency-sensitive deployment, BFM alone may provide a useful compromise, whereas the complete model favors restoration quality.

Table 3. Local model complexity and training budget. FLOPs and CPU latency use the same local 256-input profiling setting.

Method	Params (M)	GFLOPs	CPU ms	Train iters
Baseline	0.3100	0.8004	89.54	100k
BFM	0.3108	0.9042	114.51	100k
MSFF	0.3169	0.8292	100.23	150k
CoNet++	0.3177	0.9330	135.24	150k

5.4. Qualitative behavior and failure cases

The qualitative comparison in Figure 4 indicates that the complete model better preserves mid-frequency detail and local contrast in representative exposure-correction cases. The evidence is consistent with the intended roles of MSFF and BFM, but visual examples do not isolate causal contributions and should be read together with the ablation. Failure-case inspection further shows residual noise amplification in dark regions and incomplete recovery of texture in saturated highlights. Once sensor information is clipped, neither contextual aggregation nor LUT transformation can reconstruct the lost signal reliably.

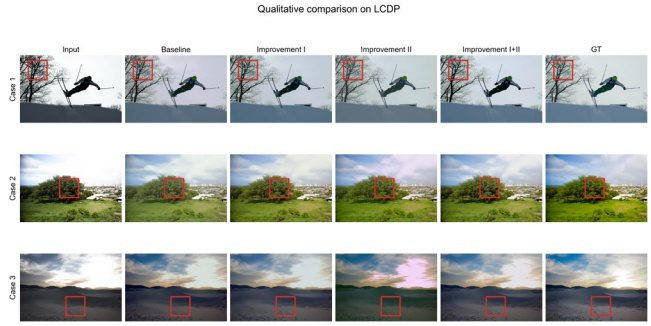

Figure 4. Representative LCDP comparison among input, reproduced baseline, single-module variants, complete CoNet++, and ground truth.

5.5. Limitations

The present evidence has four boundaries. First, the central ablation is evaluated on LCDP only; cross-dataset generalization to MSEC, SICE, and ultra-high-definition images remains unverified for CoNet++. Second, each setting is represented by a single training run, so variance and significance cannot be estimated. Third, the MSFF-containing variants use a larger training budget than the baseline and BFM-only variants, which may contribute to performance differences. Fourth, local CPU profiling does not establish mobile-NPU or GPU runtime. These limitations constrain the claim to improved local LCDP performance under the recorded implementation and training conditions.

6. Conclusion

This paper introduced CoNet++, a targeted extension of CoTF for exposure correction. MSFF aggregates local and global bottleneck context, while BFM refines spatial-channel alignment before the final collaborative interaction. On the local LCDP test set, the complete model improved PSNR from 23.9292 to 24.1819 dB and SSIM from 0.8559 to 0.8666, while reducing LPIPS from 0.1507 to 0.1430. These gains were obtained with a 2.49% parameter increase but higher computation and CPU latency. The evidence supports complementarity between the two modules on LCDP; repeated runs, hardware-matched deployment tests, and cross-dataset evaluation are required before making broader generalization or real-time claims.

Acknowledgements

The authors acknowledge the developers of the open-source CoTF implementation and the providers of the LCDP dataset. This research received no external funding.

References

[1] Huang, X., Zhang, Q., Hu, J.-F., & Zheng, W.-S. (2025). CLIP-RestoreX: Restore image structure and perception in exposure correction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 4, pp. 3760-3768). AAAI Press. https://doi.org/10.1609/aaai.v39i4.32392

[2] Huang, M., Chang, K., Qin, Q., Tang, Y., & Li, G. (2025). Conditional Laplacian pyramid networks for exposure correction. Signal Processing: Image Communication, 134, Article 117276. https://doi.org/10.1016/j.image.2025.117276

[3] Wang, Y., Peng, L., Li, L., Cao, Y., & Zha, Z.-J. (2023). Decoupling-and-aggregating for image exposure correction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 18115-18124). IEEE.

[4] Baek, J.-H., Kim, D., Choi, S.-M., Lee, H.-J., Kim, H., & Koh, Y. J. (2023). Luminance-aware color transform for multiple exposure correction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 6156-6165). IEEE.

[5] Zeng, H., Cai, J., Li, L., Cao, Z., & Zhang, L. (2022). Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4), 2058-2073. https://doi.org/10.1109/TPAMI.2020.3044140

[6] Kim, W., & Cho, N. I. (2024). Image-adaptive 3D lookup tables for real-time image enhancement with bilateral grids. In Proceedings of the European Conference on Computer Vision (ECCV). Springer. https://doi.org/10.1007/978-3-031-72967-6_6

[7] Yang, C., Jin, M., Jia, X., Xu, Y., & Chen, Y. (2022). AdaInt: Learning adaptive intervals for 3D lookup tables on real-time image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 17522-17531). IEEE.

[8] Li, Z., Zhang, F., Cao, M., Zhang, J., Shao, Y., Wang, Y., & Sang, N. (2024). Real-time exposure correction via collaborative transformations and adaptive sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2984-2994). IEEE.

[9] Yang, C., Jin, M., Xu, Y., Zhang, R., Chen, Y., & Liu, H. (2022). SepLUT: Separable image-adaptive lookup tables for real-time image enhancement. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 201-217). Springer.

[10] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7132-7141). IEEE. https://doi.org/10.1109/CVPR.2018.00745

[11] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5728-5739). IEEE. https://doi.org/10.1109/CVPR52688.2022.00564

[12] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612. https://doi.org/10.1109/TIP.2003.819861

[13] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 586-595). IEEE.

https://doi.org/10.1109/CVPR.2018.00068