Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 347–354
DOI: 10.15439/2023F978
ISSN 2300-5963 ACSIS, Vol. 35

Dual-Path Image Reconstruction: Bridging Vision Transformer and Perceptual Compressive Sensing Networks

Zakaria Bairi*, Kadda Beghdad Bey*, Olfa Ben-Ahmed†, Abdennour Amamra* and Abbas Bradai†
*Ecole Militaire Polytechnique, Bordj El Behri, Algiers, Algeria
†XLIM Research Institute, UMR CNRS 7252
University of Poitiers, France

***Abstract*—Over the past few years, notable advancements have been made through the adoption of self-attention mechanisms and perceptual optimization, which have proven to be successful techniques in enhancing the overall quality of image reconstruction. Self-attention mechanisms in Vision Transformers have been widely used in neural networks to capture long-range dependencies in image data, while perceptual optimization has been shown to enhance the perceptual quality of reconstructed images. In this paper, we present a novel approach to image reconstruction by bridging the capabilities of Vision Transformer and Perceptual Compressive Sensing Networks. Specifically, we use a self-attention mechanism to capture the global context of the image and guide the sampling process, while optimizing the perceptual quality of the sampled image using a pre-trained perceptual loss function. Our experiments demonstrate that our proposed approach outperforms existing state-of-the-art methods in terms of reconstruction quality and achieves visually pleasing results. Overall, our work contributes to the development of efficient and effective techniques for image sampling and reconstruction, which have potential applications in a wide range of domains, including medical imaging and video processing.**

I. Introduction

COMPRESSIVE Sensing (CS) is an important technique in the field of signal processing and computer vision. CS is a technique for acquiring and processing signals at a lower rate than required by the Nyquist-Shannon sampling theorem. CS is used for image reconstruction, by reconstructing a high-quality image from a set of low-quality or incomplete observations. It has emerged as an alternative to traditional image compression techniques. The potential of CS and image reconstruction in computer vision lies in their ability to enable high-quality imaging and data acquisition with minimal resources. Indeed, CS can be used to reduce the amount of data required to capture an image, making it possible to store or transmit images more efficiently. It can also be used to reduce the amount of data required for image processing, enabling real-time processing of large datasets. Most CS approaches are CNN-based models, which represent a limitation by the receptive field of the convolution kernels and their non-ability to handle long-term dependencies.

Deep learning models have shown impressive success in various computer vision tasks, but they are not always efficient at processing large and complex images. One possible solution to this problem is to incorporate visual attention mechanisms into deep learning models, inspired by the way humans selectively process relevant information and filter out distractions. Attention mechanisms allow deep learning models to focus on important regions of an image and suppress irrelevant information, leading to improved accuracy and efficiency.

In addition to visual attention, which is a selective process that allows us to focus on important information in the environment while ignoring distractions, we can also find visual perception, which refers to the process of interpreting and making sense of visual information from the environment. This process involves multiple stages of processing, where the basic features of a stimulus (e.g., color, shape, motion) are detected and encoded then integrated into meaningful objects and scenes. Overall, visual attention and visual perception are both important processes in visual recognition. Visual attention allows us to selectively focus on important information in the environment, while visual perception allows us to interpret and make sense of visual information. These processes are intricately linked and work together to support our visual experiences.

In this paper, we propose a novel CS approach for image sampling and reconstruction. Our proposed model combines self-attention and perceptual information to selectively attend to different regions of an image at multiple levels of abstraction. We evaluate the effectiveness of our proposed model using experiments on benchmark datasets and demonstrate that it outperforms existing models in terms of reconstruction quality.

Hence, the main contributions of this paper are :

- We propose a framework based on a hybrid architecture that combines the self-attention mechanism provided by vision transformers for image long-range dependencies and global context modeling with the advantages of convolutional neural networks for optimal local feature extraction.
- We propose to add a transformer-based coding path, so the model coding is done in two paths, a CNN-based CSNet sampling path, and a transformer path. These two paths are linked by a fusion layer to merge the features and produce the vector that presents the input image.
- We use a perceptual optimization in the training process, to semantically guide the model to learn long-range

and local high-frequency details of visual and contextual features.
- Finally, we run extensive experiments to evaluate our approach in term of reconstruction quality and compare it to state-of-the-art methods on different image compression benchmarks.

The remainder of this paper is organized as follows. In section II, we present and discuss previous works on image-based CS reconstruction. In section III, we explain the proposed approach. In section IV, experimental results and comparisons with State-Of-The-Art (SOTA) methods are carried out. Finally, in section V, we summarize our findings and present some opportunities for future works.

II. Related Work

In this section, we present a CS image reconstruction literature review. We first discuss the existing deep learning-based CS methods. Then we review the recent development of vision transformers for image reconstruction.

A. Deep learning-based CS approaches

Compressing Sensing theory was first proposed in 2004 by David Donoho [1]. Deep learning-based CS approaches have been proposed to solve the CS reconstruction problem through the extraction (learning) of significant features from the input signal itself. Several reconstruction algorithms based on CNNs have been proposed to overcome the complexity of traditional methods. At the outset, Kulkarni et al. [2] developed a non-iterative reconstruction model using CNN (ReconNet). Based on iterative thresholding algorithms, Zhang et al. [3] proposed the convolutional ISTA-Net model for image recovery. Afterward, Shi et al. proposed a Scalable Convolutional Neural Network (SCSNet), and right after proposed a sampling reconstruction framework called CSNet [4], which replaces the sampling model with a convolutional layer. However, these methods have limitations due to their random sampling. To address this problem, Siwang Zhou et al. in [5] have proposed a Block-Based Image Compressive Sensing (BCS-Net), which uses block correlation for sampling. Nevertheless, the model training overlooked the semantic information of the image to draw the prior knowledge. Hence, in order to improve the reconstruction quality by considering the prior knowledge, Wenxue Cui et al. [6] have proposed a non-local CSNet (NL-CSNet) based on non-local self-similarity priors.

However, all the previous methods did not consider perceptual information, which is important for visual and semantic content reconstruction of the images. Recently, in [7], Bairi et al. proposed a perceptual-optimized CS framework that uses perceptual information for image reconstruction. The model is based on an auto-encoder, which is trained using perceptual optimization. Despite its power in the reconstruction of semantic information, this model still lacks high-frequency feature extraction.

B. Transformer-based image reconstruction

The first transformer was proposed by Vaswani et al. [8] for Natural Language Processing (NLP) tasks. In the latter, the long-range dependencies were given by multi-headed self-attention and feed-forward Multi Layer Perceptron (MLP) block. Among the best-known models dealing with this type of task are BERT [9] and GPT [10]. Based on the transformer force in NLP, transformers were recently integrated into the context of image processing. For classification tasks, the innovative work of Vision Transformer (ViT) [11] divides an image into 16 by 16 patches, to use the previous multi-headed self-attention and feed-forward MLP to build a classifier. In addition to the original ViT, transformer models, with different versions and architectures were proposed for several computer vision tasks namely for classification [12], [13], [14], [15], [16], [17], [8], for object detection [18], [19]and for image segmentation tasks [20], [21], [22].

Few works have investigated transformers for image reconstruction. Indeed, this task produces images as a final output, which is more difficult than high-level vision tasks such as classification, segmentation, and object detection, whose outputs are labels or areas. For transformer-based image reconstruction, Hanting et al. [23] proposed a pre-trained model called IPT that can be used for computer image reconstruction tasks. This approach suffers from the large number of parameters and image features are still extracted from CNN. A concurrent work [24] proposed a U-shaped transformer for image reconstruction, which is built upon the UNet architecture and based on the Swin's transformer block. However, these models, based solely on pure transformers, overlook local feature identification and low-frequency information. To preserve the advantages of both CNN-based networks for the local description and the transformer for long-range dependencies handling, Liang et al. [25] proposed a SwinIR model for the restoration of compressed or noisy images based on both Swin transformer blocks and CNNs which were designed for image classification in [15], this model showed better results than those obtained by IPT. Similarly, a transformer-based image reconstruction (TIC) method is developed in [26]. The latter uses a canonical architecture of the VAE variational autoencoder in the form of convolutional layers and Swin transform blocks to capture long and short-term dependencies of the input image. Test results on the Kodark dataset show the good performance of this approach. Dongjie et al. [27] extend the technique of self-attention in compressed sensing to overcome the limitations of convolution layers in modeling global features, by a CSformer model that combines the advantages of CNNs and transformers. The model contains a sampling module as a convolution layer and a reconstruction module in the form of two branches that integrate local and global-range dependencies. Nevertheless, these architectures need the integration of perceptual information, which helps the reconstruction of semantic details of the image.

III. Proposed approach

In this section, we present the proposed PCST-Net framework by using self-attention through vision transformer for better feature extraction and visual perception to make sense of these features. Fig.1 illustrates the proposed approach architecture. Indeed, it is based on CS sampling/reconstruction autoencoder which adopts an attention mechanism to capture long-range contextual information. The learning process is guided by the image's visual content information. The proposed approach involves two neural networks, an encoder, and a decoder. The encoder network compresses the input images by projecting them into a lower-dimensional space, while the decoder network restores the original image representation from the compressed representation. The network is trained in an end-to-end manner to minimize image reconstruction error, allowing it to find the optimal parameters that enable sampling and reconstruction for any input image.

A. Sampling network

The Sampling network (Encoder) is a combination of CNN and transformer models to take advantage of the spatial locality and self-attention mechanisms. The CNN model is inspired by PSCS-Net[7] and is laid out as three Convolution/MaxPooling blocks. In the original CS framework, encoded data are the result of sampling the input image. The latter are called encoded data as they correspond to the rows of the sampled image. In the context of deep learning, the encoded data is arranged rather like an ordinary 3D tensor like any CNN feature map. Theoretically, they still correspond to CS sampled vectors, just stacked in a 3D tensor. When we apply the sampling operator S_{CNN} on the input image x, we obtain y_1, which corresponds to the encoded data obtained by the CNN sampling Network.

$$S_{CNN}(x) = W_s^1 * x \tag{1}$$

In Eq.1, the network operates on 2D image patches with the convolution operator ($*$) with the sampling Matrix W_s^1. Such an operation projects an input image $x \in R^{d_x}$ onto one of the encoded vectors $y_1 \in R^{d_y}$. The sampling matrix W_s^1 is a composition of convolutions and nonlinear activation functions f that allows for better features extraction. The obtained result y_1 can be written as:

$$y_1 = S_{CNN}(x) = f(W_3 * f(W_2 * f(W_1 * x + b_1) + b_2) + b_3) \tag{2}$$

Transformer-based encoder aims to capture long-range visual dependencies through the self-attention mechanism. It is composed of a projection layer and a transformer block which is the architecture of the ViT backbone [11]. An image projection is a lower-dimensional representation of the image. In other words, it is a dense vector representation of the image. First, the image is divided into P × P non-overlapping patches, then this feature projection layer projects the input patches having a size of (P x P x C) into a dimension of (1 x Pd) such that Pd is the projection dimension. The self-attention mechanism is an integral component of a transformer, which explicitly models the interactions between all entities in a sequence. For an input sequence of Np elements, self-attention captures the interaction between all Np entities and encodes each entity in terms of global contextual information. For this fair, three weight learning matrices are defined, *Queries* ($W^Q \in \mathbb{R}^{Pd*q}$), *Keys* ($W^K \in \mathbb{R}^{Pd*k}$), and *Values* ($W^V \in \mathbb{R}^{Pd*v}$). The input sequence X is projected onto these weight matrices to obtain:

$$\begin{aligned} Q &= XW^Q \\ K &= XW^k \\ V &= XW^V \end{aligned} \tag{3}$$

Self-attention is formulated by:

$$A = softmax(\frac{QK^T}{\sqrt{q}})V \tag{4}$$

Fig.2 shows the transformer block architecture which consists of two LN normalization layers, a multi-headed self-attention layer MSA and a MLP made up of two fully connected layers, the τ norm is inserted before MSA and MLP.

The multi-headed self-attention MSA comprises several blocks of self-attention, each block has its own set of learnable weight matrices *Query*, *key*, and *Value*. Multi-headed self-attention runs h times in parallel, such that h is the number of heads, then concatenated into a single matrix. This block takes a series of sequences I patches of size (Np x Pd) as input and globally calculates the self-attention between them. The whole process of this block can be formulated as follows:

$$\begin{aligned} F_t &= MSA(\tau(I)) + I \\ y_2 = S_{ViT}(x) &= MLP(\tau(F_t)) + F_t \end{aligned} \tag{5}$$

The transformer path is composed of four transformer blocks. Feature fusion aims to extract the most discriminating information and eliminate redundant information. The fusion function combines the global features of the transformer and the local features of the CNN by a fusion strategy, such as addition or average. The fusion of y_1 and y_2 is given by Eq.6.

$$y = Fusion(y_1, y_2) \tag{6}$$

Since the stems of the transformer and the CNN have different dimensions, we need to modify the characteristics of the transformer to match those of the CNN.

B. Reconstruction Network

The upsampling network (Decoder) is designed in [7] as a three-block de-convolutional network to learn the inverse convolution filters to reconstruct images. The decoder returns y to the input space by obtaining the feature representation in the image recovery process. The decoder represents a nonlinear mapping that is learned from measurements y to its original image x by training. The decoder is symmetric with the CNN sampling network and consists of three layers: the input layer and two hidden layers. The decoder function (Eq. 7) is used to

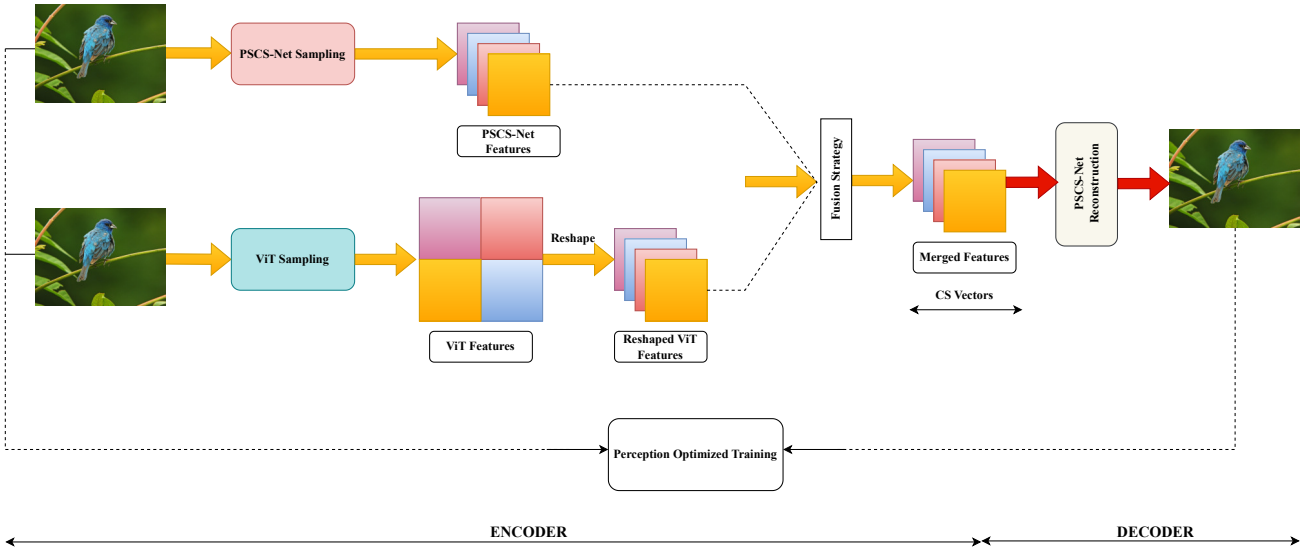


Fig. 1: Overview of the proposed image reconstruction-based framework: the model is trained using the combination of visual perception and self-attention.

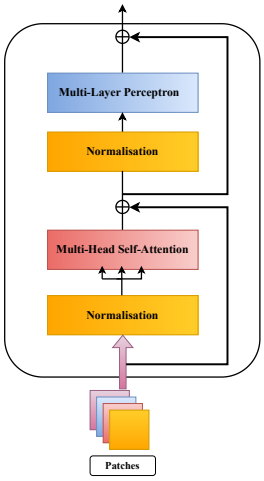


Fig. 2: Architecture of a transformer block

recover the reconstruction images $\widetilde{x}$ from measurement vector y.

$$\widetilde{x} = R(y) = f(W_6 * f(W_5 * f(W_4 * y + b_4) + b_5) + b_6) \quad (7)$$

C. Training of PCST-Net

To semantically guide our model to learn visual and contextual features, we use perceptual loss optimization in the training process as shown in [7]. The used perceptual loss measures the distance between images in high-level feature space using a pre-trained compressing sensing network [4] (CSNet). This model is originally trained on ImageNet dataset. The PCST-Net network is trained in an end-to-end fashion through the minimization of the global loss term expressed as:

$$\mathscr{L}_{total}(x, \widetilde{x}) = \alpha_1 L_p(x, \widetilde{x}) + \alpha_2 L_2(x, \widetilde{x}) + \alpha_3 L_s(W, b) \quad (8)$$

With : L_p in Eq.9 is the perceptual loss, L_s in Eq.10 is the sparsity loss, and L_2 in Eq.11 is the L_2 Norm between the original and reconstruted image. The three terms are weighted by α_1, α_2, and α_3, respectively.

$$L_p(x, \widetilde{x}) = MSE(\phi(x) - \phi(\widetilde{x})) \quad (9)$$

Where ϕ is the sampling operator of CSNet to compute the difference between the feature vector of the input image x and the predicted image $\widetilde{x}$.

$$\mathrm{L}_s(W_s^1, b) = 1/2\beta_1 \sum ||W_s^1||^2 + \beta_2 \sum_{j=1} KL(\rho||\rho_j) \quad (10)$$

The first term of L_s in Eq.10 limits the weight parameters W with L_2 norm as to penalize large weight. The second term is the sparsity regularizer. β_1 is the penalty term and KL is the Kullback-Leibler divergence for penalizing active code units. β_2 is the intensity of the sparsity, ρ is the sparse factor, and ρ_j represents the mean value of activation of the j^{th} neuron in each batch of the training set.

$$L_2(x, \widetilde{x}) = \sqrt{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (x(i,j) - \widetilde{x}(i,j))^2} \quad (11)$$

The L_2 Norm is used to profit from the qualities of pixel-wise loss functions.

The goal of training PCST-Net model is to minimize $\mathscr{L}_{total}$ as shown in Algorithm 1. First, parameters W_s^1, W^Q, W^K, and W^V are randomly initialized to serve the purpose of symmetry breaking. Then, encoded data y and the reconstruction images $\widetilde{x}$ are obtained through the encoder and decoder sub-networks, respectively.

Algorithm 1 PCST-Net training

Input:
Input original image x
Output:
Sampling Network weights W_s^1, W^Q, W^K, and W^V
Encoded data y
Reconstruction Network weights W_r
Instructions:
W_s, W_r : Randomly initialize
for epoch = 1 to number of epochs
$y_1 = S_{CNN}(x)$
$y_2 = S_{ViT}(x)$
$y = Fusion(y_1, y_2)$
$\widetilde{x} = R(y)$
Compute encoded image y
Compute perceptual loss $\mathscr{L}_{total}(x, \widetilde{x})$ (Eq.8)
Minimize final loss by gradient descent algorithm
Update W_s^1, W^Q, W^K, W^V, and W_r
end for

IV. Experiments and Results

In this section, we first introduce the dataset used for PCST-Net model training and the evaluation metrics. Second, we present the model settings for better training (Section IV-B). Next, in section IV-C, we conduct an experimental study on image compression benchmarks for model objective evaluation and compare the proposed approach with state-of-the-art methods. Finally, in Section IV-D, we evaluate the quality of PCST-Net image reconstruction with a subjective evaluation.

A. Datasets and evaluation metrics

PCST-Net is trained using a large-scale dataset which is COCO 2017 dataset [1]. 118k and 40k images have been used for training and validation respectively.

We evaluate our PCST-Net on different widely used benchmark datasets, such as Set5 [28], Set14 [29], and BSD100 [30].

To evaluate the model, two metrics are computed: Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM). PSNR measures image reconstruction quality, while SSIM, a perceptual metric, quantifies image degradation.

B. PCST-Net model settings and training

1) Model hyper-parameters selection: After an axial empirical study, the hyper-parameters of the model are set to 256 x 256 x 3, 8 x 8 x 3, and 8 x 8 x 3 for Image size, Patch size, and Block size, respectively. The perceptual loss function is optimized using the Adam optimizer with a batch size equal to 32 and a learning rate of 0.002 for 100 epochs.

[1]https://cocodataset.org/home

2) Fusion strategy selection: Our method adopts an addition strategy to merge the features of different paths. To illustrate the effectiveness of this method, we construct a variant in which the features of the CNN and the transformer are averaged rather than summed.

Fig.3 shows the PSNR results of the two models on Set5, Set14, and BSD100. The feature addition fusion operation shows superior PSNR performance with different compression ratios. The feature averaging operation achieves a close performance when the compression ratios are lower than 10%, but above this compression ratio, the addition shows its efficiency against the average.

3) Path selection: **PCST-Net** is a **Dual Path** model that aims to combine the efficiency of convolution in extracting local features with the capability of the transformer in modeling global representations. To compare the advantages of the two branches-based approach, we created a **Single Path** model called **SPCST-Net**, which uses only the transformer path for compression. The results of the tests on three datasets (Set5, Set14, and BSD100) are presented in Fig.4.

Obtained results on Set5, Set14, and BSD100 datasets confirm that PCST-Net helps in recovering more details and semantic information of the images compared to PSCS-Net (based only on CNN) or SPCST-Net (based only on transformers).

C. Objective Evaluation

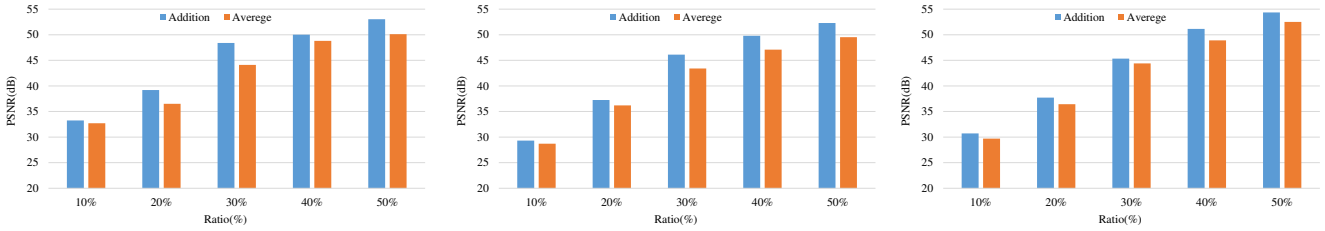
The results of the comparative study of PSNR and SSIM, between the different state-of-the-art reconstruction methods namely ISTA-Net+[3], CSNet+[4], NL-CSNet*[6], DPA-Net[31], CSFormer[27], PSCS-Net[7], and our PCST-Net, applied on the Set11, Set5, and BSD100 reconstruction datasets are shown in Table I-III, while varying the compression ratio between 0.1 and 0.5.

Our experimental results show that our approach achieves higher performance for image reconstruction compared to state-of-the-art algorithms.

TABLE I: Comparaison of PSNR(dB) and SSIM on Set5

Algorithm/Ratio	0.1	0.2	0.3	0.4	0.5
ISTA-Net+[3]	28.61 0.8315	33.12 0.9058	35.45 0.9408	36.94 0.9612	38.42 0.9804
CSNet+[4]	32.59 0.9062	36.05 0.9481	38.25 0.9644	40.11 0.9740	41.79 0.9803
NL-CSNet*[6]	33.84 0.9312	36.91 0.9589	38.86 0.9703	41.20 0.9895	43.15 0.9942
DPA-Net[31]	30.32 0.8713	-	36.17 0.9495	38.05 0.9632	39.57 0.9716
CSFormer[27]	34.20 0.9262	36.88 0.9514	39.74 0.9689	-	43.55 0.9845
PSCS-Net[7]	**33.75** **0.9422**	38.68 **0.9893**	47.10 **0.9946**	49.92 0.9950	52.27 0.9973
Ours	33.25 0.9387	**39.19** 0.9747	**48.40** 0.9899	**50.02** **0.9955**	**53.04** **0.9975**

The results obtained by PCST-Net on the different compression datasets benefited from the coupling between perception and self-attention to give the best PSNR and SSIM values compared to other state-of-the-art reconstruction methods.



(a) (b) (c)

Fig. 3: PSNR(dB) histogram for each fusion strategy on Set5(a), Set14(b), and BSD100(c).

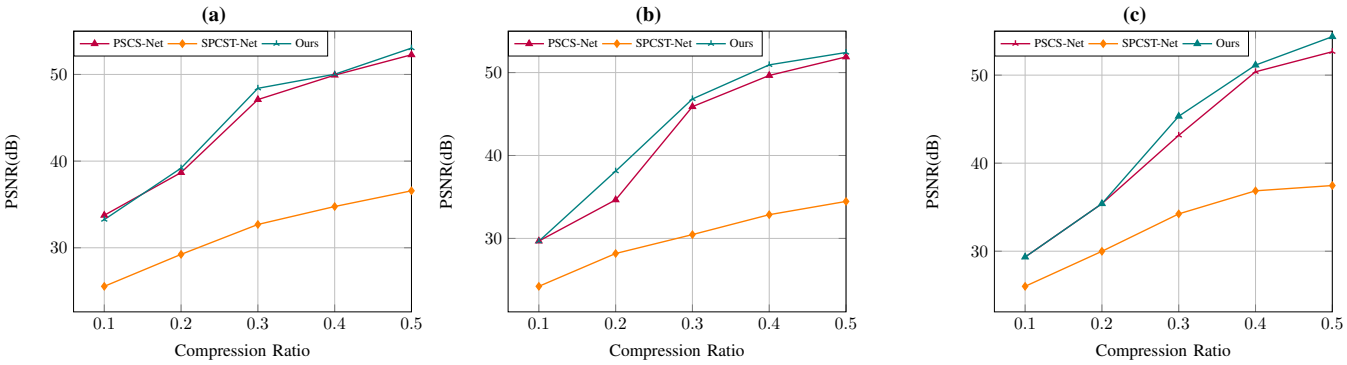


Fig. 4: Average PSNR(dB) for different Path-based methods on Set5(a), Set14(b), and BSD100(c).

TABLE II: Comparaison of PSNR(dB) and SSIM on Set14

Algorithm/Ratio	0.1	0.2	0.3	0.4	0.5
ISTA-Net+[3]	26.49 0.8010	30.79 0.8950	33.76 0.9345	36.03 0.9547	38.49 0.9127
CSNet+[4]	29.13 0.8169	32.15 0.8941	34.34 0.9297	36.16 0.9502	37.97 0.9754
NL-CSNet*[6]	30.16 0.8527	32.96 0.9150	34.88 0.9405	37.21 0.9752	40.17 0.9891
DPA-Net[31]	27.22 0.8401	31.51 0.9249	33.37 0.9395	35.91 0.9592	37.84 0.9701
CSFormer[27]	30.85 0.8515	34.02 0.9274	36.47 0.9459	-	40.41 0.9730
PSCS-Net[7]	**29.68** **0.8987**	34.65 0.9644	45.89 0.9920	49.65 0.9967	51.89 0.9981
Ours	29.31 0.8921	**37.25** **0.9720**	**46.11** **0.9929**	**49.81** **0.9967**	**52.30** **0.9985**

TABLE III: Comparaison of PSNR(dB) and SSIM on BSD100

Algorithm/Ratio	0.1	0.2	0.3	0.4	0.5
ISTA-Net+[3]	24.79 0.6726	27.64 0.7906	29.86 0.8580	31.70 0.9003	33.02 0.9513
CSNet+[4]	28.53 0.7834	31.05 0.8721	33.08 0.9171	34.91 0.9443	36.68 0.9618
NL-CSNet*[6]	28.61 0.8361	31.20 0.9141	33.30 0.9354	36.91 0.9627	39.94 0.9845
DPA-Net[31]	26.47 0.7388	29.87 0.8611	30.23 0.8894	32.70 0.9241	34.19 0.9488
CSFormer[27]	28.28 0.8078	31.62 0.9110	33.57 0.9399	-	38.01 0.9712
PSCS-Net[7]	29.34 0.8884	35.40 0.9632	43.16 0.9924	50.38 0.9969	52.66 0.9982
Ours	**30.71** **0.9044**	**37.72** **0.9680**	**45.33** **0.9932**	**51.14** **0.9971**	**54.37** **0.9989**

D. Subjective Evaluation

In this section, we describe the subjective evaluation to visualize the quality of reconstructed images. This qualitative assessment is done with the naked eye by noting the differences between images at a ratio of 0.25. We also provide PSNR and SSIM values for each image to highlight quantitative differences.

The visualization obtained by PCST-Net in Fig.5 shows again that the use of both perception and self-attention gives the best result compared to other reconstruction methods. Obtained results suggest that the combination of self-attention and perceptual optimization can provide a powerful tool for improving the quality of image reconstruction. The use of self-attention mechanisms to capture long-range dependencies in the image data can lead to better sampling performance, while the incorporation of perceptual optimization can enhance the perceptual quality of the reconstructed images.

V. Conclusion

In this paper, we proposed a novel approach for image sampling and reconstruction that combines Vision Transformer and perceptual optimization techniques. Our approach leverages the power of self-attention to capture the global context of the image and guide the sampling process while optimizing the perceptual quality of the sampled image using a perceptual loss function. We have demonstrated the effectiveness of our proposed approach through experiments on several benchmark datasets, and we have shown that it outperforms existing state-of-the-art methods in terms of reconstruction quality

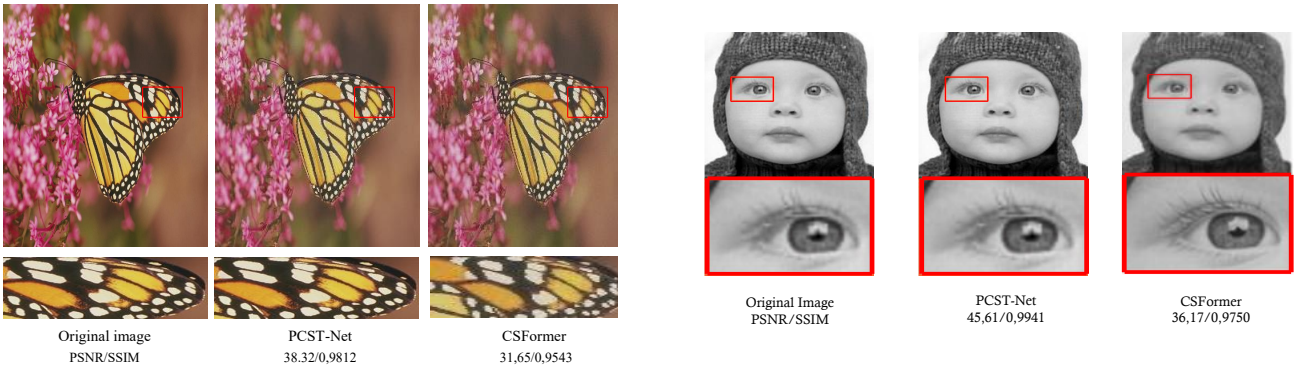


Fig. 5: Comparison of the visual quality of image reconstruction using a ratio of 0.2(a) and 0.4(b).

and visual fidelity. Our approach has potential applications in a wide range of domains, including medical imaging, video processing, and computer graphics. In conclusion, our work contributes to the development of efficient and effective techniques for image sampling and reconstruction, which are critical components in the field of multimedia processing. We believe that our proposed approach can serve as a foundation for future research in this area, and we hope that it will inspire further innovations in the field of computer vision.

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 449–458.

[3] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1828–1837.

[4] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image compressed sensing using convolutional neural network," *IEEE Transactions on Image Processing*, vol. 29, pp. 375–388, 2019.

[5] S. Zhou, Y. He, Y. Liu, C. Li, and J. Zhang, "Multi-channel deep networks for block-based image compressive sensing," *IEEE Transactions on Multimedia*, 2020.

[6] W. Cui, S. Liu, F. Jiang, and D. Zhao, "Image compressed sensing using non-local neural network," *IEEE Transactions on Multimedia*, 2021.

[7] Z. Bairi, O. Ben-Ahmed, A. Amamra, A. Bradai, and K. B. Bey, "Pscsnet: Perception optimized image reconstruction network for autonomous driving systems," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[8] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[13] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[14] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[16] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," *arXiv preprint arXiv:2104.05707*, 2021.

[17] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, and L. Van Gool, "Vision transformers with hierarchical attention," *arXiv preprint arXiv:2106.03180*, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[19] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, pp. 261–318, 2020.

[20] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[21] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

[22] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[23] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 299–12 310.

[24] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[25] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 1833–1844.

[26] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," *arXiv preprint arXiv:2111.06707*, 2021.

[27] D. Ye, Z. Ni, H. Wang, J. Zhang, S. Wang, and S. Kwong, "Csformer: Bridging convolution and transformer for compressive sensing," *arXiv preprint arXiv:2112.15299*, 2021.

[28] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[29] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, 2010, pp. 711–730.

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

[31] Y. Sun, J. Chen, Q. Liu, B. Liu, and G. Guo, "Dual-path attention network for compressed sensing image reconstruction," *IEEE Transactions on Image Processing*, vol. 29, pp. 9482–9495, 2020.