# IIEU: Rethinking Neural Feature Activation from Decision-Making

Sudong Cai
Graduate School of Informatics, Kyoto University
scai@vision.ist.i.kyoto-u.ac.jp

## Abstract

*Nonlinear **Act**ivation (**Act**) models which help fit the underlying mappings are critical for neural representation learning. Neuronal behaviors inspire basic Act functions,* e.g.*, Softplus and ReLU. We instead seek improved explainable Act models by re-interpreting neural feature Act from a new philosophical perspective of **M**ulti-**C**riteria **D**ecision-**M**aking (**MCDM**). By treating activation models as selective feature re-calibrators that suppress/emphasize features according to their importance scores measured by feature-filter similarities, we propose a set of specific properties of effective Act models with new intuitions. This helps us identify the unexcavated yet critical problem of* mismatched feature scoring *led by the differentiated norms of the features and filters. We present the **I**nstantaneous **I**mportance **E**stimation **U**nit**s** (**IIEUs**), a novel class of interpretable Act models that address the problem by re-calibrating the feature with the **I**nstantaneous **I**mportance (**II**) score (which we refer to as) estimated with the adaptive norm-decoupled feature-filter similarities, capable of modeling the cross-layer and -channel cues at a low cost. The extensive experiments on various vision benchmarks demonstrate the significant improvements of our IIEUs over the SOTA Act models and validate our interpretation of feature Act. By replacing the popular/SOTA Act models with IIEUs, the small ResNet-26s outperform/match the large ResNet-101s on ImageNet with far fewer parameters and computations.*

## 1. Introduction

Nonlinear Act models are the foundation for the unprecedented success of neural networks in pattern recognition tasks [11, 46, 43, 7]. The choice of the Act model is a decisive yet non-trivial factor in the performance of a neural network. Basic methods such as ReLU [38] and Softplus [16] are originated from neuronal behaviors [44, 27]. Based on them, past works have proposed to improve Act models with channel context (*e.g.*, FReLU [32], DyReLU [10] and ACONs [31]), statistical strategies (*e.g.*, GELU [22], Pserf [4], and SMU [5]), and task-specific

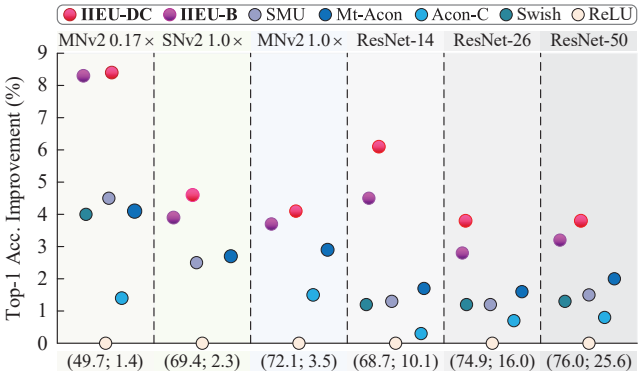



Figure 1. ImageNet Top-1 Accuracy (Acc.) relative improvements compared with the ReLU [38] baselines and SOTAs (Swish [40], ACONs [31](CVPR'21), and SMU [5](CVPR'22)) with (1) MobileNetV2 [42] (MNv2) 0.17× and 1.0×; (2) ShuffleNetV2 [33] (SNv2) 1.0×; (3) ResNet-14, -26, and 50 [21]. We show the ReLU baseline results by "(Acc.(%); parameters(M))". Our IIEUs achieve the new SOTA improvements to the ReLU baselines and outperform the SOTAs remarkably, with negligible/marginal additional parameters to ReLU (shown by the relative areas of the circular patterns, where each ReLU network denotes the unit area).

periodic functions [35, 45]. Existing methods, however, still leave critical problems in the optimal decision on Act models. As a major reason, although several past efforts [37, 2, 19] suggested extending Act models with dynamic approximators, it still lacks tailored interpretations to help specify the properties of effective Act models for pattern recognition. These specific properties, however, are difficult to be identified from pure biological intuitions.

To explore new improvements in feature Act, we rethink neural operations from MCDM (a typical problem in operational research) [41, 39, 26, 15, 9, 50]. As a core of our interpretation, we treat a nonlinear Act model as a selective re-calibrator that suppresses or emphasizes features according to their importance. Such importance, in fact, is first modeled by the feature-filter inner product which supposes to indicate the similarity of the feature to the filter. However, differentiated feature and filter norms can significantly bias the similarities modeled with feature-filter inner-products, thus likely interfering with the estimation of actual feature importance. We identify this as a critical yet

unexcavated problem, namely *mismatched feature scoring* (as discussed in Figure 2(a)), which we infer from our interpretation and otherwise invisible to past explanations.

To address the problem, we propose a set of specific properties of effective Act models with new intuitions and introduce the initial solution *i.e.*, a novel kind of explicable Act models which we refer to as the *IIEUs*, to selectively re-calibrate features with an adaptive norm-decoupled importance measure. Specifically, we first treat each feature-filter inner product (suppose without biases and normalization layers) as a **T**ransitive **I**mportance (**TI**) score, as its input feature vector is de facto determined by a series of prior learning factors (*e.g.*, the initial input, filters and Acts of the prior layers) and transmits their cues. We then estimate the corresponding norm-decoupled **I**nstantaneous **I**mportance (**II**) score with a low-cost adaptive shift term that incorporates mild learning adjustments. Finally, the feature Act is realized by multiplying each TI-score with the II-score. This feature re-calibration preserves meaningful prior learning information carried by the TI-scores yet eliminates the negative effect led by the *mismatched feature scoring* problem. Note that we formalize the *mismatched feature scoring* problem and TI-, II-scores in Section 2.

The contributions of this work are 3-fold: (1) We propose to interpret neural feature Act from MCDM, where we identify the unexplored problem of *mismatched feature scoring* and introduce a set of specific properties with our intuitions to help explain the working mechanism of Act models. (2) We present explainable IIEU(s) as the initial solution to the problem identified from (1). (3) We extensively validate the (a) effectiveness and versatility of IIEUs with various vision benchmarks, where IIEUs significantly improve the SOTA Act models; (b) our interpretation with targeted ablation studies. Code is disseminated at `https://github.com/SudongCAI/IIEU`.

## 2. Rethinking Feature Activation from MCDM

We aim to interpret neural feature Act from MCDM, find the unexplored critical problem, and propose our novel Act model by addressing the new problem. We first clarify our *Intuitions* and their induced *Properties*. We then present our IIEU constructed on them. With the preliminaries, **we formalize the *Properties* with *Definitions* and *Propositions*, where the *Deductions* and *Proofs* are detailed in the Appendix (in Supp, marked by $\ast$)**. For coherence, we discuss the related works in Section 3 with our interpretation.

### 2.1. Preliminaries

We consider the simple settings with image inputs: (1) A network has $T$ sequential learning layers indexed by $\tau = 1, 2, \ldots, T$. Let $\mathbf{X}^\tau \in \mathbb{R}^{C^\tau \times H^\tau \times L^\tau}$ which has $C^\tau$ channels and a spatial resolution of $H^\tau \times L^\tau$ denote the input feature map of the layer-$\tau$. (2) Let $x_c^{\tau+1}(h,l) \coloneqq \phi\left(\tilde{x}_c^\tau(h,l)\right)$ denote the learning of the layer-$\tau$ at a given location $(h,l) \in \Omega_{H^\tau \times L^\tau}$ with the $c$-th filter $\boldsymbol{w}^\tau(c) \in \mathbb{R}^{C^\tau}$ and feature vector $\boldsymbol{x}^\tau(h,l) \in \mathbb{R}^{C^\tau}$, **where** $\tilde{x}_c^\tau(h,l) = \langle \boldsymbol{w}^\tau(c), \boldsymbol{x}^\tau(h,l) \rangle$ **denotes the inner product** and $\Omega_{H^\tau \times L^\tau}$ is the spatial lattice of $\mathbf{X}^\tau$. Note that the layer-$\tau$ includes a total of $C_{\tau+1}$ filters. $\phi : \mathbb{R} \to \mathbb{R}$ is a given Act function and we suppose $\phi\left(\tilde{x}_c^\tau(h,l)\right) = \rho\left(\tilde{x}_c^\tau(h,l)\right)\tilde{x}_c^\tau(h,l)$, where $\rho : \mathbb{R} \to \mathbb{R}$ defines the reweighting function of $\phi$ about $\tilde{x}_c^\tau(h,l)$.

Note that **(1)** *we first leave aside normalization layers (*e.g.*, BN [25] and LN [3]) and biases for simplicity and will consider them in **Section 2.3 (Method)***. (2) for region-dependent learning with a $K \times K$ convolution, we meet the supposed settings by vectorizing the neighborhood of features/filters from size $C^\tau \times K \times K$ to $C^\tau \cdot K^2$. From MCDM, we treat (1) a filter $\boldsymbol{w}^\tau(c)$ as an updatable ideal candidate[1] of the $c$-th group of criteria (*i.e.*, the $C^\tau$ channels of $\boldsymbol{w}^\tau(c)$); (2) a feature vector $\boldsymbol{x}^\tau(h,l)$ as an alternative candidate whose importance score about a group of criteria is measured by the feature-filter similarity, *i.e.*, **A**lternative-**I**deal (**A-I**) similarity. **Following we omit the layer index $\tau$ and spatial coordinate $(h,l)$ to simplify the notations for the operations of layer-$\tau$** (*e.g.*, we denote $\boldsymbol{x}^\tau(h,l)$, $\boldsymbol{w}^\tau(c)$, and $\tilde{x}_c^\tau(h,l)$ by $\boldsymbol{x}$, $\boldsymbol{w}$, and $\tilde{x}$, respectively).

For clarity, **as for** $\tilde{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$, we suppose the influence of a candidate on the inferencing and filter updating can be quantified as $\tilde{x}$ and $\boldsymbol{x}$, where the corresponding intensities are $|\tilde{x}|$ and $\|\boldsymbol{x}\|$, respectively, as (1) the difference of two vectorial candidates in a standard neural network can be measured by Euclidean distance; (2) the influence of $\boldsymbol{x}$ on the updating of $\boldsymbol{w}$ can be controlled by $\nabla_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \boldsymbol{x}$.

In particular, the supposed settings can be extended to the case: $\phi(\tilde{x}, \blacksquare) = \rho(\tilde{x}, \blacksquare)\,\tilde{x}$, $\phi, \rho : \dot{\mathbb{D}} \to \mathbb{R}$, where $\dot{\mathbb{D}}$ denotes the extended domain of $\tilde{x}$ with other given real variables/constants (denoted by $\blacksquare$), if (1) $\phi$ and $\rho$ are still functions about $\tilde{x}$ when the values of other variables are given; (2) $\rho$ is continuous and differentiable about $\tilde{x}$ or at most has a finite number of points where the left- and right-hand limits exist but are unequal; (3) for a non-differentiable point on $\rho$, let the single-side derivative of the side where the point is defined be the derivative for calculation. In the following, we omit "$\blacksquare$" (*e.g.*, denote $\phi(\tilde{x}, \blacksquare)$ as $\phi(\tilde{x})$) if not specified.

### 2.2. IIEU: Intuitions and Properties

We begin by rethinking "nonlinearity," the foundation of neural feature Act, from MCDM:

*Intuition* 1. The "Nonlinearity" of feature Act can be interpreted as a "loose selectivity" that is necessary but not sufficient to differentiate features by their importance scores.

"Nonlinearity" is indispensable for the learning of discriminative neural representations. The mathematically ab-

[1] With the given conditions, the ideal candidate in MCDM [26, 49, 39, 41] denotes the acquirable/virtual optimal choice capable of quantitatively measuring the performance of an alternative candidate by the similarity.

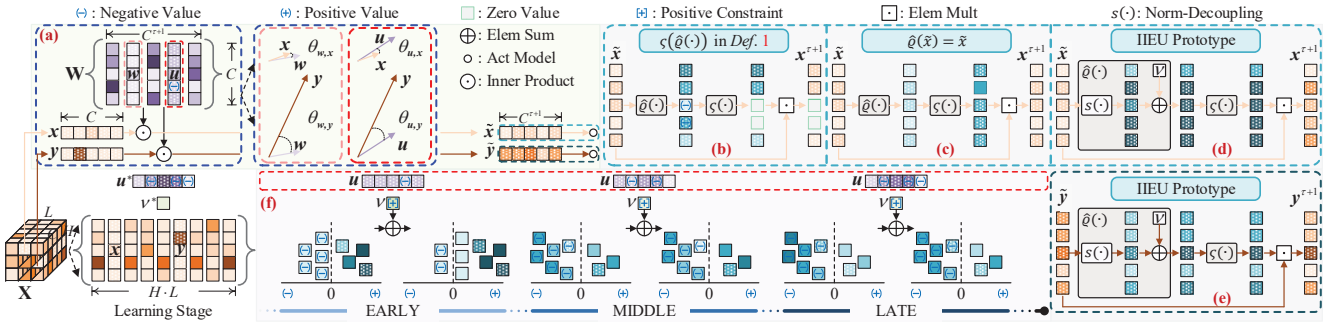



Figure 2. Illustration of the intuitions for IIEU. Suppose w/o normalization layers and biases. The shades of colors denote the intensities (the darker the higher and positive if w/o "(−)"), where "orange," "purple," "aqua," and "olive" denote features, filters, importance scores, and the parameters of the term-*B*. **(a)** *Mismatched feature scoring* problem: it is possible to find feature vectors $\boldsymbol{x}, \boldsymbol{y}$ and filters $\boldsymbol{w}, \boldsymbol{u}$ *s.t.* $\langle \boldsymbol{u}, \boldsymbol{y} \rangle \gg \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ and $\langle \boldsymbol{w}, \boldsymbol{y} \rangle \gg \langle \boldsymbol{w}, \boldsymbol{x} \rangle$, where $\boldsymbol{y}$ is far dissimilar to $\boldsymbol{u}$ and $\boldsymbol{w}$ compared with $\boldsymbol{x}$ to $\boldsymbol{w}$, due to the significant differences of the norms. **(b)** Intuition 1: a "nonlinear" Act model does not be specified to suppress/emphasize candidates with their expected importance. **(c)** An example of typical Act model, where $\tilde{x}$ is directly applied as the approximated similarity $\hat{\varrho}(\tilde{x})$ and the **(a)** is left unsolved. **(d)** and **(e)** IIEU eliminates the **(a)** by scoring feature with the adaptive norm-decoupled approximated similarity, such that the influence of $\boldsymbol{x}$ are relatively emphasized by assigning higher scores compared to $\boldsymbol{y}$. **(f)** *Properties of the term*-B: $\boldsymbol{u}^*, \boldsymbol{\nu}^*$ denote the (virtual) optimal $\boldsymbol{u}, \boldsymbol{\nu}$ for $\boldsymbol{u}, \boldsymbol{\nu}$ to approach in training, respectively. we suppose $\nu$ to be updatable, positive, and bounded since (1) the perfectness of filters as ideal candidates cannot be ensured (as discussed with Intuition 3); (2) we identify the positive translation to the codomain of the approximated similarity $\hat{\varrho}(\tilde{x})$ help to selectively suppress/emphasize the influence of targeted candidates; (3) a bounded $\nu$ ensures that the contribution of the bounded main term-*S* will not be neutralized by the auxiliary $\nu$ (as further discussed in Section 2.3 with the ablation study (4)).

solute "nonlinearity," however, can also be brought by other basic operations, *e.g.*, BN [25], LN [3], and the biases of linear layers. From MCDM, as for Act model, non-important candidates are likely to be scored with negative A-I inner products, where the candidates with intense negative inner products are possible to deteriorate the learning ($\divideontimes$). This necessitates a *selective* re-calibration to suppress/preserve the harmful/positive influence, respectively. Our Intuition 1 aims at bridging the meaning of "Nonlinearity" to "Selectivity" with the Proposition 1:

**Definition 1.** For a function $\rho : \mathbb{R} \to \mathbb{R}$, we refer to this $\rho$ as a function that holds ***Loose Selectivity*** (on $\mathbb{R}$) if: $\exists \tilde{x}, \tilde{y} \in \mathbb{R}$ while $\tilde{x}, \tilde{y} \neq 0$ and $\tilde{x} \neq \tilde{y}$ ***such that (s.t.)*** $\rho(\tilde{x}) \neq \rho(\tilde{y})$.

**Proposition 1.** $\divideontimes$ *For a given $\rho$ and $\phi$: $\phi(\tilde{x}) = \rho(\tilde{x})\tilde{x}$, then, $\rho$ satisfies Definition 1 $\iff$ $\phi$ is nonlinear about $\tilde{x}$.*

Proposition 1 helps us to identify the meaning of a nonlinear Act model as a selective re-calibrator, where its reweighting function $\rho$ can assign unequal weights to different A-I inner products. However, this "loose selectively" is not yet sufficiently specified to suppress/emphasize candidates based on their measured importance scores (Figure 2(b)), thus our goal is to suggest improved selectivity by proposing specific properties with further intuitions.

*Intuition* 2. There exists an A-I similarity measure $\varrho(\tilde{x})$ capable of completely reflecting the importance of $\tilde{x}$ about its criteria, which we refer to as the *ideal similarity*.

By assuming the existence of $\varrho$ which satisfies:

- For any given alternative and ideal candidates $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{w}, \boldsymbol{v}$, suppose $\tilde{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ and $\tilde{y} = \langle \boldsymbol{v}, \boldsymbol{y} \rangle$. If $\varrho(\tilde{x}) \geqslant \varrho(\tilde{y})$, then, $\boldsymbol{x}$ has higher/equal importance than $\boldsymbol{y}$ about their importance measure criteria,

we specify the reweighting function as $\rho(\tilde{x}) = \varsigma(\varrho(\tilde{x}))$, where $\varsigma$ is an ***adjuster*** function that casts suitable constraints on the codomain of $\varrho$ such that:

*Property* 1. $|\varsigma(\varrho(\tilde{x}))| \geqslant |\varsigma(\varrho(\tilde{y}))|$ if $\varrho(\tilde{x}) \geqslant \varrho(\tilde{y})$.

Where $\varrho(\tilde{x})$ is continuous and differentiable at $\tilde{x}, \forall \tilde{x} \in \mathbb{R}$; $\varsigma(\varrho(\tilde{x}))$ is continuous and differentiable about $\varrho(\tilde{x})$ on the domain (or at most has finite points where the left- and right-hand limits of the function exist but are unequal). Note that Property 1 is ensured by $\varsigma$, as the monotonicity of $|\varrho(\tilde{x})|$ about $\varrho(\tilde{x})$ is uncertain. Moreover, Property 1 can be met with a simplified condition, *i.e.*,

**Proposition 2.** $\divideontimes$ *Property 1 $\iff$ (1) $\varsigma(\varrho_x)$ is (monotonically) non-decreasing about $\varrho_x \wedge \varsigma(\varrho_x) \geqslant 0$ $\vee$ (2) $\varsigma(\varrho_x)$ is (monotonically) non-increasing about $\varrho_x \wedge \varsigma(\varrho_x) \leqslant 0$ ($\varrho_x$ denotes $\varrho(\tilde{x})$; $\wedge$ denotes "and;" $\vee$ denotes "or").*

In particularly, we discuss $\varsigma(\varrho(\tilde{x})) \geqslant 0$ (*i.e.*, $\varsigma(\varrho(\tilde{x}))$ is lower-bounded) without loss of generality.

As for real-world application, we identify the suitable approximation to the ideal similarity $\varrho(\tilde{x})$ (denoted by $\hat{\varrho}(\tilde{x})$) a critical problem, as $\varrho(\tilde{x})$ is difficult to be determined or may not exist, which lies in the fact that the underlying mappings of neural learning can be extremely complex. Accordingly, we propose our fundamental intuition to introduce the approximated similarity $\hat{\varrho}(\tilde{x})$ of IIEU with Property 1:

*Intuition* 3. **Mismatched feature scoring**. Typical Act functions directly apply A-I inner product $\tilde{x}$ as $\hat{\varrho}(\tilde{x})$ (*suppose w/o normalization layers and biases*, Figure 2(c)).

However, we identify A-I inner product can be largely biased by the norms of features or/and filters. This can lead to unreliable importance scoring for features. We refer to this problem as the ***mismatched feature scoring*** (Figure 2(a)).

We clarify Intuition 3 with the **Transitive** and **Instantaneous Importance** (**TI** and **II**) scores. We refer to each A-I inner-product $\tilde{x}$ as an **TI-score**, as its input feature $\boldsymbol{x}$ is in fact determined by a series of prior learning factors (*e.g.*, the initial input and filters of the prior layers). That is, TI-score which transmits prior layer information does not exactly measure the current importance of $\boldsymbol{x}$ about the criteria of $\boldsymbol{w}$, as $\tilde{x}$ can be drastically biased by the norms $\|\boldsymbol{x}\|$ and $\|\boldsymbol{w}\|$. In contrast, we suppose **II-score** measures the current importance of $\boldsymbol{x}$ with its norm-independent similarity to $\boldsymbol{w}$.

We identify the cosine similarity, *i.e.*, $\cos\theta_{\boldsymbol{w},\boldsymbol{x}} = \frac{\tilde{x}}{\|\boldsymbol{w}\|\|\boldsymbol{x}\|}$ a suitable II-score, with the prerequisite that the filter $\boldsymbol{w}$ is a perfect representative for its criteria (*i.e.*, channels). The perfectness of filters in reality, however, cannot be ensured, especially in the early/medium training stages where filters are far from being optimized. This weakens the reliability of the vanilla cosine similarity. To eliminate this critical problem, we propose ***IIEU*** equipped with an adaptive term to enable flexible II-score estimation (Figure 2(d) and 2(e)):

$$\phi\left(\tilde{x}\right) = \varsigma\left(\frac{\tilde{x}}{\|\boldsymbol{x}\|\,\|\boldsymbol{w}\|} + \nu\right)\tilde{x}\,, \tag{1}$$

where we suppose $\|\boldsymbol{x}\|\,\|\boldsymbol{w}\| > 0$; $\nu$ is an updatable bias term with specific constraints (as described in Figure 2(f)) to perform adaptive shifts, *i.e.*, we propose to estimate II-score by $\rho\left(\tilde{x}\right) = \varsigma\left(\frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|} + \nu\right)$, where the approximated similarity $\hat{\varrho}\left(\tilde{x}\right) = \frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|} + \nu$. Then, IIEU realises norm-decoupled feature Act by rectifying (*i.e.*, multiplying) the TI-score $\tilde{x}$ with the II-score adapted to the training conditions. Particularly, we refer to these $\frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|}$ and $\nu$ as the **main similarity term (term-*S*)** and **auxiliary bias term (term-*B*)**, respectively. More generally, **we let $\tilde{x} = \psi\left(\langle \boldsymbol{w}, \boldsymbol{x}\rangle\right)$ if with biases or normalization layers (denoted by $\psi$) applied to the A-I inner product**. This differs term-*S* from the cosine similarity, yet not changes its meaning of importance scoring, as we suppose the decoupling of the norms of features/filters as the essence to eliminate the transitive biases.

Next, we discuss further properties to embody the term-*B* $\nu$ and adjuster $\varsigma$ by specifying the relationships of the ideal similarity $\varrho$ and its adjuster $\varsigma$, with the preceding deductions and new intuitive assumptions, termed as ***C**onstraint on **N**egative **I**nfluence (**CNI**)*, ***P**reservation on **P**ositive **I**nfluence (**PPI**)*, and ***O**riented **D**iscriminativeness (**OD**)*:

*Intuition* 4. $^*$***CNI***: We suppose any non-important candidate have constrained influence.

*Intuition* 5. $^*$***PPI***: We suppose any important candidates $\boldsymbol{x}, \boldsymbol{y}$ with close importance scores $\varrho\left(\tilde{x}\right)$ and $\varrho\left(\tilde{y}\right)$ will have comparable influence, *i.e.*, the influence of the one with lower weight will not be covered by the higher one.

*Intuition* 6. $^*$***OD***: We suppose the core of the Act model, *i.e.*, the reweighting function $\rho$, has a sufficient capability to differentiate between important/non-important candidates.

The intuitions 4, 5, and 6 suggest three dependent constraints on the influence of negative and positive candidates, which we formalize as three *Properties*, *i.e.*, ***(CNI)***, ***(PPI)***, and ***(OD)***, and two corresponding *Propositions* that further specify the Properties for practical IIEUs, with supposing a set of simple constraints: (1) $\phi\left(-\infty\right) = 0$ (*i.e.*, we adopt the boundedness constraint for self-gated Act functions [48] to ensure the stability and convergence of training, with the pre-condition that $\rho\left(\tilde{x}\right)$ is lower-bounded; (2) $\nabla_{\tilde{x}}\varrho\left(\tilde{x}\right)$ is bounded; (3) $\varrho\left(\tilde{x}\right)^{-1}\tilde{x}$ is bounded at $\forall\varrho\left(\tilde{x}\right) \neq 0$ (as detailed in the Appendix$^*$).

Next, we present **IIEU-B** and **IIEU-DC** as two practical IIEU derivatives, built based on the suggested intuitions and properties. In particular, as II-score built upon the proposed approximation to the ideal similarity (*i.e.*, $\hat{\varrho}\left(\cdot\right)$), we suppose a loosened Property 1 to bring additional learning flexibility, *i.e.*, $\varsigma\left(\hat{\varrho}\left(\tilde{x}\right)\right)$ is possible to have small negative values and be non-monotonic about $\left|\hat{\varrho}\left(\tilde{x}\right)\right|$ at $\left|\hat{\varrho}\left(\tilde{x}\right)\right| \leqslant |\eta|$, where $\eta$ denotes a given threshold close to 0.

## 2.3. Practical Method

We present IIEU-B as the initial practical IIEU (Figure 3) and IIEU-**DC** (**D**ynamic **C**oupler) Figure 4 as a tailored enhancement to IIEU-B. In the subsequent, we introduce IIEU-B and IIEU-DC in detail.

**Formulation.** We propose IIEU-B built on the prototype of IIEU (Equation (1)) described in Section 2.2, by embodying the term-*B* $\nu$ and the adjuster $\varsigma$ with the proposed properties. Specifically, for IIEU-B, we let term-*B* be

$$\nu = \delta\left(\mathrm{LN}\left(\mathrm{avgpool}\left(\tilde{\boldsymbol{X}}_c\right)\right)\right), \tag{2}$$

where LN denotes the LayerNorm [3] to perform flexible channel-dependent scaling and shift to channel statistics with negligible cost. $\delta$ is Sigmoid function to cast upper-bounded positive constraint on channel statistics to help meet the supposed properties (Figure 2(f)). Moreover, with the prior that $\hat{\varrho}\left(\tilde{x}\right)$ of IIEU-B is bounded, we propose a suitable conditional adjuster $\varsigma$ to meet the proposed Properties:

$$\varsigma\left(\hat{\varrho}_x\right) = \begin{cases} \hat{\varrho}_x, & \hat{\varrho}_x \geqslant \eta \\ \eta\exp\left(\hat{\varrho}_x - \eta\right), & \hat{\varrho}_x < \eta \end{cases}, \tag{3}$$

where $\hat{\varrho}_x$ denotes $\hat{\varrho}\left(\tilde{x}\right)$ and $\eta$ a learnable threshold shared within each channel, initialized by a small value (0.05 by default). Note that (1) $\varsigma\left(\hat{\varrho}_x\right)$ is continuous about $\hat{\varrho}_x$ if the domain of $\hat{\varrho}_x$ is continuous, as $\lim_{\hat{\varrho}_x-\eta\to 0^-} = \lim_{\hat{\varrho}_x-\eta\to 0^+} = \eta$; (2) we suppose the right-hand derivative

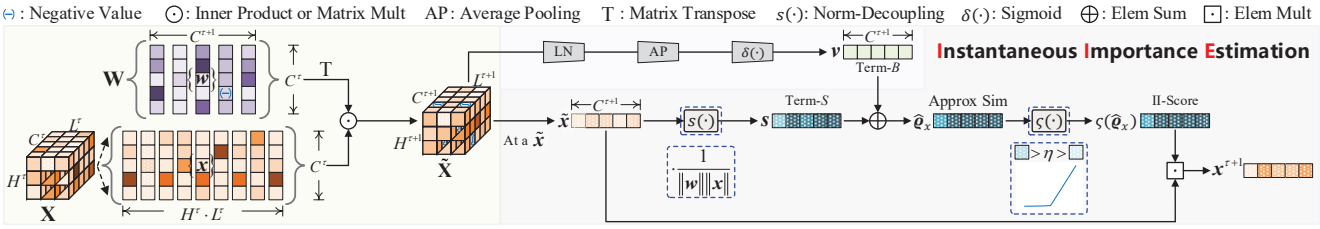



Figure 3. Operational illustration of IIEU-B. "Elem" and "Mult" denote "Element-wise" and "Multiplication," respectively.

as the derivative at $\hat{\varrho}_x = \eta$ (Section 2.1); (3) the influence of any candidate with $\hat{\varrho}_x \leqslant \eta$ will be silenced if $\eta = 0$.

∗**Boundedness of $\rho(\tilde{x})$.** We suppose the boundedness of II-score $\rho(\tilde{x})$ as a pre-condition for Intuitions 4, 5, and 6 to ensure training stability. As for IIEU-B, as $\nu$ is bounded and $\varsigma$ is conditionally linear about $\hat{\varrho}_x$ for $\hat{\varrho}_x > \eta$, the boundedness of $\rho(\tilde{x})$ is solely determined by the term-*S*. For generality, we discuss the common case that BatchNorm [25] is applied, *i.e.*, with the channel scaling and shift factors $\gamma, \beta \in \mathbb{R}$ (extensible to LayerNorm [3]). Let $E = \|\boldsymbol{x}\| \|\boldsymbol{w}\| \neq 0$, the codomain of term-*S* is calculated as:

$$-|r| + \frac{\beta - r\mu}{E} \leqslant \frac{\tilde{x}}{E} \leqslant |r| + \frac{\beta - r\mu}{E}, \tag{4}$$

where $r = \frac{\gamma}{\sigma}$; $\sigma \neq 0$ and $\mu$ denote the standard deviation and mean of $\tilde{x}$ for channel-$c$. That is, we can calculate both the upper- and lower-bound of term-*S* with the factors $\gamma$ and $\beta$ whose values are constrained by the weight-decay (*i.e.*, $\mathcal{L}_2$-regularization) in the training phase. Unlike the cosine similarity with a range $[0, 1]$, the range of term-*S* can be broader. Moreover, as the adjuster $\varsigma$ constrains $\rho(\hat{\varrho}_x) < \eta$ for $\hat{\varrho}_x < \eta$, the II-score can adaptively emphasize/suppress the informative/meaningless candidates.

∗**Analysis for term-*S* and -*B* from filter updating.** We suppose term-*B* to be bounded to prevent it from neutralizing the contribution of the term-*S* (Figure 2(f)). In this paragraph, **we first discuss the case that $\hat{\varrho}_x \geqslant \eta$, *i.e.*,** $\varsigma(\hat{\varrho}_x) = \hat{\varrho}_x$ ***s.t.*** $\phi(\tilde{x}) = \varsigma\left(\frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|} + \nu\right)\tilde{x} = \frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|}\tilde{x} + \nu\tilde{x}$. Further, we simplify the case of comparing term-*B* and -*S* by considering $\tilde{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ without loss of generality, as they share the same BN layers. Note that we discuss the derivatives about $\boldsymbol{w}$ and denote the term-*S* and -*B* by $s(\boldsymbol{w}) = \frac{\tilde{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|}$ and $\nu(\boldsymbol{w}) = \delta(\mathrm{LN}(\bar{\tilde{x}}))$, respectively, where $\bar{\tilde{x}}$ denotes the mean statistic for channel-$c$ and $\delta$ is the Sigmoid function. Moreover, we approximate the operation of LN by $\mathrm{LN}(\bar{\tilde{x}}) = \dot{\gamma}(\bar{\tilde{x}}) + \dot{\beta}$, where $\dot{\gamma}$ and $\dot{\beta}$ are the scaling and shift factors of the LN layer. Then, we can calculate the (partial) derivative about $\boldsymbol{w}$ of $s(\boldsymbol{w})$ as:

$$\nabla_{\boldsymbol{w}} s(\boldsymbol{w}) = \frac{\|\boldsymbol{w}\|^2 \boldsymbol{x} - \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{x}\| \|\boldsymbol{w}\|^3}, \tag{5}$$

where T denotes matrix/vector transpose. Correspondingly, we calculate the derivative about $\boldsymbol{w}$ for term-*B* as:

$$\nabla_{\boldsymbol{w}} \nu(\boldsymbol{w}) = \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{x}} + \dot{\beta}\right)\left(1 - \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{x}} + \dot{\beta}\right)\right)\dot{\gamma}\bar{\boldsymbol{x}}, \tag{6}$$

where $\bar{\boldsymbol{x}} = \mathrm{avgpool}(\mathbf{X}) \in \mathbb{R}^C$ denotes the vectorial channel mean statistics of the feature map $\mathbf{X}$. Particularly, we can expand the top-right term in Equation (5) as:

$$\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = \left(\sum\nolimits_{c=1}^{C} w_c x_c\right)\boldsymbol{w}. \tag{7}$$

That is, we identify term-*S* enabling each neuron to model detailed cross-channel feature-filter interactions at every spatial coordinate and leverage these informative cues to improve the filter updating. In contrast, as a control group, we calculate the derivative about $\boldsymbol{w}$ of ReLU [38] as:

$$\nabla_{\boldsymbol{w}} \mathrm{ReLU}(\tilde{x}) \,|_{\langle \boldsymbol{w}, \boldsymbol{x} \rangle > 0} = \boldsymbol{x}, \tag{8}$$

where ReLU is shown to model channel-independent information only and lacks the capability to improve filter updating with inter-channel relationships.

Next, we discuss the function of term-*B* from filter updating, which de facto realizes aligned adaptive adjustments to the term-*S* with statistical inter-channel information. As the representative of long-range channel cues, term-*B*, however, does not provide the instance details about the features, hence it may dilute the contribution of the term-*S* if has excessive derivative about the filter. We propose to eliminate this problem by casting a positive constraint on the term-*B* with Sigmoid. Specifically, as the terms $\|\boldsymbol{w}\|^2 \boldsymbol{x}$ and $\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ in Equation (5) are composed of the same member vectors (*i.e.*, $\boldsymbol{w}$ and $\boldsymbol{x}$), without loss of generality, we suppose $\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = -\alpha\|\boldsymbol{w}\|^2\boldsymbol{x}, \alpha \in \mathbb{R}$ and we have $\nabla_{\boldsymbol{w}} s(\boldsymbol{w}) = \frac{\|\boldsymbol{w}\|^2\boldsymbol{x} + \alpha\|\boldsymbol{w}\|^2\boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|^3} = \frac{1+\alpha}{\|\boldsymbol{x}\|\|\boldsymbol{w}\|}\boldsymbol{x}$. Then, we can calculate the average contribution of term-*S* to the updating of filter $\boldsymbol{w}$ as: $\left|\nabla_{\boldsymbol{w}} \bar{s}(\boldsymbol{w})\right| = \left|\frac{1+\alpha}{\|\bar{\boldsymbol{x}}\|\|\boldsymbol{w}\|}\right| |\bar{\boldsymbol{x}}|$. With the preceding conditions, we have an conditional corollary: $\left|\frac{1+\alpha}{\|\bar{\boldsymbol{x}}\|\|\boldsymbol{w}\|}\right| \geqslant \frac{1}{4}|\dot{\gamma}| \Longrightarrow \left|\nabla_{\boldsymbol{w}} \bar{s}(\boldsymbol{w})\right| \geqslant |\nabla_{\boldsymbol{w}} \nu(\boldsymbol{w})|$, because $\nabla_{\boldsymbol{w}} |\nu(\boldsymbol{w})| \leqslant \frac{1}{4}|\dot{\gamma}||\bar{\boldsymbol{x}}|$. In particular, with the two critical priors: (1) the range of value of learnable parameters are tightly constrained by the $\mathcal{L}_2$-regularization of a small weight-decay (*e.g.*, $1 \times 10^{-4}$ for ImageNet experiments); (2) $|\dot{\gamma}|$ is usually a small value fallen in $10^{-1}$ level, we suppose $\|\bar{\boldsymbol{x}}\|, \|\boldsymbol{w}\| < 1$ in common, so that $\left|\frac{1+\alpha}{\|\bar{\boldsymbol{x}}\|\|\boldsymbol{w}\|}\right| > |1 + \alpha| \geqslant$

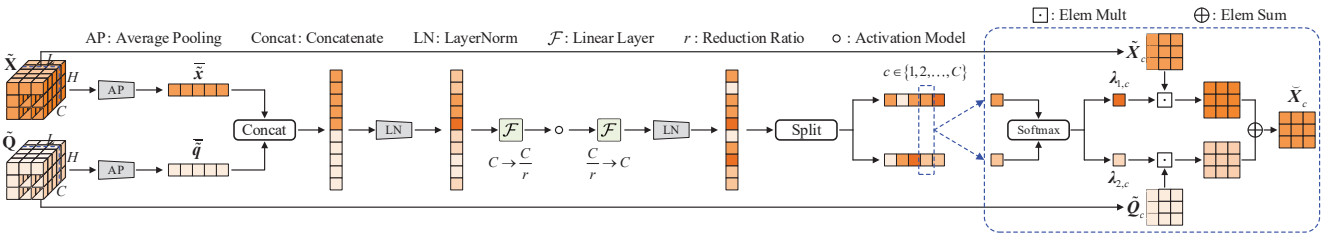



Figure 4. DC operations. $\bar{\tilde{\boldsymbol{x}}}, \bar{\tilde{\boldsymbol{q}}} \in \mathbb{R}$ denote the vectorial channel statistics of the main branch feature map $\tilde{\mathbf{X}}$ and the residual feature map $\tilde{\mathbf{Q}}$.

$\frac{1}{4}|\dot{\gamma}|$ (*i.e.*, $\left|\nabla_{\boldsymbol{w}} \bar{s}(\boldsymbol{w})\right| \geqslant |\nabla_{\boldsymbol{w}} \nu(\boldsymbol{w})|$) can be met easily. This ensures the applicability of the term-*B* to IIEU-B.

Moreover, for $\hat{\varrho}_x < \eta$, the relative relationship of the term-*S* and -*B* about any given $\boldsymbol{w}$ preserves, as both have $\frac{\partial \varsigma}{\partial \hat{\varrho}_x} = \frac{\partial\left(\eta \exp^{\hat{\varrho}_x - \eta}\right)}{\partial(\hat{\varrho}_x - \eta)} \frac{\partial(\hat{\varrho}_x - \eta)}{\partial \hat{\varrho}_x} = \eta \exp^{\hat{\varrho}_x - \eta}$, which still preserves the applicability of the term-*B* to IIEU-B.

**Dynamic Coupler.** Recent neural networks usually leverage the shortcut (*i.e.*, residual) to transmit details of the lower layers to the main branch of the current layer. The estimated II-scores of the features from the main branch and the shortcut, however, are un-calibrated before fusion for their cross-layer relationships such that they possibly meet compromised comparability in terms of importance measure as we propose IIEU mainly to score alternative candidates within the same layer. To address this problem, we propose the **D**ynamic **C**oupler (**DC**) module as a tailor-made enhancement tool for IIEU-B. DC module is a new lightweight joint-feature-gating model that dynamically rectifies features of the main branch and the shortcut with the channel contexts such that the cross-layers features can be adaptively fused with calibrated intensities. In particular, we refer to the enhanced IIEU-B as IIEU-**DC**.

DC module works at a low-cost, which only employs a joint-channel LayerNorm [3] with a small MLP (with a reduction ratio $r$ defaulted by 16) to project the global channel statistics of the input main and shortcut features to the adaptive channel weights. Specifically, DC aims to estimate the channel-wise combination weights dynamically for the effective fusion of the main and shortcut features by extending the channel attention mechanism [24] from

$$\breve{\boldsymbol{X}}_c = \lambda_{1,c} \tilde{\boldsymbol{X}}_c \oplus \tilde{\boldsymbol{Q}}_c \,, \tag{9}$$

*i.e.*, a single-side channel weights estimation without involving the contextual information of residual features, to

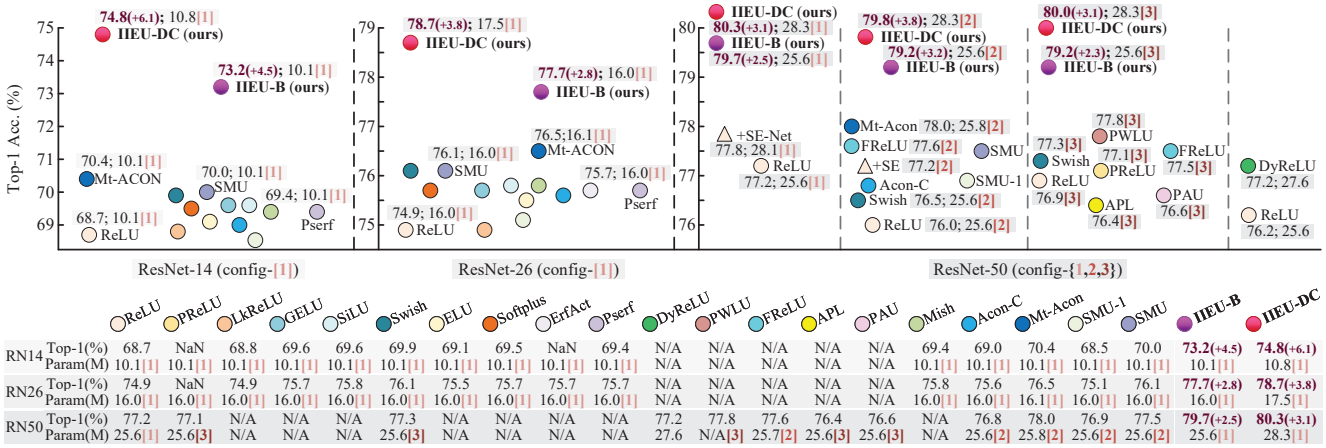
$$\breve{\boldsymbol{X}}_c = \lambda_{1,c} \tilde{\boldsymbol{X}}_c \oplus \lambda_{2,c} \tilde{\boldsymbol{Q}}_c \,, \tag{10}$$

*i.e.*, the double-side channel weights estimation that jointly exploits the dual contextual cues of the main branch and the residual features in an interactive manner, where $\oplus$ denotes the element-wise summation. $\tilde{\boldsymbol{X}}_c, \tilde{\boldsymbol{Q}}_c \in \mathbb{R}^{H \times L}$ denote the main branch and residual feature matrices of the $c$-th channel (*i.e.*, the channel slices of the corresponding feature maps), respectively. $\lambda_{1,c}, \lambda_{2,c} \in \mathbb{R}$ denote the estimated weights for the $c$-th main branch and shortcut feature matrices, respectively. $\breve{\boldsymbol{X}}_c \in \mathbb{R}^{H \times L}$ denotes the fused feature matrix of the $c$-th channel. In particular, we constrain $\lambda_{1,c} + \lambda_{2,c} = 1$ by Softmax function. Note that besides the clear differences in operations, the motivation of our DC, *i.e.*, to realize targeted dynamic weighted mixing of the main branch and shortcut features, is also different from the SK-Net [29] which generalizes SE-Net to merge multi-scale features. The DC operations is illustrated in Figure 4

## 3. Related Work

In Section 2, we explore the possible working mechanism of neural feature Act from MCDM with supposing $\phi(\tilde{x}) = \varsigma(\hat{\varrho}(\tilde{x}))\tilde{x}$. Based on it, we propose to categorize the related methods of Act models by the different adjusters $\varsigma$ or/and approximated ideal similarities $\hat{\varrho}(\tilde{x})$ they introduced. As a prevailing practice, most of the popular methods applied $k\tilde{x}$ as $\hat{\varrho}(\tilde{x})$, where $k \in \mathbb{R}$, and devoted to presenting new variants of $\varsigma$ (*i.e.*, $\phi(\tilde{x}) = \varsigma(k\tilde{x})\tilde{x}$). Inspired by neuronal behaviors, ReLU [38] is a maxout approximation to Softplus [16], whose $\varsigma$ is a binary mask of 0 and 1 for $\tilde{x} \leqslant 0$ and $\tilde{x} > 0$, respectively. LeakyReLU [34] allows slight information leakage from the negative interval to prevent *dead tensors*. PReLU [20] instead learns an adaptive slope for the negative interval. Besides, ELU [13] activates negative $\tilde{x}$ with an exponential function. Goodfellow *et al.* [18] discussed the universal function approximators with piecewise linear components. More recently, PWLU [52] suggested a learnable piecewise linear adjuster. Molina *et al.* [37] presented a versatile approximator for Act functions (*i.e.*, PAU) based on the Padé approximant [6].

ReLU also encouraged recent self-gated Act models. SiLU [17] introduced the first Sigmoid-based $\varsigma$ to enable smooth masking on $\tilde{x}$. Similarly, Swish [40] also considered a Sigmoid $\varsigma$ with an updatable slope $k$ for $\tilde{x}$ to enable flexible fitting. Mish [36] proposed a recent smooth $\varsigma$, *i.e.*, $\tanh(\text{softplus}(\cdot))$. ACON-C [31] extended Swish with the learnable upper/lower bounds for the gradient. GELU [22] introduced the first Gauss-Error-Function-based (ERF) smooth $\varsigma$. GELU also inspired a series of SOTA Act models, *e.g.*, ErfAct/Pserf [4] and Smooth Maximum Units (*i.e.*, SMU-1 and SMU) [5], which are different kinds of smooth variants/approximations to ReLU and GELU with




| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RN14 | Top-1(%) | 68.7 | NaN | 68.8 | 69.6 | 69.6 | 69.9 | 69.1 | 69.5 | NaN | 69.4 | N/A | N/A | N/A | N/A | N/A | 69.4 | 69.0 | 70.4 | 68.5 | 70.0 | **73.2**(+4.5) | **74.8**(+6.1) |
| | Param(M) | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | N/A | N/A | N/A | N/A | N/A | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.1[1] | 10.8[1] |
| RN26 | Top-1(%) | 74.9 | NaN | 74.9 | 75.7 | 75.8 | 76.1 | 75.5 | 75.7 | 75.7 | 75.7 | N/A | N/A | N/A | N/A | N/A | 75.8 | 75.6 | 76.5 | 75.1 | 76.1 | **77.7**(+2.8) | **78.7**(+3.8) |
| | Param(M) | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | 16.0[1] | N/A | N/A | N/A | N/A | N/A | 16.0[1] | 16.0[1] | 16.1[1] | 16.0[1] | 16.0[1] | 16.0[1] | 17.5[1] |
| RN50 | Top-1(%) | 77.2 | 77.1 | N/A | N/A | N/A | 77.3 | N/A | N/A | N/A | N/A | 77.2 | 77.8 | 77.6 | 76.4 | 76.6 | N/A | 76.8 | 78.0 | 76.9 | 77.5 | **79.7**(+2.5) | **80.3**(+3.1) |
| | Param(M) | 25.6[1] | 25.6[3] | N/A | N/A | N/A | 25.6[3] | N/A | N/A | N/A | N/A | 27.6 | N/A[3] | 25.7[2] | 25.6[3] | 25.6[3] | N/A | 25.6[2] | 25.8[2] | 25.6[2] | 25.6[2] | 25.6[1] | 28.3[1] |

Figure 5. Comparison of different Act models with **Res**Net (**RN**) backbones on ImageNet. **IIEU-B** and **-DC** are ours; ErfAct/Pserf (AAAI'22) [4], ACON-C/Mt-ACON (*i.e.*, Meta-ACON, CVPR'21) [31], PWLU (ICCV'21) [52], and SMU-1/SMU (CVPR'22) [5] are SOTAs. We train our and compared Act models which have the public official projects with RN-14 and -26 from scratch using ***cfg-1*** [51] and report the results by "Top-1 Acc.(%); Params.(M)[cfg]", ***where "$(+\cdot)$" show the improvements in Top-1 Acc. of our IIEUs over the ReLU baselines***. For RN-50, we report the official results for all the compared models (including the ReLU baselines w/ or w/o SE-Net [24]) and implemented results for IIEUs with ***cfg-1*** [51], ***-2*** [31], and ***-3*** [52], respectively. "NaN" denotes failed training; "N/A" means non-applicable/unknown.

new ERF-based adjusters. These works achieved clear gains to ReLU networks by introducing flexible smooth adjusters $\varsigma$. However, as discussed with Intuition 3, Act models that apply $k\tilde{x}$ as the approximated similarities $\hat{\varrho}(\tilde{x})$ will encounter the *mismatched feature scoring* problem which puts an obstacle impeding them from further improvements.

Several recent works leveraged attention to activate features, which we treat as presenting a class of approximated similarities $\hat{\varrho}(\tilde{x})$ that tune $\tilde{x}$ with content-based cues. FReLU [32] encoded local spatial cues to rectify $\tilde{x}$ with depth-wise convolutions. Dy-ReLU [10] introduced the SE-Net-based [24] channel attention to improving feature activation. Meta-ACON [31] further extended Swish by generalizing channel attention to learn a dynamic scaling factor for $\tilde{x}$. These works generalized attention to enhance feature activation, provided a promising design space, and realized SOTA gains to ReLU networks. However, as the biasing effects led by the norms occur before the attention, the *mismatched feature scoring* problem remains unsolved. In contrast, IIEU which presents the initial solution to the critical problem achieves the new SOTA improvements with fewer additional parameters.

More related to our work, Wu [48] comprehensively analyzed the convergency, stability, and feasibility of the non-monotonic self-gated Act models at a theoretical level, which laid a solid foundation for our exploration. Wu worked to explain past methods while did not present new Act methods. Concurrently, in a different but related field, Cho *et al.* [12] also found evidence from decision-making to explain how neural networks capture temporal patterns in channels. We agree with their explanations of neural operations and propose to interpret neural feature Act from the new perspective of MCDM, in which we identify the unexcavated yet critical *mismatched feature scoring* problem and present our new Act model, IIEU, as its solution, enjoying remarkable improvements to the SOTAs, based on our new intuitions and the deduced properties of effective feature Act (*i.e.*, selective re-calibration).

## 4. Experiment

We evaluate the effectiveness and versatility of our IIEUs on various vision benchmarks: ImageNet [14] and CIFAR-100 [28] image classification; COCO [30] object detection (in Supp); KITTI-Materials [8] road scene material segmentation (in Supp). We validate our IIEU-B and IIEU-DC through extensive experimental comparisons with the popular and SOTA Act models which include (1) ReLU families/derivatives: [16, 38, 34, 20]; (2) smooth/self-gated models: [13, 22, 17, 40, 36, 5, 4]; (3) attention-based models: [10, 32, 31]; (4) others: [52, 37, 1]. We conduct targeted ablation studies to validate the core components of our IIEU-B, *i.e.* the proposed approximated similarity measure $\hat{\varrho}_x$ and adjuster $\varsigma$ with the MCDM interpretation.

### 4.1. ImageNet Classification

**Implementation details**. We evaluate our IIEUs with three kinds of networks, *i.e.*, the popular ResNet [21] and lightweight MobileNetV2 (MNV2) [42] and ShuffleNetv2 (SNV2) [33] of various sizes, where the baselines use ReLU. To ensure fair comparisons with existing Act models trained with various configures, we adopt three different basic configures applied in [51], [31], and [52] **(denoted by**

Table 1. Comparison of different Act models on ImageNet using lightweight backbones. We train each of the networks with our IIEUs and popular/SOTA act models from scratch using cfg-*c*. For SOTA competitors (Pserf (AAAI'22) [4] and SMU-1/SMU (CVPR'22) [5]), we adopt their official model settings (*i.e.*, the initialization strategies for learnable parameters and values of the hyper-parameters).

| Backbone | Method | GELU[22] | Swish[40] | Mish[36] | Pserf[4] | SMU-1[5] | SMU[5] | ReLU[38] | **IIEU-B** | **IIEU-DC** |
|---|---|---|---|---|---|---|---|---|---|---|
| MobileNetV2 0.17× [42] | Top-1(%)↑ | 52.9 | 53.7 | 53.1 | 52.6 | 51.7 | 54.2 | 49.7 | **58.0**(+8.3) | **58.1**(+8.4) |
| | Params. | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.5M | 1.5M |
| ShuffleNetV2 0.5× [33] | Top-1(%)↑ | 61.5 | 61.8 | 61.5 | 60.8 | 60.2 | 61.8 | 59.9 | **65.8**(+5.9) | **66.8**(+6.9) |
| | Params. | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M | 1.4M |

Table 2. Comparison of Act models with ***cfg-2*** [31]. ***We compare IIEUs with ResNet-26 and -50 backbones*** to the official results of the popular/SOTA Act models with the large ResNet-101.

| Method | Backbone | Params. | FLOPs | Top-1(%)↑ |
|---|---|---|---|---|
| ReLU [38] | ResNet-101 [21] | 44.5M | 7.6G | 77.2 |
| PReLU [20] | | 44.5M | 7.6G | 77.3 |
| Swish [40] | | 44.5M | 7.6G | 77.3 |
| FReLU [32] | | 45.0M | 7.8G | 77.9 |
| ACON-C [31] | | 44.6M | 7.6G | 77.9 |
| Mt-ACON [31] | | 44.9M | 7.6G | 78.9 |
| **IIEU-B (ours)** | **ResNet-50** [21] | **25.6**M | **4.2**G | **79.2** |
| **IIEU-DC (ours)** | | **28.3**M | **4.2**G | **79.8** |
| **IIEU-B (ours)** | **ResNet-26** [21] | **16.0**M | **2.4**G | **77.3** |
| **IIEU-DC (ours)** | | **17.5**M | **2.4**G | **78.3** |

Table 3. Comparing IIEUs with ReLU baseline and SOTA Act models on ShuffleNetV2 (SNV2) with ***cfg-l*** [42].

| | Method | Mish | SMU-1 | SMU | Mt-Acon | ReLU | **IIEU-B** | **IIEU-DC** |
|---|---|---|---|---|---|---|---|---|
| SNV2 1.0× | Top-1(%)↑ | 70.5 | 71.2 | 71.9 | 72.1 | 69.4 | **73.3**(+3.9) | **74.0**(+4.6) |
| | Params. | 2.3M | 2.3M | 2.3M | 2.6M | 2.3M | 2.5M | 2.6M |

***cfg*-1, -2, and -3**, respectively, as detailed in Supp) to train ResNets equipped with IIEU-B and IIEU-DC, respectively. We train MobileNetV2(s) and ShuffleNetv2(s) with two different configures, where the former is a standard configure used in [23, 42, 33, 5, 32, 31] and the later replaces the linear learning rate scheduler in the former with the cosine scheduler (denoted by *cfg-l* and *–c*, respectively). This allows us to investigate the stability of Act models in different training conditions. We follow the common practice to train and test all the implemented networks with an image size of $224 \times 224$ and report our results and the official results for the compared methods by Top-1 Accuracy (Acc.) with one decimal place. Experiments are conducted on a computer with $4\times$ A6000 GPUs.

**Experimental results**. Figure 5 and Tab. 1, 2, 4, 3 report the comparative results of our and the popular/SOTA Act models with various networks on ImageNet, where we have three major observations: (1) IIEUs remarkably improve the popular and SOTA Act models on different networks with negligible/marginal additional cost to the ReLU baselines (*as detailed in Supp, where we show IIEU-B ResNets add only 0.3% parameters and 1.3% FLOPs to the ReLU counterparts*). On ResNet-14, MobileNetV2 $0.17\times$, and ShuffleNetv2 $0.5\times$, IIEU-B and -DC improve ReLU by $\{\mathbf{4.5}\%, \mathbf{8.3}\%, \mathbf{5.9}\%\}$ and $\{\mathbf{6.1}\%, \mathbf{8.4}\%, \mathbf{6.9}\%\}$, respectively. It is worth noting that the improvements of our IIEUs to the SOTAs on some of the networks (*e.g.*, MobileNetV2, ResNet-14, and ResNet-26) are more significant than the SOTAs to the ReLU baselines. (2) With IIEUs, the small ResNet-26s outperform/match the deeper ResNet-50s and -101s with the SOTA Act models and the ResNet-50s enjoy clear improvements over the large ResNet-101s, where IIEU-B and -DC achieve the high Top-1 Acc. of $\{\mathbf{79.7}\%, \mathbf{79.2}\%, \mathbf{79.2}\%\}$ and $\{\mathbf{80.3}\%, \mathbf{79.8}\%, \mathbf{80.0}\%\}$ trained with cfg-1, -2, and -3, respectively. (3) IIEUs are highly stable with different training configures and consistently outperform the SOTA activation models by a clear margin on different networks. We show the accuracy and loss curves in Supp, where IIEUs not only reach the highest Top-1 Acc. but also draw the steepest slopes of optimization. This validates our IIEU for neural feature activation.

Table 4. Comparing IIEUs with ReLU baseline and SOTA Act models on MobileNetV2 (MNV2) with ***cfg-l*** [42].

| Method | MobileNetV2 0.17× [42] | | | MobileNetV2 1.0× [42] | | |
|---|---|---|---|---|---|---|
| | Params. | FLOPs | Top-1(%)↑ | Params. | FLOPs | Top-1(%)↑ |
| PWLU [52] | N/A | N/A | N/A | N/A | N/A | 74.7 |
| ACON-C [31] | 1.5M | 48M | 51.1 | 3.6M | 350M | 73.6 |
| Mt-ACON [31] | 1.9M | 51M | 53.8 | 3.9M | 359M | 75.0 |
| ReLU [38] | 1.4M | 39M | 49.7 | 3.5M | 320M | 72.1 |
| **IIEU-B (ours)** | 1.5M | 46M | **58.1(+8.4)** | 3.6M | 344M | **75.8(+3.7)** |
| **IIEU-DC (ours)** | 1.5M | 47M | **58.7(+9.0)** | 3.6M | 345M | **76.2(+4.1)** |

## 4.2. CIFAR-100 Classification

**Implementation details**. We evaluate different Act models with the public CIFAR versions [47] of ResNets and ShuffleNetV2, which contain fewer parameters than the ImageNet networks. For fair comparisons, we train our and each compared model from scratch with the same standard training configures (as detailed in Supp) with basic data augmentations used in [29].

**Experimental results**. As shown in Table 5, our IIEUs significantly improve all the popular and SOTA Act models

Table 5. Comparison of different Act models on CIFAR-100. We train each model 8 times and report the $\mathrm{mean} \pm \mathrm{std}$ of the Top-1.

| Method | Params. | ELU[13] | PReLU[20] | GELU[22] | SiLU[17] | Swish[40] | Mish[36] | SMU[5] | SMU-1[5] | Pserf[4] | AN-C[31] | Mt-AN[31] | ReLU[38] | **IIEU-B** | **IIEU-DC** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CF-RN-29 | 0.3M | 72.6±0.2 | 70.1±0.5 | 71.4±0.3 | 72.0±0.4 | 71.5±0.3 | 72.1±0.3 | 71.1±0.4 | 70.7±0.3 | 71.6±0.2 | 70.9±0.2 | 72.2±0.3 | 70.5±0.3 | **74.7**±0.3 | **75.8**±0.4 |
| CF-RN-56 | 0.6M | 74.7±0.3 | 73.2±0.4 | 75.3±0.3 | 75.3±0.4 | 74.8±0.2 | 75.2±0.3 | 74.9±0.3 | 74.7±0.2 | 75.3±0.2 | 74.1±0.3 | 75.7±0.2 | 74.4±0.3 | **77.2**±0.3 | **78.1**±0.2 |
| CF-SNV2 | 1.4M | 71.0±0.2 | 72.4±0.3 | 75.2±0.2 | 74.5±0.4 | 74.0±0.2 | 74.8±0.2 | 74.7±0.4 | 74.9±0.3 | 74.8±0.3 | 67.7±0.5 | 71.1±0.4 | 72.9±0.3 | **76.0**±0.3 | **76.8**±0.4 |

Table 6. Ablation study on approximated similar $\hat{\varrho}_x$ and adjuster function $\varsigma$. We report mean Top-1 accuracy for each model.

| $\varsigma(\tilde{x})$ | ReLU | Act-$\varsigma$ | $\varsigma(\hat{\varrho}_x)$ | IIEU-B | (-R) | (-$\delta$) | $\hat{\varrho}_x$ | W/o $\varsigma$ |
|---|---|---|---|---|---|---|---|---|
| | 74.4 | 73.2 | | 77.2 | 77.0 | 74.5 | | 76.6 |

Table 7. Ablation study on the term-*S* and term-*B*, where we report the $\mathrm{mean} \pm \mathrm{std}$ of the Top-1 accuracy for each model.

| IIEU-B | Term-*S* | | Pos-cst on Term-*B* | | |
|---|---|---|---|---|---|
| | W/ (Raw) | ***W/o*** | (a) $\delta$ (Raw) | (b) Softplus | (c) W/o |
| | 77.2±0.3 | ***32.6***±0.4 | 77.2±0.3 | 76.8±0.3 | 75.8±0.2 |

on various networks. These experimental results are highly consistent with the ImageNet evaluations. It is worth noting that IIEUs show superior stability to the compared SOTA self-gated and attention-based Act models, as IIEUs demonstrate higher consistency for the improvements on the corresponding ImageNet and CIFAR networks. This validates the scalability of IIEUs for datasets of different sizes.

### 4.3. Ablation Study

**Approximated similarity** $\hat{\varrho}_x$. $\hat{\varrho}_x$ serves as the core of the II-score estimation for IIEU. Here we discuss $\hat{\varrho}_x$ with three targeted control groups using CIFAR-ResNet-56: (1) $\phi(\tilde{x}) = \varsigma(\tilde{x})$ (denoted by **Act-**$\varsigma$), *i.e.*, let the adjuster $\varsigma$ apply individually without the proposed $\hat{\varrho}_x$ such that IIEU-B degrades to a simpler parametric Act model; (2) replacing $\varsigma$ by ReLU (denoted by "(-R)"); (3) Without $\varsigma$. As shown in Table 6, Act-$\varsigma$ shows a significant drop in accuracy compared to the original IIEU-B. In contrast, model (-R) that preserves the approximated similarity $\hat{\varrho}_x$ shows slight accuracy decreases. Without $\varsigma$, the control group (3) still improves the ReLU baseline by a large margin. The experimental results are consistent with our interpretation, where $\hat{\varrho}_x$ of IIEU is supposed to introduce the main accuracy gains and the $\varsigma$ serves as a helper function to ensure Property 1 which is possibly met conditionally without $\varsigma$.

**Adjuster** $\varsigma$. We further discuss $\varsigma$ by replacing it with the Sigmoid function (denoted by $\delta$, *i.e.* the smooth adjuster $\varsigma$ of SiLU [17] and Swish [40]). Table 6 reports the comparative results of different control groups, where our original $\varsigma$ outperforms Sigmoid function by a large margin. It is worth noting that our $\hat{\varrho}_x$ also achieves competitive Top-1 with ReLU-based $\varsigma$. These results are in line with Intuition 6, as our $\varsigma$ and ReLU function are both conditionally linear about $\hat{\varrho}_x$ for $\hat{\varrho}_x \leqslant 0$, while the slope of Sigmoid gradually declines with the increases of $\hat{\varrho}_x$ if $\hat{\varrho}_x > 0.5$. Moreover, in contrast to ReLU, our original $\varsigma$ introduces further improvements with the adaptive threshold $\eta$.

**W/ or W/o term-*S***. We suppose that the term-*S* (Equation (1)) which serves as the main term of the II-score estimation introduces the main accuracy gains. We validate term-*S* by comparing IIEU-B to the abridged IIEU-B which removes the term-*S*. As shown in Table 7, **removing term-*S* will cause a dramatic drop in accuracy**, which is consistent with our intuition.

**Positive constraint on term-*B***. We suppose a bounded and positive term-*B* is helpful for the II-score estimation and choose Sigmoid function to cast effective positive constraint (pos-cst) on term-*B* (Section 2.3). Herein, we further investigate the selection for positive constraint by replacing Sigmoid (denoted by (a)) with two tailored control groups: (b) Softplus function; (c) identity (*i.e.*, without positive constraint). Experimental results in Table 7 demonstrate that Acc. (a) > Acc. (b) > Acc. (c), which is in line with our intuition, as we suppose the term-*B* needs to be less influential on filter updating than the term-*S* , which contributes to preventing the harmful neutralization effect. That is, (1) as term-*B* with (b) can have significantly higher gradients than (a), we suppose it to show inferior accuracy to (a); (2) we expect (c) to show relatively worst accuracy, as it likely violates the Intuition 6 with outputting negative values.

## 5. Conclusion

We propose to interpret neural feature activation from the new perspective of multi-criteria decision-making, where we identify the critical yet unsettled problem, *i.e. mismatched feature scoring* and present our explainable activation model IIEU to solve it with new intuitions and their corresponding properties for effective activation models. We validate our practical IIEUs and interpretation through comprehensive experimental analysis and extensive comparisons with popular and SOTA activation models on various vision benchmark datasets, where IIEUs achieve the new SOTA improvements to the ReLU baseline and also significantly outperform the current popular and SOTA activation models. As a limitation, IIEU-B brings relatively more throughput decrease than its theoretical additional FLOPs, which we detail in the Appendix.

## References


[1] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[2] Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general relu activations. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Koushik Biswas, Sandeep Kumar, Shilpak Banerjee, and Ashish Kumar Pandey. Erfact and pserf: Non-monotonic smooth trainable activation functions. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[5] Koushik Biswas, Sandeep Kumar, Shilpak Banerjee, and Ashish Kumar Pandey. Smooth maximum unit: Smooth activation function for deep networks using smoothing maximum technique. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Claude Brezinski and Jeannette Van Iseghem. Padé approximations. *Handbook of Numerical Analysis*, 3:47–222, 1994.

[7] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[8] Sudong Cai, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Rgb road scene material segmentation. In *Proc. Asian Conference on Computer Vision (ACCV)*, 2022.

[9] Shyi-Ming Chen, Shou-Hsiung Cheng, and Tzu-Chun Lan. Multicriteria decision making based on the topsis method and similarity measures between intuitionistic fuzzy values. *Information Sciences*, 367:279–295, 2016.

[10] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic relu. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] Sohee Cho, Wonjoon Chang, Ginkyeng Lee, and Jaesik Choi. Interpreting internal activation patterns in deep temporal neural networks by finding prototypes. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

[13] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

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

[15] Yucheng Dong, Yating Liu, Haiming Liang, Francisco Chiclana, and Enrique Herrera-Viedma. Strategic weight manipulation in multiple attribute decision making. *Omega-International Journal of Management Science*, 75:154–164, 2018.

[16] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

[17] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[18] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *Proc. International Conference on Machine Learning (ICML)*, 2013.

[19] Mohit Goyal, Rajan Goyal, and Brejesh Lall. Learning activation functions: A new paradigm for understanding neural networks. *arXiv preprint arXiv:1906.09529*, 2019.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[23] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[24] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(8):2011–2023, 2020.

[25] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proc. International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[26] Deepa Joshi and Sanjay Kumar. Interval-valued intuitionistic hesitant fuzzy choquet integral based topsis method for multi-criteria group decision making. *European Journal of Operational Research*, 248(1):183–191, 2016.

[27] Minjoon Kouh. *Toward a more biologically plausible model of object recognition*. PhD thesis, MIT, 2007.

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[29] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019.

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in

context. In *Proc. European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[31] Ningning Ma, Xiangyu Zhang, Ming Liu, and Jian Sun. Activate or not: Learning customized activation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8032–8042, 2021.

[32] Ningning Ma, Xiangyu Zhang, and Jian Sun. Funnel activation for visual recognition. In *Proc. European Conference on Computer Vision (ECCV)*, pages 351–368. Springer, 2020.

[33] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proc. European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.

[34] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML Workshop*, 2013.

[35] Lassi Meronen, Martin Trapp, and Arno Solin. Periodic Activation Functions Induce Stationarity. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[36] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. In *Proc. British Machine Vision Conference (BMVC)*, 2020.

[37] Alejandro Molina, Patrick Schramowski, and Kristian Kersting. Padé Activation Units: End-to-end Learning of Flexible Activation Functions in Deep Networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[38] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[39] Jindong Qin, Xinwang Liu, and Witold Pedrycz. An extended todim multi-criteria group decision making method for green supplier selection in interval type-2 fuzzy environment. *European Journal of Operational Research*, 258(2):626–638, 2017.

[40] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. In *Proc. Workshop Track of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[41] Jafar Rezaei. Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega-International Journal of Management Science*, 64:126–130, 2016.

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[43] Gabriel Schwartz and Ko Nishino. Recognizing Material Properties from Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(8):1981–1995, 2020.

[44] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *PNAS*, 104(15):6424–6429, 2007.

[45] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[46] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[47] Weiaicunzai. pytorch-cifar100. `https://github.com/weiaicunzai/pytorch-cifar100`.

[48] Lei Wu. Learning a Single Neuron for Non-monotonic Activation Functions. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2022.

[49] Ye Xu, Ye Li, Lijun Zheng, Liang Cui, Sha Li, Wei Li, and Yanpeng Cai. Site selection of wind farms using gis and multi-criteria decision making method in wafangdian, china. *Energy*, 207:118222, 2020.

[50] Hong-Bin Yan, Tieju Ma, and Van-Nam Huynh. On qualitative multi-attribute group decision making and its consensus measure: A probability based perspective. *Omega-International Journal of Management Science*, 70:94–117, 2017.

[51] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled Dynamic Filter Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Yucong Zhou, Zezhou Zhu, and Zhao Zhong. Learning specialized activation functions with the piecewise linear unit. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 12095–12104, 2021.