This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# A contrast-robust trainable ConvNet layer for image classification using quaternion monogenic signal

**E. ULISES MOYA-SÁNCHEZ[1], (Member, IEEE), SEBASTIÀ XAMBÓ-DESCAMPS[2], ABRAHAM, SÁNCHEZ[3], SEBASTIÁN SALAZAR-COLORES [4] AND ULISES CORTÉS [2]**

[1] Dirección de Inteligencia Artificial, Gobierno de Jalisco and Universidad Autónoma de Guadalajara, Guadalajara, Jalisco, Mexico
[2] UPC-Barcelona Tech and Barcelona Supercomputing Center, Barcelona, Spain
[3] Dirección de Inteligencia Artificial, Gobierno de Jalisco and Technology, Guadalajara, Jalisco, Mexico
[4] Centro de Investigaciones en Óptica, León, Guanajuato, Mexico

Corresponding author: E. Ulises Moya-Sánchez (e-mail: eduardo.moya@{jalisco.gob.mx, edu.uag.mx}).

This work was supported in part by CONACYT Mexico and Barcelona Supercomputing Center.

**ABSTRACT** At present, Convolutional Neural Networks (ConvNets) achieve remarkable performance in image classification tasks. However, current ConvNets cannot guarantee the capabilities of mammalian visual systems such as invariance to contrast and illumination changes. Some ideas for overcoming the illumination and contrast variations must usually be tuned manually and tend to fail when tested with other types of data degradation. In this context, a new bio-inspired entry layer is presented in this work, M6, which detects low-level geometric features (lines, edges, and orientations) similar to those patterns detected by the V1 visual cortex. This new trainable layer is capable of dealing with image classification tasks even with large contrast variations. The explanation for this behavior is due to the use of monogenic signal geometry, which represents each pixel value in a 3D space using quaternions, a fact that confers a degree of explainability to the networks. The M6 was compared to conventional convolutional layer (C) and a deterministic quaternion local phase layer (Q9). The experimental setup is designed to evaluate the robustness of this M6 enriched ConvNet model and includes three architectures, four datasets, and three types of contrast degradation (including non-uniform haze degradations). The numerical results reveal that the models with M6 are the most robust in front of any kind of contrast variations. This amounts to a significant enhancement of the C models, which usually have reasonably good performance only when the same training and test degradation are used, except for the case of maximum degradation. Moreover, the Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR) are used to analyze and explain the robustness effect of the M6 feature maps under any kind of contrast degradations.

**INDEX TERMS** Bio-Inspired models, ConvNet, Robust deep-learning, Monogenic Signal.

## I. INTRODUCTION

ONE important feature of the mammalian visual cortex is its built-in capacity to recognize objects independently of size, contrast, illumination, angle of view, or brightness, among other transformations [1]. Achieving an equivariance or invariance response to these transformations is an important goal in Deep Learning (DL) [2], [3]. In fact, an increasing number of studies have found various weaknesses in the generalization capacity of ConvNets models [2], [3] related to large contrast changes. In addition, previous evidence demonstrates that the deployment of DL models in the real world could be affected significantly by day to night light changes, haze, or the effects of glare, such as self-driving cars [4], surface glazes in medical images [5], or 24-hour surveillance.

Data augmentation is one idea for overcoming illumination and the contrast variations in image classification problems [6], [7]. However, previous work indicates that learning of an invariant response may fail even with large data augmentations in the training process [8]. In addition, data augmentation techniques have three main problems [9], [10]: i) they must be tuned (manually) by human experts, which causes large variances in the DL-model performance in practice; ii) because of a lack of analytic tools (even for



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



simple models), it is not well-understood how training on augmented data affects the learning process; and iii) data augmentation approaches focus on improving the overall performance of a model, and it is often imperative to have a finer-grained perspective. This means that data augmentation techniques are required to mitigate weak performance when dealing with under-represented classes.

Another common approach to tackling contrast and illumination problems is data normalization, such as local response normalization [11]. However, the main problem of these approaches appears when the change of illumination is non-uniform across the images in the dataset [12]. Rad et al. proposed using an adaptive local contrast normalization based on a window (region) difference of Gaussians for image detection. Although their approach is novel, the parameters of the Gaussian functions are based on dataset illumination. As a result, overall detection performance will be reduced for a new image with very different contrast.

Generative Adversarial Networks (GANs) are capable of restoring contrast affected images [2], [13]. Nevertheless, these architectures are primarily designed for visual restoration, not for image classification.

The limitations of these approaches reveal opportunities to explore novel methods. This work proposes a new strategy to progress toward achieving contrast-illumination invariance in image classification tasks. Specifically, a new bio-inspired trainable layer, M6, is presented which detects low-level geometric features (such as lines, edges, and orientations), which are not unlike similar patterns detected by the V1 visual cortex [14].

The M6 layer is based on the 2D extension of the analytic signal, called the *monogenic signal*. As a result, each pixel value of an image $I(x, y) \in \mathbf{R}$ is mapped to a Hamilton quaternion (see Figure 1). The geometry of this approach has the remarkable simple property that the quality of the representation is not affected by large changes of the pixel intensities, a fact that confers a degree of explainability to the networks.

On the experimental side, to evaluate the predictive performance and robustness of M6, contrast changes were simulated in three different ways using four datasets: MNIST [15], Fashion MNIST [16], CIFAR-10 [17], and Dogs and Cats [18], with three architectures: A1 (shallow), A2 (medium), and A3 (very deep). The performance of M6 was compared to a 2D conventional ConvNets layer (C) and the Q9 layer [19], which follows a similar approach to handle contrast changes. To evaluate the robustness response the models were trained with a specific data degradation and tested with other types of data degradations. The numerical results in the test set confirm that M6 achieves a remarkable resilient response to contrast variations when compared to standard convolution (layers) networks and to the Q9 approach.

The rest of the paper is organized as follows: Section II, acknowledges previous related works. Section III summarizes background materials concerning the local phase computation, the monogenic signal, and bio-inspired tools. Section IV, presents the computational aspects of the monogenic layer M6, and Section V the data and experimental arrangements. The Section VI describes the experimental outcomes and their analysis and Section VII presents the authors' conclusions and future work.

## II. RELATED WORK

Promising approaches were advanced in the bio-inspired papers [1] (which introduces **VisNet**) and [20]. These introduce and study, via quite distinct approaches, interesting hypotheses about how the cortex achieves various invariant representations. However, testing is performed on relatively small datasets, and although lighting invariance is considered, contrast invariance is not.

The mathematical tools used in [21]–[23] involve only complex numbers, but are otherwise somewhat akin to those used here. Their goals, however, are quite different, and in particular, they do not seek a robust response to large contrast or illumination changes. The first presents (complex) harmonic networks (**H-nets**) and shows, by means of complex circular harmonic functions, that they exhibit equivariance to translations and to in-plane rotations. The aim of the second is to present a complex-valued learnable Gabor-modulated network which features orientation robustness. The third reference applies the *analytic signal* (defined in section III) using the Hilbert-Huang transform to decompose the ranked pooling features into finite and often few data-dependent functions. This approach can deal with occlusion or heavy camera movement in action recognition. However, the scope of this work do not include the occlusion problem.

Recent developments dealing with contrast robustness have been reported. In [24], for instance, the authors present an entry layer (**VOneNet**) which uses Gabor functions and report the results of testing for different data degradations and perturbations. The layer is deterministic (non-trainable) and consequently requires fine tuning several parameters (related to input size, normalization, and random parameters) to set the Gabor function. For contrast degradation, the test accuracy drops from 75.6% to 28.5% on the Imagenet dataset.

Another deterministic approach is presented in [25]. The authors propose combining the chromatic component of a perceptual color-space to improve image segmentation. Nevertheless, tests in this work deal only with outdoor images with normal degradations, and their parameters must tuned according to the dataset.

A previous version of the M6 layer still deterministic was presented in [26], but already with quite robust responses for illumination or contrast variations. In another recent work [19], the authors proposed a bio-inspired approach to leverage quaternion Gabor filters to improve classification even with contrast degradations. An adaptive local contrast normalization was proposed in [12]. Although this approach was more effective than conventional normalizations, the trainable parameters are based on the dataset illumination. As a result, overall detection performance will be reduced for images with different degradations.



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



To assess the performance and robustness of our M6 layer in more concrete terms, it was compared primarily to conventional 2D convolutional layers C, which are the most popular and provide the architecture for many ConvNet models. Because of this, the authors looked for studies that satisfied the following conditions: i) robustness to contrast or illumination changes in image classification tasks, ii) available and implementable code for different architectures and datasets. The condition selected for that purpose, already cited above, are [12], [19], [24], [25]; and on closer inspection mentioned above, it was determined that only [19] satisfies the four criteria. As we mention before, the main strength of M6 with respect to [19] is that the latter is based on a deterministic unit, while M6 is trainable, making possible to improved robustness and performance in image classification tasks.

## III. BACKGROUND

This section explains background and notations used in this work. These include the analytic and monogenic signal properties. The bio-inspired connection to the proposed layer is also explained.

### A. SIGNALS

We define 1D (resp. 2D) *multivectorial signals* as $C^1$ maps $U \to \mathcal{G}$ from an interval $U \subset \mathbf{R}$ (a region $U \subset \mathbf{R}^2$) into a *geometric algebra* $\mathcal{G}$ (see [27] for detailed definition). For $\mathcal{G} = \mathbf{R}$ ($\mathcal{G} = \mathbf{C}$, $\mathcal{G} = \mathbf{H}$) we say that the signal is *scalar* (*complex, quaternionic*). For technical reasons, it is also assumed that signals are in $L^2$ (that is, their modulus is square-integrable). For more information about quaternions and geometric algebra, please see the Appendix and [27].

### B. ANALYTIC SIGNAL

In 1946, D. Gabor proposed a complex-valued function (signal) called the *analytic signal* for removing negative frequencies [28]. Using the analytic signal instead of the original real-valued signal has mitigated estimation biases and eliminated cross-term artifacts due to the interaction of negative and positive frequencies [29]. However, the most interesting property of the analytic signal is its phase representation. In contrast to amplitude-based computer vision techniques, the phase-based methods are not sensitive to smooth shading and lighting variations [30], [31]. Moreover, beyond the global Fourier phase (not localized), the analytic signal encodes both local space and frequency characteristics of a signal simultaneously. Phase-based feature detection has been investigated extensively in the classic computer vision approach, as in [30], [32]–[34]. For a 1D real signal (function) $f(x)$, its analytical signal $f_A(x)$ is defined as follows [31]:

$$f_A(x) = f(x) - \boldsymbol{i} f_{\mathcal{H}}(x), \tag{1}$$

where $\boldsymbol{i} = \sqrt{-1}$ and $\mathcal{H}(f(x)) = f_{\mathcal{H}}(x)$ is the Hilbert transform of $f(x)$, namely

$$f_{\mathcal{H}}(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{\tau - x} d\tau. \tag{2}$$

The amplitude $A$ and the phase $\varphi$ of $f_A$ are defined by the following expressions:

$$A(x) = \sqrt{f^2(x) + f^2_{\mathcal{H}}(x)} \tag{3}$$

$$\varphi(x) = \arctan\left(\frac{f(x)}{f_{\mathcal{H}}(x)}\right). \tag{4}$$

The local phase computation needs an additional operator to enhance the localization of the features [30] which can be achieved by computing a filtered version of the signal $f(x)$ with an even function $f_e(x)$, as follows [30]:

$$f'_A(x) = f_e(x) * f(x) - \boldsymbol{i}\mathcal{H}(f_e(x) * f(x)), \tag{5}$$

where $*$ represents the convolution operator. and $\mathcal{H}$ is the *Hilbert transform*. According to [30], [31] the approximation filter $f_e$ must be a symmetric band pass filter. In practice, the approximation of the local phase and the local amplitude uses a pair of band-pass quadrature filters such as log-Gabor.

### C. MONOGENIC SIGNAL

The most accepted 2D generalization of the analytic signal, is the *monogenic signal*. Felsberg and Sommer in [35], proposed the monogenic signal, which satisfies the generalized Cauchy-Riemann equations of Clifford analysis using the $\mathcal{G}$ framework.

A monogenic signal $I_M = I_M(x, y) \in \langle 1, \boldsymbol{i}, \boldsymbol{j} \rangle \subset \mathbf{H}$ associated to an image $I = I(x, y) \in \mathbf{R}$ (where $x, y \in U$, $U$ a region of $\mathbf{R}^2$) is defined as follows [35]:

$$I_M = I' + I_R, \quad I_R = \boldsymbol{i} I_1 + \boldsymbol{j} I_2, \tag{6}$$

where the signals $I_1$ and $I_2$ are the *Riesz transforms* of $I$ in the $x$ and $y$ directions. In this work, a quadrature filter approximation was applied, by using $I' = g * I$, where $*$ is the convolution operator and $g = g(x, y)$ is a log-Gabor function (radial, isotropic, bandpass filter).

Rewriting the monogenic signal equations (adding quadrature filters) in the Fourier domain returns:

$$I_M = \mathcal{F}^{-1}(J' + J_R), \quad J_R = \boldsymbol{i} J_1 + \boldsymbol{j} J_2, \tag{7}$$

where $\mathcal{F}^{-1}$ is 2D inverse Fourier transform, $J = \mathcal{F}(I)$, $J' = J \cdot G$, $J_1 = J \cdot H_1 \cdot G$ , $J_2 = J \cdot H_2 \cdot G$, with:

$$H_1(u_1, u_2) = \frac{u_1}{\sqrt{u_1^2 + u_2^2}}, \tag{8}$$

$$H_2(u_1, u_2) = \frac{u_2}{\sqrt{u_1^2 + u_2^2}}, \tag{9}$$

$$G(u_1, u_2) = \exp\left(-\frac{\log\left(\frac{\sqrt{u_1^2+u_2^2}}{\omega^s}\right)^2}{2\log(\sigma)^2}\right), \tag{10}$$

$$\omega^s = \frac{1}{\omega f^{s-1}}. \tag{11}$$

Here, $u_1, u_2$ are the frequency components, $\sigma$ is the variance of the log-Gabor, $\omega$ is the minimum wavelength, $f$ is a scale factor, and $s$ is the current scale.



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



The *local amplitude* $A_M = A_M(x, y)$ is defined by the expression [35]:

$$A_M = \sqrt{I'^2 + I_1^2 + I_2^2}. \tag{12}$$

The *local phase* $I_\phi$ and the *local orientation* $I_\theta$ associated to $I'$ are defined, again following [35], by the relationships

$$I_\phi = \text{atan2}\left(\frac{I'}{|I_R|}\right), \tag{13}$$

$$I_\theta = \text{atan}\left(\frac{-I_2}{I_1}\right), \tag{14}$$

where $|I_R| = \sqrt{I_1^2 + I_2^2}$ and the signal quotients are taken pointwise. The geometric interpretation of the monogenic signal is depicted in Figure 1. Note how changes in the pixel value intensity $I'$ do not affect the local phase $I_\theta$ and the local orientation $I_\phi$ values. This theoretically invariant response to large illumination changes to the local phase is in line with that reported in [19], [30]

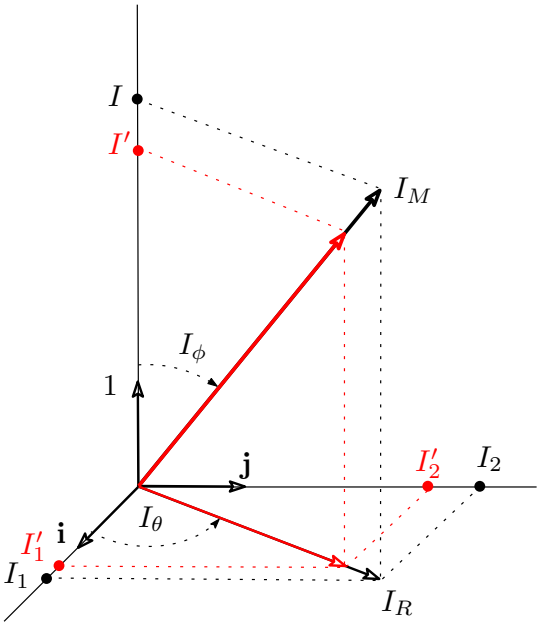



**FIGURE 1:** Geometry of the monogenic signal. The changes in $I'$ representing one pixel value intensity doesn't affect the value of $I_\theta$ and $I_\phi$.

### D. BIO-INSPIRED PROPERTIES AND TOOLS

This subsection highlights the main properties of the V1 cortex layer and how they are reflected in the functionality of M6. The reasons for choosing two bio-inspired technical tools as key ingredients as guides in its design and construction are also outlined.

V1 cells form the first layer of the hierarchical cortical processing [36]. The analogy for M6, quite literally, is that it is meant to be the first layer of a ConvNet.

Moreover, V1 neurons respond vigorously only to edges (odd-signals) and lines (even-signals) at a particular spatial direction through their orientation columns [31]. The counterpart of this in M6 is realized by computing the local phase and orientation, as these have the capability of detecting lines and edges and their orientations. This also justifies placing M6 as the first layer, as the features in question are the most primitive, and hence, it would be the not very effective use the M6 layer in a deeper position.

In [37], the authors reported a cortical adaptation to brightness and darkness in the primary visual cortex V1 of a macaque. In relation to this, it is previously mentioning in [26], [30], [31] that the local phase and local orientation should be robust (invariant) with respect to contrast changes due to its geometry representation.

Daugman [38] discovered that the Gabor functions resembled the experimental findings of Hubel and Wiesel [14] on orientation selectivity of the visual cortical neurons of cats. However, Gabor functions (filters) may cause fairly poor pattern retrieval accuracy in certain applications (see [39]) because they have, for certain bandwidths, an undesirable non-zero value of the so-called DC component (cf. [39]). For pattern recognition applications, this DC component entails that a feature can change with the average value of the signal. Fortunately, this weakness can be overcome with log-Gabor filters, as explained in [39], and this is why they are use them henceforth.

The other bio-inspired tool is the HSV color space. Although the main virtue of this space is that it best fits the human perception of color [40], it is used in this work as a means to geometrically code the phase and orientation in the Hue channel. This is quite natural, as the purpose of this channel is to hold a phase. The use of this transformation in the M6 layer results in an increase of classification accuracy (see Figure 3).

## IV. MONOGENIC LAYER M6

Monogenic signal characteristics (contrast invariance based on its geometry, feature extraction in frequency domains, and its similarities to the V1 properties) are the main stimuli for the design of our M6 trainable (top) layer.

A scheme of the computational flow of the M6 unit on a one-channel image as input is displayed in Figure 2. For color images ($I_c$) the mean value of the channels were computed and used as the input value. The convolution operations are carried out in the Fourier domain, that is, on $\mathcal{F}(I)$, where $I$ is the input image. This is to allow a straightforward implementation of the monogenic signal and avoid the problem of having to select a convolution kernel size proceeding instead to with a band-pass filter.

It is fundamental to note that M6 has only four trainable parameters $(s, f, \omega, \sigma)$, which are the log-Gabor function parameters (eq. (11)). Remember that $s$ is the current scale indicator; $\sigma$, the variance of the log-Gabor; $\omega$, the minimum wavelength; and $f$ a scaling factor.

Although the local phase $I_\phi(x, y)$ (Eq. 13) and local orientation $I_\theta(x, y)$ (Eq. 14) are theoretically invariant to illumination changes, it was found that inserting additional procesing operations is beneficial to better mimic the V1 behavior and increase classification performance, as explained



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



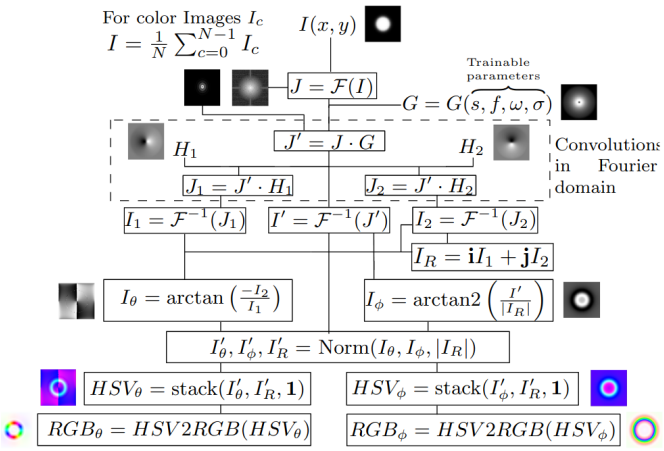



**FIGURE 2:** Computational flow of an M6 unit layer and examples (images) of the computation outputs using a white circle as input images. See the text below for details.

below. Figure 3 presents an example of how performance increases by leveraging the RGB phases specially for shallow architectures such as A1 (see Figure 10 for the ConvNet architecture definition).

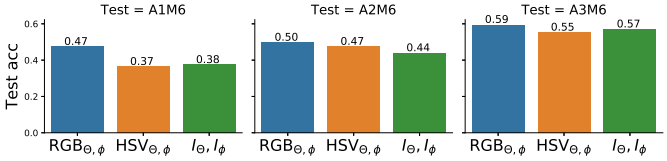



**FIGURE 3:** Test accuracy values over the three models with CIFAR-10 using different elements of M6: RGB output, HSV output and the phases.

After the local phase is computed, a normalization step, $I'_\theta, I'_\phi, I'_R = \text{Norm}(I_\theta, I_\phi, I_R)$ is added, where Norm is generically defined by the following expression:

$$\text{Norm}(I) = \frac{I(x,y) - \min(I(x,y))}{\max(I(x,y)) - \min(I(x,y))}. \qquad (15)$$

Next, using the polar geometry of the HSV color space, the normalized local phase $I'_\phi$ and local orientation $I'_\theta$ are stacked in a hue channel (H) of a HSV color space with $I'_R$ as the saturation (S) and the constant matrix $\mathbf{1}_{[m,n]}$ as the value component (V). Finally, the HSV images are converted by the standard function $HSV2RGB$ into RGB images (see [41, page 304]).

As a result, an M6 unit produces six output feature maps per channel image input. Figure 4 depicts an example of M6 outputs using a white circle as the input image.

For backpropagation, the automatic differentiation and weight updating implemented by Tensorflow 2 (TF) were used, although their symbolic expression is not reproduced here because it lies beyond the scope of this paper. Instead, two graphics that document the learning process are provided. Figure 5 illustrates how the M6 trainable parameters are adjusted during the training process using the MNIST dataset. Note that the major changes in the parameters take place in the first fifty epochs. Figure 6 displays the accuracy

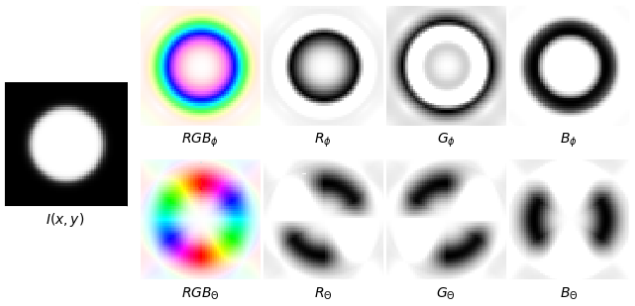



**FIGURE 4:** The white circle is the input image ($I(x,y)$) and $RGB_\phi$ and $RGB_\theta$ are the M6 outputs. The grey images are the corresponding RGB components.

and loss in the training and validation processes corresponding to the same training job. The validation loss (val loss) and validation accuracy (val acc) undergo major changes before the fiftieth epoch, thus matching what was found in Figure 5. Validation loss rises slightly after the fiftieth epoch, a sign that the training has entered the overfitting regime. In addition, Figure 7 shows an example from another perspective, namely a CIFAR-10 image and the associated pre and post-training feature maps, revealing that the post training feature maps are sharper. This behavior is expected inasmuch as the trainable parameters define the band pass size of the log-Gabor filter.

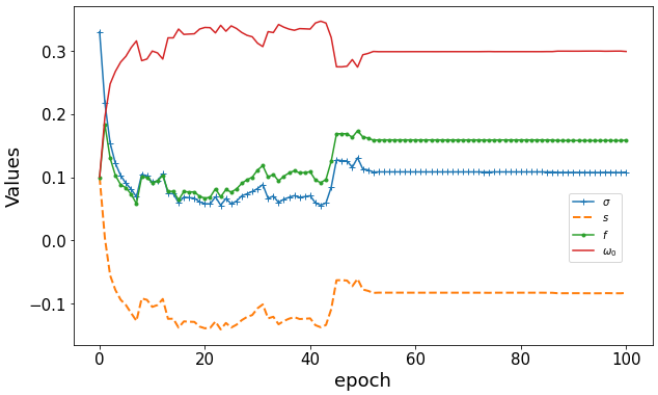



**FIGURE 5:** Example of the evolution of the M6 trainable parameters $(s, f, \omega, \sigma)$ on the MNIST dataset.

### *A. M6 PROPERTIES AND COMPARISON WITH A REGULAR CONVOLUTION LAYER*

Table 1 summarizes a comparison of some characteristics of a 2D convolutional layer with those of the proposed layer $M6$. An important differences between a 2D conventional ConvNet (C) and the proposed layer M6 is that the convolutions in the latter are carried out in the Fourier domain. This feature helps to avoid the problem of selecting a convolution kernel size depending on the input-image size, as the convolution in Fourier domain is a pointwise multiplication and the bandpass filter size is learned with the log-Gabor parameters. Furthermore, the number of trainable M6 parameters (four) is significantly lower than the number of C parameters



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



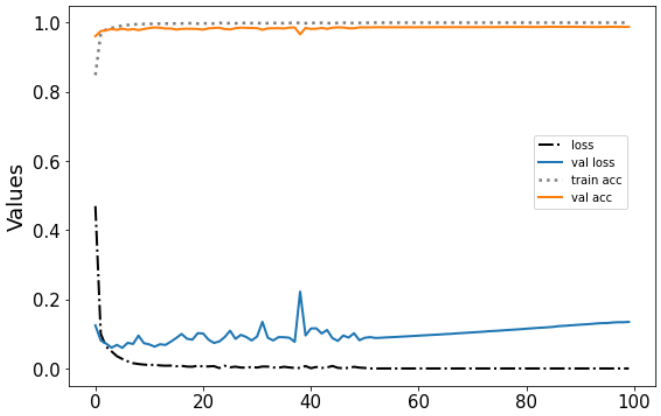



**FIGURE 6:** Example of the evolution of the loss, validation loss, accuracy and validation accuracy on the MNIST dataset and the M6 layer.

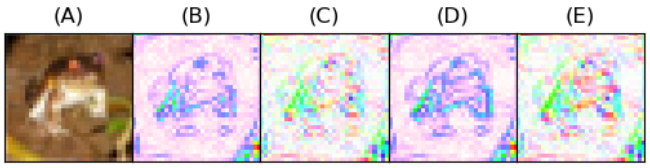



**FIGURE 7:** Example of the activations of the M6 unit before and after training. (A) Input image; (B) and (C), activations $RGB_\phi$ and $RGB_\theta$ before training; (D) and (E), the same activations after training. Note that (D) and (E) are shaper than (B) and (C), respectivetly.

(weights), which depends on the number and size of their filters. For example, for six C units with $3 \times 3$ filters, the number of parameters is $6 \times 3^2 = 54$. On the other hand, if both C and M6 can detect lines (l) and edges (e), C can usually detect more features, such as corners (c) or some irregular shapes, whereas M6 is also crucially sensitive to orientations (thus resembling V1). The orientation response is important to obtain a robust performance even with rotated images. Another important difference is the activation function. The ReLU (Rectifying Linear Unit) function is often used in C instances [42], while in M6 use $\arctan$ and $\arcsin$ as activation function. This notwithstanding, both M6 and C behave similarly with respect to all the tested optimizers (for instance, SGD, ADAM, and NADAM). An important remark is that M6's trainability sets it apart from deterministic (pre-processing) layers used in other systems, as stressed in Section II.

**TABLE 1:** Comparison of a regular 2D convolutional layer C, and the M6 layer. The abbreviations: l, e,c stand for lines, edges, and corners, respectively

| Characteristics | C | M6 |
|---|---|---|
| Convolution domain | Space | Frequency |
| Parameters | (3@3)@6=54 | 4 |
| Feature elements | l, e, c, etc | l and e |
| Oriented response | No | Yes |
| Nonlinear function | ReLU | $\arctan$, $\arcsin$ |
| Layer position | Any | First |
| Trainable | Yes | Yes |

## V. EXPERIMENTAL SETUP

The experimental setup is designed to evaluate the robustness and classification performance under different datasets (simple to complex task, different input shapes), architectures (shallow to deep) and degradations (three different types of contrast).

### A. DATASETS

Four datasets were used: MNIST [43], Fashion-MNIST (f-MNIST) [16], CIFAR-10 [17], and Dogs vs Cats (DvsC) [18]. All datasets are available online through the Tensorflow datasets (tfds) [44] package. The haze degradation dataset is available through the Gitlab link provided below. Table 2 shows how the datasets were split and their main characteristics.

**TABLE 2:** Split size and input shape of the datasets.

| | MNIST | f-MNIST | CIFAR-10 | DvsC |
|---|---|---|---|---|
| Training set | 48,000 | 48,000 | 40,000 | 16,284 |
| Validation set | 12,000 | 12,000 | 10,000 | 3,489 |
| Test set | 10,000 | 10,000 | 10,000 | 3,489 |
| Total | 70,000 | 70,000 | 60,000 | 23,262 |
| Input shape | [28x28x1] | [28x28x1] | [32x32x3] | [128x128x3] |

### B. DEGRADING PROCEDURES

Three contrast transformations were used to degrade images $I$. The max-min scale transformation, $C_S I$; The TF contrast transformation, $C_{TF} I$; and the haze transformation, $C_H I$. Figure 8 displays an example of three degradation levels $d_j$ ($j = 1, 2, 3$) for each degradation procedure applied to the image in column $d_0$ (no degradation).

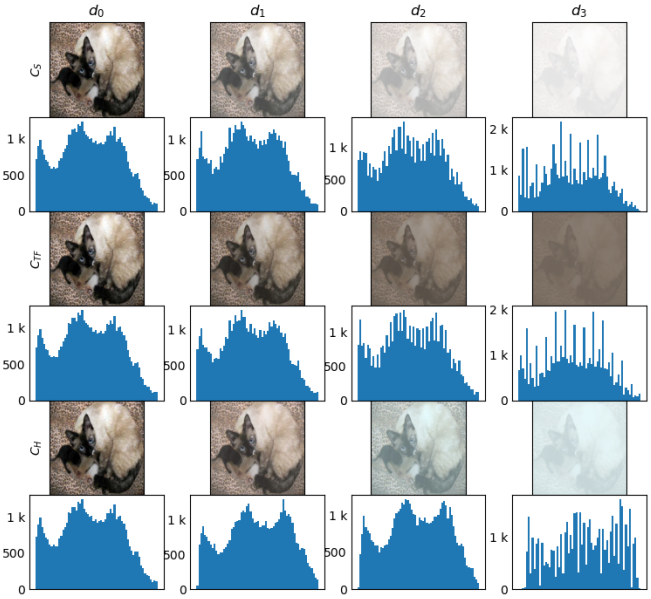



**FIGURE 8:** Examples of the three contrast degradation procedures $C_S, C_{TF}, C_H$ and the corresponding image histograms. The original image (column $d_0$, no degradation) and the rest of the columns represent each degradation level $d_j$ ($j = 1, 2, 3$).



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



*Max-min Scale transformation ($C_S$).* Applying this to an image $I(x, y)$ with respect to an interval $S = [a, b]$, it produces an image $C_S I(x, y)$, as follows:

$$C_S I(x, y) = a + \frac{(I(x, y) - \min(I(x, y))(b - a)}{\max(I(x, y)) - \min(I(x, y))}. \quad (16)$$

The four degradation levels $d_j$ corresponding to the intervals $S$ defined by $a = 0.3, 0.7, 0.9$ and $b = 1$, respectively.

*Contrast degradation using $C_{TF}$.* This is defined by

$$C_{TF} I(x, y) = \mu + F(I(x, y) - \mu), \quad (17)$$

where $\mu$ is the mean value of the input image $I$ and $F \in [0, 1]$ is a contrast factor. The four degradation levels $d_j$ correspond, respectively, to the values $F \in \{0.7, 0.3, 0.1\}$.

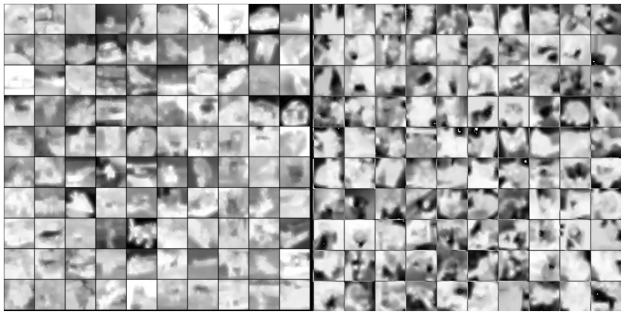
*Contrast degradation by haze $C_H$.* This method works by adding non-uniform haze to a given image. The main basis for gauging haze and fog in images stems from the atmospheric scattering model proposed by [45], which can be summarized, following [46], by the equation

$$C_H I(x, y) = t(x, y)I(x, y) + (1 - t(x, y))A(x, y). \quad (18)$$

In this expression, $I(x, y)$ denotes what the image would be if haze were removed and $C_H I(x, y)$ is the measured (observed) hazed image. The rationale behind (18) is that $I(x, y)$ undergoes an attenuation $t(x, y)I(x, y)$ caused by the *medium transmission* $t(x, y) \in [0, 1]$, which measures the fraction of light reaching the camera from the $(x, y)$ direction. Its value is 1 for perfectly transparent air and 0 when no light from $(x, y)$ reaches the camera. The term $A(x, y)$ denotes the *total atmospheric light*, and hence $(1 - t(x, y))A(x, y)$ measures the fraction of light contributing to $C_H I(x, y)$ not originating from the source $(x, y)$, usually produced by scattering and reflection processes.

Now the main point of contrast degrading $C_H$ is using equation (18), which means generating various values of $A$, and using these $I(x, y)$ to estimate $t(x, y)$ using the dark channel prior proposed in [46]. In the experiments, the degradation levels $d_j$ are defined by choosing, respectively, $S = [a = 0.5, b = 0.8], [a = 0.3, b = 0.5], [a = 0, b = 0.15]$.

It remains to be seen how to generate the various $A$s. The components of an RGB image $I$ ($I^r$, $I^g$, $I^b$) allow handling $I(x, y)$ as a 3-vector, $I(x, y) = [I^r(x, y), I^g(x, y), I^b(x, y)]$. In this case, $A(x, y)$ must also be a 3-vector, say $A(x, y) = [A^r(x, y), A^g(x, y), A^b(x, y)]$, and as a result, $C_H I(x, y)$ can be treated in the same way. The generation of $A$ vectors is done here by choosing each channel $^c$ independently and randomly in the interval $[0.8, 1]$, $A^c \in [0.8, 1]$. Note that the contribution of $A$ is not limited to the light intensity, as it produces changes in color, a fact that is in accordance with the physical effects of the atmospheric light. See in Figure 9 and example of the transmission maps estimation of CIFAR-10 and DvsC.



(a) (b)

**FIGURE 9:** Examples of transmission map estimation $t(x, y)$. (a) Transmission maps of 100 images from CIFAR-10; (b) Transmission maps of 100 images from DvsC dataset.

### C. ARCHITECTURES, TRAINING AND TEST

The three architectures used in this work are labeled A1, A2, and A3. A1 is a shallow ConvNet; A2, a medium depth ConvNet; and A3, a deep ConvNet that which a ResNet20 [47] as a subnet or backbone. Each of these architectures actually stands for tree: one, AjC, topped by a standard 2D convolutional layer (C), other, AjQ9, topped by a Q9 layer defined at [19] and AjM6, topped by the M6 layer. See Figure 10 for details on each of them. The classification robustness of the M6 layer was compared against the conventional ConvNet layer (kernel $3 \times 3$, with six output channels and no data augmentation in the training process). The $K$ in the last layer (with a softmax function), means the size of the output layer (number of classes), which is 10 in for CIFAR-10, MNIST, and f-MNIST and 2 for DvsC. Note that the large depth of A3 does not allow processing small size images such as those from MNIST and f-MNIST.

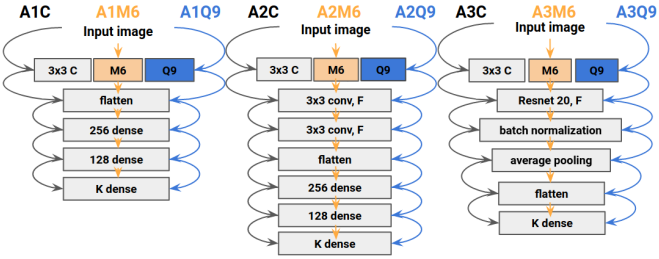



**FIGURE 10:** Triple architectures (AjC, AjM6, AjQ9) used for training, validation, and testing. The input image shapes are described in Table 2. The terminology used follows the conventions of the TF framework. For more details, see the source code available through the link provided below.

To evaluate the response robustness, the models were trained with a specific data degradation and tested not only with the same data degradation but also with three additional data degradations. This kind of experimental setup is common in other works to evaluate robustness, such as [2], [19], [24]. The experimental setup concerns the aforementioned four primary datasets, the nine networks summarized in Figure 10, and the three contrast degradation methods. For each degradation method and each primary dataset, three additional datasets were constructed by applying the degradation



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



method to the primary dataset with degradation levels $d_1$, $d_2$, and $d_3$. Then, one of the networks was trained four times, one for each degraded datasets, including the primary degradation level $d_0$. Note that all training was done from scratch. Finally, each of the four trained classifiers was tested on the four degraded datasets. These arrangements are sketched in Table 3. As indicated above, the $A3$ networks can be used only on CIFAR-10 and DvsC data.

**TABLE 3:** Sketch of the experimental setup, where Tr denotes training.

| $C_{TF}$ | | $C_S$ | | $C_H$ | |
|---|---|---|---|---|---|
| Tr | Test | Tr | Test | Tr | Test |
| $d_0$ | $d_0, d_1, d_2, d_3$ | $d_0$ | $d_0, d_1, d_2, d_3$ | $d_0$ | $d_0, d_1, d_2, d_3$ |
| $d_1$ | $d_0, d_1, d_2, d_3$ | $d_1$ | $d_0, d_1, d_2, d_3$ | $d_1$ | $d_0, d_1, d_2, d_3$ |
| $d_2$ | $d_0, d_1, d_2, d_3$ | $d_2$ | $d_0, d_1, d_2, d_3$ | $d_2$ | $d_0, d_1, d_2, d_3$ |
| $d_3$ | $d_0, d_1, d_2, d_3$ | $d_3$ | $d_0, d_1, d_2, d_3$ | $d_3$ | $d_0, d_1, d_2, d_3$ |

The hyperparameters used in the experimental setup are a learning rate of 0.001, 100 epochs, a batch-size of 128, ReLU as an activation function for C, and ADAM as an optimizer. Keras-turner (Bayesian optimization) [48] was employed to choose the kernel size of the first 2D convolutional layer and learning rate of the ConvNet. The initial parameters of M6 layer are $s = 1$, $f = 1$, $\sigma = 0.33$, and $\omega = 1$ based on the previous version of the layer [26]. All datasets were normalized to one by dividing each pixel value by 255. The TF 2.1 deep learning framework, run on a V-100 Nvidia-GPU for all experiments, was used. Its reproducibility is supported by the supplemental material uploaded at the Gitlab link M6 project.

### D. METRICS FOR ANALYSIS

A Structural Similarity Index Measure (SSIM) index was chosen in order to compare the effects of the degradation procedures on the M6 feature maps (in Section VI). This index was selected because it allowed numerically comparing the feature maps changes made by the degradations, taking into account some key aspects of human perception [49]. Moreover, SSIM could quantify the image quality as the *perceived* changes in the structural information (SSIM *map*), and at the same time, SSIM takes the luminance and contrast changes into account. The SSIM index formula is defined as follows [49] :

$$\text{SSIM}(x_1, y_1) = \frac{(2\mu_{x_1}\mu_{y_1} + c_1)(2\sigma_{x_1 y_1} + c_2)}{(\mu_{x_1}^2 + \mu_{y_1}^2 + c_1)(\sigma_{x_1}^2 + \sigma_{y_1}^2 + c_2)}, \quad (19)$$

where $x_1$ and $y_1$ are two arrays of size $N \times N$, $\mu_{x_1}$, $\mu_{y_1}$ are the averages, and $\sigma_{x_1}^2$ and $\sigma_{y_1}^2$ the variances, of $x_1$ and $y_1$, respectively, while $\sigma_{x_1 y_1}$ is the covariance of $x_1$ and $y_1$, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$, with $L$ being the dynamic range, $k_1 = 0.01$ and $k_2 = 0.03$. The SSIM value belongs to the interval $[0, 1]$, with SSIM$= 1$ corresponding to maximum similarity and SSIM$= 0$ to minimum similarity.

In addition, the Peak Signal to Noise Ratio (PSNR) was used to evaluate the robustness of the proposed layer. The PSNR is used to calculate the ratio between the maximum possible signal (image) power and the power of corrupted or degraded image [50]. PSNR could be defined via the Mean Squared Error (MSE). Given a $m \times n$ monochrome image $I$ and its corrupted approximation $K$, The MSE is defined as:

$$\text{MSE} = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \quad (20)$$

The PSNR (in decibel) is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (21)$$

where the $\text{MAX}_I^2$ is the maximum possible pixel value of the image; for instance, when the pixels are represented using 8 bits per sample, this is 255. In general, the higher the PSNR value, the better. Note that for two equal images, the MSE will be zero and the PSNR value will be infinite.

## VI. RESULTS AND ANALYSIS

Figures 11, 12, and 13 present a grid of *boxen* plots of the classification testing-accuracy values according to the experimental setup. Each box represents five numbers: minimum (botton whisker), first quartile $(Q_1)$, median $(Q_2)$, third quartile $(Q_3)$, and maximum (top whisker). C and Q9 were compared against M6 (grid-columns) using four datasets (box-color), three architectures (grid-rows), and four degradation levels (using $C_S$, $C_{TF}$, and $C_H$, respectively). Each box represents four test accuracy values associated with $d_0, d_1, d_2, d_3$. For instance, the A1C model, trained with $d_0$ for MNIST, has four testing accuracy values using $C_S$: $d_0 = 0.986$, $d_1 = 0.953$, $d_2 = 0.225$, and $d_3 = 0.089$. Their box plot reflects the computed (quartile) values $Q_0 = 0.089$, $Q_1 = 0.191$, $Q_2 = 0.590$, $Q_3 = 0.961$ and $Q_4 = 0.986$. The dashed lines represent the maximum test accuracy value over all layer-models (C, Q9, or M6) for each dataset. The missing boxes for MNIST and f-MNIST with $A3$ are due to their input sizes are incompatible with that architecture's greater depth. The numerical results can be found in the tables included in the *Jupyter notebooks* at M6 project/example results.

On a wider angle, the analysis of these results can be divided into two major classes: i) evaluating the robustness using the size of the box and whiskers, and ii) evaluating the maximum performance with the top whisker. For i), these tests highlight that the ConvNets with the M6 present the smallest boxes for all cases. These results provide indisputable evidence for the robustness effect of M6 under different contrast degradations. Please note that training was performed on each degradation level and testing with all degradation levels. According to these results, M6 can be trained with any degraded level and yet achieving almost the same performance. This is in marked contrast to the C models, which have quite low performance when trained with $d_3$ degraded images. For ii) dashed lines were added to facilitate view of the maximum values for each model type (A1, A2, and A3) and each dataset. Maximum performance



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



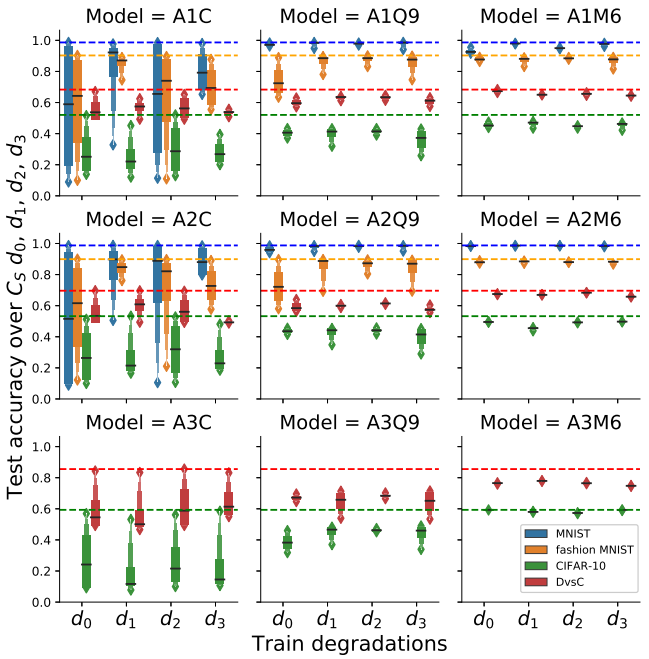



**FIGURE 11:** Grid of box plots representing test accuracy over $C_S$ degradations for all models and dataset combinations. The size of each box (and whiskers) evaluate robustness against contrast degradations. The dashed lines represent the maximum test accuracy value over all layer models for each dataset. See the first paragraph of section VI for more details.

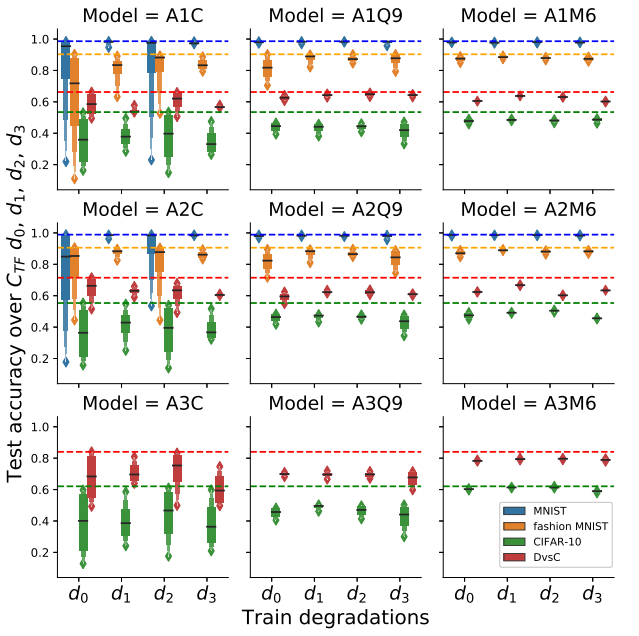



**FIGURE 12:** Grid of box plots representing test accuracy over $C_{TF}$ degradations for all models and datasets combinations. The size of each box and whiskers evaluate robustness against contrast degradations. The dashed lines represent the maximum test accuracy value over all layer models for each dataset. See the first paragraph of section VI for more details.

occurs at 17/26 for C, at 3/26 for Q9, and at 6/26 for the M6 models. However, ConvNets with M6 have the closest performance to the maximum test accuracy value, easily confirmed this by computing the square of the difference with respect to each maximum over all models, resulting in 29.3, 4.7, and 1.0 for the C, Q9, and M6 architectures, respectively. It is important to note that the maximum values from the C models are generally achieved when the training uses the same data degradation as the test dataset. However, for the maximum degradation level $d_3$, the classification performance is significantly lower, even with the same data degradation level as that used for training.

The results also reveal parallel behavior patterns between the C and M6 models. For instance, all models have higher test accuracy performance with simple datasets (MNIST and f-MNIST). In addition, performance tends to increase with the depth of the architecture. However, it is important to remark that the single M6 unit, which has only four weights, can extract enough features for different types of datasets, in contrast to the C layer with six convolution units and fifty-four weights. Moreover, the increase in test accuracy of M6 in relation to the increase of the number of layers can be ascribed to the fact that the output features of M6 can be learned and combined by the later layers, similar to how they are processed by a conventional convolutional layer. Combining the robustness and the maximum classification performance of M6 leads to boosted efficacy of those layers when fed with the M6 output rather than having to elicit it from a raw input. In other words, M6 delivers sharper features for image classification.

As stressed in the description of M6, the robust performance stems from the geometry behind its design. One way to evaluate (numerically) its resilience to contrast changes is by using the SSIM index and PSNR. The main idea is to compute the SSIM and PSNR for $d_0$ and $d_j$ images ($j = 1, 2, 3$) and compare it to the two SSIM indexes of the image transformed by M6 (which is regarded as two images, $RGB_\theta$ and $RGB_\phi$). The SSIM index comparison is even more forceful if the indices are displayed together with the SSIM maps. In the top row of of Figure 14, for example, SSIM is applied to a $d_0$ image and its $d_3$ degradation by $C_S$, and it is not surprising to find a low index, 0.31, and a discrepancy map that is quite close to the original image. In the bottom row, SSIM is applied to the same images after being transformed by M6, and it can be seen that the two indices are high (close to 1) and that the discrepancy maps have very small values, which means that M6 drastically reduces the discrepancy between M6($d_0$) and M6($d3$). This behavior is a compelling evidence of M6's robustness capability and suggests the potential of layers designed on similar, possibly more general principles.

Figure 15 shows a box plot of the results of a similar computation, but this time involving 1000 random images, three degradation methods ($C_S, C_{TF}, C_H$), and three degradation levels $d_j$ ($j = 1, 2, 3$). It is clear that the box plot's M6 transformation values are more compact and quite close to one. This confirms and explains the findings the in the preceding example, namely that M6 tends to see the $d_j$ degraded images as a $d_0$ image. PSNR results are presented in



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



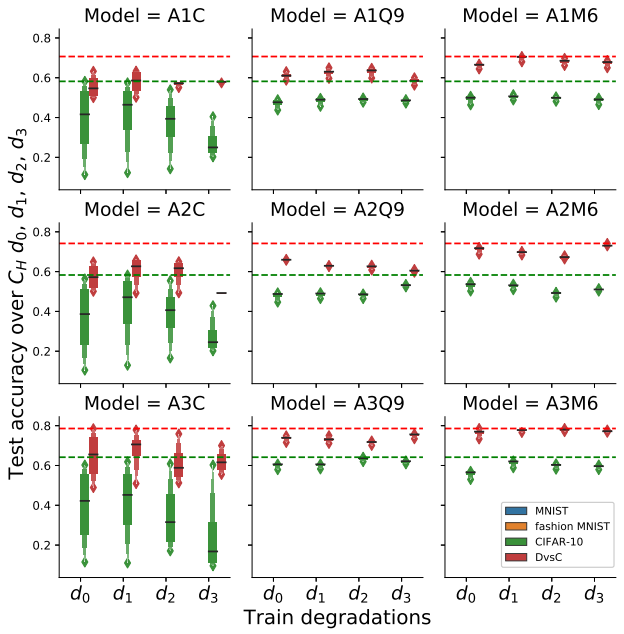



**FIGURE 13:** Grid of box plot representing test accuracy over $C_H$ degradation for all model and datasets combination. The size of each box (and whiskers) evaluate robustness in front of the contrast degradations. The dashed lines represent the maximum test accuracy value over all layer models for each dataset. See the first paragraph of section VI for more details.

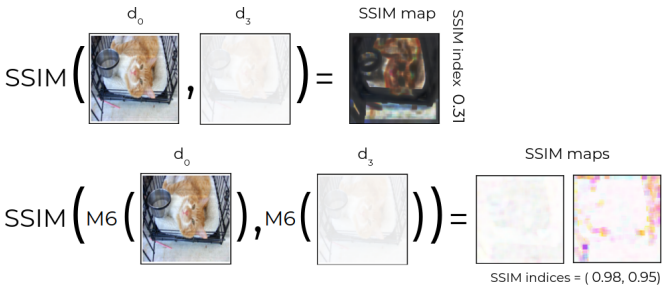



**FIGURE 14:** Example of SSIM map and index when applied to $d_0$ and $d_3$ degradation (top row) and the SSIM maps and indexes of M3($d_0$) and M3($d_3$). See the text above for the interpretation of these data.

Figure 16. The higher values of PSNR represent more similar (more robust) images, and for M6's feature maps, the PSNR values are significantly higher than the degraded images.

In order to compare the performance analysis of M6 against C, the time and memory consumption in training and testing were averaged, showing that M6 achitectures spend 7% more time on training and 26% more time on testing and consume five times as much memory. This difference could be attributed to the fast Fourier transform computation. Further work will be needed to optimize memory consumption. Note also, its performance cannot be compared to that of the Q9 performance because its code is only available for CPU, while the M6 runs on GPU.

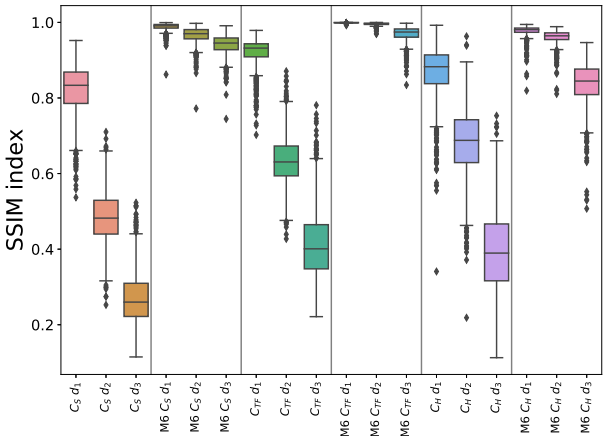



**FIGURE 15:** SSIM index of a thousand $d_0$ images with their degradations $d_j$ $(j = 1, 2, 3)$ using $C_S, C_{TF}, C_H$, together with similar computations after being transformed by M6. Note that SSIM$= 1$, corresponding to maximum similarity, and SSIM$= 0$, to minimum similarity.

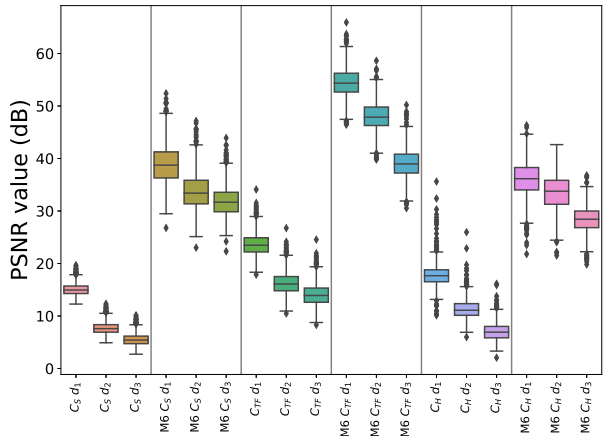



**FIGURE 16:** PSNR value of thousand $d_0$ images with their degradations $d_j$ $(j = 1, 2, 3)$ using $C_S, C_{TF}, C_H$, together with similar computations after being transformed by M6. Note that in general, the higher PSNR the better.

## VII. CONCLUSIONS AND FUTURE WORK

A new trainable bio-inspired front-end layer for ConvNets has been designed and presented. This new layer generates a 3D geometric representation of each pixel value by computing the quaternion monogenic signal in the Fourier domain. As a result, it is possible to leverage the local phase and the local orientation to elicit low-level geometric features, such as oriented lines or edges. Couplingaaa the proposed layer with a regular ConvNet, or with dense networks, achieves an image classification that is little affected by severe contrast degradings and which, on the whole, has better accuracy. The experimental results are consistent with the SSIM and PSNR results and the geometrical observation that the local phase and the local orientation are invariant to variable contrast conditions.

Concerning the impact of this work, it is to be searched



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



in situations where an invariant response to contrast alterations is required. Among the possible scenarios we count self-driving cars under haze conditions, surface glazes in medical images (biopsies), or day-round autonomous video surveillance. The authors hope that this research will serve as one of the lines for future studies on equivariant and invariant representations by ConvNets. In addition, we will try to compare or combine this layer with other approaches such as deformable ConvNets [51] or depthwise ConvNets [52], among others. Moreover, further work is needed to compare the functionality of this approach to other methods and techniques, such as object detection or segmentation.

.

## APPENDIX A QUATERNION ALGEBRA

Hamilton's quaternions have many representations in geometric algebra }. It is possible to represent the Hamilton quaternions as the even subalgebra of $\mathcal{G}(3,0)$ with basis $\{1, e_2e_3, e_3e_1, e_1e_2\}$ and may be seen by identifying $\boldsymbol{i} \mapsto -e_2e_3$, $\boldsymbol{j} \mapsto -e_3e_1$ and $\boldsymbol{k} \mapsto -e_1e_2$. This work uses the most widely-known definition: quaternion algebra $\mathbf{H}$ is a four dimensional real vector space with basis $1, \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$,

$$\mathbf{H} = \mathbf{R}1 \oplus \mathbf{R}\boldsymbol{i} \oplus \mathbf{R}\boldsymbol{j} \oplus \mathbf{R}\boldsymbol{k} \tag{22}$$

endowed with the bilinear product (multiplication) defined by Hamilton's relations, namely

$$\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1. \tag{23}$$

As it is easily seen, these relations imply that

$$\boldsymbol{ij} = -\boldsymbol{ji} = \boldsymbol{k}, \quad \boldsymbol{jk} = -\boldsymbol{kj} = \boldsymbol{i}, \quad \boldsymbol{ki} = -\boldsymbol{ik} = \boldsymbol{j}. \tag{24}$$

The elements of $\mathbf{H}$ are named quaternions, and $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$, quaternionic units. By definition, a quaternion $q$ can be written in a unique way in the form

$$q = a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}, \quad a, b, c, d \in \mathbf{R}. \tag{25}$$

Its conjugate, $\bar{q}$, is defined as

$$\bar{q} = a - (b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}). \tag{26}$$

Note that $(q+\bar{q})/2 = a$, which is called the real part or scalar part of $q$, and $(q - \bar{q})/2 = q - a = b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}$, the vector part of $q$.

Since the conjugates of $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are $-\boldsymbol{i}, -\boldsymbol{j}, -\boldsymbol{k}$, the relations (23) and (24) imply that the conjugation is an antiautomorphism of $\mathbf{H}$, which means that it is a linear automorphism such that $\overline{qq'} = \bar{q}'\bar{q}$. Using Hamilton's relations again, we easily conclude that

$$q\bar{q} = a^2 + b^2 + c^2 + d^2. \tag{27}$$

This allows to define the modulus of $q$, $|q|$, as the unique non-negative real number such that

$$|q|^2 = q\bar{q}. \tag{28}$$

Observe that $|qq'| = |q||q'|$. Indeed, $|qq'|^2 = qq'\overline{qq'} = qq'\bar{q}'\bar{q} = q|q'|^2\bar{q} = |q|^2|q'|^2$.

Finally, for $q \neq 0$, $|q| > 0$ and $q(\bar{q}/|q|^2) = 1$, which shows that any non-zero quaternion has an inverse and therefore that $\mathbf{H}$ is a (skew) field.

## ACKNOWLEDGMENT

The authors would like to thank to CONACYT and Barcelona supercomputing Center. Sebastián Salazar-Colores (CVU 477758) would like to thank CONACYT (Consejo Nacional de Ciencia y Tecnología) for the financial support of his PhD studies under Scholarship 285651.

## REFERENCES

[1] E. T. Rolls and S. M. Stringer, "Invariant visual object recognition: a model, with lighting invariance," Journal of Physiology-Paris, vol. 100, no. 1-3, pp. 43–62, 2006.

[2] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019. https://arxiv.org/pdf/1903.12261.pdf.

[3] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, pp. 1–6, IEEE, 2016.

[4] Y. Wang, F. Fu, F. Lai, W. Xu, J. Shi, and J. Wang, "Efficient road specular reflection removal based on gradient properties," Multimedia Tools and Applications, vol. 77, no. 23, pp. 30615–30631, 2018.

[5] M. Halicek, H. Fabelo, S. Ortega, J. V. Little, X. Wang, A. Y. Chen, G. M. Callico, L. Myers, B. D. Sumer, and B. Fei, "Hyperspectral imaging for head and neck cancer detection: specular glare and variance of the tumor margin in surgical specimens," Journal of Medical Imaging, vol. 6, no. 3, p. 035004, 2019.

[6] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in International Conference on Document Analysis and Recognition (ICDAR), p. 958, IEEE Computer Society, 2003.

[7] A. Hernández-García and P. König, "Data augmentation instead of explicit regularization," 2018. https://arxiv.org/pdf/1806.03852.pdf.

[8] T. Cohen and M. Welling, "Group equivariant convolutional networks," in International conference on machine learning, pp. 2990–2999, 2016.

[9] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," Advances in neural information processing systems, vol. 30, p. 3239, 2017.

[10] T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Ré, "A kernel theory of modern data augmentation," in International Conference on Machine Learning, pp. 1528–1537, PMLR, 2019.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proceedings of the Conference Neural Information Processing Systems (NIPS 2011), pp. 1097–1105, 2012. http://www.image-net.org/.

[12] M. Rad, P. M. Roth, and V. Lepetit, "ALCN: Adaptive Local Contrast Normalization," Computer Vision and Image Understanding, vol. 194, p. 102947, 2020.

[13] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3238–3247, 2017.

[14] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," The Journal of physiology, vol. 160, no. 1, pp. 106–154, 1962.

[15] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in Neural networks: Tricks of the trade, pp. 9–50, Springer, 1998.

[16] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," CoRR, vol. abs/1708.07747, 2017.

[17] A. Krizhevsky, G. Hinton, et al., "Learning multiple layers of features from tiny images," 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf.

[18] J. Elson, J. J. Douceur, J. Howell, and J. Saul, "Asirra: A captcha that exploits interest-aligned manual image categorization," in Proceedings of 14th ACM Conference on Computer and Communications Security (CCS), Association for Computing Machinery, Inc., October 2007.



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access



[19] E. U. Moya-Sánchez, S. Xambó-Descamps, A. S. Pérez, S. Salazar-Colores, J. Martínez-Ortega, and U. Cortés, "A bio-inspired quaternion local phase cnn layer with contrast invariance and linear sensitivity to rotation angles," Pattern Recognition Letters, vol. 131, pp. 56–62, 2020.

[20] J. Z. Leibo, Q. Liao, F. Anselmi, and T. Poggio, "The invariance hypothesis implies domain-specific regions in visual cortex," PLoS Computational Biology, vol. 11, no. 10, p. e1004390, 2015.

[21] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5028–5037, 2017.

[22] F. Richards, A. Paiement, X. Xie, and P.-A. Duc, "Learnable Gabor modulated complex-valued networks for orientation robustness," 2020. https://arxiv.org/pdf/2011.11734.pdf.

[23] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "First-person action recognition with temporal pooling and hilbert–huang transform," IEEE Transactions on Multimedia, vol. 21, no. 12, pp. 3122–3135, 2019.

[24] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. Cox, and J. J. DiCarlo, "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations," Advances in Neural Information Processing Systems, vol. 33, 2020.

[25] N. Alshammari, S. Akcay, and T. P. Breckon, "On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding," in 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1027–1032, IEEE, 2018.

[26] E. U. Moya-Sánchez, S. Xambó-Descamps, S. Salazar Colores, A. Sánchez Pérez, , and U. Cortés, "A Quaternion Deterministic Monogenic CNN Layer for Contrast Invariance," in Systems, Patterns and Data Engineering with Geometric Calculi (A. Delshams and S. Xambó-Descamps, eds.), ICIAM2019 SEMA SIMAI Springer Series, pp. 131–149, Springer, 2021.

[27] S. Xambó-Descamps, Real spinorial groups—a short mathematical introduction. SBMA/Springerbrief, Springer, 2018.

[28] J. O. Smith, Mathematics of the discrete Fourier transform (DFT): with audio applications. Julius Smith, 2007.

[29] S. Kay, "Maximum entropy spectral estimation using the analytical signal," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 5, pp. 467–469, 1978.

[30] D. Boukerroui, J. A. Noble, and M. Brady, "On the choice of band-pass quadrature filters," Journal of Mathematical Imaging and Vision, vol. 21, no. 1-2, pp. 53–80, 2004.

[31] G. H. Granlund and H. Knutsson, Signal processing for computer vision. Springer, 1995.

[32] E. U. Moya-Sánchez and E. Vázquez-Santacruz, "A geometric bio-inspired model for recognition of low-level structures," in International Conference on Artificial Neural Networks, pp. 429–436, Springer, 2011.

[33] E. Bayro-Corrochano, E. Vazquez-Santacruz, E. Moya-Sanchez, and E. Castillo-Muñis, "Geometric bioinspired networks for recognition of 2-D and 3-D low-level structures and transformations," IEEE transactions on neural networks and learning systems, vol. 27, no. 10, pp. 2020–2034, 2016.

[34] E. U. Moya-Sánchez and E. Bayro-Corrochano, "Symmetry feature extraction based on quaternionic local phase," Advances in Applied Clifford Algebras, vol. 24, no. 2, pp. 333–354, 2014.

[35] M. Felsberg and G. Sommer, "The monogenic signal," IEEE Transactions on Signal Processing, vol. 49, no. 12, pp. 3136–3144, 2001.

[36] J. Bigun, Vision with direction. Springer, 2006.

[37] D. Xing, C.-I. Yeh, J. Gordon, and R. M. Shapley, "Cortical brightness adaptation when darkness and brightness produce different dynamical states in the visual cortex," Proceedings of the National Academy of Sciences USA, vol. 111, no. 3, pp. 1210–1215, 2014.

[38] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," J. Opt. Soc. Am. A, vol. 2, pp. 1160–1169, jul 1985.

[39] R. Nava, B. Escalante-Ramírez, and G. Cristóbal, "Texture image retrieval based on log-gabor features," in Iberoamerican Congress on Pattern Recognition, pp. 414–421, Springer, 2012.

[40] A. R. Smith, "Color gamut transform pairs," ACM Siggraph Computer Graphics, vol. 12, no. 3, pp. 12–19, 1978.

[41] M. K. Agoston, Computer graphics and geometric modeling, vol. 1. Springer, 2005.

[42] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[43] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, vol. 2, 2010.

[44] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[45] E. J. McCartney and F. F. Hall, "Optics of the atmosphere: scattering by molecules and particles," Physics Today, vol. 30, no. 5, pp. 76–77, 1977.

[46] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 12, pp. 2341–2353, 2011.

[47] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in Proceedings of the 25th International Conference on World Wide Web, pp. 507–517, International World Wide Web, 2016. http://jmcauley.ucsd.edu/data/amazon/.

[48] F. Chollet et al., "Keras." https://github.com/keras-team/keras, 2015.

[49] Y. Gao, A. Rehman, and Z. Wang, "Cw-ssim based image classification," in 2011 18th IEEE International Conference on Image Processing, pp. 1249–1252, IEEE, 2011.

[50] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnr—a comparative study," Journal of Computer and Communications, vol. 7, no. 3, pp. 8–18, 2019.

[51] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 764–773, 2017.

[52] Z. Y. Khan and Z. Niu, "Cnn with depthwise separable convolutions and combined kernels for rating prediction," Expert Systems with Applications, vol. 170, p. 114528, 2021.

E. ULISES MOYA-SÁNCHEZ was honored with Fulbright fellowship and he was a postdoc of High-Performance Artificial Intelligence group at Barcelona Supercomputing Center. He received his PhD degree from CINVESTAV unidad Guadalajara. He is a member of Sistema Nacional de Investigadores (SNI-1), CONACyT, México. He is currently the artificial intelligence director of the Jalisco Government and researcher Universidad Autónoma de Guadalajara.

SEBASTIÀ XAMBÓ-DESCAMPS is an Emeritus Full Professor at the Department of Mathematics of the Universitat Politècnica de Catalunya (UPC). He holds a Ph.D. in Mathematics from the University of Barcelona, and an M.Sc. degree in Mathematics from Brandeis University, USA. Has been Full Professor at the Department of Algebra of the Universidad Complutense of Madrid, President of the Catalan Mathematical Society, Dean of the Faculty of Mathematics and Statistics of the UPC and President of the Spanish Conference of Deans of Mathematics. Led various R+D+I projects, including the development of the Wiris mathematical platform. Authored Block Error-Correcting Codes–A Computational Primer and Real Spinorial Groups. Co-authored An invitation to Geometric Algebra through Spacetime Physics, Robotics and Molecular Geometry. Co-edited Cosmology, Quantum Vacuum and Zeta Functions.

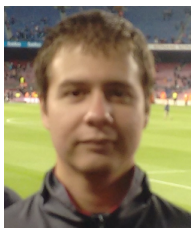
SEBASTIÁN SALAZAR-COLORES received his B. S. degree in Computer Science from Universidad Autónoma Benito Juárez de Oaxaca, received his M. S. degree in Electrical Engineering at Universidad de Guanajuato. He is a PhD candidate in Computer Science at the Universidad Autónoma de Querétaro. His research interest are image processing and computer vision.



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128552, IEEE Access





ABRAHAM SÁNCHEZ PÉREZ is an artificial intelligence analyst at Jalisco government. Received his M. S. degree in Computer Sciences at Universidad Autónoma de Guadalajara.

ULISES CORTÉS is a Full-Professor and Researcher of the Universitat Politècnica de Catalunya (UPC). He is the scientific coordinator of the High-Performance Artificial Intelligence group at Barcelona Supercomputing Center (BSC). He is a member of Sistema Nacional de Investigadores (SNI-III), CONACyT, México.



This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/