Multi-Scale Dual-Attention Unfolding Network for Compressed Sensing Image Reconstruction

Liangjun Wang[*] ljwang0819@ujs.edu.cn

Meixin Wang 2222208036@stmail.ujs.edu.cn

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Deep Unfolding Networks have emerged as a prominent strategy in compressed sensing image reconstruction, effectively merging optimization techniques with deep learning through end-to-end training of truncated inferences. Despite their advantages, these algorithms generally require extensive iterations and parameters, potentially limited by storage capacity. Additionally, the image-level transmission at each iterative step does not optimally harness the inter-scale feature information available. To address these issues, we introduce a novel approach in this paper: the **M**ulti-**S**cale **D**ual-**A**ttention **U**nfolding **N**etwork (**MSDAUN**) for compressed sensing image reconstruction. We propose a cross-stage multi-scale deep reconstruction module **D** as an iterative process, which is composed of multiple attention submodules. These include Cross Attention Transformer(**CAT**) Modules that enhance the reconstruction with multi-channel inertia, thereby facilitating feature-level transmission and robust information exchange. Concurrently, Texture Attention Transformer(**TAT**) Modules are designed to meticulously extract salient reconstruction information, subsequently channeling it into the texture path to effectuate the precise prediction of textural regions, thereby contributing to the meticulous restoration of textural details. Our comprehensive experimental evaluation across diverse datasets confirms that MSDAUN surpasses existing state-of-the-art methods. This work presents significant potential for further advancements and applications in inverse imaging problems and optimization models.

Keywords: Compressed Sensing; Image Reconstruction; Deep Unfolding Network; Deep Learning

1. Introduction

Compressed Sensing (CS) theory posits that signals which demonstrate sparseness in some specific space can be reconstructed with high probability from a substantially reduced set of measurements, as compared to the quantity prescribed by the Shannon-Nyquist sampling theorem (Candès et al., 2006). This strategy in signal acquisition is particularly advantageous for hardware constraints, enabling the capture of visual data at sub-Nyquist rates. By exploiting the intrinsic redundancy inherent in signals, CS performs sampling and compression concurrently, thereby markedly reducing the demand for substantial transmission bandwidth and storage infrastructure. CS has been widely used in a range of practical applications, including medical imaging (Jiang et al., 2024; Hong et al., 2024), single-pixel imaging systems (Huang et al., 2024) and snapshot compression techniques (Zhang et al., 2022).

[*] Corresponding Author

Mathematically, compressed sensing reconstruction aims to deduce the original signal, denoted as $\mathbf{x} \in \mathbb{R}^N$, from a set of random linear measurements represented by the vector $\mathbf{y} = \mathbf{\Phi}\mathbf{x} \in \mathbb{R}^M$. This process involves a matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$, which encapsulates the linear random projection. The notation $M \ll N$ signifies that the problem's inverse is generally ill-conditioned due to the disparity in the dimensions of M and N. The concept of a compression ratio is pivotal in compressed sensing and is expressed as the ratio $\frac{M}{N}$.

Traditional Compressed Sensing method (Kim et al., 2010) is designed to reliably reconstruct the original image $\mathbf{x}$, by solving an optimization problem that leverages the given linear measurements $\mathbf{y}$:

$$\arg\min_{\mathbf{x}} \ \frac{1}{2} \left\| \mathbf{\Phi}\mathbf{x} - \mathbf{y} \right\|_2^2 + \lambda \mathcal{R}(\mathbf{x}), \tag{1}$$

where $\frac{1}{2} \left\| \mathbf{\Phi}\mathbf{x} - \mathbf{y} \right\|_2^2$ is instrumental in modeling the potential for signal degradation, capturing the closeness of the reconstructed signal to the observed measurements. $\lambda\mathcal{R}(\mathbf{x})$ refers to the prior term, which includes a regularization parameter λ to enforce certain structural assumptions on the solution. In conventional model-based CS methods, the prior term is often linked to sparsifying operators that are related to predefined transform domains, such as the Discrete Cosine Transform (DCT) (Zhao et al., 2014) or wavelet (Zhao et al., 2016), which promote sparsity. A variety of iterative optimization techniques, such as the Iterative Shrinkage-Thresholding Algorithm (ISTA) (Zhang and Ghanem, 2018; You et al., 2021a) and Alternating Direction Method of Multipliers (ADMM) (Afonso et al., 2010), have been introduced to address the reconstruction problem. Although these methods are known for their robust convergence properties and theoretical backing, they are frequently limited by their high computational requirements and lack of flexibility. The advent of deep learning has led to the adoption of Convolutional Neural Networks (CNNs) in CS, with neural network architectures largely falling into two categories: deep non-unfolding networks (DNUNs) and deep unfolding networks (DUNs). DNUNs are designed to directly learn the inverse mapping from the CS measurement domain to the original signal domain (Iliadis et al., 2018), often employing a supervised learning paradigm. DUNs integrate networks with optimization algorithms to train an unrolled inference through end-to-end optimization of a loss function (Zhang and Ghanem, 2018; Zhang et al., 2020a; You et al., 2021b; Song et al., 2021). This latter approach has emerged as the predominant method for CS, offering a more flexible and iterative framework for signal reconstruction.

The emergence of DUNs represents a pivotal advancement in CS, enhancing interpretability by seamlessly integrating optimization with end-to-end truncated inference training (Zhang et al., 2020a,b). These networks have rapidly ascended to become the predominant methodology within the field. Nonetheless, DUNs encounter challenges stemming from their substantial computational requirements, typified by numerous iterations and a vast array of parameters (Song et al., 2021, 2023). These factors can lead to limitations imposed by storage limitations, consequently impacting their performance. Moreover, the transmission of image-level details during iterative processes fails to fully capture and apply the rich inter-scale feature information present in the data.

To address the challenges outlined, we introduce a novel **M**ulti-**S**cale **D**ual-**A**ttention **U**nfolding **N**etwork (**MSDAUN**) specifically designed for image CS, as shown in Fig. 1. Our approach centers on a cross-stage multi-scale deep reconstruction module, denoted as

D, which employs strided convolutions (SConv$(\cdot)$) and transposed convolutions (TConv$(\cdot)$) to construct a three-scale **W**-shaped architecture for effective scale transformation. This module ensures seamless integration of deep features from previous stages with those of the current stage, thereby preserving essential information across iterations of DUNs. The integration process is driven by **C**ross **A**ttention **T**ransformer (**CAT**) modules and **T**exture **A**ttention **T**ransformer (**TAT**) modules. The CAT module is designed to optimize the flow of information within the feature space. The TAT module enhances the precision of texture region estimation by the Texture Attention(TA) module, focusing on salient texture areas and effectively restoring intricate details. Furthermore, we have innovated the Gradient Descent Block (GDB) module by incorporating generalized physical operators, significantly enhancing reconstruction capabilities of the network. The main contributions are summarized as follows:

- We introduce a novel deep unfolding network (MSDAUN) for CS image reconstruction. At the heart of MSDAUN lies a series of multiple attention modules. The proposed TAT module enhances the transfer of textural details from reconstruction data to the dedicated texture path, thereby allowing the system to concentrate on the accurate recovery of these fine details.

- We introduce a cross-stage multi-scale deep reconstruction module, denoted as **D**, which is characterized by its generalized multi-scale perception and fully activated physical injection. This design effectively reduces the need for extensive iterations and a large number of parameters.

- Integrating the cross-stage multi-scale descent module with texture attention, our proposed MSDAUN has yielded remarkable outcomes, especially at extremely low CS ratios, as demonstrated by extensive experimentation.

2. Related Work

2.1. Deep Unfolding Network

Deep Unfolding Networks (DUNs) represent a class of neural network architectures that have been advanced for a diverse array of image processing tasks, including but not limited to image demosaicing (Kokkinos and Lefkimmiatis, 2018), image fusion (Zhao et al., 2021), and image denoising (Lefkimmiatis, 2017). The foundational principle of DUNs is the transformation of conventional iterative optimization techniques into a sequence of trainable recurrent units, designed to tackle a spectrum of inverse imaging problems. This paradigm is often articulated as a bi-level optimization framework, underpinned by a dataset comprising paired observations $\{(\mathbf{y}_j, \mathbf{x}_j)\}_{j=1}^{N_a}$ and characterized by the totality of training instances denoted by N_a, as follows:

$$
\begin{aligned}
&\min_{\mathbf{\Theta}} \sum_{j=1}^{N_a} \mathcal{L}(\hat{\mathbf{x}}_j, \mathbf{x}_j), \\
&\mathbf{s.t.}\ \hat{\mathbf{x}}_j = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| \mathbf{\Phi}\mathbf{x} - \mathbf{y}_j \right\|_2^2 + \lambda \mathcal{R}(\mathbf{x}),
\end{aligned}
\tag{2}
$$

where $\mathbf{\Theta}$ represents the learnable parameters and $\mathbf{\Phi}$ denotes the measurement matrix. $\mathcal{R}(\cdot)$ represents a general nonlinear transformation function.

In the domain of CS, DUN methodologies frequently amalgamate sophisticated CNN denoising components with a spectrum of optimization strategies. This integration spans techniques such as Proximal Gradient Descent (PGD) (Chen and Zhang, 2022; You et al., 2021b; Zhang and Ghanem, 2018), Approximate Message Passing (AMP) (Zhang et al., 2020b), and the Inertial Proximal Non-Convex Optimization (iPiano) (Su and Lian, 2020), each contributing to a unique set of optimization heuristics within DUN frameworks. Nonetheless, these emerging solutions exhibit a paucity of flexibility in managing channel information and are encumbered by the complexity of the models they propose. In addition, the above methods have made substantial progress in compressed sensing image reconstruction, but there are still challenges in accurately preserving texture nuances, especially at low measurement rates.

2.2. Vision Transformer

Vision Transformers (Vaswani et al., 2017) have become a powerful tool in the realm of computer vision and image processing, originally gaining prominence in natural language processing and later proving its efficacy in complex visual tasks. The self-attention mechanism (Dosovitskiy et al., 2021), a cornerstone of the Vision Transformer, excels at capturing long-range pixel dependencies by aggregating features based on their similarities, a crucial concept in the recent progress of the field. Cai et al. (2022) proposed the DAUF, which utilizes the DUN structure for spectral compressed imaging and employs self-attention to construct robust Transformer blocks. Similarly, Shen et al. (2022) developed a Transformer architecture tailored for CS, based on the ISTA method. A common challenge in DUNs is that the input and output at each iteration are inherently images, which can limit information transfer and representational capacity. To overcome these limitations, our approach integrates Transformers with DUNs to create an efficient framework for CS image reconstruction.

3. Proposed Method

3.1. Overall Architecture

Sun et al. (2020) introduced a dual-path attention network designed to effectively disentangle structural and textural features within images. This network reconstructs the image by summing the outputs from two distinct paths: one capturing structural information and the other, textural details. This approach allows for a more flexible representation of image content. Nonetheless, the complexity of the model, exemplified by a large number of parameters, engendered formidable challenges. The reliance on meticulous tuning for optimal learning performance complicates both the training process and theoretical analysis.

Our approach integrates the proximal gradient descent (PGD) algorithm (Combettes and Wajs, 2005), where we meticulously balance the gradient descent component with inertial forces to optimize the performance of the model. To address the issue of information loss during image transmission, we have introduced a Cross Attention Transformer(CAT) module. This module is specifically designed to enhance the interaction of information

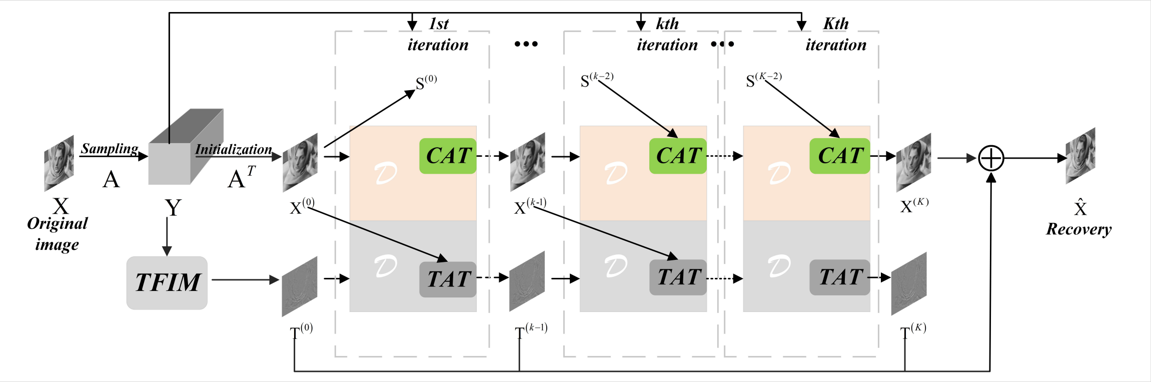

Figure 1: Architecture of our MSDAUN, which consists of K iterations. $\mathbf{X}$ denotes the full-sampled image for training, $\mathbf{Y}$ is the under-sampled data and $\mathbf{X}^{(0)}$ denotes the initialization, $\mathbf{T}^{(0)}$ denotes the initialization of the texture path from the Texture Feature Integration Module (TFIM). $\mathbf{D}$ is the cross-stage multi-scale unfolding network that is the kth iterative process. The features $\mathbf{X}^{(k-1)}$ and $\mathbf{S}^{(k-2)}$ are the inputs of our Cross Attention Transformer (CAT) module, the features $\mathbf{X}^{(k-1)}$ and $\mathbf{T}^{(k-1)}$ are the inputs of our Texture Attention Transformer (TAT) module and $\hat{\mathbf{x}}$ is the recovered result gotten from the sum of the output $\mathbf{X}^{(K)}$ and $\mathbf{T}^{(K)}$in the Kth iteration.

across different images. Furthermore, to improve the capture of textural details, we have developed a Texture Attention Transformer (TAT) module, which links the reconstruction information with the textural path information.

To counteract the performance decline and saturation that can result from merely increasing the unfolding iterations, network depth, or feature channels in the DUN, we have designed a novel multi-scale module, referred to as module $\mathbf{D}$. A comprehensive description of module D is provided in Sec. 3.2.

In the k-th iteration of our MSDAUN, CAT and TAT modules are applied in a manner that can be mathematically expressed as ($k \in \{1, 2, \cdots, K\}$):

$$\mathbf{X}^{(k)} = \mathcal{D}(\mathcal{H}_{\text{CAT}}(\mathbf{X}^{(k-1)}, \mathbf{S}^{(k-2)})), \tag{3}$$

$$\mathbf{T}^{(k)} = \mathcal{D}(\mathcal{H}_{\text{TAT}}(\mathbf{T}^{(k-1)}, \mathbf{S}_{\text{T}}^{(k-1)})), \tag{4}$$

where $\mathbf{X}^{(k)}$, $\mathbf{T}^{(k)} \in \mathbb{R}^{\frac{H}{r} \times \frac{W}{r} \times r^{(2)}C}$ are the outputs in the feature domain, and $\mathbf{S}^{(k-2)} \in \mathbb{R}^{\frac{H}{r} \times \frac{W}{r} \times r^{(2)}(C-1)}$ is obtained by clipping latter $C-1$ channels from $\mathbf{X}^{(k-2)}$. For the first iteration, the input $\mathbf{X}^{(0)}$ is generated by a 3×3 convolution ($\text{Conv}_0\,(\cdot)$) on the initialization $\mathbf{x}^{(0)}$. And for the first iteration of the texture path, the initial input $\mathbf{T}^{(0)}$ is provided by the output initialization of the Texture Feature Integration Module(TFIM). The TFIM begins by employing a 3×3 convolution to extract preliminary features. Subsequently, it utilizes texture residual blocks, which encompass five sub-layers. Each sub-layer consists of a block that includes two convolutions. Then, residual dense block(RDB) blocks (Zhang et al., 2018) are applied, which contain multiple stacked residual units designed to capture local features within the image. Each RDB block possesses its own capability for local feature extraction,

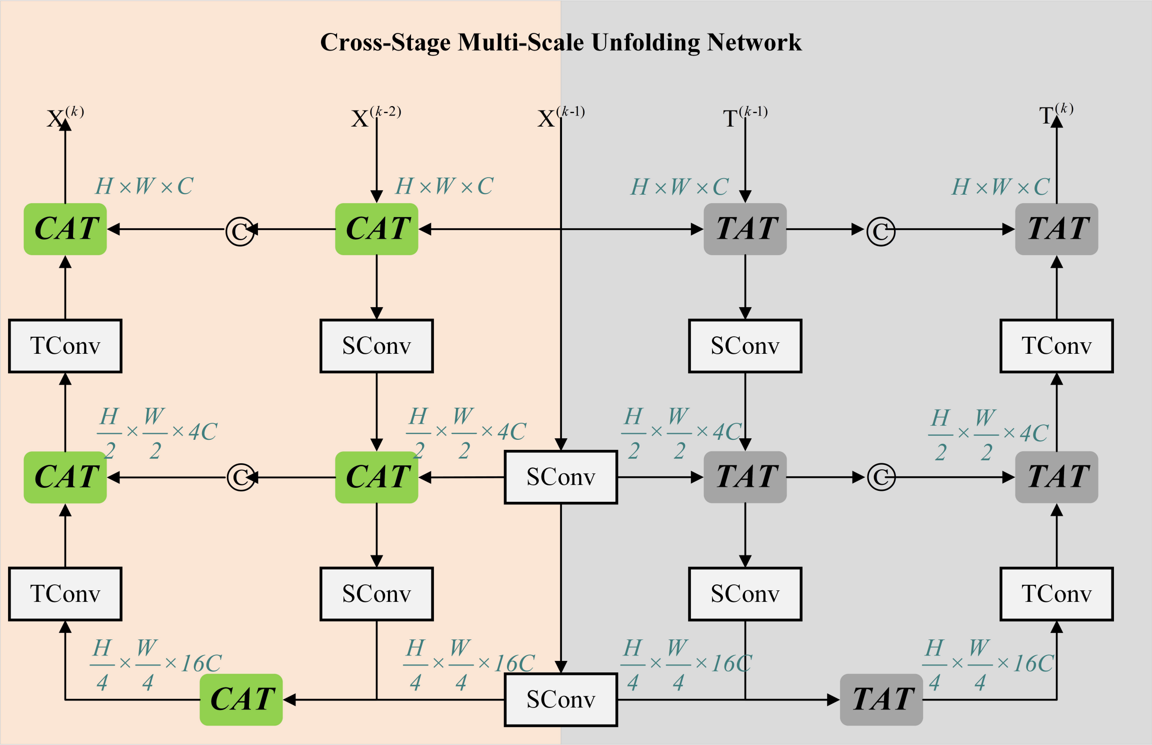
Cross-Stage Multi-Scale Unfolding Network
$X^{(k)}$
$X^{(k-2)}$
$X^{(k-1)}$
$T^{(k-1)}$
$T^{(k)}$
$H \times W \times C$
CAT
TAT
TConv
SConv
$\frac{H}{2} \times \frac{W}{2} \times 4C$
$\frac{H}{4} \times \frac{W}{4} \times 16C$
©

Figure 2: Architecture of our cross-stage multi-scale unfolding network **D**, which consists of a three-scale **W**-shaped backbone, in which the last two scales are equipped with skip connections. **D** is built by multiple CATs and TATs. The strided convolution(SConv $(\cdot)$) and the transposed convolution(TConv $(\cdot)$) are adopted as the downscaling and upscaling operators.

which aids in capturing the texture information of the image. Afterwards, SubpixelConv $(\cdot)$ (Shi et al., 2016) is utilized for upsampling, increasing the dimensions of the feature map, which facilitates the extraction of higher-level texture features. Ultimately, the recovered result $\hat{\mathbf{x}}$ is obtained by splitting the first channel from the sum of $\mathbf{X}^{(K)}$ and $\mathbf{T}^{(K)}$, as shown in Fig. 1.

3.2. Cross-Stage Multi-Scale Unfolding Network

DUNs adopt cross-stage feature fusion as their foundational principle, which decomposes the image reconstruction process into multiple manageable stages to progressively refine the image. We apply gradient descent projection in the image domain and complement it with deep denoising techniques in the feature domain during each stage. However, enhancing the recovery capability is not as simple as just increasing the number of stages, or expanding the feature capacity. We have explored the strategy of augmenting these parameters in previous iterations of DUNs (Sun et al., 2016; Zhang and Ghanem, 2018; Zhang et al., 2020a; Song et al., 2021), but this often leads to a performance plateau and a significant extension of inference time. To avoid such undesirable outcomes that negate the benefits of the tech-

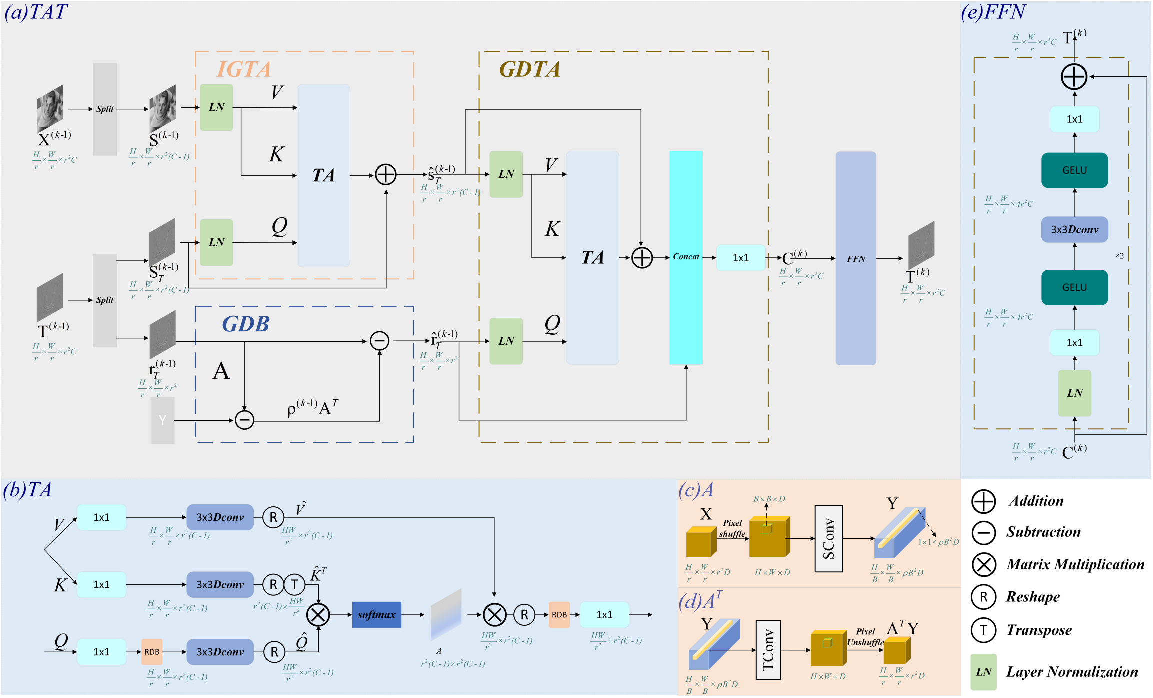

Figure 3: The architecture of the Texture Attention Transformer (TAT) module. (a) The TAT module consists of an IGTA block, a GDB block, a GDTA block and a Feed-Forward Network (FFN) sub-module. (b) The Texture Attention (TA) block, which is the basic component of two attention blocks. (c) The physical forward operator $\mathbf{A}$ by PixelShuffle with a $\mathbf{B}\times\mathbf{B}$ SConv $(\cdot)$. (d) The physical transposed forward operator $\mathbf{A}^\top$ by a $\mathbf{B}\times\mathbf{B}$ TConv $(\cdot)$ with PixelUnshuffle and $\mathbf{A}$. (e) FFN sub-module is composed of two sets of LN and Feed-Forward Block (FFB).

nology, we have developed a more efficient network architecture. Our approach simplifies network design by utilizing a set of elementary modules and generalized operations, thereby enhancing its overall efficiency and effectiveness.

The structure $\mathbf{D}$, as illustrated in Fig. 2, employs a three-layer $\mathbf{W}$-shaped architecture that integrates multiple CAT and TAT modules, which are detailed in Sec. 3.3. This design notably includes skip connections in the last two tiers to facilitate information flow. The feature channel counts are strategically set to $\mathbf{D}$, $\mathbf{4D}$, and $\mathbf{16D}$ for the first, second, and third tiers, respectively, while spatial scales are designated as $\times1$, $\times2$, and $\times4$ to maintain consistent capacity across tiers. Each layer is composed of two groups and employs 2×2 strided convolution(SConv $(\cdot)$) and 2×2 transposed convolution(TConv $(\cdot)$) as the primary scale transformation operators to manage size adjustments. After passing through the initial SConv $(\cdot)$ layer, the input $\mathbf{X}^{(k-1)}\in\mathbb{R}^{H\times W\times C}$ is transformed into $\mathbf{X}^{(k-1)}\in\mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$, and after the second SConv $(\cdot)$ layer, it becomes $\mathbf{X}^{(k-1)}\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 16C}$.

3.3. Convolutional Dual-Attention

Our dual-attention is composed of two integral components: Texture Attention Transformer (TAT) and Cross Attention Transformer (CAT). TAT is specifically tailored with the Inertia-Guided Texture Attention (IGTA) module and the Gradient Descent Texture Attention (GDTA) at its core, which will be explicated in detail within this section, as shown in Fig. 3. The architecture of the CAT, is similar to our TAT. The Feed-Forward Network (FFN) submodule, described extensively in Fig. 3 (e), is intricately designed with two Layer Normalizations (LN) and Feed-Forward Blocks (FFB), integrated with global skip connections. The architecture of the FFB is similar to Song et al. (2023).

Our TAT is shown in Fig. 3 (a). To effectively capture both local and global information when processing input images, we split $\mathbf{T}^{(k-1)} \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times r^2C}$ into two parts, which include $\mathbf{r}_T^{(k-1)} \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times r^2}$ and $\mathbf{S}_T^{(k-1)} \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times r^2(C-1)}$. Here, $\mathbf{r}_T^{(k-1)}$ is obtained by separating the first channel from $\mathbf{T}^{(k-1)}$, and $\mathbf{S}_T^{(k-1)}$ is the remaining $(\mathbf{C}-\mathbf{1})$ channels cropped from $\mathbf{T}^{(k-1)}$. They are then used as inputs for the Gradient Descent Block (GDB) and the IGTA module respectively. Additionally, our IGTA module also takes $\mathbf{S}^{(k-1)}$ as input, fully leveraging the inertial information of multiple channels. Therefore, our designed TAT and CAT sub-module can be formulated as follows:

$$\mathbf{C}^{(k)} = \mathcal{H}_{\mathrm{GDTA}}(\mathcal{H}_{\mathrm{GDB}}(\mathbf{r}_T^{(k-1)}, \mathbf{Y}), \mathcal{H}_{\mathrm{IGTA}}(\mathbf{S}^{(k-1)}, \mathbf{S}_T^{(k-1)})). \tag{5}$$

Among IGTA and GDTA modules, Texture Attention (TA) plays a crucial role as a fundamental building block. In the following part of Sec. 3.3, we first introduce the Texture Attention block (TA), and then proceed to describe the IGTA and GDTA blocks respectively.

Texture Attention. TA block is meticulously designed to enhance the operational performance of the network, thereby refining its computational efficiency and performance metrics,as shown in Fig. 3 (b). The input $\mathbf{Q}$ comes from a different component than $\mathbf{V}$ and $\mathbf{K}$. They are first embedded by a 1×1 convolution ($\mathrm{Conv}_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\cdot)$) to obtain feature with the size being $\frac{H}{r}\times\frac{W}{r}\times r^2(C-1)$. Furthermore, We employ a RDB block to capture textural information in the image. Then a 3×3 depth-wise convolution ($\mathrm{Dconv}_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\cdot)$) is used to encode channel-wise spatial context. Finally, a reshape operation ($\mathrm{R}(\cdot)$) reformulates $\mathbf{V}$, $\mathbf{K}$, and $\mathbf{Q}$ into tokens $\{\hat{\mathbf{V}}, \hat{\mathbf{K}}, \hat{\mathbf{Q}}\} \in \mathbb{R}^{\frac{HW}{r^2}\times r^2(C-1)}$. Therefore, this process can be defined as the following function:

$$\begin{cases} \hat{\mathbf{V}} = \mathrm{R}(\mathrm{Dconv}_{\mathbf{V}}(\mathrm{Conv}_{\mathbf{V}}(\mathbf{V}))), \\ \hat{\mathbf{K}} = \mathrm{R}(\mathrm{Dconv}_{\mathbf{K}}(\mathrm{Conv}_{\mathbf{K}}(\mathbf{K}))), \\ \hat{\mathbf{Q}} = \mathrm{R}(\mathrm{Dconv}_{\mathbf{Q}}(\mathrm{RDB}(\mathrm{Conv}_{\mathbf{Q}}(\mathbf{Q})))). \end{cases} \tag{6}$$

Next, the **Softmax** function is utilized to reweight the matrix multiplication $\hat{\mathbf{K}}^\top\hat{\mathbf{Q}}$, thereby generating a transposed attention map $\mathbf{A} \in \mathbb{R}^{r^2(C-1)\times r^2(C-1)}$, yielding

$$\mathbf{A} = \mathrm{Softmax}(\hat{\mathbf{K}}^\top\hat{\mathbf{Q}}), \tag{7}$$

where $\hat{\mathbf{K}}^\top$ denotes the transposed matrix of $\hat{\mathbf{K}}$. The aggregation result is calculated as $\hat{\mathbf{V}}\mathbf{A}$, which is reshaped into the features of size $\mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times r^2(C-1)}$. Afterward, a layer of

RDB module is engaged for the acquisition of textural features. Finally, we apply a 1×1 convolution $\mathrm{Conv}_{\mathbf{A}}(\cdot)$ to enhance the feature extraction. Overall, the Texture Attention block is defined as:

$$\mathcal{G}_{\mathrm{TA}}(\mathbf{V},\mathbf{K},\mathbf{Q}) = \mathrm{Conv}_{\mathbf{A}}(\mathrm{RDB}(\mathrm{R}(\hat{\mathbf{V}}\mathbf{A}))). \tag{8}$$

Texture Attention block helps to extract useful information via channel-wise similarity with low computational cost.

Inertia-Guided Texture Attention. To enhance the information exchange within inertial terms, we have integrated a multi-channel inertial component and have put forward an Inertia-Guided Texture Attention (IGTA) block. This block merges reconstruction information with texture path information, which aids in the efficient calculation of texture attention and enables the capture of a greater array of textural nuances. Comprising an LN function and a TA block, the IGTA module is illustrated in Fig. 3(a). Specifically, we set the $(k-1)$th iteration output $\mathbf{S}^{(k-1)}$ as *value* ($\mathbf{V}^{(k)}_{\mathrm{IGTA}}$) and *key* ($\mathbf{K}^{(k)}_{\mathrm{IGTA}}$), and we set the $(k-1)$th iteration output of texture path $\mathbf{S}_T^{(k-1)}$ as *query* ($\mathbf{Q}^{(k)}_{\mathrm{IGTA}}$), pass through TA block after normalization by LN function, so $\hat{\mathbf{S}}_T^{(k-1)} = \mathcal{H}_{\mathrm{IGTA}}(\mathbf{S}^{(k-1)}, \mathbf{S}_T^{(k-1)})$ as:

$$\begin{aligned}
&\mathbf{V}^{(k)}_{\mathrm{IGTA}},\ \mathbf{K}^{(k)}_{\mathrm{IGTA}},\ \mathbf{Q}^{(k)}_{\mathrm{IGTA}} = \mathrm{LN}(\mathbf{S}^{(k-1)}),\ \mathrm{LN}(\mathbf{S}^{(k-1)}),\ \mathrm{LN}(\mathbf{S}_T^{(k-1)}),\\
&\hat{\mathbf{S}}^{(k-1)} = \mathcal{G}_{\mathrm{TA}}(\mathbf{V}^{(k)}_{\mathrm{IGTA}}, \mathbf{K}^{(k)}_{\mathrm{IGTA}}, \mathbf{Q}^{(k)}_{\mathrm{IGTA}}) + \mathbf{S}_T^{(k-1)}.
\end{aligned} \tag{9}$$

Gradient Descent Texture Attention. Similar to the IGTA block, the Gradient Descent Texture Attention (GDTA) block captures rich texture feature information based on channel-wise similarity. Specifically, given $\mathbf{T}^{(k-1)}$, the input of the gradient descent term is gotten by its first channel (*i.e.*, $\mathbf{r}_T^{(k-1)}$). So, the calculation of the term has the following expression:

$$\hat{\mathbf{r}}_T^{(k-1)} = \mathbf{r}_T^{(k-1)} - \rho^{(k-1)}\mathbf{A}^\top(\mathbf{A}\mathbf{r}_T^{(k-1)} - \mathbf{Y}). \tag{10}$$

where $\rho^{(k-1)}$ represents the learnable stride. We implement the physical forward operator and its transpose, we use PixelShuffle with a $\mathbf{B}\times\mathbf{B}$ SConv$(\cdot)$, and a $\mathbf{B}\times\mathbf{B}$ TConv$(\cdot)$ with PixelUnshuffle and share all Conv$(\cdot)$ weights with the sampling matrix $\mathbf{A}$, as shown in Fig. 3(c) and (d).

Next, $\hat{\mathbf{r}}_T^{(k-1)}$ and the IGTA output $\hat{\mathbf{S}}_T^{(k-1)}$ pass through the LayerNorm function and TA block, Finally, $\hat{\mathbf{r}}_T^{(k-1)}$ is concatenated, reshaped to match the original channel dimensions and mixed with a 1×1 convolution ($\mathrm{Conv}_{\mathbf{O}}(\cdot)$):

$$\begin{aligned}
&\mathbf{V}^{(k)}_{\mathrm{GDTA}},\ \mathbf{K}^{(k)}_{\mathrm{GDTA}},\ \mathbf{Q}^{(k)}_{\mathrm{GDTA}} = \mathrm{LN}(\hat{\mathbf{S}}_T^{(k-1)}), \mathrm{LN}(\hat{\mathbf{S}}_T^{(k-1)}), \mathrm{LN}(\hat{\mathbf{r}}_T^{(k-1)}),\\
&\mathbf{C}^{(k)} = \mathrm{Conv}_{\mathbf{O}}(\mathrm{Concat}(\mathcal{G}_{\mathrm{TA}}(\mathbf{V}^{(k)}_{\mathrm{GDTA}}, \mathbf{K}^{(k)}_{\mathrm{GDTA}}, \mathbf{Q}^{(k)}_{\mathrm{GDTA}}) + \hat{\mathbf{S}}_T^{(k-1)}, \hat{\mathbf{r}}_T^{(k-1)})).
\end{aligned} \tag{11}$$

3.4. Loss Function

To obtain the train data pairs $\{(\mathbf{y}_j, \mathbf{x}_j)\}_{j=1}^{N_a}$ for the MSDAUN network, compressed measurements are acquired using fully sampled images $\{\mathbf{x}_j\}_{j=1}^{N_a}$ and sampling pattern $\mathbf{A}$, where $\mathbf{A}$ is used in place of $\mathbf{\Phi}$ in $\mathbf{y}_j = \mathbf{\Phi}\mathbf{x}_j$. Specifically, our MSDAUN model takes $\mathbf{y}_j$ as the input

Table 1: Average PSNR(dB)/SSIM performance comparisons of recent deep network-based CS methods on Set11 (Kulkarni et al., 2016b) and CBSD68 (Martin et al., 2001) dataset with different CS ratios. The best and second-best results are highlighted in red and blue colors, respectively.

Dataset	Set11						CBSD68						Times(ms)	#Param.
CS Ratio	1%	4%	10%	25%	50%	Average	1%	4%	10%	25%	50%	Average	/GFLOPs	(M)
ReconNet (CVPR 2016)	17.43 /0.4017	20.93 /0.5897	24.38 /0.7301	28.44 /0.8531	32.25 /0.9177	24.69 /0.6985	18.27 /0.4007	21.66 /0.5210	24.15 /0.6715	26.04 /0.7833	29.86 /0.8951	24.00 /0.6543	2.69 /1.31	0.23
ISTA-Net$^+$ (CVPR 2018)	17.34 /0.4131	21.31 /0.6240	26.58 /0.8066	32.48 /0.9242	38.06 /0.9706	27.15 /0.7477	19.14 /0.4158	22.17 /0.5486	25.32 /0.7022	29.36 /0.8525	34.04 /0.9424	26.01 /0.6923	5.65 /35.17	0.47
DPA-Net (TIP 2020)	18.05 /0.5011	23.50 /0.7205	27.66 /0.8530	32.38 /0.9311	36.80 /0.9685	27.68 /0.7948	20.25 /0.4267	23.50 /0.6096	25.47 /0.7372	29.01 /0.8595	32.55 /0.9386	26.16 /0.7143	36.49 /563.27	9.31
AMP-Net (TIP 2020)	20.55 /0.5638	25.14 /0.7701	29.40 /0.8779	34.63 /0.9481	40.34 /0.9804	30.01 /0.8281	22.18 /0.5207	25.47 /0.6534	27.79 /0.7853	31.37 /0.8749	36.59 /0.9620	28.68 /0.7593	27.36 /47.93	0.86
OPINE-Net$^+$ (TIP 2020)	20.15 /0.5340	25.69 0.7920	29.81 /0.8884	34.86 /0.9509	40.17 /0.9797	30.14 /0.8290	22.11 /0.5140	25.20 /0.6825	27.82 /0.8045	31.51 /0.9061	36.35 /0.9660	28.60 /0.7746	17.32 /36.29	0.62
MADUN (ACMMM 2021)	20.28 /0.5572	25.71 /0.8042	30.20 /0.9016	35.76 /0.9616	41.00 /0.9837	30.59 /0.8417	22.28 /0.5247	25.36 /0.6985	28.18 /0.8219	32.27 /0.9219	37.23 /0.9733	29.02 /0.7881	92.15 /390.03	3.12
CASNet (TIP 2022)	21.97 /0.6140	26.41 /0.8153	30.36 /0.9014	35.67 /0.9591	40.93 /0.9826	31.07 /0.8545	22.49 /0.5520	25.73 /0.7079	28.41 /0.8231	32.31 /0.9196	37.48 /0.9728	29.28 /0.7951	97.37 /1294.75	16.97
TransCS (TIP 2022)	20.22 /0.5431	25.41 /0.7883	29.54 /0.8877	35.06 /0.9548	40.49 /0.9815	30.14 /0.8311	22.28 /0.5318	25.28 /0.6881	27.86 /0.8086	31.74 /0.9121	36.81 /0.9699	28.79 /0.7821	22.72 /489.21	2.13
DGUNET$^+$ (CVPR 2022)	22.15 /0.6113	26.82 /0.8230	30.92 /0.9088	36.18 /0.9616	41.24 /0.9837	31.46 /0.8578	22.13 /0.5215	25.45 /0.6986	28.13 /0.8165	31.97 /0.9158	37.04 /0.9718	28.94 /0.7848	247.31 /98.41	37.81
OCTUF$^+$ (CVPR 2023)	22.07 /0.6235	26.84 /0.8221	30.70 /0.9030	36.10 /0.9604	41.31 /0.9838	31.40 /0.8586	22.78 /0.5413	25.65 /0.6999	28.28 /0.8177	32.24 /0.9185	37.41 /0.9729	29.27 /0.7901	94.74 /287.39	0.82
MSDAUN (Ours)	22.37 /0.6735	27.67 /0.8240	31.12 /0.9077	36.54 /0.9627	41.27 /0.9831	31.79 /0.8702	23.04 /0.5458	26.07 /0.7043	29.25 /0.8251	33.61 /0.9240	38.59 /0.9738	30.11 /0.7946	38.48 /101.77	7.92
MSDAUN$^+$ (Ours)	22.48 /0.6741	27.72 /0.8247	31.15 /0.9105	36.54 /0.9629	41.54 /0.9842	31.89 /0.8713	23.24 /0.5467	26.19 /0.7047	29.29 /0.8254	33.69 /0.9245	38.67 /0.9743	30.22 /0.7951	69.56 /342.18	10.71

and generates the reconstructed result $\hat{\mathbf{x}}_j$ as the output. To minimize the difference between $\mathbf{x}_j$ and $\hat{\mathbf{x}}_j$, we employ the mean squared error (MSE) as the loss function, as follows:

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{NN_a}\sum_{j=1}^{N_a}\|\mathbf{x}_j - \hat{\mathbf{x}}_j\|_2^2, \tag{12}$$

where N_a and N represent the number of training images and the size of each image respectively. $\mathbf{\Theta}$ denotes the learnable parameter set of our proposed MSDAUN and can be formulated as $\mathbf{\Theta} = \{\mathbf{A}, \mathrm{Conv}_0(\cdot), \mathrm{TFIM}(\cdot)\} \bigcup \{\mathcal{D}^{(k)}(\cdot), \mathcal{H}_{\mathrm{CAT}}^{(k)}(\cdot), \mathcal{H}_{\mathrm{TAT}}^{(k)}(\cdot),\ \mathcal{H}_{\mathrm{FFN}}^{(k)}(\cdot)\}_{k=1}^{K}$.

4. Experiments

4.1. Implementation Details

For training, we use 400 images from the training and test dataset of the BSD500 dataset (Arbelaez et al., 2010). Two benchmarks: Set11 (Kulkarni et al., 2016b)(nine 256×256 and two 512×512 grayscale images, each 512×512 image is considered into four 256 × 256 images) and CBSD68 (Martin et al., 2001)(68 color images with 321 × 481 pixels) are used for evaluation. For the network parameters, the block size B = 33, the default feature map C = 16, the default batch size K = 5 and the learnable parameter $\rho^{(k)}$ is initialized to 0.5. All experiments were conducted on an NVIDIA GeForce RTX4090, using PyTorch as the deep learning framework. Two commonly used image assessment criteria, Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), are adopted to evaluate the reconstruction results.

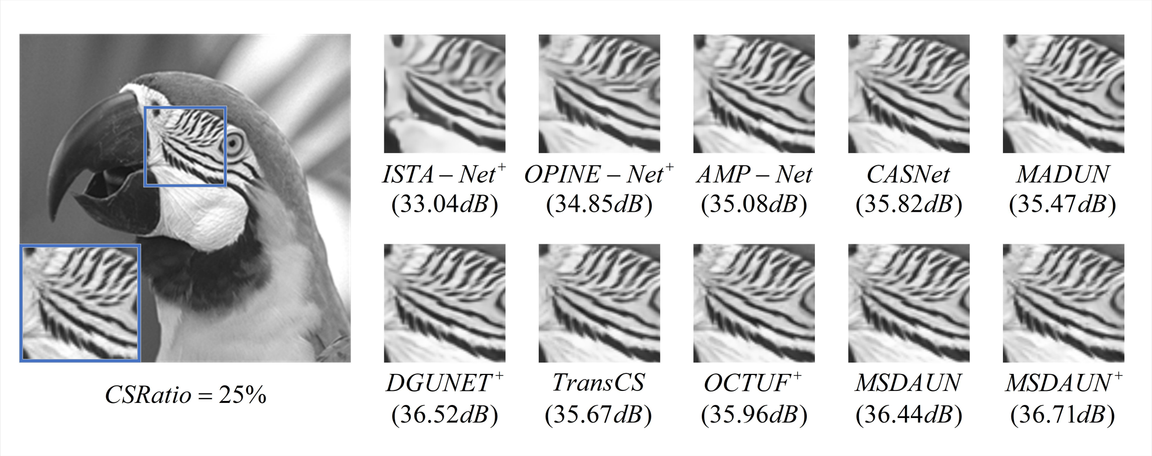

Figure 4: Comparisons on recovering an image from the Set11 dataset (Kulkarni et al., 2016b) in the case of CS ratio = 25%.

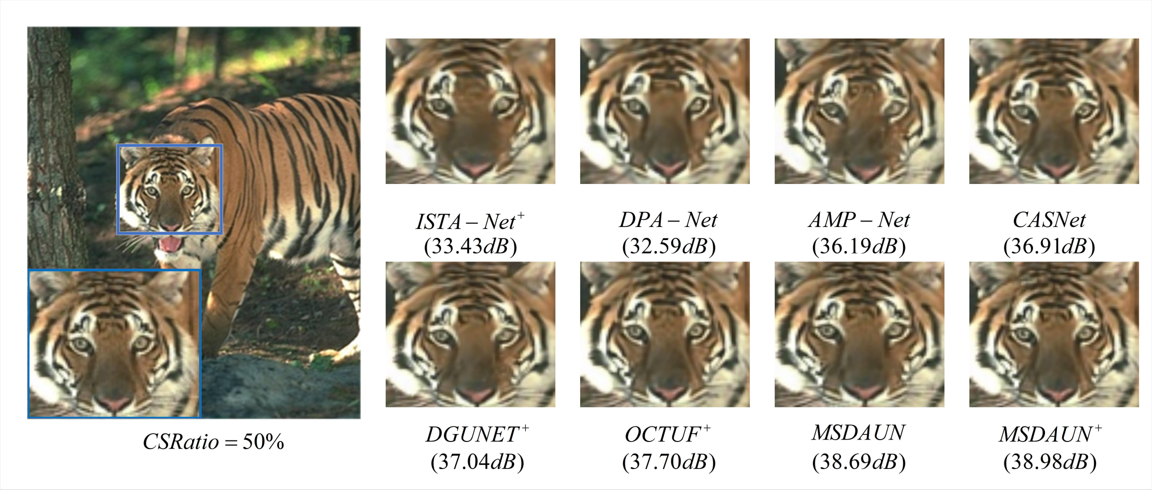

Figure 5: Comparisons on recovering an image from the CBSD68 dataset (Martin et al., 2001) in the case of CS ratio = 50%.

4.2. Qualitative Evaluation

We compare our proposed MSDAUN with nine SOTA CS reconstruction methods, including ReconNet (Kulkarni et al., 2016b), ISTA-Net$^+$ (Zhang and Ghanem, 2018), DPANet (Sun et al., 2020), AMP-Net (Zhang et al., 2020b), OPINE-Net$^+$ (Zhang et al., 2020a), MADUN (Song et al., 2021), CASNet (Chen and Zhang, 2022), TransCS (Shen et al., 2022), DGUNET$^+$ (Mou et al., 2022) and OCTUF$^+$ (Song et al., 2023). We summarize the average PSNR/SSIM reconstruction performance for five different CS ratios on the Set11 (Kulkarni et al., 2016a) and CBSD68 (Martin et al., 2001) dataset, as detailed in Table. 1. In our MSDAUN, we set the number of iterations to 5 and the initial learning rate to 5×10^{-4}. To enhance model performance, we also proposed an enhanced version, MSDAUN$^+$, with 7 iterations and an initial learning rate of 2×10^{-4}. The results indicate that both MSDAUN

Table 2: PSNR(dB) results on the Set11 dataset (Kulkarni et al., 2016b) in the case of CS ratio = 50% and the inference time of our network with three different types of architectures. The best and second-best results are highlighted in red and blue colors, respectively.

Stage Number K	10	20	30	40	feature Domain Dimensionality d
Plain-MSDAUN	40.71/32.569	40.87/61.514	40.96/84.691	41.09/124.547	16
	40.88/74.822	40.98/128.086	41.24/181.694	41.29/232.529	32
Group Stage Number K	1	3	5	7	feature Domain Dimensionality d
MSDAUN*	40.51/49.652	40.89/137.956	41.16/194.546	41.33/267.487	8
	40.60/95.546	41.01/269.533	41.29/408.561	41.47/578.244	16
MSDAUN	40.58/12.245	41.05/29.587	41.27/49.584	41.34/57.848	8
	40.91/16.545	41.23/36.249	41.39/58.462	41.54/81.524	16

Table 3: Ablation of TAT sub-module on Set11 dataset (Kulkarni et al., 2016b) when the CS ratio is 25%. The best PSNR(dB) is labeled in bold.

Cases	FFN	GDB	IF	FD	IGTA	GDTA	PSNR
(a)	√	-	-	-	-	-	33.72
(b)	√	√	-	-	-	-	35.06
(c)	√	√	-	√	-	-	36.15
(d)	√	√	√	√	-	-	36.16
(e)	√	√	-	√	-	√	36.46
(f)	√	-	-	√	√	-	36.41
MSDAUN	√	√	-	√	√	√	**36.54**

and MSDAUN$^+$ achieve a better reconstruction quality at all sampling rates, particularly at low CS ratios.

Fig. 4 and Fig. 5 show the visual comparisons of challenging images when a CS ratio of 25% and 50%, respectively. The images generated by our MSDAUN and MSDAUN$^+$ are visually superior and more consistent with the original images. Our MSDAUN provides higher fidelity and better detail compared to these methods. The experimental results demonstrate that, compared to similar methods, our MSDAUN has the best performance under low measurement rates, accurately reconstructing fine textures and better meeting application requirements.

4.3. Ablation Study

Effect of Multi-scale Architecture. We first conducted an ablation experiment on the multi-scale architecture with a CS ratio of 50%. Our architecture addresses the performance saturation and the difficulty (Zhang et al., 2020a; You et al., 2021a) of simple expansion by employing a deep multi-scale unfolding approach. We have constructed both regular and single-scale **W**-shaped variants, where the former adopts a conventional architecture, and the latter is denoted as MSDAUN*, replacing the SConv $(\cdot)$ and TConv $(\cdot)$ in Fig. 3(c) with regular conv layers to expand the feature capacity $\frac{H}{r} \times \frac{W}{r} \times r^2 C$, as multi-scale awareness is lost.

Table. 2 investigates the comparable networks with different stage numbers Ks and feature dimensions $C = D = d$. It is observed that the single-scale MSDAUN* outperforms the Plain-MSDAUN variant of the ordinary FD, but with a significant increase in time complexity. Our MSDAUN, which is highly flexible and comparable to MSDAUN*, significantly reduces inference time and enhances accuracy by focusing on feature refinement at smaller scales. When $K \geq 40$ and $d \geq 32$, the performance of the ordinary variant tends to saturate, but MSDAUN still has room for expansion. It should be noted that even the lightest MSDAUN variant with $K = 1$ and $2.9M$ parameters can achieve higher accuracy and is 11 times faster than the MADUN (Song et al., 2021) with 3.1M parameters, thereby validating the necessity of multi-scale expansion generalization.

Ablation of TAT. We conducted a carefully designed ablation study on the components of the TAT sub-module in the case of CS ratio = 25% with the results presented in Table. 3. where "IF" denotes the inertial force implemented through a simple method akin to Ochs et al. (2014), and "FD" indicates that the entire iterative process is conducted in the feature domain. Compared to Case (a), Case (b) achieved a 1.34 dB improvement, demonstrating the superiority of DUN over scenarios where only a single neural network is incorporated within the structure. Subsequently, the performance could be significantly enhanced by another 1.09 dB upon employing "FD" (as shown in Case (c)). We also conducted a detailed comparative experiment of the inertial force across Cases (c), (d), and (e), proving that our IGTA block can more fully leverage the effect of the inertial force. As illustrated in Cases (e) and (f), the application of IGTA and GDTA blocks yields superior performance. Our proposed TAT submodule effectively combines the gradient descent algorithm and texture features, and fully exploits structural characteristics.

5. Conclusion

This paper introduces a novel Multi-Scale Dual-Attention Unfolding Network (MSDAUN) for Compressed Sensing (CS), which employs a cross-stage multi-scale deep reconstruction module **D** in each iteration, comprising multiple attention modules. Specifically, we propose a Texture Attention Transformer (TAT) module, which consists of dual texture attentions: Inertia-Guided Texture Attention (IGTA) and Gradient Descent Texture Attention (GDTA). The IGTA module connects reconstruction information with texture path information, aiding in the effective computation of texture attention. The GDTA module utilizes gradient descent steps and inertial terms to guide the fine integration of channel features, allowing the texture path to focus on the recovery of textural details. Finally, the texture details are fused into the final iterative reconstruction structure, enabling more precise image reconstruction. Extensive experiments demonstrate that our MSDAUN exhibits lower complexity and superior performance compared to existing SOTA techniques. In the future, we aim to extend our MSDAUN to other image inverse problems and video applications.

References

Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE transactions on image processing*, 20(3):681–695, 2010.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2010.

Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle Transformer for spectral compressive imaging. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

Bin Chen and Jian Zhang. Content-aware scalable deep compressed sensing. *IEEE Transactions on Image Processing*, 31:5412–5426, 2022.

Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Tao Hong, Xiaojian Xu, Jason Hu, and Jeffrey A Fessler. Provable preconditioned plug-and-play approach for compressed sensing mri reconstruction. *arXiv preprint arXiv:2405.03854*, 2024.

Guancheng Huang, Yong Shuai, Yu Ji, Xuyang Zhou, Qi Li, Wei Liu, Bin Gao, Shutian Liu, Zhengjun Liu, and Yutong Li. Compressed hermite–gaussian differential single-pixel imaging. *Applied Physics Letters*, 124(11), 2024.

Michael Iliadis, Leonidas Spinoulas, and Aggelos K Katsaggelos. Deep fully-connected networks for video compressive sensing. *Digital Signal Processing*, 72:9–18, 2018.

Donghua Jiang, Nestor Tsafack, Wadii Boulila, Jawad Ahmad, and JJ Barba-Franco. Asbcs: Adaptive sparse basis compressive sensing model and its application to medical image encryption. *Expert Systems with Applications*, 236:121378, 2024.

Yookyung Kim, Mariappan S Nadar, and Ali Bilgin. Compressed sensing using a Gaussian scale mixtures model in wavelet domain. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2010.

Filippos Kokkinos and Stamatios Lefkimmiatis. Deep image demosaicking using a cascade of convolutional residual denoising networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. ReconNet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a.

Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 449–458, 2016b.

Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17399–17410, 2022.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Minghe Shen, Hongping Gan, Chao Ning, Yi Hua, and Tao Zhang. TransCS: A Transformer-based hybrid architecture for image compressed sensing. *IEEE Transactions on Image Processing*, 2022.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

Jiechong Song, Bin Chen, and Jian Zhang. Memory-augmented deep unfolding network for compressive sensing. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021.

Jiechong Song, Chong Mou, Shiqi Wang, Siwei Ma, and Jian Zhang. Optimization-inspired cross-attention transformer for compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6174–6184, 2023.

Yueming Su and Qiusheng Lian. iPiano-Net: Nonconvex optimization inspired multi-scale reconstruction network for compressed sensing. *Signal Processing: Image Communication*, 89:115989, 2020.

Jian Sun, Huibin Li, Zongben Xu, et al. Deep admm-net for compressive sensing mri. *Advances in neural information processing systems*, 29, 2016.

Yubao Sun, Jiwei Chen, Qingshan Liu, Bo Liu, and Guodong Guo. Dual-path attention network for compressed sensing image reconstruction. *IEEE Transactions on Image Processing*, 29:9482–9495, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Di You, Jingfen Xie, and Jian Zhang. ISTA-Net^{++}: Flexible deep unfolding network for compressive sensing. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2021a.

Di You, Jian Zhang, Jingfen Xie, Bin Chen, and Siwei Ma. COAST: Controllable arbitrary-sampling network for compressive sensing. *IEEE Transactions on Image Processing*, 30: 6066–6080, 2021b.

Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jian Zhang, Chen Zhao, and Wen Gao. Optimization-inspired compact deep compressive sensing. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):765–774, 2020a.

Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17532–17541, 2022.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

Zhonghao Zhang, Yipeng Liu, Jiani Liu, Fei Wen, and Ce Zhu. AMP-Net: Denoising-based deep unfolding for compressive image sensing. *IEEE Transactions on Image Processing*, 30:1487–1500, 2020b.

Chen Zhao, Siwei Ma, and Wen Gao. Image compressive-sensing recovery using structured laplacian sparsity in DCT domain and multi-hypothesis prediction. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2014.

Chen Zhao, Siwei Ma, Jian Zhang, Ruiqin Xiong, and Wen Gao. Video compressive sensing reconstruction via reweighted residual sparsity. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1182–1195, 2016.

Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. 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, 2021.