Top-Down Networks: A coarse-to-fine reimagination of CNNs

Ioannis Lelekas Nergis Tomen Silvia L. Pintea Jan C. van Gemert

Computer Vision Lab
Delft University of Technology, Netherlands

Abstract

Biological vision adopts a coarse-to-fine information processing pathway, from initial visual detection and binding of salient features of a visual scene, to the enhanced and preferential processing given relevant stimuli. On the contrary, CNNs employ a fine-to-coarse processing, moving from local, edge-detecting filters to more global ones extracting abstract representations of the input. In this paper we reverse the feature extraction part of standard bottom-up architectures and turn them upside-down: We propose top-down networks. Our proposed coarse-to-fine pathway, by blurring higher frequency information and restoring it only at later stages, offers a line of defence against adversarial attacks that introduce high frequency noise. Moreover, since we increase image resolution with depth, the high resolution of the feature map in the final convolutional layer contributes to the explainability of the network's decision making process. This favors object-driven decisions over context driven ones, and thus provides better localized class activation maps. This paper offers empirical evidence for the applicability of the top-down resolution processing to various existing architectures on multiple visual tasks.

1. Introduction

In human biological vision, perceptual grouping of visual features is based on Gestalt principles, where factors such as proximity, similarity or good continuation of features generate a salient percept [42]. Salient objects are rapidly and robustly detected and segregated from the background in what is termed the "pop-out" effect [7, 22]. This initial detection and grouping of salient features into a coherent percept, leads to preferential processing by the visual system, described as stimulus-driven attention [52]. For relevant visual stimuli, the exogenously directed attention is sustained, and results in a more detailed visual evaluation of the object. This typical pipeline of perception and attention allocation in biological vision represents an efficient, coarse-to-fine processing of information [14]. In contrast, modern CNNs (Convolutional Neural Networks) do not incorporate this perspective [12, 23, 38, 40].

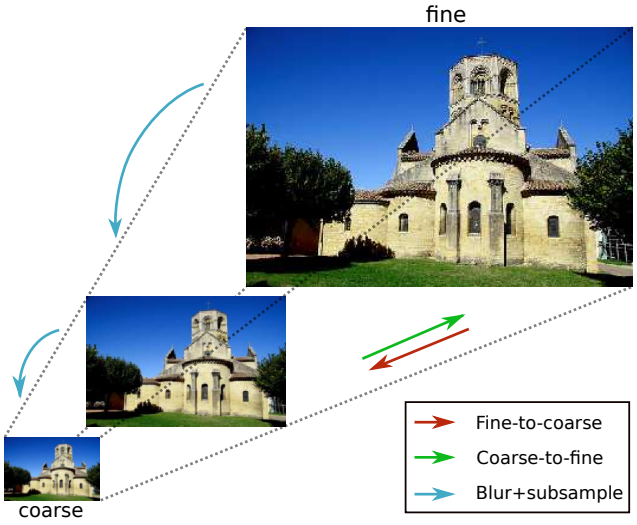

Figure 1. A coarse-to-fine versus fine-to-coarse processing pathway. The conventional fine-to-coarse pathway in a CNN sacrifices localization for semantically richer information. The opposite path, proposed in this paper, starts from the coarsest input and focuses on the context: given the sky, grass and building, it is clearly a landscape scene of a building. Moving to finer representations of the input, the focus shifts to local information. Architectural aspects of the building, and the cross on the top, are now the most informative for classifying the image as a church. Our proposed coarse-to-fine pathway is in line with human biological vision, where detection of global features precedes the detection of local ones, for which further processing of the stimuli is required.

Standard CNNs begin with the high resolution input, and propagate information in a fine-to-coarse pathway. Early layers learn to extract local, shareable features, whereas deeper layers learn semantically rich and increasingly invariant representations. In this paper we propose the reversal of the conventional feature extraction of standard CNNs, as depicted in Figure 1. More specifically, we suggest the adoption of a coarse-to-fine processing of the input, which

can be interpreted as gradual focusing of visual attention. The top-down hierarchy first extracts the gist of a scene, starting from a holistic initial representation, and subsequently enhances it with higher frequency information.

A growing body of literature since the seminal work of [10, 41] shows that adversarial perturbations with high-frequency components may cause substantial misclassifications. Suppressing higher frequencies in the input image, as proposed in our top-down paradigm, can provide a first line of defence. At the same time, explainability of the decision making process of CNNs has recently emerged as an important research direction [36, 54]. In this context, our coarse-to-fine processing scheme, having feature maps with higher spatial resolution at deeper layers, favors object-driven decisions over context-driven ones, and provides better localized class activation maps.

We make the following contributions: (i) We propose biologically inspired top-down network architectures, obtained by reversing the resolution processing of conventional bottom-up CNNs; (ii) We analyze various methods of building top-down networks based on bottom-up counterparts as well as the difference in resolution-processing between these models, providing a versatile framework that is directly applicable to existing architectures; (iii) We compare our proposed model against the baseline on a range of adversarial attacks and demonstrate enhanced robustness against certain types of attacks. (iv) We find enhanced explainability for our top-down model, with potential for object localization tasks. Trained models and source code for our experiments are available online: https://github.com/giannislelekas/topdown.

2. Related work

Coarse-to-fine processing. Coarse-to-fine processing is an integral part of efficient algorithms in computer vision. Iterative image registration [30] gradually refines registration from coarser variants of the original images, while in [16] a coarse-to-fine optical flow estimation method is proposed. Coarse-to-fine face detection is performed by processing increasingly larger edge arrangements in [8], and coarse-to-fine face alignment using stacked auto-encoders is introduced in [50]. Efficient action recognition is achieved in [44] by using coarse and fine features coming from two LSTM (Long Short-Term Memory) modules. In [34] coarse-to-fine kernel networks are proposed, where a cascade of kernel networks are used with increasing complexity. Existing coarse-to-fine methods consider both coarse input resolution, as well as gradually refined processing. Here, we also focus on coarse-to-fine image resolution, however we are the first to do this in a single deep neural network, trained end-to-end, rather than in an ensemble.

Bottom-up and top-down pathways. Many approaches exploit high spatial resolution for finer feature localization, which is crucial for semantic segmentation. The U-net [33] and FPN (Feature Pyramid Networks) [29] merge information from bottom-up and top-down pathways, combining semantically rich information of the bottom-up with the fine localization of the top-down stream. Similarly, combinations of a high-resolution and a low-resolution branch were proposed for efficient action recognition [5], for face hallucination [25], and depth map prediction [3]. Top-down signals are also used to model neural attention via a back-propagation algorithm [49], and to extract informative localization maps for classification tasks in Grad-CAM [36]. Similarly, we also focus on top-down pathways where we slowly integrate higher levels of detail, however our goal is biologically-inspired resolution processing, rather than feature-map activation analysis.

Multi-scale networks. Merging and modulating information extracted from multiple scales is vastly popular [15, 21, 47, 48, 46]. In [48] feature maps are resized by a factor to obtain cascades of multiple resolutions. Incremental resolution changes during GAN (Generative Adversarial Network) training are proposed in [20]. Convolutional weight sharing over multiple scales is proposed in [1, 47]. Similarly [6] performs convolutions over multiple scales in combination with residual connections. In [21] convolutions are performed over a grid of scales, thus combining information from multiple scales in one response, and [39] combines responses over multiples scales, where filters are defined using 2D Hermite polynomials with a Gaussian envelope. Spatial pyramid pooling is proposed in [11] for aggregating information at multiple scales. In this work, we also extract multi-resolution feature maps, in order to start processing from the lowest image scale and gradually restore high frequency information at deeper layers.

Beneficial effects of blurring. Suppressing high frequency information by blurring the input can lead to enhanced robustness [43, 53]. Models trained on blurred inputs exhibit increased robustness to distributional shift [19]. The work in [9] reveals the bias of CNNs towards texture, and analyzes the effect of blurring distortions on the proposed Stylized-ImageNet dataset. Anti-aliasing by blurring before downsampling contributes to preserving shift invariance in CNNs [51]. By using Gaussian kernels with learnable variance, [37] adapts the receptive field size. Rather than changing the receptive field size, works such as [27, 26, 31] use spatial smoothing for improved resistance to adversarial attacks. Similarly, we also rely on Gaussian blurring before downsampling the feature maps to avoid aliasing effects, and as a consequence we observe improved robustness to adversarial attacks.

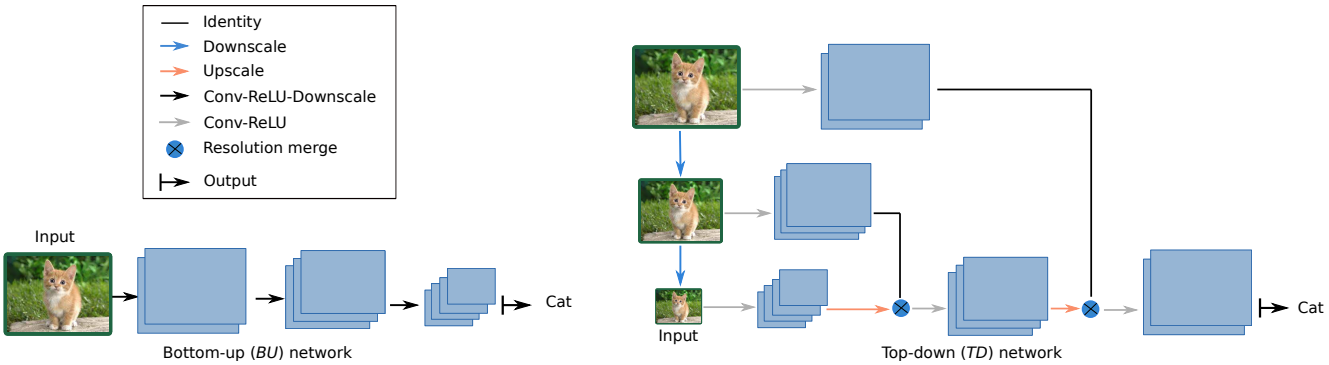

Figure 2. **Left:** The bottom-up (BU) baseline network. Feature maps decrease in spatial resolution with network depth. **Right:** The proposed top-down (TD) network. The TD model reverses the feature extraction pathway of the baseline network. It employs three inputs from highest to lowest scale, starts processing from the lowest resolution and progressively adds high resolution information.

3. Top-down networks

Top-down (TD) networks mirror the baseline bottom-up (BU) networks, and reverse their feature extraction pathway. Information flows in the opposite direction, moving from lower to higher resolution feature maps. The initial input of the network corresponds to the minimum spatial resolution occurring in the BU baseline network. Downscaling operations are replaced by upscaling, leading to the coarse-to-fine information flow. By upscaling, the network can merely "hallucinate" higher resolution features. To restore the high frequency information, we use resolution merges, which combine the hallucinated features with higher frequency inputs, after each upscaling operation. Figure 2 depicts the difference between the BU architecture and our proposed TD architecture.

3.1. Input and feature map resizing

To avoid artifacts hampering the performance of the network [51], we blur the inputs before downsampling. For the upsampling operation we use interpolation followed by convolution. We have experimented with both nearest neighbour and bilinear interpolation, and have noticed improved robustness against adversarial attacks for nearest neighbor interpolation. We have also considered the use of transpose convolutions, however we did not adopt these due to detrimental checkerboard artifacts.

3.2. Merging low and high resolution

Figure 3 depicts the considered method for merging the high resolution input with the low resolution information. We first upsample the low resolution input via a 1×1 convolution and use an element-wise addition with the high-resolution branch. This information is then concatenated with the original high resolution information on the channel dimension. We subsequently use a 3×3 convolution to expand the receptive field of the filters. The proposed merging of information slightly increases the number of parameters, while being effective in practice.

ERF (effective receptive field) size computation. Neurons in each layer i of a typical bottom-up network has a single ERF size r_i determined by the kernel size k_i and the cumulative stride m_i (given stride s_i at layer i).

$$\begin{aligned} r_i &= r_{i-1} + (k_i - 1)\, m_{i-1} \\ m_i &= m_{i-1} \cdot si \end{aligned} \tag{1}$$

Assuming only 3×3 convolutions with stride 1, the example BU architecture in Figure 2 will have an ERF size of 3 pixels, and 18 pixels in each direction after the first and final convolutional layers, respectively. In contrast, for the TD network, considering a Gaussian blurring window of width 6σ, the lowest resolution branch will already have an ERF size of $12\sigma+2$ at the input level and of $12\sigma+10$ after the first convolutional layer (comparable to the final layer of a BU network already with $\sigma = 2/3$ pixels). Furthermore, in contrast to BU, output from neurons with varying ERFs are propagated through the merging points. To get a lower bound on the TD ERF sizes, we consider resolution merging methods which do not provide RF enlargement (e.g. as depicted in fig. 3, but without the 3×3 convolution at the end). Thus, at the final merging point of the TD architecture, ERF sizes of 3 pixels and $12\sigma+14$ pixels are merged together. In conclusion, already from the first layer, TD has the ERF size that the BU only obtains at the last layer.

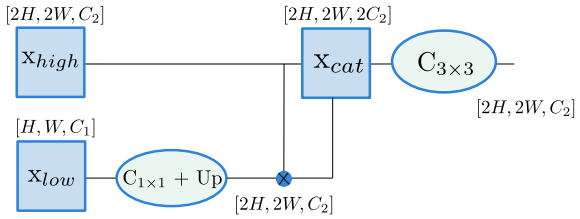

Figure 3. Merging low and high-frequency feature maps: we use a 1×1 convolution followed by an element-wise addition; this information is concatenated with the high-resolution input and followed by a 3×3 convolution that expands the receptive field size.

3.3. Filter arrangement

Feature extraction pathway of the TD network reverses the BU: information propagates from lower to higher spatial dimensions in a TD network, while the number of filters shrinks with increasing depth. The choice of expanding the number of filters at deeper layers in the BU network is efficiency-oriented. As the feature map resolution decreases, the number of channels increases, retaining the computational complexity roughly fixed per layer. Typically, in standard architectures the filters are doubled every time dimensions are halved [12, 38].

In our method we consider three options for deciding the number of filters per layer: the TD model which is exactly the opposite of the BU in that the number of channels are reduced with depth; the *uniform* model (TD_{uni}) where the layers have a uniform number of filters; and the *reversed* model (TD_{rev}) which follows the BU filter arrangement, with channel dimension widened with depth.

4. Experiments

In **Exp 1** we evaluate the three different filter arrangement options proposed for the top-down model. We compare these model variations with the bottom-up baseline on the MNIST, Fashion-MNIST and CIFAR10 classification tasks. In **Exp 2** we evaluate the robustness of our proposed model against various adversarial attacks applied on the same datasets. Finally, in **Exp 3** we illustrate the explainability capabilities of our top-down model when compared to the bottom-up, and demonstrate its benefits for a small object localization task.

Experimental setup. We compare our TD proposal with its respective BU baseline on MNIST, Fashion-MNIST and CIFAR10. For the simpler MNIST tasks we consider as baselines the "LeNetFC", a fully-convolutional variant of LeNet [24] and following [28], a lightweight version of the NIN (Network-In-Network) architecture, namely "NIN-light" with reduced filters. The original architecture was used for the CIFAR10 task, along with the ResNet32 introduced in [12] incorporating the pre-activation unit of [13]. Batch Normalization [17] is used in all the networks prior to the non-linearities. The corresponding TD networks are defined based on their BU baselines. Table 1 depicts the number of parameters of different models. For TD we consider three variants: TD – which is mirroring the BU architecture also in terms of filter depth; TD_{uni} using uniform filter depth; and TD_{rev} where the filter depth of the TD is reversed, thus following the filter depth of BU. There is an increase in the number of parameters for the TD networks, because we need additional convolutional layers for merging the high and low resolution information.

We abide by the setup found in the initial publications for the BU models. For the TD networks we performed a linear search for learning rate, batch size, and weight decay. For all cases we train with a 90/10 train/val split, using SGD with momentum of 0.9 and a 3-stage learning rate decay scheme, dividing the learning rate by 10 at 50% and 80% of the total number of epochs. For the CIFAR10 dataset we test with and without augmentation—employing horizontal translations and flips. We repeat runs four times, with dataset reshuffling and extracting new training and validation splits, and report mean and standard deviation of the test accuracy.

Model	#parameters			
	BU	TD	TD_{uni}	TD_{rev}
LeNetFC	8k	14k	23k	58k
NIN-light	62k	213k	215k	214k
ResNet32	468k	528k	320k	563k
NIN	970k	3,368k	3,397k	3,388k

Table 1. Number of trainable parameters for the different architectures considered. Different rows correspond to the different baseline architectures and columns indicate the bottom-up model and the three top-down variants with different filter arrangements (section 3.3). There is an increase in the number of parameters for the TD networks, because they merge the high and low resolution information using additional convolutional layers.

4.1. Exp. 1: Bottom-up versus top-down

Figure 4 shows the test accuracy of the considered models across datasets. The TD networks are on par with, and in some cases surpassing the corresponding baseline BU performance. When considering the different filter depth configurations, TD_{rev} performs best due to increased representational power at higher scales, coming though at cost of increased complexity. The NIN architecture adopts a close to uniform filter arrangement, hence the three TD variants reach roughly the same performance. We adopt the TD variants henceforth, on account of the small gap in performance and reduced complexity. This experiment provides empirical evidence of the applicability of the proposed pipeline to different network architectures.

4.2. Exp. 2: Adversarial robustness

We evaluate the robustness of BU versus TD against various attacks, where we attack the test set of each dataset using the Foolbox [32]. For all the attacks, the default parameters were used. To make the attack bound tighter, we repeat each attack three times and keep the worst case for each to define the minimum required perturbation for fooling the network.

Figure 6 provides for each attack, plots of loss in test accuracy versus the $L2$ distance between the original and the perturbed input. TD networks are visibly more re-

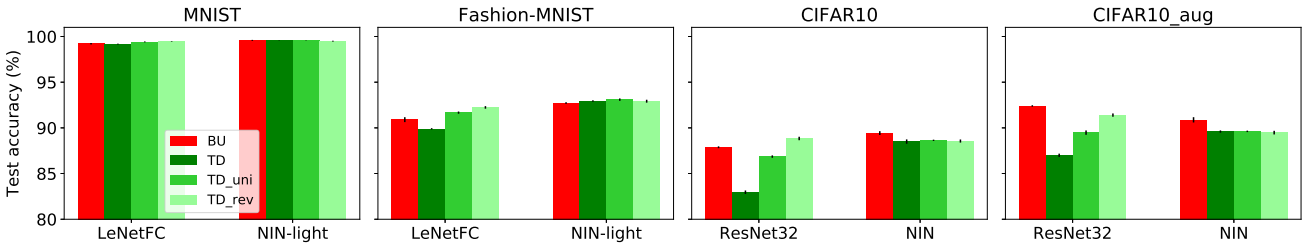

Figure 4. **Exp 1:** Comparison of MNIST, Fashion-MNIST, CIFAR10, and CIFAR10_aug (with augmentation) mean test accuracies between BU and the three different configurations of TD proposed in subsection 3.3. TD networks perform on par with, and at times surpassing, the baseline performance of its respective BU. Regarding filter depth configurations, TD_{rev} displays the highest performance, at the cost of increased parameters. Considering the small gap in performance and the increased cost for TD_{rev}, we henceforth adopt the TD configuration.

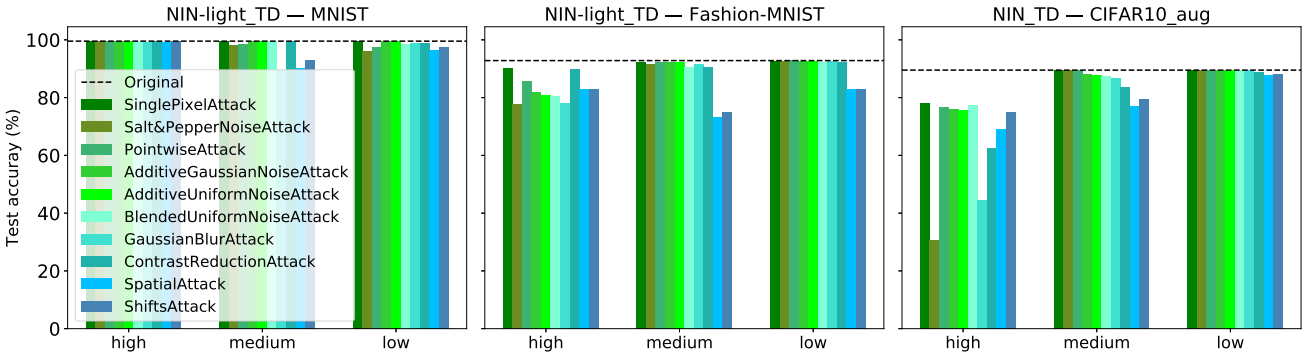

Figure 5. **Exp 2:** Test accuracy when extracted adversarial perturbations are fed to either the highest, medium, or lowest scale input of the TD network (refer to figure 2), using the NIN-light baseline on MNIST and Fashion-MNIST, and NIN on CIFAR10. The remaining two inputs are fed the original, unperturbed samples. As the dataset becomes more challenging, the highest vulnerability moves from the medium input to the highest scale input. This is attributed to the absence of information in the high frequency region for the simpler cases: i.e. MNIST. (See supplementary material for additional results.)

silient against attacks introducing uncorrelated noise, due to the coarse-to-fine processing adopted, with downscaled inputs diminishing the noise. For attacks introducing correlated noise such as the "Pointwise" attack [35], the perturbed pixels tend to lie in smooth regions of the image. Thus each single pixel value of 0 (or 1) in a region of 1s (or 0s) essentially acts as a Dirac delta function. Based on the convolutional nature of CNNs this type of attack "pollutes" the input with imprints of the learned filters[1], which gradually span a greater part of the feature map as more convolutions are applied. Due to the highly correlated nature of the perturbation, the blurred downsampling can not completely eradicate the noise, but helps decrease the introduced pollution. On the contrary, for BU networks, the noise is directly propagated down the network. Additionally, the blurred downsampling wired in the network architecture offers enhanced robustness against blurring attacks, as the network encounters the input image at multiple scales during training, and is, thus, more resilient to resolution changes. Since anti-aliasing before downsampling is suggested to better preserve shift-invariance [51], we expected our networks to also be more robust against the "Spatial" attack [4]. However, no enhanced robustness is reported for TD networks; a substantial difference in robustness is observed for ResNet32, which could be due to the performance gap measured in **Exp 1** between the TD and its BU baseline. We also tested with the TD_{uni} and TD_{rev} variants of the ResNet32 architecture, with respective results provided in the supplementary material.

To get a better insight on TD robustness, we introduce the generated attacks to a single resolution branch of the TD networks using the NIN-light architecture on MNIST and Fashion-MNIST, and NIN on CIFAR10. This is displayed in figure 5. We feed the extracted perturbations to either the low, medium or high resolution input branch, as illustrated in the model architecture in figure 2. For the simpler MNIST task, the medium-resolution input of the network is the most vulnerable, which is mainly attributed to

[1]For imperfect delta function, this yields blurred versions of the filters.

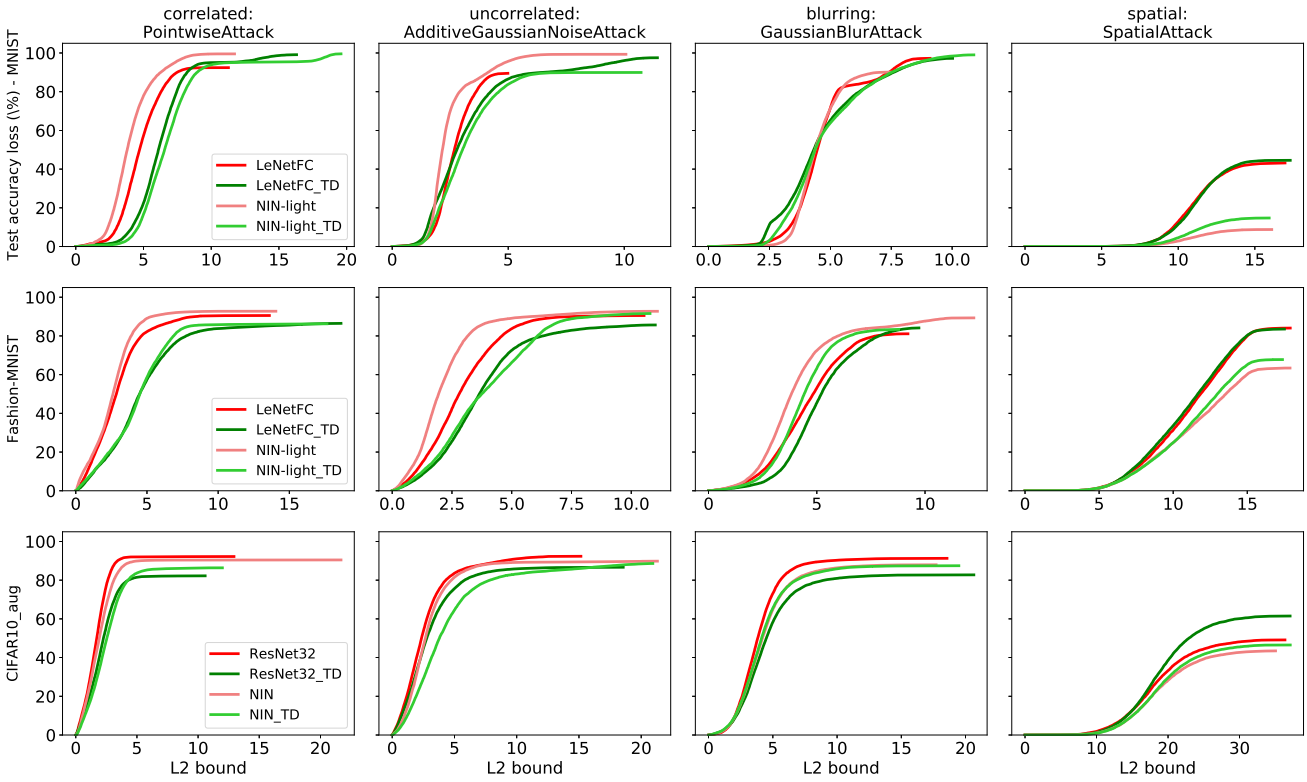

Figure 6. **Exp 2:** Comparison of adversarial robustness considering different datasets, models and attacks. The x-axis of each figure corresponds to the $L2$ distance between the original and the perturbed image and the y-axis is the introduced loss in test accuracy. A lower curve suggests increased robustness. Green curves corresponding to TD are consistently underneath the respective red curves of the BU networks, for most attacks. The TD networks are more robust against both correlated and uncorrelated noise attacks due to the coarse-to-fine processing, suppressing high frequency information on earlier stages. Additionally, the blurred downsampling offers enhanced robustness against blurring attacks. For spatial attacks, we see no increased robustness. (See supplementary material for additional results.)

the absence of information in the high frequency region of the input's spectrum. Moving to more challenging Fashion-MNIST and CIFAR10 tasks, the high frequency input becomes the easiest path for fooling the network. Please see the supplementary material for additional results when perturbing two inputs simultaneously.

4.3. Exp 3: Explainability and localization

(a) Grad-CAM heatmap visualizations. Grad-CAM [36] provides class-discriminative localization maps, based on the feature maps of a convolutional layer, highlighting the most informative features for the classification task. Here, we use the features of the last convolutional layer. The extracted heatmap is restored to the original image scale, thus producing a coarser map in the case of the BU whose feature map size at the final layer is smaller. On the contrary, for TD the corresponding scale of the feature maps matches the scale of the input, hence Grad-CAM outputs a finer map.

The Grad-CAM heatmaps corresponding to a BU and TD network are provided in figure 7. These are obtained from various layers of a ResNet18 architecture [12] trained on the Imagenette dataset [18]. For further information about the setup please refer to supplementary material. "Layer 1" corresponds to the activation of the input to the first group of residual blocks, and "Layer 2" to "Layer 5" to the activations of the output of each of these four groups, each one corresponding to different spatial resolution. The visualizations demonstrate that TD follows an opposite, coarse-to-fine path starting from a coarser representation and gradually enriching it with higher frequency information. Hence, TD networks do not only mirror the BU solely in the architectural design, but also in their learning process.

Additional heatmaps corresponding to correctly classified images, taken from the last convolutional layer of the networks are visualized in figure 8. The figures depict the coarse localization in BU versus the fine localization in TD. We selected intentionally images with multiple objects. The TD networks recognize objects based on fine-grained information: such as the spots on the dog, the cross on the church or boundary information of various objects.

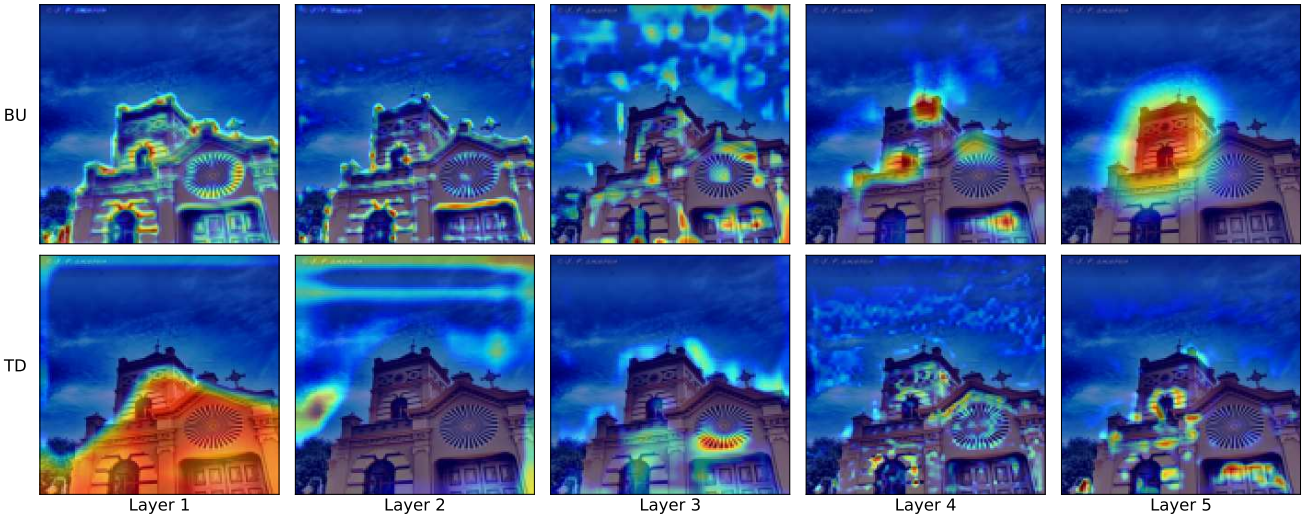

Figure 7. **Exp 3.(a):** Fine-to-coarse versus coarse-to-fine processing. We show Grad-CAM heatmaps for ResNet18 BU versus its respective TD, trained on the Imagenette dataset [18] for a random validation image. Higher layer index means increased depth in the architecture: "Layer 1" corresponds to the activation of the input to the first group of residual blocks, and "Layer 2" to "Layer 5" to the activations of the output of each of these four groups, each one corresponding to different spatial resolution. **Top**: the BU network, employing fine-to-coarse processing. **Bottom**: the respective TD network following the opposite path, starting with a holistic representation and gradually adding higher frequency information in deeper layers.

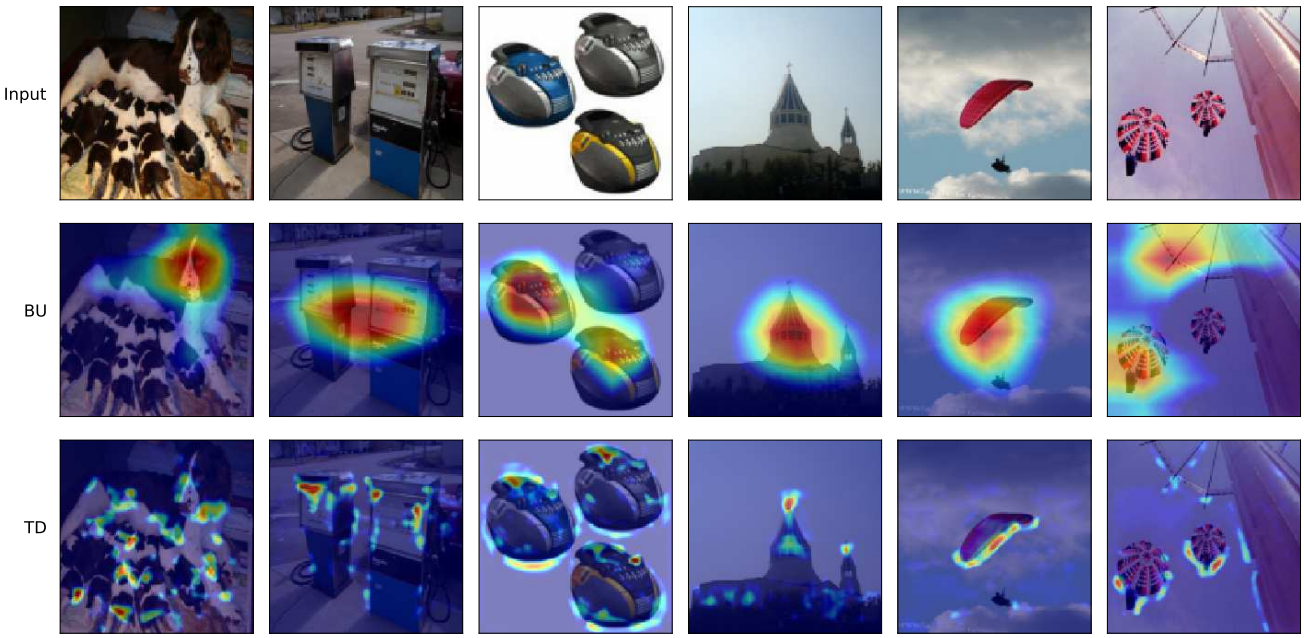

Figure 8. **Exp 3.(a):** Grad-CAM heatmaps corresponding to the last convolutional layer in the network. **Top**: The original input image, randomly selected from the validation set. **Middle**: Corresponding Grad-CAM heatmaps for the BU ResNet18. **Bottom**: Grad-CAM heatmaps for the TD ResNet18. Contrary to the coarse output of the BU, the TD network outputs high frequency feature maps, based on which the final classification is performed. TD recognized objects based on their fine-grained attributes: such as the spots on the dogs, or the cross on the church, or shape information. (See supplementary material for additional results.)

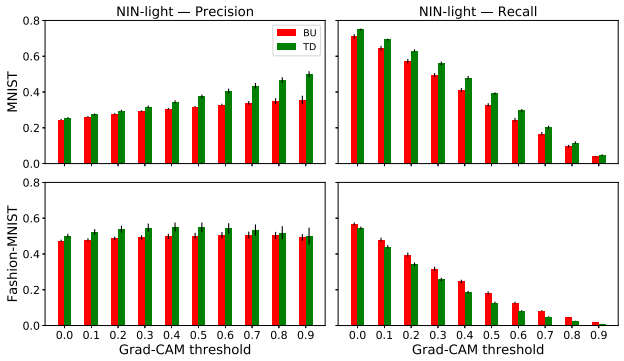

Figure 9. **Exp 3.(b):** Precision and recall for the MNIST and Fashion-MNIST datasets using the NIN-light architecture. The numbers are reported over four runs and we also plot standard deviations. For each run, models are trained from scratch and the set of TP (true positive), FP (false positive), FN (false negative) is computed, between the Grad-CAM heatmaps and the segregated objects. The TD model has higher precision on both MNIST and Fashion-MNIST due to more accurate object localization, while having slightly lower recall than BU on the Fashion-MNIST.

(b) Weakly-supervised object localization. For a quantitative evaluation of the localization abilities of TD, we used the MNIST and Fashion-MNIST datasets and the NIN-light model as a backbone architecture. Figure 9 shows mean precision and recall scores for the TD and BU models over four runs. For each run models were trained from scratch, then TP (true positive), FP (false positive), FN (false negative) values were computed between the Grad-CAM heatmaps and the thresholded objects, corresponding to the test set of the considered task. We used a threshold empirically set to $t = 0.2$. Based on the computed values precision and recall scores were extracted and aggregated over the four runs. For a fair comparison only the samples correctly classified from both TD and BU were considered.

The TD models report higher precision for both tasks considered, suggesting finer object localization. The lower recall scores for the Fashion-MNIST is attributed to the higher number of FN compared to the BU model. The larger object sizes of the Fashion-MNIST task, along with the coarse output of the BU model, being able to capture a greater extent of them, leads to fewer FN. On the contrary, the TD models focus on finer aspects of the objects, which are informative for the classification task. Considering the fine-grained focus in the Grad-CAM outputs and the potential for weakly-supervised object localization, our proposed TD networks comprise a promising direction for future research.

5. Discussion

The current work aims at providing a fresh perspective on the architecture of CNNs, which is currently taken for granted. The coarse-to-fine pathway is biologically inspired by how humans perceive visual information: first understanding the context and then filling in the salient details.

One downside of our proposed TD networks is that expanding dimensions at increased network depth leads to memory and computational bottlenecks. This is due to the feature map size being larger at higher depths. Moreover, for the same reason, adding fully-connected layers before the output layer of the TD architectures leads to a vast increase in the number of model parameters. Hence, fully convolutional networks are preferable. This increase in memory is also more visible with large-scale datasets such as ImageNet [2]. A simple workaround requiring no architectural adaptations would be to employ mixed-precision training, which would decrease the memory requirements, but would increase the computational complexity. Instead of increasing the spatial resolution of the feature maps at later depths, we could use patches of the input of limited sizes. The selection of these informative patches could be defined using the Grad-CAM heatmaps by selecting the high-activation areas of the heatmap, or considering self-attention mechanisms [45]. In addition to addressing the aforementioned limitations, we find the weakly-supervised setting to be a promising area of future research.

6. Conclusion

In the current work, we revisit the architecture of conventional CNNs, aiming at diverging from the manner in which resolution is typically processed in deep networks. We propose novel network architectures which reverse the resolution processing of standard CNNs. The proposed TD paradigm adopts a coarse-to-fine information processing pathway, starting from the low resolution information, providing the visual context, and subsequently adding back the high frequency information. We empirically demonstrate the applicability of our proposed TD architectures when starting from a range of baseline architectures, and considering multiple visual recognition tasks. TD networks exhibit enhanced robustness against certain types of adversarial attacks. This resistance to adversarial attacks is induced directly by the network design choices. Additionally, the high spatial dimensions of the feature maps in the last layer significantly enhance the explainability of the model, and demonstrate potential for weakly-supervised object localization tasks.

References

[1] Shubhra Aich, Masaki Yamazaki, Yasuhiro Taniguchi, and Ian Stavness. Multi-scale weight sharing network for image recognition. *Pattern Recognition Letters*, 131:348–354, 2020.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. *CoRR*, 2017.

[5] Quanfu Fan, Chun-Fu Richard Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *Advances in Neural Information Processing Systems*, pages 2261–2270, 2019.

[6] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. Scale-wise convolution for image restoration. *Association for the Advancement of Artificial Intelligence (AAAI*, 2020.

[7] David J Field, Anthony Hayes, and Robert F Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision research*, 33(2):173–193, 1993.

[8] Francois Fleuret and Donald Geman. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41(1-2):85–107, 2001.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2019.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[12] 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.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on Computer Vision*, pages 630–645, 2016.

[14] Jay Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Progress in neurobiology*, 2008.

[15] Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5743–5752, 2016.

[16] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, 2015.

[18] FastAI Jeremy Howard. The imagenette dataset. `https://github.com/fastai/imagenette`.

[19] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, 2017.

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018.

[21] Tsung-Wei Ke, Michael Maire, and Stella X Yu. Multigrid neural architectures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6665–6673, 2017.

[22] I Kovács and B Julesz. A closed curve is much more than an incomplete one: effect of closure in figure-ground segmentation. *PNAS*, 1993.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[25] Mengyan Li, Yuechuan Sun, Zhaoyu Zhang, and Jun Yu. A coarse-to-fine face hallucination method by exploiting facial prior knowledge. In *International Conference on Image Processing (ICIP)*, pages 61–65, 2018.

[26] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5772, 2017.

[27] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial examples in deep networks with adaptive noise reduction. *CoRR*, 2017.

[28] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *International Conference on Learning Representations*, 2014.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[30] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[31] Ravi Raju and Mikko Lipasti. Blurnet: Defense by filtering the feature maps. *CoRR*, 2019.

[32] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *CoRR*, 2017.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[34] Hichem Sahbi. Coarse-to-fine deep kernel networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1131–1139, 2017.

[35] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *CoRR*, 2018.

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International conference on Computer Iision*, 2017.

[37] Evan Shelhamer, Dequan Wang, and Trevor Darrell. Blurring the line between structure and learning to optimize and adapt receptive fields. *CoRR*, 2019.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[39] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *International Conference on Learning Representations*, 2020.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

[42] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological bulletin*, 2012.

[43] Haohan Wang, Xindi Wu, Pengcheng Yin, and Eric P Xing. High frequency component helps explain the generalization of convolutional neural networks. *CoRR*, 2019.

[44] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. In *Advances in Neural Information Processing Systems*, pages 7778–7787, 2019.

[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*, 2015.

[46] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *CoRR*, 2014.

[47] Taojiannan Yang, Sijie Zhu, Shen Yan, Mi Zhang, Andrew Willis, and Chen Chen. A closer look at network resolution for efficient network design. *CoRR*, 2019.

[48] Chengxi Ye, Chinmaya Devaraj, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Evenly cascaded convolutional networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4640–4647, 2018.

[49] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, 2016.

[50] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*, pages 1–16, 2014.

[51] Richard Zhang. Making convolutional networks shift-invariant again. *International Conference on Machine Learning*, 2019.

[52] Xilin Zhang, Li Zhaoping, Tiangang Zhou, and Fang Fang. Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron*, 2012.

[53] Zhendong Zhang, Cheolkon Jung, and Xiaolong Liang. Adversarial defense by suppressing high-frequency components. *CoRR*, 2019.

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.