Perception Over Time: Temporal Dynamics for Robust Image Understanding

Maryam Daniali Edward Kim
Drexel University
Philadelphia, PA
{maryam.daniali,edward.kim826}@drexel.edu

Abstract

While deep learning surpasses human-level performance in specific vision tasks, it is fragile and overconfident in its classification. For example, minor transformations in perspective, illumination, or object deformation in the image space can result in drastically different labeling. This is especially apparent when adversarial perturbations are present. Conversely, human visual perception is orders of magnitude more robust to input stimulus changes. Neuroscience research suggests that biological perception is a dynamic process that converges over time, even for static images and scenes. Almost all perception frameworks lack this convergence property, which makes them vulnerable to minor perturbations. Motivated by our human task results, we introduce a novel framework for incorporating temporal dynamics into static image understanding. We demonstrate a biologically plausible model that decomposes a single image into a series of coarse-to-fine images, mimicking the integration of visual information in the human brain. Our model utilizes this information "over time", resulting in significant improvements in its accuracy, robustness, and cost-effectiveness over standard CNNs. We explicitly quantify the adversarial robustness properties of our coarse-to-fine framework through multiple studies. Our quantitative and qualitative results convincingly demonstrate exciting and transformative improvements over standard architectures.

1. Introduction

Human visual perception is remarkably slow. The moment an input stimulus is presented to the time of recognition takes several hundred milliseconds [3, 19]. Neuroscience experiments and recordings offer an explanation; the visual system is performing *recognition over time*, even for static and simple visual scenes [16, 41]. In humans, perceptual clarity increases as bottom-up signals and top-down feedback mechanisms compete over time and converge to a confident agreement. In fact, it is precisely this "slowness" that makes human perception robust and accurate.

In contrast, standard deep learning classifiers typically only implement a single feed-forward pass and can be optimized in hardware to be orders of magnitude faster than human recognition. But this speed comes at a cost, i.e., there is no top-down feedback loop nor any notion of predictive coding feedback (expectation guiding perception) and lateral competition. Thus, the idea of processing static images and scenes holistically and over time is missing in state-of-the-art deep learning models. We believe a more biologically inspired model of "slowness" will ultimately provide mechanisms and solutions that are robust in the general classification and especially effective against adversarial examples. Thus, in this work, we present a novel architecture that utilizes the idea of classification over time for the robust classification of static images. In essence, the model "sees" a gradual progression of the input signal over a generated time series of increasing perceptual clarity extracted from a single input stimulus. The final classification is the culmination of all the information integrated over time. Our contributions are as follows:

- We demonstrate a coarse-to-fine perception framework that integrates visual information over time to perform more accurate and robust image classification.
- We propose a bio-inspired sparse model that reflects the dynamic properties of human perception and outputs a set of decompositions that capture the gradual progression of static image resolution.
- To the best of our knowledge, this is the first study that considers temporal dynamics in the form of coarse-to-fine flow by incorporating low-, intermediate-, and high-level feedback in the perception of static images.
- Our simple yet computationally-efficient perception framework achieves superior performance compared with the state-of-the-art methods in object classification and is more robust to perturbations and adversarial attacks.

2. Related Work

Our perception of a visual scene changes rapidly over time, even if the scene remains unchanged [16]. Although

we are far from fully understanding the changes in human visual perception over time, some studies provide considerable evidence of the existence of temporal dynamics in visual recognition [19, 26]. Psychophysical studies show that around 150 ms after the stimulus onset, humans acquire the "gist" of complex visual scenes, even when the stimulus is presented very briefly. They require longer processing to identify individual objects, and it may even take longer for a more comprehensive semantic understanding of the scene to be encoded into short-term memory [16]. Consistent with the timing of perceptual understanding, many studies have suggested that the visual system integrates visual input in a coarse-to-fine (CtF) manner [3]. The CtF hypothesis states that low-frequency information is processed quickly first, which then projects to high-level visual areas. Critically, the high-level areas *generate a feedback signal* that guides the processing of the high-frequency input [32]. David Marr's work on a functional model of the visual system also emphasizes several levels of computation, e.g., primal sketch to 2.5D to 3D representation, mimicked by the cortical areas of the primate visual system [28].

Processing visual input in a CtF manner helps humans achieve robust and accurate perception. Indeed, our own experiences demonstrate that small changes in input do not change our understanding dramatically (Section 3.4). After some amount of information, there is a certain point in time when our brains can detect and identify objects with high certainty, but prior to this "aha" moment that occurs hundreds of milliseconds after stimulus onset, we are (justifiably) neither confident nor accurate in object recognition.

At the other end of the recognition spectrum, deep learning is very sensitive and fragile to small changes in the input stimuli and overly confident in classification. Our study will elucidate these points further (Section 3.4). Many studies attempt to protect deep learning models from transformations, perturbations, and adversarial attacks by augmenting training data [27], adding stochasticity to the hidden layers [8], and applying preprocessing techniques [10, 14, 22, 46]. While such techniques can improve image classification models on specific tasks and data [38, 42], even at the cost of heavy computations, research has shown every defense against adversarial attacks has eventually been found to be vulnerable. Furthermore, solutions that are dependent on providing a massive amount of data, attempting to reflect true distribution, are not feasible in all studies, as true distribution is unknown or at least very expensive to achieve in many applications. In another line of work, attention mechanisms are employed in the vision domain to simulate feedback flows in the human brain [10, 13, 45, 47]. While such mechanisms are great in incorporating semantic feedback with conventional feature extraction, they skip intermediate-level feedback and their role in perception. More precisely, such models have no "gist" understanding, nor any gradual perception. Thus, even in attention-based models, the convergence property is still missing as defined for human perception. As such, there still is a monumental gap between human perception and the current state of machine vision.

3. Methods

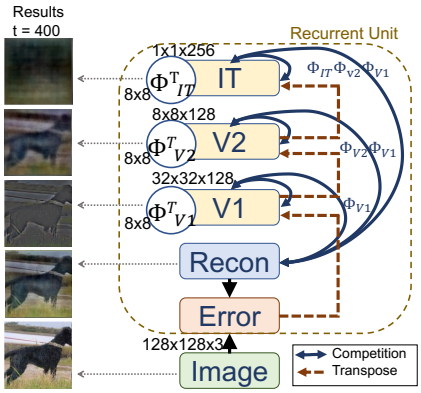
In this work, we take a neuro-inspired approach to robust vision understanding. We simulate a series of reflections from an input image and then demonstrate a model that processes them over time (Figure 2b). The first step is to simulate the coarse-to-fine structure of a visual scene by generating components that represent the changes in visual perception over time. Furthermore, by generating these components, we can investigate the robustness of available architectures and design a model inspired by the psychophysical findings on dynamic components in perception over time.

3.1. Coarse-to-Fine (CtF) Decomposition

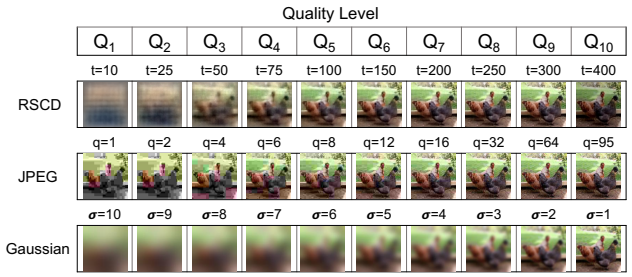
Image decomposition is the general process of separating an input stimulus into a combination of the generators (or causes) of the data. Decomposition methods have been used in various computer vision applications such as background subtraction [17] and moving object detection [37]. They also have applications in image smoothing and deblurring [12, 44]. In this paper, we introduce a sparse coding model that can faithfully mimic CtF decomposition over time. We also describe two other decomposition approximators of minimal biological fidelity, but more readily available to the general public, e.g. JPEG and Gaussian decompositions. While studies such as [21] show that biological models such as sparse coding are more robust to perturbations than JPEG Compression ([8]) and Gaussian Smoothing ([46]), we believe these methods can serve as baselines and may be preferred over sparse coding in applications with computational constraints.

Sparse Coding. Sparse coding provides a class of algorithms for finding sparse representations of stimuli, input data. [30] introduced sparse coding to explain the sparse and recurrent neural representations in the primary visual cortex. Given only unlabeled data, sparse coding looks for generating a minimal set of components that can reconstruct each input signal as accurately as possible, resulting in having a high representative capacity that surpasses the capabilities of dense networks on pairing inputs and outputs. Also, it can leverage the availability of unlabeled data. Unlike some other unsupervised learning techniques such as [18, 34, 43], sparse coding can be applied to learning overcomplete basis sets, where the number of bases is greater than the input dimension [25].

Sparse coding can be defined using the following objective function, where $x^{(n)}$ represents the input signal, and



(a) RSCD model.



(b) Sample decomposed images.

Figure 1. (a) Schematic of our sparse model (RSCD) for image decomposition over time. (b) Sample decomposed images with different qualities using RSCD (1st row), JPEG compression (2nd row), and Gaussian smoothing (3rd row).

$a^{(n)}$ is the sparse representation, activation, that can reconstruct the input $x^{(n)}$.

$$\min_{\Phi} \sum_{n=1}^{\mathcal{N}} \min_{a^{(n)}} \frac{1}{2} \|x^{(n)} - \Phi a^{(n)}\|_2^2 + \lambda \|a^{(n)}\|_1 \qquad (1)$$

Here, Φ is an overcomplete dictionary containing all components that share features to reconstruct the input, and $\hat{x}^{(n)} = \Phi a^{(n)}$ is the reconstructed form. λ balances the sparsity versus the reconstruction quality. n is a training element, and there are a total of $\mathcal{N}$ training elements.

There are different solvers for Equation 1, and, among them, there are some systems of nonlinear differential equations, including but not limited to Fast Iterative Shrinkage and Thresholding Algorithm (FISTA) [4] and Locally Competitive Algorithm (LCA) [33]. Here, we select the Locally Competitive Algorithm, a bio-inspired technique that evolves the dynamical variables, the membrane potential of the neuron, when an input signal is presented. In this model, the activations of neurons compete and inhibit other units from firing. The neuron's excitatory potential is proportional to the match between the input signal and the dictionary element of that neuron. The inhibitory strength is proportional to the similarity of elements/convolutional patches between the current neuron and other competing neurons, forcing it to decorrelate.

In the LCA algorithm, the active coefficients for a neuron, m, with the membrane potential, i.e., the internal state, u^m, can be defined as:

$$a^m = T_\lambda(u^m) = H(u^m - \lambda)u^m \qquad (2)$$

where T is a soft-threshold function with threshold parameter, λ, and H is the Heaviside function [1].

The differential equation below determines the dynamics of a neuron, m, with an input signal, I.

$$\dot{u}^m = \frac{1}{\tau}\Big[-u^m + (\Phi^T I) - (\Phi^T \Phi a - a^m) \Big] \qquad (3)$$

where τ is the time constant, $-u^m$ is the internal state leakage term, a is the activation vector of all neurons, $(\Phi^T I)$ is the driver that charges up the state by the match between the dictionary element and the input signal, here calculated by the inner product between them. $(\Phi^T \Phi a - a^m)$ shows the competition between the set of active neurons proportional to the inner product between dictionary elements, which applies as a lateral inhibition signal. $-a^m$ excludes self-interactions, including self-inhibition.

In short, using LCA, neurons that are selective to the input stimulus charge up faster, then pass a threshold of activation. Once they pass the threshold, they begin to compete with other neurons to claim the representation. Thus, sparse coding with LCA creates a sparse representation of selective neurons that compete to represent stimuli [20, 23, 31].

Recurrent Sparse Coding Decomposition (RSCD). We developed a biologically inspired recurrent model that uses selectivity through competition, holistic processing, and top-down feedback to generate image decomposition over time using sparse coding. We call our model Recurrent Sparse Coding Decomposition (RSCD) and use it to decompose images in a CtF manner over $t = 400$ time steps. The interactions between layers and how they contribute to the final reconstruction are presented in Equation 4,

$$\hat{x}^{(n)} = \sum_{k=1}^{K} (\prod_{l=1}^{k} \Phi_l) a_k^{(n)}, \qquad (4)$$

where $\hat{x}^{(n)}$ represents the final reconstruction / decomposition of input $x^{(n)}$, which we can substitute into Equation 1 and dynamically solve using Equation 3. K is the number

of layers in the sparse model. For RSCD, we have three layers, and $k \in \{V1, V2, IT\}$ which simulates the brain areas involved in the ventral pathway of the cortex used for form recognition and object representation (Figure 1a).

The input images to the RSCD model have been resized to $128\times128\times C$, where C is the number of channels, 3. In Φ_{V1}, dictionary of layer $V1$, there are 128 elements of size $8\times8\times C$. We set the same dictionary size for layer $V2$, and expanded the number of neurons to 256 at the top layer, IT. The receptive field of neurons increases by a factor of 4 at each layer since we stride by 4 over the hierarchy. More precisely, the receptive filed of neurons in $V1$ is 8×8, 32×32 in $V2$, and 128×128 (the whole image) in IT. Thus, the size of layers $V1$, $V2$, and IT are $32\times32\times128$, $8\times8\times128$, and $1\times1\times256$, respectively (Figure 1a). For evaluation purposes, we empirically selected a subset of 10 decomposed images generated by RSCD, namely, $t \in \{10, 25, 50, 75, 199, 150, 200, 250, 300, 400\}$, where t is the timestep (Figure 1b, first row).

3.2. Approximate CtF Decomposition

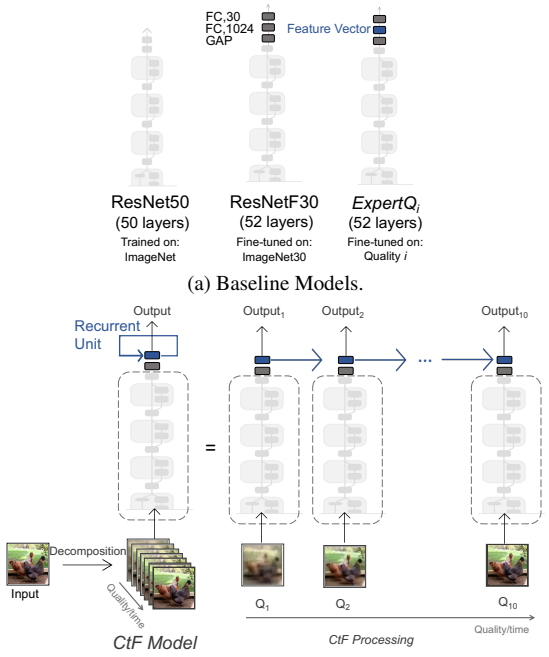
While approximation methods do not decompose input stimuli over time and are not biologically plausible, they provide a reasonable CtF approximation and are generally fast. Among them, JPEG Compression and Gaussian Smoothing have various image processing applications and can be used as points of reference for our RSCD method.

JPEG Compression. We used the JPEG compression technique in [7] to generate 95 different quality level images, in which scales 1 and 95 are the lowest and the highest quality levels, respectively. We then selected a subset of 10 qualities for each image to match the CtF samples used in RSCD, empirically based on our human subjects' results. More specifically, we selected $q \in \{1, 2, 4, 6, 8, 12, 16, 32, 64, 95\}$, where q is the quality scale (Figure 1b, second row).

Gaussian Smoothing. Gaussian smoothing is widely used in image processing applications to reduce noise and other high-frequency details [46]. We applied 10 different values for the standard deviation of the Gaussian kernel, σ, starting from 10 to 1, to match the CtF samples used in RSCD and create 10 decomposed images (Figure 1b). We would like to emphasize that although the mapping was selected empirically based on the first set of volunteers, we focus our experiments and results on the performance pattern rather than minute differences in these decompositions.

3.3. Dataset

In this study, we focus on image classification since it has become the leading task with a broad range of applications in machine learning and computer vision. Among popular datasets available for image classification, ImageNet has been widely used in evaluating cutting-edge models. Also,



(a) Baseline Models.

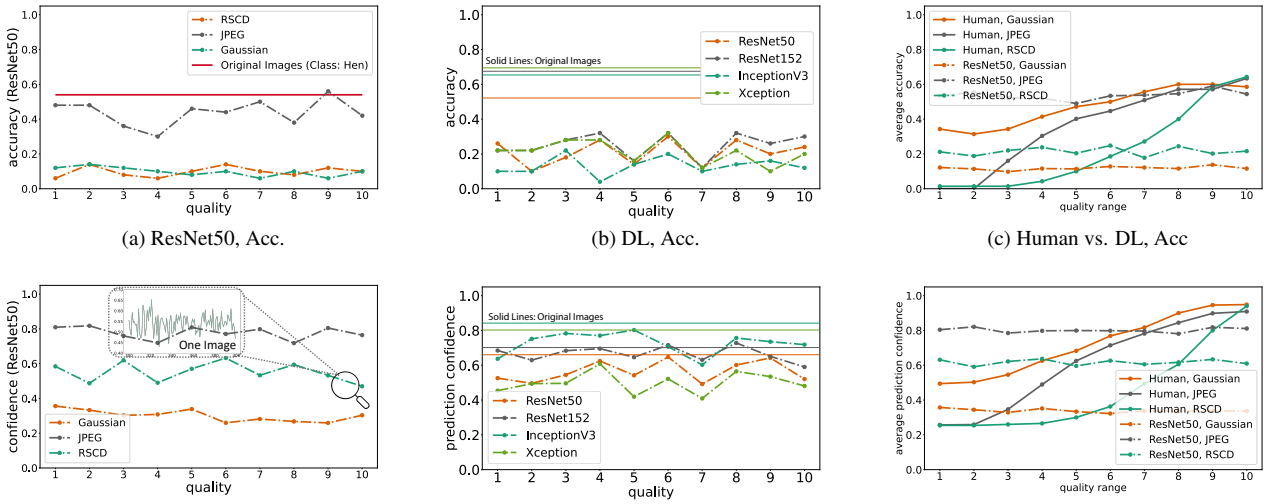
(b) Our Proposed CtF Framework.

Figure 2. Baselines and CtF models' architecture. Cells in light gray use transfer learning from ResNet50 and are frozen during training/fine-tuning. The CtF model takes $ExpertQ_i$ as its backbone and processes each input image in a CtF manner using a recurrent unit. ResNet50 illustration is borrowed from [36].

some prior studies have presented algorithms that could surpass human-level performance on this dataset [35]. We used two datasets sub-sampled from ImageNet for our experiments. This approach allowed us to investigate the models' behavior with high resolution and diverse data compared to standard and smaller datasets such as CIFAR10 [24]. For the first set of experiments, including the off-the-shelf models' comparison and human subjects' results, we handpicked 10 classes from ImageNet that were visually distinctive, referred to as ImageNet10. We also randomly chose 20 classes of ImageNet, and added them to the 10 previously chosen classes, leading to 30 unique classes, referred to as ImageNet30. To study the scalability of our proposed algorithms and challenge existing models, the majority of our experiments were carried out on ImageNet30. More specifically, we used a subset of the ILSVRC-2012 validation set, [35], which contains 50 images per class resized to 128×128. We generate 10 different versions of each image for our analyses (Section 1 Supplemental).

3.4. Motivation: Deep Learning vs. Human

Overconfident Deep Models. As a part of our motivation experiment, we selected 4 off-the-shelf deep learning models with outstanding results on ImageNet, namely ResNet50 [15], ResNet152 [15], InceptionV3 [40], and Xception [6], and examined their performance on decomposed images of



(a) ResNet50, Acc. (b) DL, Acc. (c) Human vs. DL, Acc

(d) ResNet50, Conf. (e) DL, Conf. (f) Human vs. DL, Conf.

Figure 3. [**Motivation Experiment**] Deep learning seems fragile and overconfident as it receives better-quality input. In contrast, humans seem to have a systematic improvement in their perception while receiving better-quality images. (a),(d): ResNet50 performance averaged on one class, "hen". The magnified area in (d) shows ResNet50's spiky confidence on an individual image even in a high-quality range ($t \in [300, 400]$). (b),(e): Baseline models' performance using the RSCD decomposition on one class, "hen". (c),(f): Deep learning vs. human: performance averaged over all classes of ImageNet10.

ImageNet10 over time. Based on the available neuroscience studies, we initially expected the deep models to perform with very low certainty and chance-level accuracy on the first quality levels and achieve higher confidence and accuracy over time as the image quality improves. However, we observed that the deep models performed completely unexpectedly in the following cases:

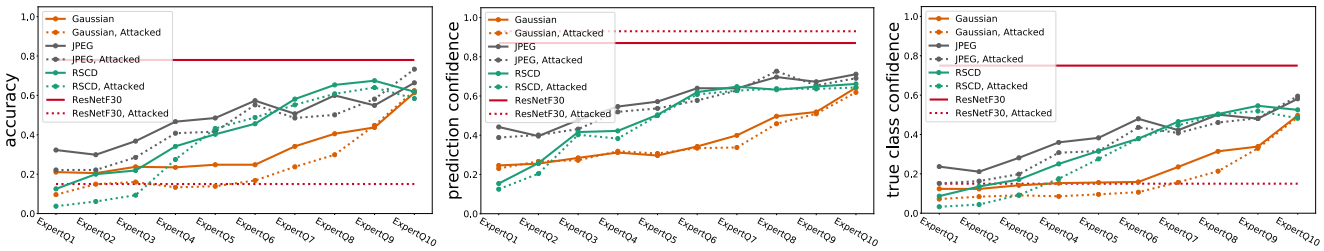
- At early timesteps, low-quality and unintelligible images achieve confident but incorrect predictions as shown in Figures 3a and 3d. Also, no steady increase is seen in the accuracy or confidence on higher quality images.
- The classification accuracy, even on Q_{10}, is significantly lower than the original images. Figures 3a and 3b show deep models' dependability to high-frequency information than the actual concepts.
- Spiky confidence, even on high-quality images, due to models' sensitivity to unnoticeable perturbations. The magnified area in Figure 3d shows ResNet50 unstable behavior on almost fully reconstructed images ($t \in [300 - 400]$). Such a behavior is noticeable even when models' performance is averaged over all images in one class (Figures 3a, 3b, 3d, and 3e). Occasionally, the models' performance drops at a higher quality (Figure 3b Q_6 and Q_7).

Different Visual Trajectory in Humans. As the second part of our motivation experiment, we conducted a similar task on human participants to verify our original hypothesis—confidence and accuracy increase over time—and compared humans' performance with that of deep learning models. In doing so, we asked seven volunteers to look at the images and type the main object they recognize in each image and their confidence level in their recognition. For each decomposition method, one image was randomly selected from each class of ImageNet10. All different qualities of each image were shown to the participants, in order, starting from the lowest quality (Section 2 Supplemental). We post-processed the participants' answers to the basic level category [5], employing the same structure used in defining labels and hierarchies in creating ImageNet using the WordNet database [11]. We then used the same categories to determine the models' performance.

Our results show humans' confidence and accuracy increase as the image quality increases. Unlike the deep learning models, there is almost no sudden change in human accuracy or confidence; instead, both accuracy and confidence increase gradually over time and quality. Figures 3c and 3f compare participants' performance averaged over ImageNet10 classes with that of ResNet50 for all three decomposition methods. These results motivated us to design a perception model that can close the gap between human results and deep models by taking advantage of CtF information. We also expect such a perception model to be more robust to minor perturbation and high-frequency data.

3.5. Our Solution to Perception Over Time

Sequential models such as 1-D convolutional neural networks (1-D CNNs) and recurrent neural networks (RNNs) have shown interesting results in computer vision tasks that



(a) Accuracy. (b) Prediction Conf. (c) True Class Conf.

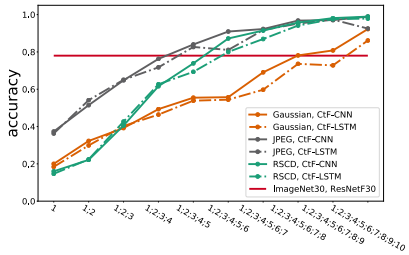
Figure 4. Baseline models performance on different quality decomposed images. While we see an increasing performance pattern on higher-quality Experts, they still have a gap with ResNetF30. There is also a gap between prediction and true-class confidence.

involve sequences such as video recognition. Motivated by [9, 39] and the different visual trajectories we saw in humans and machines in Section 3.4, we designed two sequential models, namely CtF-CNN and CtF-LSTM, that can take in information from static images over time and perform a more robust and accurate perception of them. In doing so, our framework utilized the decomposition methods in Section 3.1 to generate a time component for static images and then process them in a sequential manner.

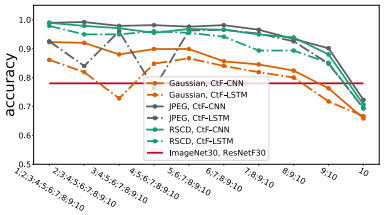
Models. We transferred the weights from ImageNet ResNet50 to a model and only fine-tuned the last 3 layers on ImageNet30, keeping other layers frozen. We refer to this model as ResNetF30. More specifically, we removed the ResNet50 classification layer, added a global average pooling layer (GAP), and a fully connected layer (FC) with 1024 nodes, followed by a classification layer containing 30 nodes (Figure 2a). Additionally, for each decomposition quality level, we create a model based on ResNetF30 and similarly fine-tune it on that quality level, leading to 10 baseline models. We call these models $ExpertQ_i$, where i represents the quality level, $i \in [1, 10]$. Furthermore, we take trained $ExpertQ_i$ as the backbone of our CtF-CNN and CtF-LSTM models, to process static images in a CtF manner over time (Figures 2a and 2b).

Our proposed coarse-to-fine architecture has two main components. The first component generates decompositions for the temporal components, and the second part processes Expert-extracted features which lead to a classification output, recurrently. This architecture allowed our model to extract meaningful features from the individual decompositions and "calibrate" the final classification decision based on all information over time. In Section 4, and Section 3 of Supplemental, we show how such architectures make the classification models more accurate and robust.

Adversarial Robustness. Existing defense mechanisms for adversarial attacks in deep neural networks try to defend against the adversaries by augmenting the training data with adversarial examples [27], or adding some level of stochasticity to the hidden layers [8]. Some studies apply some preprocessing techniques to defend against such attacks [14, 22, 46]. The main problem in deep neural networks is their tendency to learn surface-level predictability of the data rather than extracting concepts and meaningful information. While such techniques improve the models' performance to some extent, they do not help models learn concepts. Thus, such models are still prone to attacks that may not have been discovered yet. However, in our sequential models, we incorporate CtF visual processing observed in human perception with the hope of overcoming a broader set of high-frequency dependencies. We evaluate the robustness of the models by a gradient-based method, Projected Gradient Descent (PGD) [29], which was motivated by [8,21,27] as a universal adversary that guarantees against first-order attacks. We attack ResNetF30 using PGD to the level that its prediction accuracy drops from 0.78 to 0.15. In addition, we evaluated the models' performance within the context of black-box attacks with Square attack [2]. We collect the attacked images and run the decomposition methods to generate CtF attacked images. Our results show the effectiveness of our CtF models in the context of adversarial examples; however, we would like to emphasize their im-



(a) Gaining Data (X-axis shows quality range).



(b) Losing Data (Y-axis starts from 0.5).

Figure 5. CtF models performance on "gaining" and "losing" data.

plications extend to robust perception in general.

4. Experiments and Results

Classification Accuracy. To study the performance of the models, we compare their classification accuracy on the decomposed test images using all three decomposition methods. We refer to the performance of the models on unperturbed decomposed images as the standard performance, i.e., Acc., and refer to their performance on the adversarial perturbed images as "attacked", i.e., Att.

Baseline. For the baseline analysis, we looked at the performance of Fine-tuned ResNet, i.e., ResNetF30, as well as Expert Models. See solid lines in Figure 4. For all decomposition methods, $ExpertQ_i$ generally performed better than $ExpertQ_j$, where $j < i$. The increasing pattern in the accuracy is in line with previous studies using JPEG compression and Gaussian smoothing [8, 21]. Note that: 1. We did not see such an increasing pattern when we examined deep learning models trained on the original images; Figures 3b and 3e. 2. The decomposed images of quality Q_{10} are not fully reconstructed and have a visually noticeable difference from the original images, especially in RSCD and Gaussian smoothing. Thus, the standard accuracy of ResNetF30 is higher than $ExpertQ_{10}$. However, $ExpertQ_{10}$ is relatively less overconfident of its detection. Furthermore, Expert Models perform significantly better on the adversarial test images (the dashed lines in Figure 4). Similar to the standard performance, we see an increasing pattern in the attacked performance of Expert Models. We also see that Expert Models are significantly less overconfident on the adversarial images compared to ResNetF30.

CtF Models. We are interested in evaluating the key role of CtF processing in visual perception. In doing so, we compare the accuracy of the standard CNN models discussed in Section 3.4, ResNetF30, and $ExpertQ_{10}$ (EQ10), with that of our proposed CtF models on unperturbed (Acc.) and adversarial attacked (Acc. Att.) images which led to the following impactful results. **(1)** Our results demonstrate transformative accuracy improvements, almost perfect accuracy, and >20% jump in accuracy over our most accurate baseline, ResNetF30, on the same dataset. **(2)** The use of our proposed models results in no penalty to the accuracy on unperturbed images. CtF models on PGD attacks using RSCD (76.8%) and JPEG (83.1%), and Square attacks using Gaussian (85.8%) perform on par (at times better) than ResNetF30 even when tested on non-attacked images. Note that here, CtF-CNN and CtF-LSTM have access to all level decomposed images for their classification (Table 1).

Data Augmentation Ablation Experiments. To study the effects of data augmentation and control data quantity and quality, we evaluated the performance of the ResNet model using two training scenarios.

Table 1. Accuracy comparisons between baselines and our proposed CtF framework. CtF-CNN is the strongest model. Our results suggest that the decomposition (approx.) method of choice depends on the image conditions and potential attacks.

	Model	Acc.	Att. PGD [29]	Att. Square [2]
	ResNet50 [15]	0.655	0.357	0.283
	ResNet152 [15]	0.698	0.41	0.377
	InceptionV3 [40]	0.658	0.466	0.363
	Xception [6]	0.652	0.461	0.386
	ResNetF30	0.78	0.15	0.434
Gaussian	EQ10	0.613	0.624	0.396
	CtF-CNN	0.922	0.6	**0.858**
	CtF-LSTM	0.861	0.547	0.594
JPEG	EQ10	0.663	0.733	0.306
	CtF-CNN	**0.989**	**0.831**	0.773
	CtF-LSTM	0.861	0.547	0.641
RSCD	EQ10	0.613	0.583	0.429
	CtF-CNN	**0.989**	0.768	0.768
	CtF-LSTM	0.978	0.663	0.636

Exp.1 Effects of data quantity: Fine-tuning ResNetF30 on all quality levels and original images, and testing its performance on original images (i.e., ResNet-[Q_1-Q_{10}],Org).

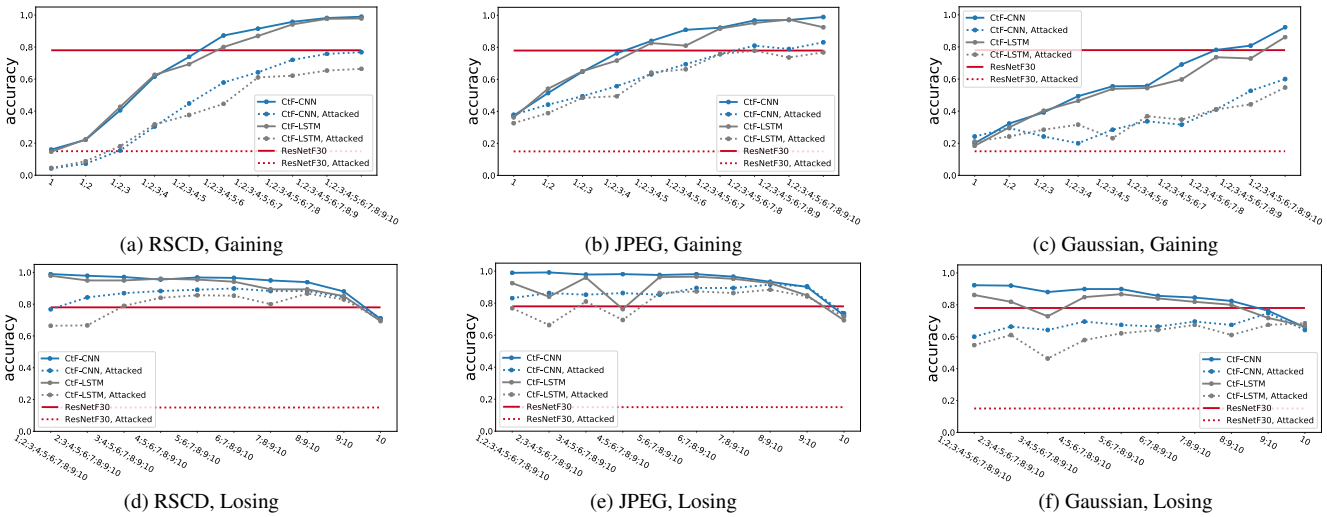
Exp.2 Effects of data quality: Fine-tuning ResNetF30 only on higher quality (Q_9, Q_{10}) and original images, and testing its performance on original images (i.e., ResNet-[Q_9,Q_{10}],Org).

Our results in Table 2 demonstrate different quality images can easily fool ResNet, making it perform even worse than standard training, i.e., on original images only (78%). While using only higher quality images in Exp.2 results in higher accuracy than Exp.1, the accuracy is still significantly lower than the standard training, on original images only (78%). We also evaluated the effects of data augmentation on ResNet-Expert models by training them over larger epochs (doubled epochs, to be precise), however, we did not see any improvements. These results suggest that the way in which a perception model processes data is more important than the quantity and quality of the provided training data.

Table 2. The reason CtF outperforms ResNet does not lie in data augmentation. Even when ResNet is trained on all or high quality decomposed images, CtF framework outperforms.

	RSCD	JPEG	Gaussian
ResNet-[Q_1-Q_{10}],Org (Exp.1)	0.513	0.455	0.359
CtF-CNN[Q1-Q10]	**0.989**	**0.989**	**0.922**
CtF-LSTM[Q1-Q10]	0.978	0.861	0.861
ResNet-[Q_9,Q_{10}],Org (Exp.2)	0.619	0.488	0.550
CtF-CNN[Q9,Q10]	0.88	0.901	0.762
CtF-LSTM[Q9,Q10]	0.85	0.847	0.717

We examined the role of each quality level and evaluated CtF-CNN and CtF-LSTM in two main categories: (1) "Gaining data": receiving better quality images, one quality level at a time, until all qualities are seen. (2) "Losing



(a) RSCD, Gaining (b) JPEG, Gaining (c) Gaussian, Gaining

(d) RSCD, Losing (e) JPEG, Losing (f) Gaussian, Losing

Figure 6. Adversarial robustness of the CtF models on "gaining" and "losing" data. Unstable behavior of JPEG and Gaussian on "losing".

data": skipping the lowest quality images, until only Q_{10} is left. These categories allow us to discover the role of each quality level while interacting with other qualities in visual perception. Our results demonstrate that the performance of CtF-CNN and CtF-LSTM improve while "gaining" data (following our observations in the human study), outperforming ResNetF30 using any of the decomposition methods (Figure 5a). We demonstrate the importance of coarse-level data in visual perception and see a general decreasing pattern in the accuracy of both models on all decomposition methods, even when losing Q_1 data (Figure 5b). We observe a sudden accuracy drop at Q_3 for Gaussian smoothing and at Q_2 and Q_4 for JPEG compression on the adversarial images. We believe this is caused by the poor ability of these methods in approximating decomposition with no artifacts. Another key finding is that the CtF models outperform ResNetF30 even by using limited quality level data.

Adversarial Robustness. In addition to the adversarial study in Table 1, we evaluated our CtF models on the adversarial images while "gaining" and "losing" data. Our results show an increasing pattern in the accuracy on all decomposition methods when gaining data, even on adversarial attacked images, dashed lines in Figures 6a, 6b, and 6c. We also see that our RSCD decomposition is more robust to changes in the quality range and contrary to JPEG compression and Gaussian smoothing, there is no sudden drop in the accuracy when "losing" one quality data (Figures 6d vs. 6e, 6f). See Section 5 Supplemental for additional details.

Limitations. Similar to other sparse coding algorithms, the two main limitations of our RSCD method are increased computation and memory costs. However, using neuromorphic hardware, sparse coding achieves orders of magnitude computational and energy efficiency over its standard von Neumann implementations. Furthermore, our CtF framework has minimal trainable layers, which makes it computationally efficient (Section 3 Supplemental).

5. Conclusions

Static image classification forms the basis of many computer vision problems such as machine vision, medical imaging, and autonomous vehicles to name a few. Many of these areas are not immune to perturbations and require better trust and safety guarantees. Many studies have proposed defense mechanisms that were eventually found to be vulnerable, and our work may not be an exception to that. However, our goal is to get inspiration from human perception to build a more robust static image understanding. We find processing over time missing in many applications and see that as a key to misleading immunity.

In this work, we created a sequential coarse-to-fine visual processing framework that is inspired by findings on human perception and incorporates temporal dynamics in static image understanding. We also proposed a novel biology-inspired decomposition method, RSCD, to generate images in a CtF manner, and show such processing helps the framework be accurate and robust even on adversarial attacks. In addition, we evaluate the performance of our CtF framework on approximated decomposition using JPEG and Gaussian reconstruction. While these methods do not decompose input over time and have noticeable artifacts, they are computationally efficient and may be preferred in various applications.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1954364.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964. 3

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. 6, 7

[3] Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, 103(2):449–454, 2006. 1, 2

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 3

[5] Maureen A Callanan. How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, pages 508–523, 1985. 5

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4, 7

[7] Alex Clark. Pillow (pil fork) documentation. *Readthedocs. https://buildmedia. readthedocs. org/media/pdf/pillow/latest/pillow. pdf*, 2015. 4

[8] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017. 2, 6, 7

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 5

[12] D Firsov and SH Lui. Domain decomposition methods in image denoising using gaussian curvature. *Journal of Computational and Applied Mathematics*, 193(2):460–473, 2006. 2

[13] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 2

[14] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 2, 6

[15] 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. 4, 7

[16] Jay Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Progress in neurobiology*, 84(4):405–439, 2008. 1, 2

[17] Sajid Javed, Seon Ho Oh, JunHyeok Heo, and Soon Ki Jung. Robust background subtraction via online robust pca using image decomposition. In *Proceedings of the 2014 conference on research in adaptive and convergent systems*, pages 105–110, 2014. 2

[18] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991. 2

[19] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004. 1, 2

[20] Edward Kim, Maryam Daniali, Jocelyn Rego, and Garrett T Kenyon. The selectivity and competition of the mind's eye in visual perception. *arXiv preprint arXiv:2011.11167*, 2020. 3

[21] Edward Kim, Jocelyn Rego, Yijing Watkins, and Garrett T Kenyon. Modeling biological immunity to adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4666–4675, 2020. 2, 6, 7

[22] Edward Kim, Jessica Yarnall, Priya Shah, and Garrett T Kenyon. A neuromorphic sparse coding defense to adversarial images. In *Proceedings of the International Conference on Neuromorphic Systems*, pages 1–8, 2019. 2, 6

[23] Edward Kim, Jessica Yarnall, Priya Shah, and Garrett T. Kenyon. A neuromorphic sparse coding defense to adversarial images. In *Proceedings of the International Conference on Neuromorphic Systems*, ICONS '19, New York, NY, USA, 2019. Association for Computing Machinery. 3

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[25] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808. Citeseer, 2007. 2

[26] Wei Ji Ma, Fred Hamker, and Christof Koch. 16neural mechanisms underlying temporal aspects of conscious visual perception. 2006. 2

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 6

[28] David Marr and Tomaso Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1156):301–328, 1979. 2

[29] Maria-Irina Nicolae, Mathieu Sinn, Tran Ngoc Minh, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Ian M Molloy, and Benjamin Edwards. Adversarial robustness toolbox v0. 2.2. 2018. 6, 7

[30] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 2

[31] Dylan M Paiton, Charles G Frye, Sheng Y Lundquist, Joel D Bowen, Ryan Zarcone, and Bruno A Olshausen. Selectivity and robustness of sparse coding networks. *Journal of vision*, 20(12):10–10, 2020. 3

[32] Kirsten Petras, Sanne Ten Oever, Christianne Jacobs, and Valerie Goffaux. Coarse-to-fine information integration in human vision. *NeuroImage*, 186:103–112, 2019. 2

[33] Christopher Rozell, Don Johnson, Richard Baraniuk, and Bruno Olshausen. Locally competitive algorithms for sparse approximation. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV–169. IEEE, 2007. 3

[34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 2

[35] 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. 4

[36] Ludwig Schubert. Openai microscope, Sep 2020. 4

[37] Moein Shakeri and Hong Zhang. Moving object detection under discontinuous change in illumination using tensor low-rank and invariant sparse decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7221–7230, 2019. 2

[38] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9661–9669, 2021. 2

[39] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 6

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4, 7

[41] Steven Vanmarcke and Johan Wagemans. Rapid gist perception of meaningful real-life scenes: Exploring individual and gender differences in multiple categorization tasks. *i-Perception*, 6(1):19–37, 2015. 1

[42] Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33:2135–2146, 2020. 2

[43] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 2

[44] Jing Xu, Hui Bin Chang, and Jing Qin. Domain decomposition method for image deblurring. *Journal of computational and applied mathematics*, 271:401–414, 2014. 2

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[46] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 2, 4, 6

[47] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9483–9492, 2020. 2