Shallow yet Perceptual Decoding for Neural Image Compression through Minimal Nonlinearity

Jaekyung Ryu* and Nam Ik Cho*†
* IPAI, Seoul National University, Korea
E-mail: jaekyung0204@snu.ac.kr
† Department of ECE, INMC, Seoul National University, Korea
E-mail: nicho@snu.ac.kr

***Abstract*—Recent neural image compression (NIC) models demonstrate superior rate-distortion performance compared to traditional codecs. However, their high computational complexity on the decoder side limits their practical use in real-time and resource-constrained settings. In response, recent research concentrates on shallow decoder designs, based on the observation that learned synthesis transforms often display quasi-linear behavior. This suggests that deep nonlinear decoders are not strictly necessary for accurate image reconstruction. While these shallow designs effectively reduce complexity and maintain strong distortion performance, they often struggle to preserve perceptual quality, which is a key focus in modern compression research.**

In this paper, we argue that a shallow decoder can be sufficient for high-fidelity reconstruction, but achieving perceptual alignment requires minimal nonlinearity. We propose two improved shallow decoder architectures, named Res-Synthesis and RA-Synthesis, which incorporate lightweight residual and attention modules. These additions introduce just enough nonlinearity for perceptual refinement while keeping decoding efficient.

Our models have achieved up to a 90% reduction in decoder-side multiply-accumulate (MAC) operations compared to deep NIC decoders and consistently outperform existing shallow designs in both distortion-based metrics (such as PSNR and MS-SSIM) and perceptual metrics (such as FID and LPIPS). Furthermore, they outperform traditional codecs across all tested bit rates, with particularly notable gains in perceptual quality. These results show that carefully designed shallow decoders, enhanced with lightweight nonlinearity, provide an effective and efficient solution for perception-aware image compression.

I. Introduction

Neural image compression (NIC) emerges as a promising alternative to traditional codecs, such as JPEG [1], BPG [2], and VVC [3], offering improved rate-distortion performance through data-driven optimization. By jointly training the encoder, decoder, and entropy model, NIC systems effectively learn complex image statistics, allowing for flexible bit rate control. However, many leading NIC models rely on deep and computationally intensive decoder architectures, which pose significant challenges for real-time use on mobile or resource-constrained devices.

To address this limitation, recent work introduces shallow decoder designs that significantly lower decoder-side complexity while preserving competitive reconstruction accuracy. Notably, Yang and Mandt [4] observed that synthesis transforms in NIC often show quasi-linear behavior, meaning that linear interpolations in the latent space correspond to nearly linear transitions in the image space. Motivated by this, they proposed a fully linear decoder architecture called *JPEG-like Synthesis*, demonstrating that linear transforms alone can generate visually plausible reconstructions. This research suggests that deep nonlinear decoders are not strictly necessary for accurate image reconstruction, and that shallow architectures can be sufficient in many cases.

Although these shallow and linear designs significantly improve computational efficiency, their effectiveness is primarily evaluated using distortion-based metrics such as PSNR. However, recent advancements in perceptual compression emphasize the importance of perceptual quality, which involves capturing high-level structures, textures, and semantics. This quality is measured using learned metrics like LPIPS [5] and FID [6]. From this perspective, purely linear or overly shallow decoding transformations may be inadequate, as they lack the necessary nonlinearity to align the reconstructed images with human visual perception.

In this paper, we argue that while shallow decoder architectures are sufficient for preserving distortion fidelity, they must incorporate minimal yet meaningful nonlinearity to achieve perceptual quality. Building on the quasi-linear behavior observed in prior work, we propose two enhanced shallow decoder architectures—*Res-Synthesis* and *RA-Synthesis*—that retain computational efficiency while introducing lightweight residual and attention modules as minimal sources of nonlinearity. Our designs are tailored to strike a balance between decoding complexity and perceptual reconstruction performance, making them well-suited for real-world NIC applications.

Through extensive experiments, we have shown that our models reduce decoder MACs by up to 90% compared to deep NIC decoders, while surpassing existing shallow designs and traditional codecs in both distortion and perceptual metrics. These results emphasize the importance of including lightweight nonlinearity in shallow decoder design and set a new direction for efficient, perception-aware neural image compression.

The main contributions of this work are:

- We revisit the role of nonlinearity in NIC decoders and argue that while deep nonlinear models are not strictly necessary, purely linear or overly shallow decoders cannot achieve high perceptual quality.
- We propose two shallow decoder architectures—Res-

979-8-3315-7206-8/25/$31.00

Synthesis and RA-Synthesis—that integrate minimal residual and attention-based nonlinear modules into an otherwise shallow decoding structure.
- Our models have achieved up to 90% reduction in decoder-side MACs compared to deep NIC decoders, and consistently outperform existing shallow and linear decoders in both distortion (PSNR, MS-SSIM) and perceptual (FID, LPIPS) metrics.
- We demonstrate that shallow decoder designs with lightweight nonlinearity can surpass traditional codecs and recent linear decoding baselines across all tested bit rate ranges, thereby showing practical and perceptual viability.

II. Related Work

Traditional Image Codecs. Traditional image compression standards, such as JPEG, BPG, and VVC, are widely used due to their handcrafted transforms and mature entropy coding methods. JPEG uses the discrete cosine transform (DCT) and simple quantization, while BPG and VVC employ more advanced block-based prediction and transform coding techniques. Despite their effectiveness, these codecs struggle to adapt to varying image statistics and often perform poorly at low bit rates.

Neural Image Compression (NIC). Recent advances in neural image compression demonstrate that learned approaches can outperform traditional codecs in terms of rate-distortion performance. NIC frameworks jointly optimize the encoder, decoder, and entropy model using end-to-end training. Notable models include the mean-scale hyperprior architecture [7], [8], and later variants such as ELIC [9], which incorporate residual and attention mechanisms to enhance reconstruction quality. Yang and Mandt [4] proposed a two-layer synthesis decoder that significantly reduces decoder-side complexity while maintaining acceptable visual quality, motivating further exploration into lightweight decoder design.

Shallow Decoder Design. Shallow decoder architectures attract increasing attention as a promising approach to alleviate the high computational load of NIC models, particularly in real-time and edge device applications. These methods reduce decoder depth or simplify upsampling steps to boost efficiency, often without significantly compromising reconstruction quality. Yang and Mandt [4] introduced *two-layer synthesis* and *JPEG-like synthesis* decoders, which significantly reduce the depth and complexity of decoder networks. They empirically observed that learned synthesis transforms in NIC exhibit quasi-linear behavior, where linear interpolations in the latent space correspond to approximately linear transitions in the image space. Moreover, they showed that the decoder manifold is nearly flat, resembling classical orthogonal transforms. These observations suggest that linear or shallow decoding transforms may be sufficient for distortion-based reconstruction, and raise the question of whether deep, nonlinear decoding is strictly necessary.

Although the above-stated shallow decoders are highly efficient, they exhibit a trade-off in rate-distortion performance, especially at low bit rates. For example, their two-layer and JPEG-like synthesis methods, while efficient, perform worse than traditional codecs like BPG in PSNR and BD-rate evaluations. Additionally, their main focus is on distortion metrics, leaving perceptual quality largely unaddressed.

Importantly, shallow decoder designs allow for allocating more representational and computational resources to the encoder. Since the decoder is often the deployment bottleneck in real-time or on-device scenarios, simplifying its complexity enables the encoder to use deeper or more expressive architectures, which can enhance the overall rate-distortion performance of the model within the same computational limits. Our work builds upon this by proposing improved shallow decoders that maintain the computational advantages of shallow architectures while adding lightweight nonlinear components through residual and attention modules. This results in better perceptual quality without significantly increasing complexity, bridging the gap between efficiency and visual fidelity.

Decoder Complexity in NIC. Decoder-side complexity, including multiply-accumulate operations (MACs), parameter count, and inference latency, is a critical factor in NIC deployment. High-performance models such as ELIC [9] and EVC [10] achieve strong reconstruction quality, but this comes at the cost of deep and computationally expensive decoder paths. These models serve as important reference points for evaluating the trade-offs between complexity and performance. In contrast, our proposed decoders are designed to operate at a significantly lower computational cost, enabling practical deployment while maintaining competitive reconstruction quality.

III. Methodology

Recent studies observe that learned synthesis transforms often exhibit quasi-linear behavior in high bit rate regimes. For example, Yang and Mandt [4] showed that linear interpolations in the latent space lead to approximately linear transitions in pixel space. Duan et al. [11] further found that synthesis transforms share characteristics with classical orthogonal transforms, such as separability and localized basis-like patterns.

While these findings highlight an emerging quasi-linear structure, we suggest that this behavior is a result of model over-parameterization rather than an intentional design goal. When decoder complexity needs to be strictly constrained, relying exclusively on implicit linearity may not deliver sufficient perceptual accuracy. Therefore, we propose incorporating minimal nonlinearity to achieve a better balance between reconstruction quality and model efficiency.

A. Analysis of Nonlinear Transform

Linear decoder and its limitations. While linear synthesis transforms offer computational efficiency, they lack the expressiveness needed to reconstruct semantically faithful and perceptually aligned images, especially when paired with powerful nonlinear encoders. In the context of learned image compression, linear decoders tend to perform suboptimally in modeling the complex distribution of natural images. As perceptual quality relies heavily on high-level semantic cues, such

as object boundaries, texture continuity, and color consistency, purely linear decoders may fail to project reconstructions onto the perceptual manifold $\mathcal{M}_{\text{percep}}$, which defines the space of perceptually plausible images [5], [12]. Since linear transformations are restricted to affine subspaces, they require excessive parameterization to approximate $\mathcal{M}_{\text{percep}}$, contradicting the design goal of lightweight and efficient decoders.

Nonlinearity and perceptual alignment. Nonlinear transforms are essential for enabling neural networks to align their outputs with perceptual similarity as judged by humans. Perceptual metrics such as LPIPS and FID evaluate image realism in deep feature spaces derived from networks like VGG [13] or Inception [14], which are inherently nonlinear. Specifically, the LPIPS score is computed by measuring the ℓ_2-distance between feature activations. Since these features are nonlinear mappings of the input space, matching them requires decoders to possess nonlinear transformations.

Perceptual quality and nonlinear modeling. The importance of nonlinear synthesis is further evidenced by the success of deep convolutional architectures in perceptual tasks. Blau & Michael [15] demonstrated that deeper networks with hierarchical nonlinearities significantly outperform linear baselines in image similarity judgment tasks. Similarly, Dosovitskiy and Brox [16] showed that generating perceptually realistic images from embeddings demands multi-layer nonlinear transformations. These results underscore that perception-driven reconstruction is inherently a nonlinear problem.

Perceptual manifold. The concept of a nonlinear perceptual manifold is formalized in the work by Isola et al. [12], where the distribution of natural images in the deep feature space is shown to lie on a complex, curved manifold rather than a linear subspace. Thus, projecting latent codes z to perceptual outputs $\hat{x}$ requires nonlinear functions $f : z \mapsto \hat{x}$ such that $\hat{x} \in \mathcal{M}_{\text{percep}} \subset \mathbb{R}^{H\times W\times 3}$. Linear decoders cannot adequately traverse this manifold.

Gradient flow and training stability. Beyond just expressiveness, nonlinear components also improve gradient flow and training dynamics. Zhu et al. [17] and Ledig et al. [18] found that even shallow nonlinear modules, such as residual blocks or attention gates, significantly enhance convergence and lead to perceptual gains in image generation and restoration.

Minimal Nonlinearity. While prior work demonstrates that deep nonlinear decoders are effective in enhancing perceptual quality [5], [17], our approach is guided by a different hypothesis: nonlinearity is essential, but its quantity can be minimal. We suggest that a single nonlinear module, such as a residual block or attention gate, is sufficient to steer the reconstruction toward the perceptual manifold $\mathcal{M}_{\text{percep}}$, especially when paired with a highly expressive encoder that already encodes rich semantic structure. In this configuration, the decoder's role is not to extract high-level features from scratch, but to apply lightweight perceptual refinement over a semantically meaningful latent space.

To align with our design philosophy of minimal nonlinearity, we place a lightweight nonlinear module at the very first stage of the decoder, where the spatial resolution begins to match that of the final output. This strategic placement aims to refine features at a scale that is more perceptually meaningful, such as textures and structural boundaries, which are crucial for visual quality.

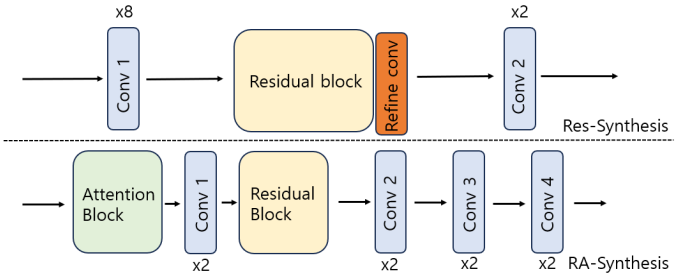

Fig. 1: Architecture of Res-Synthesis (top)/RA-Synthesis (bottom): IGDN is applied for Res-Synthesis after the first convolution. Residual blocks and two 1×1 convolutions refine the features before final upsampling. Attention modules and residual blocks are inserted between transposed convolutions in RA-Synthesis to enhance spatial modeling.

B. Proposed Synthesis Architecture

We develop two different shallow decoder architectures, *Res-Synthesis* and *RA-Synthesis*, that build on the concept of minimal nonlinearity. Both aim to improve not only perceptual quality but also distortion fidelity, delivering strong rate-distortion performance under constrained decoder complexity. Each architecture is intentionally crafted to preserve high information capacity early in the decoding process. In Res-Synthesis, this is accomplished by aggressively upsampling the latent representation in the first layer with an $8\times$ stride, enabling the network to process spatially richer features from the start. In RA-Synthesis, the attention module and residual blocks are placed at the very beginning to refine and selectively improve the latent features before spatial expansion. This design approach maximizes the effective use of latent information for both distortion fidelity and perceptual quality.

Res-Synthesis. We design *Res-Synthesis* as a shallow decoder that strikes a balance between reconstruction quality and computational efficiency. The architecture consists of two transposed convolution layers, $\texttt{conv}_1$ and $\texttt{conv}_2$, with strides $(8, 2)$, kernel sizes $(13, 5)$, and output channels (N, C_{out}), respectively. The intermediate feature channels follow a descending order, e.g., $(48, 3)$.

After the first transposed convolution $\texttt{conv}_1$, a nonlinear activation $\xi(\cdot)$ implemented as *Inverse Generalized Divisive Normalization (IGDN)* is applied. This activation is carefully chosen to introduce only minimal nonlinearity at the early decoding stage, thereby maximizing the representational richness of the initial feature map while maintaining efficiency.

The activated feature is then passed through a stack of residual blocks to improve representational capacity. The residual block adopts a bottleneck-style structure and is defined as:

$$R(\mathbf{x}) = \mathbf{x} + \text{Conv}^{(1\times1)} \circ \sigma \circ \text{Conv}^{(3\times3)} \circ \sigma \circ \text{Conv}^{(1\times1)}(\mathbf{x}) \quad (1)$$

where $\sigma(\cdot)$ denotes the ReLU activation.

The refined feature is then processed by two sequential 1×1 convolution layers, first with ReLU activation and then linear, which we denote as $\phi(\cdot)$, before being passed into the final transposed convolution $\texttt{conv}_2$. The overall transformation is given by:

$$\hat{\mathbf{x}} = \text{conv}_2\left(\phi\left(R\left(\xi(\text{conv}_1(\mathbf{z}))\right)\right)\right) \tag{2}$$

where $\hat{\mathbf{x}}$ is the reconstructed output image. This architecture achieves efficient yet expressive decoding, making it particularly well-suited for low-bit-rate scenarios with stringent decoder complexity constraints. The full structure is illustrated in Figure 1.

RA-Synthesis. We design *RA-Synthesis* by extending Res-Synthesis with additional attention and residual modules for enhanced feature modeling. The architecture consists of four transposed convolution layers $\texttt{conv}_1$, $\texttt{conv}_2$, $\texttt{conv}_3$, and $\texttt{conv}_4$, with strides $(2,2,2,2)$, kernel sizes $(5,5,5,5)$, and output channels $(192,160,128,3)$, respectively. The intermediate feature channels follow a descending order, enabling progressive refinement and upsampling.

At the start of the decoder, we use a simplified attention block inspired by [8], where the attention map is derived from the input feature through three residual blocks followed by a 1×1 convolution and a channel-wise sigmoid activation. The resulting attention map then modulates the output of a parallel trunk branch, and the final output combines the modulated feature with the original input via a skip connection. As with Res-Synthesis, early nonlinearity is minimized to preserve high information capacity in the initial stage.

The attention-weighted feature is then decoded by $\texttt{conv}_1$, followed by a stack of residual blocks that enhance representation without significantly increasing complexity. The resulting features are then sequentially refined by $\texttt{conv}_2$, $\texttt{conv}_3$, and finally $\texttt{conv}_4$, which reconstructs the output image.

To maintain decoder efficiency despite using four convolutional layers, we replace the computationally expensive IGDN activation used in Res-Synthesis with a simpler ReLU activation in all nonlinear units. This choice reduces decoding complexity while still providing enough nonlinearity for perceptual refinement.

The overall transformation is formulated as:

$$\hat{\mathbf{x}} = \texttt{conv}_4\left(\texttt{conv}_3\left(\texttt{conv}_2\left(R\left(\texttt{conv}_1\left(\text{Attn}(\mathbf{z})\right)\right)\right)\right)\right) \tag{3}$$

where $\texttt{conv}_i$ denotes the i-th transposed convolution, $R(\cdot)$ is a residual block stack, and $\text{Attn}(\cdot)$ is the attention module. This design maintains shallow depth while leveraging lightweight attention and residual connections, achieving a strong balance between perceptual quality and decoder efficiency. See Figure 1 for a schematic overview.

IV. Experimental Setup

A. Datasets and Metrics

All models are trained on the COCO dataset and tested on the entire Kodak dataset. For distortion-oriented evaluation, we use PSNR and MS-SSIM, while perceptual quality is measured using LPIPS (AlexNet-based) and FID (Inception-V3-based).

Decoder complexity is computed using an input size of $1\times16\times16\times192$. We report four complexity indicators: FLOPs, the number of parameters, decoder GMACs per pixel, and inference latency (in milliseconds). The latency is measured on a single NVIDIA RTX 2080 Ti GPU using TensorFlow.

B. Implementation Details

To ensure a fair comparison, all models share the same encoder architecture based on ELIC and adopt the mean-scale hyperprior entropy model for latent coding. The proposed *Res-Synthesis* and *RA-Synthesis* decoders are fully retrained for all experiments. For fair comparison, we also retrain the *Two-Layer Synthesis* decoder [4] under the same experimental setup. In contrast, the performances of ELIC and EVC are taken from their original publications, while their decoder complexities are re-evaluated under our unified measurement setting.

We train each model with eight different rate-distortion trade-off weights:

$$\lambda \in \{0.00125,\ 0.003,\ 0.005,\ 0.008,\ 0.0125,\ 0.02,\ 0.05,\ 0.08\}.$$

For each λ, training is conducted with a batch size of 16, an initial learning rate of 1×10^{-4}, and the Adam optimizer. Early stopping is applied if no PSNR improvement is seen on the validation set for eight consecutive evaluations, starting from step 400,000.

Validation is performed every 10,000 steps, saving the checkpoint with highest PSNR. Training uses TensorFlow with automatic logging.

V. Results and Analysis

A. Rate-Distortion Performance

We evaluate all models using both distortion-oriented metrics (PSNR, MS-SSIM) and perceptual metrics (FID, LPIPS). Figure 2 presents the rate-distortion curves for the proposed *Res-Synthesis* and *RA-Synthesis* decoders, in comparison with the *Two-layer Synthesis* baseline [4] and the traditional BPG codec.

Across all metrics and bit rate levels, the proposed decoders consistently outperform both BPG and the linear decoder baseline. In particular, *RA-Synthesis* achieves the best performance in PSNR and MS-SSIM, while also producing the lowest FID and LPIPS values, demonstrating its strong perceptual reconstruction ability. *Res-Synthesis* also offers an excellent balance between distortion and perception, achieving significantly better perceptual scores than the baseline while maintaining comparable PSNR.

It is noteworthy that the original *Two-layer Synthesis* decoder, when paired with its default encoder, performs worse than BPG in both distortion and perceptual metrics. However, after retraining the decoder in our unified setup with the ELIC encoder, a more expressive encoder architecture, it shows noticeable improvements and has outperformed traditional codecs

TABLE I: Decoder-side complexity and PSNR BD-rate saving (anchor: BPG) across models.

Decoder	FLOPs (M)	Params	GMACs/pix	Latency (ms)	PSNR BD-rate (↓)
EVC_L	90796.29	1,238,019	177.336	20.94	-
EVC_M	51785.01	643,683	101.143	15.65	-
EVC_S	13812.38	170,099	26.977	7.48	-22.56
ELIC	19285.70	4,268,419	37.667	5.72	-26.98
TwoLayerRes (Baseline)	433.91	779,691	0.847	0.54	-1.49
Res-Synthesis (Proposed)	1392.97	1,573,491	2.721	0.88	-4.16
RA-Synthesis (Proposed)	7190.12	3,090,083	14.043	2.26	-9.30

Note: We compare decoder-side computational costs and PSNR BD-rate savings (↓) relative to BPG. Complexity was computed with input size $1 \times 16 \times 16 \times 192$, and latency measured on an NVIDIA RTX 2080 Ti GPU using TensorFlow.

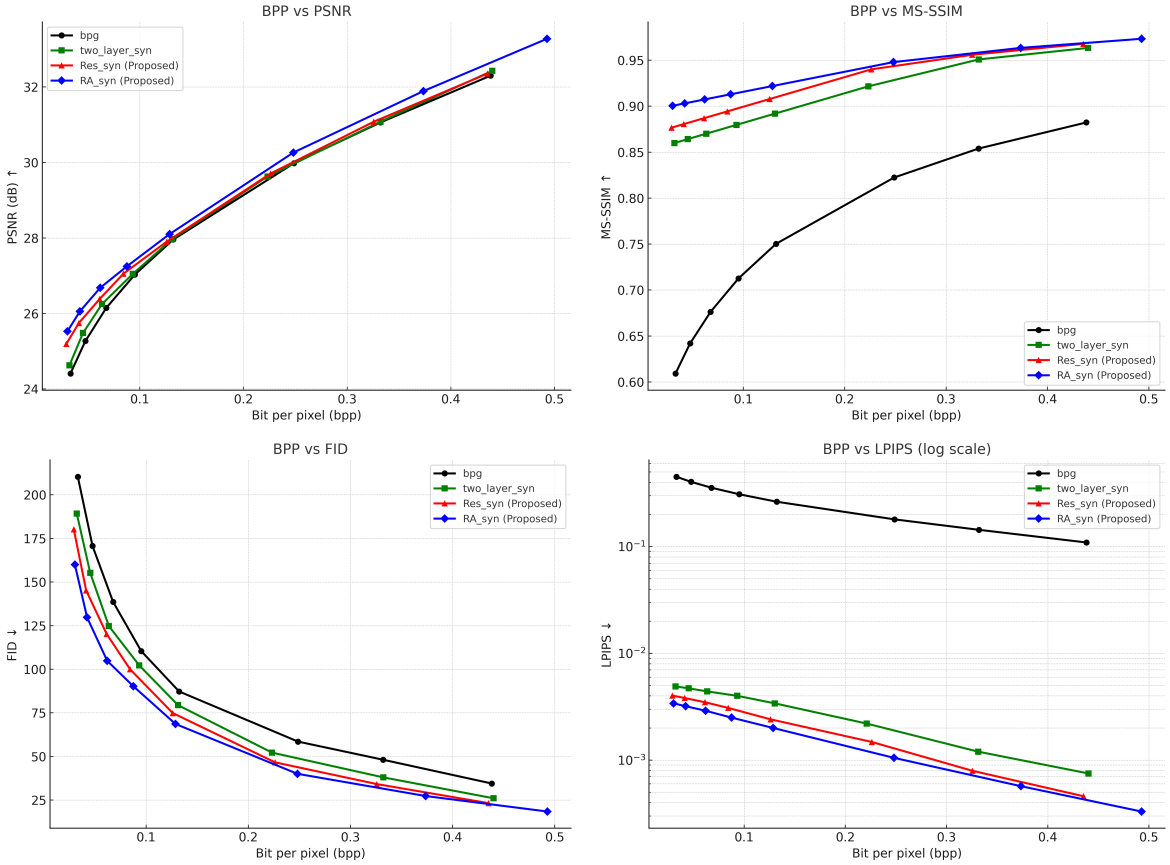

Fig. 2: Rate-distortion and perceptual quality comparisons on the Kodak dataset. The proposed Res-Synthesis and RA-Synthesis models consistently outperform both BPG and the linear decoder baseline across all metrics.

across all bit rates. This supports the idea that shallow decoders can be effective when combined with powerful encoders.

Interestingly, while shallow decoders with nonlinear transforms (such as our *Res-Synthesis* and *RA-Synthesis*) show only marginal gains or parity in PSNR compared to linear baselines, they achieve significantly better performance in perceptual metrics FID, LPIPS, and perception-correlated MS-SSIM. This substantial difference suggests that even minimal nonlinearity in the synthesis transform plays a crucial role in enhancing perceptual quality, which linear decoders and traditional codecs often fail to capture.

At lower bit rates, our proposed methods exhibit especially strong perceptual gains. *RA-Synthesis*, in particular, achieves significant improvements in FID and LPIPS compared to both the baseline and BPG, underscoring the benefit of incorporating minimal nonlinearity in the decoder design. These results highlight the importance of carefully balancing model depth and decoder-side nonlinearity to maximize both efficiency and perceptual quality.

B. Decoder Complexity Comparison

We compare the decoder-side complexity of each model in terms of FLOPs, parameter count, MACs per pixel, and inference latency. Table I summarizes the results. The proposed *Res-Synthesis* achieves more than **93%** MAC reduction com-

pared to ELIC, while *RA-Synthesis* also maintains low latency and computational cost despite using a deeper structure. Notably, both models are significantly more efficient than EVC and ELIC, while providing competitive or better perceptual quality.

These results demonstrate that our decoders balance efficiency and quality, making them suitable for low-latency deployment.

VI. Conclusion

In this paper, we revisit the decoder design in neural image compression by questioning the necessity of deep nonlinear synthesis transforms. While earlier studies suggest that shallow or even linear decoders can achieve reasonable distortion performance, especially when paired with strong encoders, we find that such designs often fall short in maintaining perceptual quality. To address this, we propose two lightweight decoder architectures, *Res-Synthesis* and *RA-Synthesis*, which preserve the efficiency of shallow decoding while adding minimal but important nonlinearity through residual and attention modules. These designs balance computational efficiency with perceptual fidelity.

Extensive experiments across various bit rate levels have shown that our models achieve competitive or better rate-distortion performance, significantly reduce the complexity of deep neural decoders, and outperform both traditional codecs and existing shallow decoders in perceptual metrics such as FID and LPIPS. Our results indicate that decoder nonlinearity does not need to be deep or complex but can be strategically minimal, enabling practical and perceptually aware neural compression models suitable for real-world deployment.

Acknowledgment

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025, and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]

References

[1] G. K. Wallace, "The jpeg still picture compression standard," in *Communications of the ACM*, vol. 34, ACM, 1991, pp. 30–44.

[2] F. Bellard, *Bpg image format*, https://bellard.org/bpg/, 2014.

[3] B. Bross, J. Chen, S. Liu, Y. Wang, and Y. Ye, "Overview of the versatile video coding (vvc) standard and its applications," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, IEEE, 2021, pp. 3736–3764.

[4] Y. Yang and S. Mandt, "Computationally-efficient neural image compression with shallow decoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 10 249–10 259.

[5] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[7] D. Minnen, J. Balle, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[8] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7939–7948. [Online]. Available: https://arxiv.org/abs/2001.01568.

[9] X. He, G. Lu, D. Wu, W. Ouyang, and D. Xu, "Elic: Efficient learned image compression with cascaded residual compression and enhanced attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5714–5723.

[10] G.-H. Wang, J. Li, B. Li, and Y. Lu, "Evc: Towards real-time neural image compression with mask decay," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2302.05071.

[11] Z. Duan, M. Lu, Z. Ma, and F. Zhu, "Opening the black box of learned image coders," in *2022 Picture Coding Symposium (PCS)*, IEEE, 2022, pp. 73–77.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[15] Y. Blau and T. Michaeli, "The perception–distortion tradeoff," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6228–6237.

[16] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *NeurIPS*, 2016.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[18] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *CVPR*, 2017.