TESA: Tensor Element Self-Attention via Matricization

Francesca Babiloni[1], Ioannis Marras[1], Gregory Slabaugh[1], Stefanos Zafeiriou[2]

[1]Huawei Noah's Ark Lab [2]Imperial College London

Abstract

Representation learning is a fundamental part of modern computer vision, where abstract representations of data are encoded as tensors optimized to solve problems like image segmentation and inpainting. Recently, self-attention in the form of a Non-Local Block has emerged as a powerful technique to enrich features, by capturing complex interdependencies in feature tensors. However, standard self-attention approaches leverage only spatial relationships, drawing similarities between vectors and overlooking correlations between channels. In this paper, we introduce a new method, called Tensor Element Self-Attention (TESA) *that generalizes such work to capture interdependencies along all dimensions of the tensor using matricization. An order R tensor produces R results, one for each dimension. The results are then fused to produce an enriched output which encapsulates similarity among tensor elements. Additionally, we analyze self-attention mathematically, providing new perspectives on how it adjusts the singular values of the input feature tensor. With these new insights, we present experimental results demonstrating how TESA can benefit diverse problems including classification and instance segmentation. By simply adding a TESA module to existing networks, we substantially improve competitive baselines and set new state-of-the-art results for image inpainting on CelebA and low light raw-to-rgb image translation on SID.*

1. Introduction

Deep Convolutional Neural Networks (DCNNs) represent the state-of-the-art method in a variety of computer vision problems but, in their standard implementation, they are limited to compute only local regions of the input. This innate characteristic makes long-range dependencies, which are a key aspect in a variety of tasks, hard to capture without the use of circumventing techniques. The use of deeper stacks of convolutional layers, for instance, increases neurons' receptive fields [31] at the cost of optimization difficulties [20] and higher complexity [38]. Recently, more sophisticated layers (e.g. *Non-Local Blocks*) have been proposed, which directly leverage these interdependencies as a means to enrich the intermediate CNN representations [41, 47, 39, 49, 28]. These blocks have been proven useful, beating competitive baselines in video action recognition, classification and instance segmentation. At the same time,

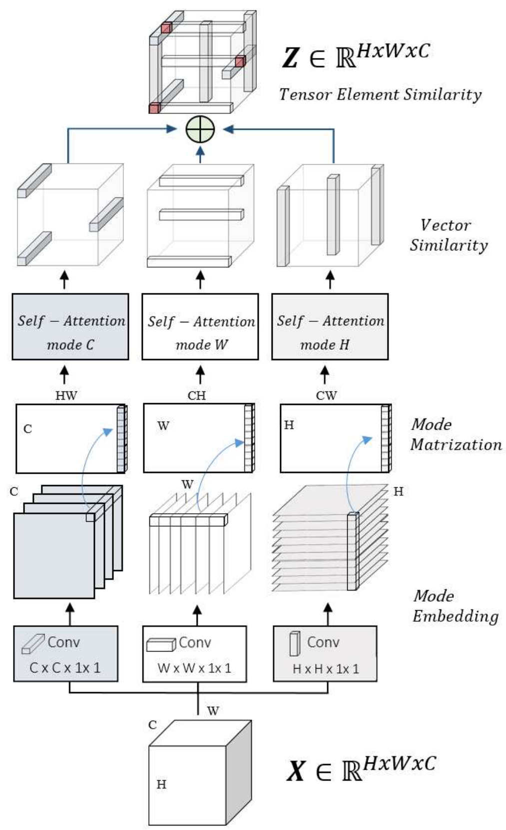


Figure 1: The input 3rd-order tensor is viewed as a combination of its three mode-matricizations. Combining their outputs allows the method to make use of inter- and intra-channel correlations. Blue-gray, white and gray 3D boxes represent similar vectors within each one of the three matricizations. Red cubes in the Z tensor represent similar elements in Z.

the majority of these methods try to estimate only spatio-temporal correlations among positions of the input tensor [49, 41, 28] or overlook its complex topology [47, 3]. In this paper, we build upon the aforementioned line of research and extend its scope to the goal of mining *tensor element interactions* from the input. Our three main contributions can be summarized as follows:

- We propose a new self-attention block (TESA) able to leverage correlations in all possible directions of the input tensor to take advantage of channel information without losing the topology of the input tensor. Instead of completely flattening the elements of the input in a single vector and facing intractable complexity, we propose to use tensor matricizations as a way to extract complex interactions (Figure 1).
- We provide a statistical interpretation of the proposed family of non-local blocks. In particular, we demonstrate that our block can be seen as an operator acting as a regulariser of the spectrum (i.e. the variance) of the various matricizations of the feature tensor. We prove from a theoretical and empirical perspective how TESA adjusts the relative importance of the singular values. This is achieved implicitly without the need to compute an expensive singular value decomposition (SVD) in a direct way.
- We demonstrate the power of TESA in a battery of heterogeneous computer vision tasks. Our method shows a consistent improvement in large-scale classification, detection, instance segmentation and puzzle-solving. It also achieves state-of-the-art performance in the two dense image-to-image translation problems of inpainting and short-exposure-raw to long-exposure-rgb.

2. Related work

Self-similarites The concept of similarity among image parts or video frames is pivotal in many computer vision applications. Therefore, a long-lasting trend in the community has been understanding how to properly define and exploit self-similarity. The idea of relating features to each other (i.e. CNN's channels or classical descriptors) has inspired various pooling methods [23, 27, 12, 6, 25, 4] where correlation is used as a higher order representation for the image and fed to a classifier. Simultaneously, a complementary line of research proposed techniques to relate an image part to its context using both classical methods [36, 13, 8] and CNN models [9, 21, 17, 50, 24].

Self-Attention The key idea of an attention mechanism is to steer the model focus on particular portions of the data, considered useful to solve the given task. Its initial formulation can be traced back to the 60's [42, 10] and has recently received interest in various applications of machine learning. In machine translation, self-attention vectors assess how strongly each element attends (i.e. is correlated) to all the others and estimate the target as the sum of all elements in a sentence, weighted by their attention values [2, 40]. Variants of self-attention have been used in computer vision to solve a variety of problems ranging from inpainting [45] to zero shot-learning [44] and visual question answering [35, 37]. Noticeable examples can be found in classification, where it has been used to estimate attention-masks for intermediate CNN features [32] or to learn re-calibration of features given global channels descriptors [19]. Recently, the *Non-Local Block* has been proposed as a plug-and-play extension to existing architectures. The purpose of this block is to enrich features using spatial-temporal interaction, considering all position at once [41, 49, 39] or a single position and its neighborhood [28]. This formulation inspired new deep learning architectures [48, 11, 7] and has been extended in recent works to the scope of integrating with the input compact global descriptor of feature maps [47, 3, 43].

3. Method

In this section, we introduce the notation used throughout the paper, give an overview of the concept of self-attention and describe in detail the proposed method. At first, we study the spatial version of our non-local block. Next, we generalize to the scope of capturing more complex tensor interdependencies. Finally, we relate our method to other existing non-local blocks.

3.1. Notation

In the rest of the paper, we adopt the notation of Kolda *et al.* in [22]. Tensors are denoted using calligraphic letters (e.g. $\mathcal{X}$) and matrices by bold upper-case letters (e.g. $\mathbf{X}$). The i^{th} row of a matrix $\mathbf{X}$ is a vector denoted using lower-case bold letters as $\mathbf{x}_i$. The order N of a tensor corresponds to the number of its dimensions and can be also called mode. A mode-n-fiber of a tensor is the vector obtained by fixing all indices of $\mathcal{X}$ except for the n^{th} dimension and can be seen as a generalization of matrix's rows and columns. The mode-n-matricization of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n}$ is a case of matricization denoted as $\mathbf{X}_{(n)}$ and arranges its mode-n-fibers to be the columns of the resulting matrix. More formally, the tensor elements $(i_1, i_2, \ldots, i_N)$ are rearranged into the matrix element (i_n, j) where $j = 1 + \sum_{k=1,k\neq n}^{N}(i_k - 1)J_k$ and $J_k = \prod_{m=1,m\neq n}^{k-1} I_m$.

3.2. Overview of self-attention

Given an input matrix $\mathbf{X}$, an attention mechanism weights $\mathbf{X}$ with an attention matrix $\mathbf{A}$ to highlight the relevant parts of the input. Different ways of computing $\mathbf{A}$ entail different variants of attention mechanisms. This paper focuses on self-attention, where weights are only a function of input $\mathbf{X}$. In particular, we consider a *pairwise function* f, which can be used to capture interdependencies between each $\mathbf{x}_i$ and every $\mathbf{x}_j$. A self-attention block is a variation of a residual block [18] which sums the output of a self-attention mechanism to the original input $\mathbf{X}$. The output $\mathbf{Y}$ of the self-attention block is expressed as follows:

$$\mathbf{Z} = \mathbf{X} + \mathbf{A}\mathbf{X} = \mathbf{X} + f(\mathbf{X}, \mathbf{X}). \quad (1)$$

3.3. Capturing spatial correlation

Let a 3^{rd} order tensor $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ be the feature map output of one of the layers of a CNN. Let $\mathcal{X}$ be rearranged, using its mode-c matricization, as a $\mathbf{X}_{(c)} \in \mathbb{R}^{WH\times C}$ where each spatial position is described by its C features. Let's assume $\mathbf{X}_{(c)}$ is mean normalized. In the linear version of the proposed block, we choose as attention matrix the covariance $\mathbf{X}_{(c)}\mathbf{X}_{(c)}^\top \in \mathbb{R}^{WH\times HW}$, which expresses the correlations between each i^{th} position and every j^{th} position. Thus, the output of the spatial self-attention block using this mechanism can be written as follows:

$$\mathbf{Z} = \alpha_c \mathbf{X}_{(c)} + \beta_c \mathbf{X}_{(c)}\mathbf{X}_{(c)}^\top \mathbf{X}_{(c)} \quad (2)$$

where α_c and β_c are learnable scalars modulating the contribution of each term. In Eq. 2, the global covariance term modulates the feature's representation with spatial similarities. The residual term, together with the two learnable scalars allows the implicit regularization of the spectrum via a polynomial function. In the following, we draw a connection between the input and the output of the self-attention block, omitting the subscripts to simplify the notation. The matrix $\mathbf{X}$ and its positive, semi-definite covariance matrix have the following singular-value and eigen decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{X}\mathbf{X}^\top = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{U}\big(\mathbf{\Sigma^2}\big)\mathbf{U}^\top \quad (3)$$

where $\mathbf{Q} = \mathbf{U}$ is the eigenvectors matrix, $\mathbf{V}^\top$ and $\mathbf{U}$ are the right and left singular vectors, $\mathbf{\Lambda}$ is the eigenvalue matrix and $\mathbf{\Sigma}$ its corresponding singular value diagonal matrix. Notably, $\mathbf{\Lambda} = \mathbf{\Sigma^2}$. Thus, the β parameter learns to modulate the contribution of the following term:

$$\mathbf{X}\mathbf{X}^\top\mathbf{X} = \mathbf{U}\mathbf{\Sigma^2}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma^3}\mathbf{V}^\top \quad (4)$$

From the above, it is evident that using the proposed self-attention block changes the spectrum of $\mathbf{X}$ as

$$\alpha\mathbf{X} + \beta\big(\mathbf{X}\mathbf{X}^\top\big)\mathbf{X} = \mathbf{U}(\alpha\mathbf{\Sigma} + \beta\mathbf{\Sigma^3})\mathbf{V}^\top \quad (5)$$

Hence, the self-attention block described in Equation (2) learns the coefficients of a polynomial function of the singular values, without operating on the input's orthogonal vectors $\mathbf{U}$ and $\mathbf{V}^\top$ and it can be seen as an operator that modifies the singular values of the input matrix $\mathbf{X}$ without the need of a direct and expensive SVD computation. Only two learnable parameters, α and β, are used for this purpose. Since α and β can be either positive or negative, the method performs an algebraic sum of two functions and, therefore, has the flexibility to regularise the spectrum by performing shrinkage or whitening.

3.4. Capturing tensor elements interdependencies

In the previous subsection, the choice of unfolding the tensor as a matrix $\mathbf{X}_{(c)} \in \mathbb{R}^{WH\times C}$ drives the focus of the attention mechanism to capture only spatial similarities. In the following, we introduce a generalization that leverages both spatial and channel-based correlations, while keeping intact the module's effect on the spectrum. As depicted in Figure 1, the proposed generalisation represents the feature tensor $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ using its three mode matricizations $\mathbf{X}_{(c)}$, $\mathbf{X}_{(h)}$ and $\mathbf{X}_{(w)}$, each embedded in a different subspace via a weight matrix $\mathbf{W}$, followed by a non-linear function σ:

$$\mathbf{Y}_{(n)} = \phi(\mathbf{X}_{(n)}) = \sigma(\mathbf{X}_{(n)}\mathbf{W}_{(n)}) \quad {\scriptstyle n\in\{C,H,W\}}. \quad (6)$$

In our implementation, σ is a $ReLU$ activation function [33] and the weight matrices $\mathbf{W}_{(c)} \in \mathbb{R}^{C\times C}$, $\mathbf{W}_{(h)} \in \mathbb{R}^{H\times H}$, $\mathbf{W}_{(w)} \in \mathbb{R}^{W\times W}$ correspond to 1×1 convolutions in the tensor space over each respective dimension.

Then, a self-attention block is applied independently to each $\mathbf{Y}_{(n)}$ and the three contributions are reshaped and combined via summation to generate the final output $\mathcal{Z}$.

$$\mathcal{Z} = \sum_{n}^{C,H,W} \Psi_{(n)}(\alpha_n \mathbf{Y}_{(n)} + \beta_n \mathbf{Y}_{(n)}\mathbf{Y}_{(n)}^\top \mathbf{Y}_{(n)}), \quad (7)$$

where $\Psi_{(n)}$ is a reshape function which rearranges the matrices as tensors of dimension $H\times W\times C$. In the above equation, each embedded matricization represents a different point of view on the input tensor: $\mathbf{Y}_{(c)}$ accounts for spatial interactions, $\mathbf{Y}_{(w)}$ for interactions between rows and channel activations and $\mathbf{Y}_{(h)}$ for interactions between columns and channels. Our method processes each $\mathbf{Y}_{(n)}$ in its own space, modulating its representation with a self-attention block as described in Eq. 2. Thus, it is not limited to capture only correlations among positions but is also capable of capturing correlations across channels. In order to be considerated simultaneously, the three contributions are fused in tensor space (i.e. overlayed in the same coordinate-space). The fusion through summation ensures i) the same dimensionality of input and output and ii) equal contribution for each term. As depicted in Figure 1, although the output of each self-attention block encapsulates similarity between pairs of vectors, their summation allows the method to directly relate *tensor elements to each other*. We call our method Tensor Element Self-Attention or TESA.

3.5. Relation with Other Self-Attention Blocks

In this section, we connect TESA with other self-attention works. The non-local block [41] and its variants [49, 39, 28] explore the introduction of non-local information in a neural network and can be framed as self-attention methods investigating spatial correlation. In the formulation that is closest to ours, the non-local block applies three learnable weights matrices ($\mathbf{W}_\theta$, $\mathbf{W}_\phi$ and $\mathbf{W}_g$) on the same input $\mathbf{X}$ [41]. The first two matrices are in charge of extracting spatial long-range dependencies using a dot-product similarity, while the third one embeds the input. Given $\mathbf{X} \in \mathbb{R}^{WH \times C}$, the output $\mathbf{Z}$ of the original non local block is: $\mathbf{Z} = \mathbf{X} + softmax(\mathbf{X}\mathbf{W}_\theta\mathbf{W}_\phi^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_g$. Our goal is to generalize the self-attention mechanism to more complex interactions without overlooking channel information. Therefore, our block embeds each tensor mode separately and aims to extract different correlations from each embedding. Recent works propose to leverage channel information by estimating a scalar global description of each channel [43] or tensor feature maps [47]. In the case closest to ours in [47], the method divides the input in G separate groups $\mathbf{X}_i \in \mathbb{R}^{WH \times \frac{C}{G}}$ and extracts a global representation ($\mathbb{R}^1$) for each of them. On the contrary, we tackle the problem from a complementary perspective, with a formulation that focuses on the explicit computation of tensor elements' pairwise correlation and provides interpretability regarding the self-attention effect on the features.

4. Illustrative experiment

One of the goals of self-attention mechanisms is to equip a model with the capacity to reason about the whole input representation at one glance. We first test this property in a controlled scenario, designing a new "puzzle MNIST" experiment. We used the MNIST dataset and a four-layer fully convolutional encoder-decoder architecture. To test the ability of our self-attention method to make use of available but scattered information, we attempted the reconstruction of an image given its shuffled version. To obtain an input puzzle, each image is split into 16 tiles of equal size. These tiles are then randomly rotated and mirrored before being stitched back together. Input and output samples can be seen in Figure 2e and 2a, respectively. Between the encoder and the decoder part of the network, the self-attention module integrates information about positions or tensor self-similarity.

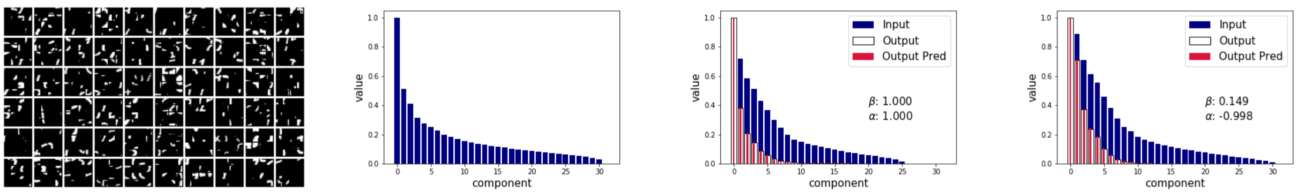
4.1. Capturing spatial correlations

We start by analyzing the spatial self-attention block as formulated in Equation 2. To highlight the effect of self-attention in the latent space, we compared a model trained without any attention ($\alpha = 1$, $\beta = 0$) and two variants of our block: one where α and β are fixed to be equal to unity and the other where they are treated as learnable scalars. The first row of Figure 2 shows a qualitative overview of our comparisons. The baseline is limited to process the input locally and performs worse than models trained with self-attention. The second row shows comparisons on the empirical distribution of singular values for the puzzle MNIST test set. Given a sample of the test-set, we extracted features before and after the self-attention block, returning for each image two matrices $\mathbf{X}_{in}$ and $\mathbf{X}_{out}$. As explained in Section 3, the left and right singular vectors are left untouched by the method. Consequently, the relation between input and output can be computed using only the α and β parameters and the effect of self-attention can be captured plotting the singular value spectrum of input and output side-by-side. Figures 2g and 2h show the singular values of the input $\mathbf{X}_{in}$ (in blue) and the singular values of the output $\mathbf{X}_{out}$ (in white) plotted in descending order. The red bars depict the prediction for $\mathbf{\Sigma}_{out}$ obtained using Equation 5. Comparing input and output in each plot shows how the attention block shrinks the spectrum of the input, automatically choosing which information (i.e. components) is highlighted and which is suppressed to simplify the subsequent decoding task. Moreover, it shows how the direct SVD computation matches closely the theoretical prediction. The comparison between the two plots shows how the possibility to learn the spectrum transformation (Figure 2h) retains more expressive components compared to the fixed contribution of self-attention and input (Figure 2g). For example, in Figure 2h the drop between the first and second singular value is substantially smaller (30% drop) than what occurs in the case of α and β fixed to one (60% drop).

4.2. Capturing tensor elements interdependencies

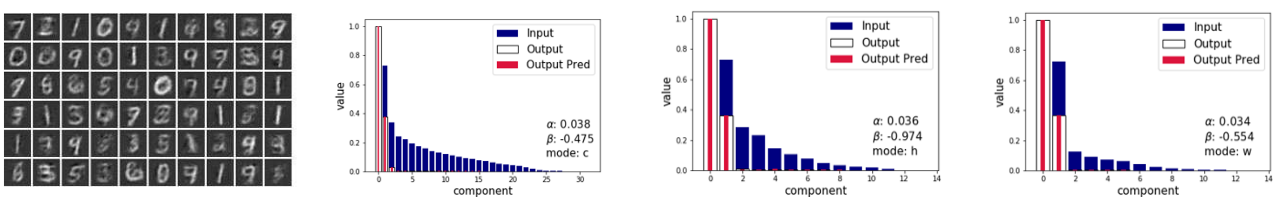
The same logic can be used to analyze the generalized case of Equation 7. As a first step, we extended the linear spatial case to consider channel based interdependencies. This case is equivalent to substitute $\mathbf{Y}_n$ with $\mathbf{X}_n$ in Equation 7. This allows a comparison in the same latent space for input ($\mathcal{X}$) and output ($\mathcal{Z}$) tensors, and gives the possibility to inspect directly their matricization's spectrum. In 3b, 3c, 3d plots, each mode matricization (H, C, W) is treated separately, showing the comparison between input and output of each self-attention. The figures depict how self-attention produces a shrinkage effect on all mode matricizations. The possibility to correlate channels with rows and columns patterns modifies the role of $\mathbf{X}_{(c)}$. Its spectrum is drastically reduced to have only two meaningful components, accounting for more than 99% of the whole variance. Figure 3a shows sample outputs for our method, which is not limited to spatial similarity but can leverage multiple views on the original tensor. It produces considerably sharper outputs

(a) Groundtruth (b) No Attention $\alpha = 1, \beta = 0$ (c) Equal Contribution $\alpha = \beta = 1$ (d) Learned Contribution $\alpha = 0.1, \beta = -1$



(e) Input (f) Singular Values $\alpha = 1, \beta = 0$ (g) Singular Values $\alpha = \beta = 1$ (h) Singular Values $\alpha = 0.1, \beta = -1$

Figure 2: **Spatial Self-Attention Overview**. The first row shows reconstructed digits. The baseline with no attention is outperformed using spatial-correlation. The best quality is achieved by α and β learnable scalars. Features' singular values show how self-attention drives the first principal components to account for the majority of the variance in the matrix. Computing the output spectrum empirically (white bars) or using Equation 5 (red bars) yields very close results. Blue plots differ due to the different embeddings learned by the architectures.



(a) TESA - w/o Embeddings (b) Singular Values Matricization mode C (c) Singular Values Matricization mode H (d) Singular Values Matricization mode W

Figure 3: **Tensor Element Self Attention Overview**. The qualitative comparisons showcase the benefits of our method, which makes use of channel information to produce more defined output.

when compared with previous cases.

To discuss the case described in Equation 7, we have to extend the analysis to consider the embedded mode matricizations, $\mathbf{Y}_c$ $\mathbf{Y}_h$ $\mathbf{Y}_w$. In this case, the input $\mathcal{X}$ and $\mathcal{Z}$ tensors live in different subspaces, due to the learnable parameters of the projection matrices $\mathbf{W}_c$ $\mathbf{W}_h$ $\mathbf{W}_w$. Each mode matricization is embedded separately, but each self-attention operates directly on its input without any additional transformation. Thus, the input/output pairs of each self-attention still share the same orthogonal vectors and their spectrum can be still compared and used to highlight the impact of the self-attention module on each latent space. In the next section, we will report this effect on different problems and datasets. On "shuffle MNIST", the use of embeddings produce shrinking trend and qualitative output similar to those reported in Figure 3.

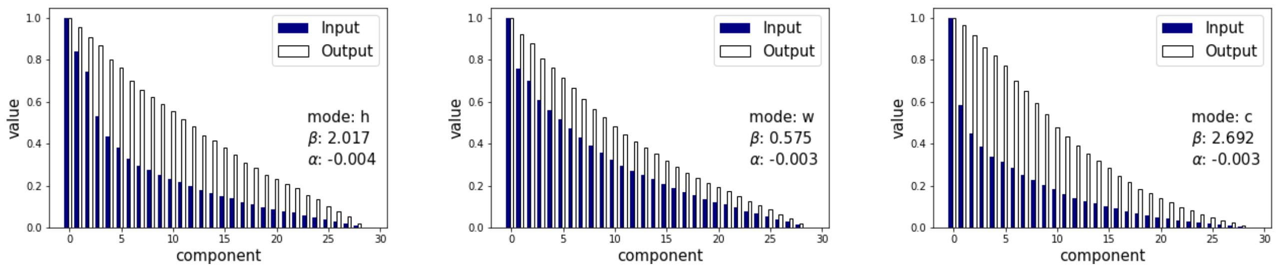
5. Experiments

We evaluated TESA on a series of computer vision problems, ranging from dense image-to-image translation to detection. This section starts by presenting results on two dense tasks based on an encoder-decoder architecture, where the self-attention block is used to enrich the encoded features representation. Then, our analysis is extended to the case of a ResNet architecture used for classification and as the backbone for instance segmentation.

5.1. Short exposure Raw to Long exposure rgb

Initially, we address the task of reconstructing a high-quality long-exposure rgb image given a noisy short-exposure sensor raw image captured in low-light conditions. In digital photography, an Image Signal-processing Pipeline (ISP) transforms raw data collected by an image sensor into

(a) Input (b) SID [5] (c) SID + NL [41] (d) SID + TESA



(e) Singular Values Matricization mode H (f) Singular Values Matricization mode W (g) Singular Values Matricization mode H

Figure 4: **Short Exposure Raw to Long Exposure rgb Overview.** The qualitative comparison shows how the method of [5] can be improved by the use of self-attention. Our method is able to recover cleaner patterns and generate outputs without strong color artifacts. The second row shows the singular value plots of the input/output pair for each mode matricization. The trend depicts the whitening effect that the self-attention has on the input spectrum. Images visible better zoomed in electronically.

a high-quality rgb image. Traditionally, an ISP relies on classical methods and is strongly dependent on the noise distribution and camera sensor. Modern deep learning approaches [1, 5, 15] replace the traditional ISP with one convolutional network, achieving good performance especially in challenging cases of low signal-to-noise ratio (SNR). The backbone in our experiments is the Unet architecture proposed in Learning to See in the Dark (SID) [5], trained and implemented as described in the original paper[1]. To investigate the capacity of the attention module to reason on the whole image representation, we inserted the self-attention block between the encoder and the decoder part of the network where units benefit from the largest receptive field. We then compare it against the plain version of the architecture and versions where alternative self-attention blocks [41, 47] are used[2][3].

The experiments report results for the SID-Sony dataset [5], which consists of short-exposure raw and long-exposure rgb pairs of high-resolution images (4240x2832). The pairs are captured in low-light conditions ranging from 0.03 to 5 lux.

[1] https://github.com/cchen156/Learning-to-See-in-the-Dark

[2] https://github.com/facebookresearch/video-nonlocal-net

[3] https://github.com/KaiyuYue/cgnl-network.pytorch

The Sony camera uses a Bayer sensor pattern to capture a single raw frame with short exposure. Simultaneously, the camera shot a reference rgb image increasing the exposure factor of 100 or 300 times used as ground-truth by the network. Table 1b reports the reference metrics PSNR and SSIM obtained by the different methods. New state-of-the-art performance is achieved by powering the architecture with our self-attention block. Qualitatively, our method shows better details and color recovery when compared with the competitors (Fig. 4). Figures 4e, 4f, 4g depict the effect of TESA on the singular values. In this case, the input spectrum of $\mathbf{Y}_{(h)}$, $\mathbf{Y}_{(w)}$, and $\mathbf{Y}_{(c)}$ is whitened; note that the input (dark blue bars) singular values fall off quickly (e.g. exponentially), whereas self-attention with TESA rebalance their intensities and produces an output (white bars) where the singular values fall off more gradually (e.g. approximately linearly).

5.2. Inpainting - CelebA

Image inpainting requires missing pixels in the input image to be filled in. An inpainting algorithm hallucinates the missing image pixels and blends them in with the surrounding regions in a coherent manner, producing a real-

(a) Input (b) Groundtruth (c) PConv [29] (d) PConv + NL [41] (e) PConv + TESA

Figure 5: **Inpainting Qualitative Comparison**. The Partial Convolution baseline (Pconv) creates blurry and artificial outputs that are partially improved by the use of self-attention (e.g. PConv + NL implementing the non-local block). Our method (right), leverages similarities across multiple dimensions and produces consistent colors and realistic details.

Method	PSNR	SSIM	MS-SSIM
PConv [29]*	25.36 / 25.10	0.877 / 0.872	0.928 / 0.922
+Wang [41]*	25.41 / 25.24	0.881 / 0.878	0.928 / 0.924
+Yue [47]*	25.66 / 25.57	0.888 / 0.885	0.931 / 0.927
+Ours	**26.00 / 25.81**	**0.895 / 0.891**	**0.936 / 0.931**

(a) **Inpainting - CelebA. (Random and Center Crop Evaluation)**

Method	PSNR	SSIM
SID [5]	28.88	0.787
SID [5]*	28.57	0.884
+Wang [41]*	29.54	0.888
+Yue [47]*	29.62	0.889
+Ours	**29.79**	**0.891**

(b) **Raw-to-rgb - SID Sony**

Table 1: **Quantitative Comparisons: Unet for inpainting and Raw-to-rgb.** Reconstruction metrics for the inpainting and short-exposure-raw to long-exposure-rgb tasks. Experiments employ different variants of the Unet architecture: 1 attention block is added for raw-to-rgb and 3 attention blocks for inpainting. State-of-the-art performance can be achieved by using our method. Asterisks '*' indicate results produced using software provided by the authors[1][2][3][4].

istic output image. Impressive results have been achieved in this area, with the latest architectures revolving around an encoder-decoder network with or without skip connections [46, 29, 45]. In these experiments, the baseline architecture is the Unet architecture with partial convolution[4], as proposed in [29]. We investigated the capacity of the self-attention block to work in a multi-scale fashion. In the encoder, the resolution of the input is downscaled multiple times with the goal to concentrate on different aspects of the image at different layers. We incorporated self-similarity information at different scales by inserting the self-attention block of Section 3.4 at layers 2, 4 and 6. We compared our architecture against a variant where our block is replaced by another version [41, 47] and against the original Unet architecture, where no attention mechanism is used. For training procedures and implementation details, please refer to [29].

The CelebA dataset [30], which consists of more than 202K samples, was used in our experiments. The training data were generated by randomly cropping a 128x128 patch from each training sample (a quarter of the input image). Table 1a reports PSNR, SSIM, MS-SSIM for the methods. TESA achieved the best results for all the evaluated metrics and generates convincing images with rich details and reduced artifacts (e.g. more defined wrinkles).

5.3. Instance Segmentation - MS-COCO

The task of image instance segmentation requires the detection and segmentation of each item in the input image, differentiating among instances. It outputs a per-pixel mask that identifies both the category and the instance for each object. The baseline model for these experiments is the two stages Mask R-CNN [16]. The first Region Proposal stage (RPN) uses a network that serves as "attention" for the entire pipeline: it takes an image as input and outputs a set of rectangular object proposals. The second stage addresses in parallel the tasks of classification and regression of the bounding-box regions. We tested the capacity of the self-attention block to enrich the representation of RPN features. Following the implementation of related work, we added a self-attention block right before the last residual block of the ResNet50 feature extractor, reducing the channel dimension while embedding the modes' matricizations. To bring back the overall output to the original channel dimension, we used one extra convolution and a weighted global skip connection. We compared against the original implementation of [16], trained end-to-end, and its non-local block extension [41, 49]. Please refer to the original papers and code[2][5]

[4] https://github.com/NVIDIA/partialconv

[5] https://github.com/latentgnn/LatentGNN-V1-PyTorch

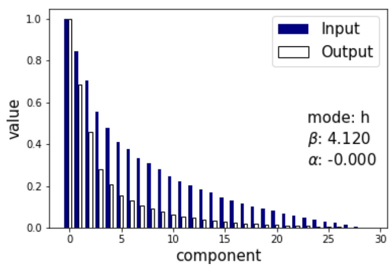
Method	Top 1	Top 5
ResNet50 [47]	76.15	92.87
+ Yue [47]	**77.69**	**93.64**
ResNet50*	75.78	92.76
+ Wang[41]*	76.09	93.00
+ Zhang[49]*	75.28	92.33
+ Ours	76.49	93.05

(a) **Classification - Imagenet**

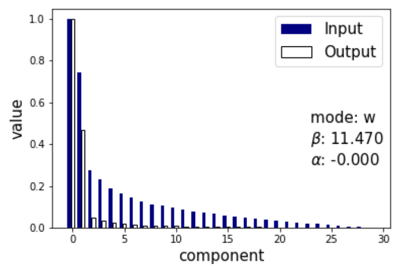
Method	AP_{box}	AP_{box50}	AP_{box75}	AP_{mask}	AP_{mask50}	AP_{mask75}
MaskR-CNN[49]	37.8	59.1	41.2	34.2	55.8	36.3
+ Zhang[49]	39.0	60.7	42.5	35.5	57.6	37.6
MaskR-CNN*	38.1	59.4	41.2	34.6	55.9	36.8
+ Wang[41]*	39.0	61.1	41.9	35.5	**58.0**	37.4
+ Zhang[49]*	39.1	60.7	42.5	35.5	57.6	37.6
+ Ours	**39.5**	**61.2**	**43.0**	**35.7**	57.9	**37.9**

(b) **Object detection and Instance Segmentation - MS-COCO**

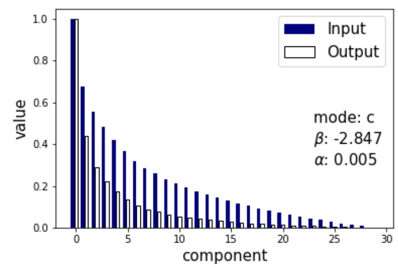
Table 2: **Quantitative Comparisons: ResNet.** Performance metrics for the task of classification on the Imagenet dataset [34] and object detection and instance segmentation on COCO [26]. Results are based on ResNet50 and Mask R-CNN with a ResNet50-FPN backbone. Both use one single attention block. Asterisks '*' indicate results obtained using software provided by the authors[235]. Yue *et al*. did not converge during our training and is not reported in table (b).



(a) Singular Values
Matricization mode H



(b) Singular Values
Matricization mode W



(c) Singular Values
Matricization mode C

Figure 6: **Tensor Element Self-Attention Spectrum on instance segmentation (MS-COCO)**. The same shrinking trend can be tracked in all three plots. The block compresses the input singular values and outputs feature maps where the meaningful components (i.e. which cumulative count for 80% of the total variance) are reduced.

for implementation details and to [14] for the training procedure. Table 2b reports results on the Microsoft Common Objects in COntext dataset [26]. We used the 2017 version of the dataset and reported the standard metrics of AP (averaged over multiple IoU thresholds) for segmentation and detection tasks. The results show how TESA outperforms its competitors. Figures 6a, 6b, 6c show the singular values for each mode-matricization input/output pairs. In this case, the self-attention block learns to shrink the singular values of the input. In other words, it implicitly performs a choice on which feature's information (i.e. singular vectors) the network should pay attention to during the feature extraction process.

5.4. Classification - Imagenet

Lastly, we evaluate our method on a large-scale classification task using the 1000 categories and the 1.2 million training images of the ImageNet dataset [34]. The backbone for our experiments is a Resnet50 architecture trained following the protocol in [14]. We extended this architecture with one TESA self-attention block or its different variants [41, 49], as described in the previous paragraph. Table 2a reports Top1 and Top 5 accuracy for the evaluated methods. The use of global descriptors achieves the best performance, but TESA makes full use of the interactions among tensor elements, producing competitive results and outperforming methods which use only spatial correlations.

6. Conclusion

In this paper, we introduced a new family of non-local blocks, framed mathematically as operators acting on the features' spectrum and proposed TESA, which generalizes earlier non-local spatial correlations to tensor-elements interactions. We demonstrated its capacity to consistently improve results over competitive and state-of-the-art baselines on diverse tasks. Finally, we showcased the distinctive characteristic of our method to single out the interesting data components, adapting its behavior to different applications. We illustrated how this can entail shrinking, where dominant components are chosen to summarize the data in a compact way, or whitening, where components are balanced and decorrelated to simplify the subsequent tasks. Next, we aim to combine TESA with the orthogonal contribution of global descriptors.

References

[1] Deepisp: Toward learning an end-to-end image processing pipeline. 4326

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2015. 4322

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019. 4322

[4] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*, pages 430–443. Springer, 2012. 4322

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 4326, 4327

[6] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017. 4322

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 4322

[8] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. Bm3d frames and variational image deblurring. *IEEE Transactions on Image Processing*, 21(4):1715–1728, 2011. 4322

[9] Thomas Deselaers and Vittorio Ferrari. Global and efficient self-similarity for object classification and detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1633–1640. IEEE, 2010. 4322

[10] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012. 4322

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 4322

[12] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 4322

[13] Stephen Gould, Tianshi Gao, and Daphne Koller. Region-based segmentation and object detection. In *Advances in neural information processing systems*, pages 655–663, 2009. 4322

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4328

[15] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2511–2520, 2019. 4326

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4327

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 4322

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4323

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4322

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 4321

[21] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6560–6569, 2017. 4322

[22] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 4322

[23] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 365–374, 2017. 4322

[24] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3587–3596, 2017. 4322

[25] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2078, 2017. 4322

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4328

[27] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 4322

[28] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018. 4321, 4322, 4324

[29] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for

irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 4327

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4327

[31] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016. 4321

[32] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 4322

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4323

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4328

[35] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 4322

[36] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR*, volume 2, page 3. Minneapolis, MN, 2007. 4322

[37] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016. 4322

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4321

[39] Yunzhe Tao, Qi Sun, Qiang Du, and Wei Liu. Nonlocal neural networks, nonlocal diffusion and nonlocal modeling. In *Advances in Neural Information Processing Systems*, pages 496–506, 2018. 4321, 4322, 4324

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4322

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4321, 4322, 4324, 4326, 4327, 4328

[42] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964. 4322

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 4322, 4324

[44] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019. 4322

[45] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 4322, 4327

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 4327

[47] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. In *Advances in Neural Information Processing Systems*, pages 6510–6519, 2018. 4321, 4322, 4324, 4326, 4327, 4328

[48] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019. 4322

[49] Songyang Zhang, Xuming He, and Shipeng Yan. Latentgnn: Learning efficient non-local relations for visual recognition. In *International Conference on Machine Learning*, pages 7374–7383, 2019. 4321, 4322, 4324, 4327, 4328

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4322