IIEU: Rethinking Neural Feature Activation from Decision-Making
SUPPLEMENTARY DOCUMENT

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

A. Discussion on The Negative Neutralization Effect

Our Intuition 1 (Section 2.2 of the main paper) aims to bridge the meaning of nonlinear feature activation to selective feature re-calibration. Specifically, we suppose a meaningless feature (as for the concerned filter) is possible to deteriorate the updating of the filter if they have an intense negative inner product. This necessitates a selective feature re-calibration to suppress/emphasize the influence of the meaningless/meaningful features, which clarifies the significance of activation models. In this Appendix, we discuss this problem in detail.

A.1. Preliminaries

Our Intuition 1 qualitatively proposes the possible relationship of "Nonlinearity" and "(the loose) Selectivity" for feature activation based on the influence of a feature on the given filter. Further, to quantitatively discuss our idea, as for the single neuron learning in layer-τ, we use $\boldsymbol{x}$ and $\|\boldsymbol{x}\|$ as the simple measures for the influence and the intensity of the influence of the feature $\boldsymbol{x}$ (*i.e.*, alternative candidate) on updating of the filter $\boldsymbol{w}$ (*i.e.*, ideal candidate), respectively, *when discussing the process* $\tilde{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ *independently w/o the activation function and normalization layers/biases*, as $\nabla_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \boldsymbol{x}$ is a controlling factor to the updating of $\boldsymbol{w}$ with $\boldsymbol{x}$. Note that we omit the layer index τ for simpler notations.

A.2. Discussion

We investigate whether a meaningless feature to a given filter with an intense negative feature-filter inner produce is possible to lead harmful effect to the filter updating by neutralizing/covering the positive effect of a meaningful feature. We formalize this problem as follows with the assumed settings:

- Suppose that for $C \in \mathbb{Z}^+$, $\boldsymbol{w} \in \mathbb{R}^C, \boldsymbol{w} \neq \mathbf{0}$ is a given vector (*i.e.*, the ideal candidate) and $\mathbf{x} = [\mathrm{x}_c], \mathbf{y} = [\mathrm{y}_c] \in \mathbb{R}^C, c \in \{1, \ldots, C\}$ are two vector-valued random variables (*i.e.*, the alternative candidates); $\mathbf{x}, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, |_{\forall c, \Sigma_{c,c} \neq 0}$ denotes a multivariate normal distribution; $\forall \mathbf{x}, \mathbf{y}$, they satisfy the condition $|\langle \boldsymbol{w}, \mathbf{x} \rangle| = \kappa_x \leqslant \kappa_y = |\langle \boldsymbol{w}, \mathbf{y} \rangle|, \langle \boldsymbol{w}, \mathbf{x} \rangle > 0, \langle \boldsymbol{w}, \mathbf{y} \rangle < 0$, where κ_x and κ_y are given (*i.e.*, observed) values. In particular, we use the norm of the expectation $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbb{E}[\mathbf{x}^2]}$ to represent the influence of a random variable candidate $\mathbf{x}$ to the given filter $\boldsymbol{w}$ based on the Appendix A.1.

As such, we show it is possible to have $\|\mathbf{y}\| \geqslant \|\mathbf{x}\|$, *i.e.*, a meaningless feature $\mathbf{y}$ can deteriorate the updating of filter $\boldsymbol{w}$ by neutralizing the positive effect of a meaningful feature $\mathbf{x}$. Note that we change to denote feature vectors by $\mathbf{x}, \mathbf{y}$ in this discussion as now we assume them to be vector-valued random variables.

a. For dimension $C > 1$. As $\boldsymbol{w} \neq \mathbf{0}$, we can find a set of Householder matrices $\{\boldsymbol{H}_c\}$ *s.t.* $\boldsymbol{H}_c \boldsymbol{w} = \lambda \boldsymbol{e}_c$, where $\lambda = |\boldsymbol{w}| \in \mathbb{R}^+$. Specifically, the c-th Householder matrix $\boldsymbol{H}_c$ is computed as:

$$\boldsymbol{H}_c = \boldsymbol{I}_{C \times C} - 2 \boldsymbol{h}_c \boldsymbol{h}_c^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{I}_{C \times C}$ is a C-dimensional identity matrix and $\boldsymbol{h}_c$ is the corresponding normal vector of $\boldsymbol{H}_c$ which can be computed as:

$$\boldsymbol{h}_c = \frac{\boldsymbol{w} - |\boldsymbol{w}|\, \boldsymbol{e}_c}{|\boldsymbol{w} - |\boldsymbol{w}|\, \boldsymbol{e}_c|}. \tag{2}$$

As such, each $\boldsymbol{H}_c$ is an orthogonal matrix that preserves the norm and inner-product of a random vector, *i.e.*, $\forall \mathbf{x}, \|\boldsymbol{H}_c\mathbf{x}\| = \|\mathbf{x}\|$ and $\langle \boldsymbol{H}_c\boldsymbol{w}, \boldsymbol{H}_c\mathbf{x}\rangle = \langle \boldsymbol{w}, \mathbf{x}\rangle$. Then, with the given condition $|\langle \boldsymbol{w}, \mathbf{x}\rangle| = \kappa_x \leqslant \kappa_y = |\langle \boldsymbol{w}, \mathbf{y}\rangle|, \langle \boldsymbol{w}, \mathbf{x}\rangle > 0, \langle \boldsymbol{w}, \mathbf{y}\rangle < 0$, we have:

$$|\langle \boldsymbol{H}_c\boldsymbol{w}, \boldsymbol{H}_c\mathbf{x}\rangle| = |\langle \lambda\boldsymbol{e}_c, \boldsymbol{H}_c\mathbf{x}\rangle| = \lambda \left|(\boldsymbol{H}_c\mathbf{x})_c\right| = \kappa_x \,, \tag{3}$$

$$|\langle \boldsymbol{H}_c\boldsymbol{w}, \boldsymbol{H}_c\mathbf{y}\rangle| = |\langle \lambda\boldsymbol{e}_c, \boldsymbol{H}_c\mathbf{y}\rangle| = \lambda \left|(\boldsymbol{H}_c\mathbf{y})_c\right| = \kappa_y \,, \tag{4}$$

$$\left|(\boldsymbol{H}_c\mathbf{x})_c\right| = (\boldsymbol{H}_c\mathbf{x})_c = \frac{\kappa_x}{\lambda} \leqslant \frac{\kappa_y}{\lambda} = \left|(\boldsymbol{H}_c\mathbf{y})_c\right| = -\left(\boldsymbol{H}_c\mathbf{y}\right)_c \,. \tag{5}$$

That is, we use $\boldsymbol{H}_c$ to rotate the given filter $\boldsymbol{w}$ to the direction of the base vector $\boldsymbol{e}_c$ such that $\forall \mathbf{x}, \mathbf{y}$, $\boldsymbol{H}_c$ preserves the projections of $\mathbf{x}, \mathbf{y}$ on $\boldsymbol{w}$ after the rotations. As such, we can calculate the conditional expectations of the rotated random vectors by $\mathbb{E}\left[\boldsymbol{H}_c\mathbf{y} \mid_{(\boldsymbol{H}_c\mathbf{y})_c = -\frac{\kappa_y}{\lambda}}\right]$ and $\mathbb{E}\left[\boldsymbol{H}_c\mathbf{x} \mid_{(\boldsymbol{H}_c\mathbf{x})_c = \frac{\kappa_x}{\lambda}}\right]$, respectively. Moreover, as $\boldsymbol{H}_c$ preserves the norms, we have the following corollary for the problem we discuss:

Corollary A.1. $\|\mathbf{y}\| \geqslant \|\mathbf{x}\| \Longleftrightarrow \sqrt{\mathbb{E}\left[(\boldsymbol{H}_c\mathbf{y})^2 \mid_{(\boldsymbol{H}_c\mathbf{y})_c = -\frac{\kappa_y}{\lambda}}\right]} \geqslant \sqrt{\mathbb{E}\left[(\boldsymbol{H}_c\mathbf{x})^2 \mid_{(\boldsymbol{H}_c\mathbf{x})_c = \frac{\kappa_x}{\lambda}}\right]}$.

In particular, we first consider $\boldsymbol{H}_c = \boldsymbol{H}_C$ without loss of generality because $\forall i, j$ where $i \neq j$, the swap of the axis-i and -j will not change the norm of a vector. As such, after applying the linear transformations with $\boldsymbol{H}_C$, we have:

$$\boldsymbol{H}_C\mathbf{x}, \boldsymbol{H}_C\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{\Sigma}'\right) \,, \tag{6}$$

where $\boldsymbol{\mu}' = \boldsymbol{H}_C\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}' = \boldsymbol{H}_c\boldsymbol{\Sigma}\boldsymbol{H}_c^{\mathrm{T}}$. For clarity, following we denote $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$ as:

$$\boldsymbol{\mu}' = \begin{bmatrix} \boldsymbol{\mu}'_P \\ \mu'_C \end{bmatrix}, \boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma}'_{P,P} & \boldsymbol{\Sigma}'_{\boldsymbol{P,C}} \\ \boldsymbol{\Sigma}'_{\boldsymbol{C,P}} & \Sigma'_{C,C} \end{bmatrix}, \tag{7}$$

where the index P denotes "from index 1 to $C-1$". Note that $\boldsymbol{\mu}'_P \in \mathbb{R}^{C-1}$ (a column vector), $\mu'_C \in \mathbb{R}$, $\boldsymbol{\Sigma}'_{P,P} \in \mathbb{R}^{C-1\times C-1}$, $\boldsymbol{\Sigma}'_{\boldsymbol{P,C}} \in \mathbb{R}^{C-1}$ (a column vector), $\boldsymbol{\Sigma}'_{\boldsymbol{C,P}} \in \mathbb{R}^{C-1}$ (a row vector), and $\Sigma'_{C,C} \in \mathbb{R}$. Then, with the calculation rules for conditional multivariate norm distribution, for $\boldsymbol{H}_C\mathbf{y}$, we have:

$$\begin{aligned}
\boldsymbol{\mu}_P^y &= \boldsymbol{\mu}'_P + \boldsymbol{\Sigma}'_{\boldsymbol{P,C}}\left(\Sigma'_{C,C}\right)^{-1}\left(-\frac{\kappa_y}{\lambda} - \mu'_C\right) \\
&= \begin{bmatrix} \mu'_1 \\ \mu'_2 \\ \vdots \\ \mu'_{C-1} \end{bmatrix} + \begin{bmatrix} \Sigma'_{1,C} \\ \Sigma'_{2,C} \\ \vdots \\ \Sigma'_{C-1,C} \end{bmatrix}\left(\Sigma'_{C,C}\right)^{-1}\left(-\frac{\kappa_y}{\lambda} - \mu'_C\right) \\
&= \begin{bmatrix} \mu'_1 \\ \mu'_2 \\ \vdots \\ \mu'_{C-1} \end{bmatrix} + \begin{bmatrix} \sigma_1\left(\kappa'_y - \mu'_C\right) \\ \sigma_2\left(\kappa'_y - \mu'_C\right) \\ \vdots \\ \sigma_{C-1}\left(\kappa'_y - \mu'_C\right) \end{bmatrix} \\
&= \left[\mu'_c + \sigma'_c\left(\kappa'_y - \mu'_C\right)\right]^{\mathrm{T}} \,,
\end{aligned} \tag{8}$$

where $\boldsymbol{\mu}_P^y = \boldsymbol{\mu}'_P \mid_{(\boldsymbol{H}_C\mathbf{y})_C = -\frac{\kappa_y}{\lambda}} \in \mathbb{R}^{C-1}$ denotes the conditional mean vector of $\boldsymbol{\mu}'_P$ with the condition $(\boldsymbol{H}_C\mathbf{y})_C = -\frac{\kappa_y}{\lambda}$; for simplicity, we use σ'_c and κ'_y to denote $\Sigma'_{c,C}\left(\Sigma'_{C,C}\right)^{-1}$ and $-\frac{\kappa_y}{\lambda}$, respectively. Similarly, for $\boldsymbol{H}_C\mathbf{x}$, we have:

$$\begin{aligned}
\boldsymbol{\mu}_P^x &= \boldsymbol{\mu}'_P + \boldsymbol{\Sigma}'_{\boldsymbol{P,C}}\left(\Sigma'_{C,C}\right)^{-1}\left(\frac{\kappa_x}{\lambda} - \mu'_C\right) \\
&= \left[\mu'_c + \sigma'_c\left(\kappa'_x - \mu'_C\right)\right]^{\mathrm{T}} \,,
\end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_P^x = \boldsymbol{\mu}'_P \mid_{(\boldsymbol{H}_C\mathbf{x})_C = \frac{\kappa_x}{\lambda}} \in \mathbb{R}^{C-1}$ and κ'_x denotes $\frac{\kappa_x}{\lambda}$.

With the above deductions, we have the following deductions for the observed projections κ'_x, κ'_y and conditional mean vectors $\boldsymbol{\mu}_P^x, \boldsymbol{\mu}_P^y$ of the random vector variables $\mathbf{x}, \mathbf{y}$ to ensure the Corollary A.1:

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{H}_C\mathbf{y})^2 \,|_{(\boldsymbol{H}_C\mathbf{y})_C=\kappa'_y}\right] = |\boldsymbol{\mu}_P^y|^2 + \left(\kappa'_y\right)^2 \geqslant |\boldsymbol{\mu}_P^x|^2 + \left(\kappa'_x\right)^2 = \mathbb{E}\left[(\boldsymbol{H}_C\mathbf{x})^2 \,|_{(\boldsymbol{H}_C\mathbf{x})_C=\kappa'_x}\right] \\
&\Longrightarrow \left(\left(\kappa'_y\right)^2 - \left(\kappa'_x\right)^2\right) + \sum_{c=1}^{C-1} \left(\left(\mu'_c + \sigma'_c\left(\kappa'_y - \mu'_C\right)\right)^2 - \left(\mu'_c + \sigma'_c\left(\kappa'_x - \mu'_C\right)\right)^2\right) \geqslant 0 \\
&\Longrightarrow \left(\left(\kappa'_y\right)^2 - \left(\kappa'_x\right)^2\right) + \sum_{c=1}^{C-1} \sigma'_c\left(\kappa'_y - \kappa'_x\right)\left(2\mu'_c + \sigma'_c\left(\kappa'_y + \kappa'_x - 2\mu'_C\right)\right) \geqslant 0\,.
\end{aligned} \tag{10}$$

As $\forall i, j$ where $i \neq j$, the swap of the axis-i and -j does not change the norm of a vector, we can directly replace the axis-C with an axis-c without changing the conclusion. As such, the above deductions can be extended to the general case of $\forall c$: $c = 1, 2, \ldots, C$. Based on the above deductions, we identify a simple condition to ensure Corollary A.1: $\forall \sigma'_c = 0$, *i.e.*, the transformed covariance matrix $\boldsymbol{\Sigma}'$ is a diagonal matrix such that all of the elements of $\forall \boldsymbol{H}_c\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ are independent. Besides, a particular case is that if the given $\boldsymbol{w}$ and a $\boldsymbol{e}_c$ has the same direction such that it does not require Householder transformations, then, the Corollary A.1 is ensured when $\boldsymbol{\Sigma}$ is a diagonal matrix (*i.e.*, the elements of $\forall \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are independent).

b. For dimension $C = 1$. The condition $C = 1$ ensures $\boldsymbol{w}$ to have the same direction with $\boldsymbol{e}_1$. Then, as $\boldsymbol{\Sigma} \in \mathbb{R}^{1\times 1}$ is a single-value diagonal matrix, the Corollary A.1 is ensured according to the preceding deductions.

Summary. Our discussions of the cases **a** and **b** show that the non-important features with intense negative feature-filter inner produce is possible to neutral/cover the positive contribution of important features if without selective feature re-calibrations to cast oriented suppressions/emphasises on the non-important/meaningful features, respectively. This clarifies the meaning of feature activation models from the perspective of MCDM.

B. Discussions, Deductions, and Proofs for Section 2.2

B.1. Proof of Proposition 1

In the main paper, we introduce Proposition 1 based on Intuition 1 and Definition 1.

Definition 1. For a function $\rho : \mathbb{R} \to \mathbb{R}$, we refer to this ρ 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 $\tilde{x} \neq \tilde{y}$ *s.t.* $\rho(\tilde{x}) \neq \rho(\tilde{y})$.

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

Proposition $\Longrightarrow$. ρ satisfies Definition 1 $\Longrightarrow$ ϕ is nonlinear about $\tilde{x}$.

Proof. $\because \exists \tilde{x}, \tilde{y} \in \mathbb{R}$ where $\tilde{x} \neq \tilde{y}$, $\tilde{x} \neq 0$, and $\tilde{y} \neq 0$ *s.t.* $\rho(\tilde{x}) \neq \rho(\tilde{y})$, then, without loss of generality, let (1) ρ_x, ρ_y denote $\rho(\tilde{x}), \rho(\tilde{y})$, respectively; (2) $\Delta\tilde{x} = \tilde{y} - \tilde{x}$ and $\Delta\tilde{x} = k\tilde{x}, k \in \mathbb{R}$. As such, we have:

$$\phi(\tilde{y}) = \rho_y\tilde{y} = \rho_y(\tilde{x} + \Delta\tilde{x}) = \rho_y(\tilde{x} + k\tilde{x}) = \rho_y(1 + k)\,\tilde{x}\,. \tag{11}$$

As our goal is to prove the nonlinearity of ϕ for $\tilde{x}, \tilde{y} \in \mathbb{R}$, we first assume that ϕ is linear about $\forall \tilde{x}, \tilde{y} \in \mathbb{R}$ (*i.e.*, the contradictory of the conclusion) and derive the paradox with this assumption, such that the nonlinearity of ϕ can be proved.

That is, with assuming that ϕ is linear about $\forall \tilde{x}, \tilde{y} \in \mathbb{R}$, we have:

$$\begin{aligned}
\phi(\tilde{y}) &= \phi(\tilde{x} + \Delta\tilde{x}) = \phi(\tilde{x}) + \phi(\Delta\tilde{x}) \\
&= \phi(\tilde{x}) + \phi(k\tilde{x}) = \phi(\tilde{x}) + k\phi(\tilde{x}) \\
&= \rho_x\tilde{x} + k\rho_x\tilde{x} = \rho_x(1 + k)\,\tilde{x}\,.
\end{aligned} \tag{12}$$

That is,

$$\rho_y(1 + k)\,\tilde{x} = \rho_x(1 + k)\,\tilde{x}\,. \tag{13}$$

Then, as we have the conditions: $\tilde{x} \neq \tilde{y}$, $\tilde{x} \neq 0$, and $k\tilde{x} = \tilde{y} - \tilde{x}$, we have $k \neq 0$. As such, to ensure the assumed conclusion, we have the following possible corollary: $\rho_y = \rho_x \vee \tilde{x} = 0 \vee k = -1$ (note that "$\vee$" denotes logical "or"). Further, as we have the primary prerequisites: $\rho_y \neq \rho_x$, $\tilde{x} \neq 0$, and $\tilde{y} \neq 0$, we find only the conclusion $k = -1$ is possible as for this corollary. However, with $k = -1$, we have $\tilde{y} = \tilde{x} + \Delta\tilde{x} = \tilde{x} + k\tilde{x} = \tilde{x} - \tilde{x} = 0$, which still violates the prerequisite: $\tilde{y} \neq 0$. Therefore, the assumption: ϕ is linear about $\forall \tilde{x}, \tilde{y} \in \mathbb{R}$ leads to a paradox under the given prerequisite condition: ρ satisfies Definition 1.

This completes the proof. ■

Proposition $\Longleftarrow$. ρ satisfies Definition 1 $\Longleftarrow$ ϕ is nonlinear about $\tilde{x}$.

Proof. $\because$ ϕ is nonlinear $\forall \tilde{x} \in \mathbb{R}$, we have: $\exists \tilde{x}, \tilde{y} \in \mathbb{R}, k \in \mathbb{R}$ where $\tilde{x} \neq \tilde{y}$ and $k \neq 0$ *s.t.* $\phi(k\tilde{x} + \tilde{y}) \neq \phi(k\tilde{x}) + \phi(\tilde{y})$ where $k\tilde{x}, k\tilde{x} + \tilde{y} \in \mathbb{R}$. As our goal is to prove that ρ satisfies Definition 1, we first assume that a nonlinear function ρ about $\tilde{x} \in \mathbb{R}$ can violate the conditions of Definition 1 and derive the paradox with this assumption, such that the proposition can be proved.

Without loss of generality, we let $\tilde{z} = k\tilde{x}, \tilde{z} \in \mathbb{R}$. Then, suppose that $\forall \tilde{u}, \tilde{v} \in \mathbb{R}, \rho(\tilde{u}) = \rho(\tilde{v}) = m, m \in \mathbb{R}$, we have:

$$\begin{aligned} \phi(k\tilde{x} + \tilde{y}) &= \phi(\tilde{z} + \tilde{y}) = \rho(\tilde{z} + \tilde{y})(\tilde{z} + \tilde{y}) \\ &= \rho(\tilde{z} + \tilde{y})\tilde{z} + \rho(\tilde{z} + \tilde{y})\tilde{y} = m\tilde{z} + m\tilde{y} \\ &= \rho(\tilde{z})\tilde{z} + \rho(\tilde{y})\tilde{y} = \phi(\tilde{z}) + \phi(\tilde{y}) \ . \end{aligned} \tag{14}$$

That is,

$$\phi(k\tilde{x} + \tilde{y}) = \phi(\tilde{z}) + \phi(\tilde{y}) = \phi(k\tilde{x}) + \phi(\tilde{y}) \ , \tag{15}$$

which violates the prerequisite: $\phi(k\tilde{x} + \tilde{y}) \neq \phi(k\tilde{x}) + \phi(\tilde{y})$. Therefore, the assumption: the reweighting function of ϕ, *i.e.*, ρ is an identity function about any $\tilde{x} \in \mathbb{R}$ leads to a paradox under the given prerequisite condition: ϕ is a nonlinear function on $\mathbb{R}$.

This completes the proof. ■

Summary. We complete the proofs for both the partial propositions (*i.e.*, directions "$\Longrightarrow$" and "$\Longleftarrow$") of Proposition 1, which ensures Proposition 1.

B.2. Proof of Proposition 2

In the main paper, we introduce Proposition 2 based on Intuition 2 and Property 1.

For simpler notations, in the following, we denote $\varrho(\tilde{x})$ as $\varrho_x, \forall \tilde{x} \in \mathbb{R}$ such that $\varsigma(\varrho(\tilde{x}))$ can be denoted as $\varsigma(\varrho_x)$.

Property 1. $\forall \tilde{x}, \tilde{y} \in \mathbb{R}, |\varsigma(\varrho_x)| \geqslant |\varsigma(\varrho_y)|$ if $\varrho_x \geqslant \varrho_y$. Note that $\varrho(\tilde{x})$ is continuous and differentiable at $\tilde{x}, \forall \tilde{x} \in \mathbb{R}$.

Where $\varsigma(\varrho_x)$ is continuous and differentiable about ϱ_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 ς, as the monotonicity of $|\varrho_x|$ about ϱ_x is uncertain. Moreover, Property 1 can be met with the more specific conditions, *i.e.*,

Proposition 2. *Property 1 $\Longleftrightarrow$ (1) $\varsigma(\varrho_x)$ is monotonically increasing (*i.e.*, non-decreasing) about $\varrho_x \wedge \varsigma(\varrho_x) \geqslant 0 \vee$ (2) $\varsigma(\varrho_x)$ is monotonically decreasing (*i.e.*, non-increasing) about $\varrho_x \wedge \varsigma(\varrho_x) \leqslant 0$ ($\wedge$ and $\vee$ denote logical "and" and "or," respectively).*

In particular, as for the cases $\varsigma(\varrho_x) \geqslant 0$ and $\varsigma(\varrho_x) \leqslant 0$ which are symmetrical about $\varsigma(\varrho_x) = 0$ and mutually exclusive with each other excluding $\varsigma(\varrho_x) = 0$, the former can be easily extended to the latter once proven and vice versa.

Proposition $\Longrightarrow$. Property 1 $\Longrightarrow$ (1) $\varsigma(\varrho_x)$ is monotonically increasing (*i.e.*, non-decreasing) about $\varrho_x \wedge \varsigma(\varrho_x) \geqslant 0 \vee$ (2) $\varsigma(\varrho_x)$ is monotonically decreasing (*i.e.*, non-increasing) about $\varrho_x \wedge \varsigma(\varrho_x) \leqslant 0$.

Proof.

First, we assume that we can find a $\varsigma(\varrho_x) > 0$ and $\varsigma(\varrho_y) < 0$, simultaneously. As such, our goal is to find a paradox with this assumption.

With the prerequisite condition: $\varsigma(\varrho_x)$ is continuous about $\varrho_x, \forall \varrho_x$ and the assumed condition: $\exists \varrho_x, \varrho_y$ *s.t.* $\varsigma(\varrho_x) > 0, \varsigma(\varrho_y) > 0$, suppose $(\varrho_z, \varsigma(\varrho_z)) : \varrho_x > \varrho_z > \varrho_y$ is a moving point between $(\varrho_y, \varsigma(\varrho_y))$ and $(\varrho_x, \varsigma(\varrho_x))$, then, $(\varrho_z, \varsigma(\varrho_z))$ traverses through the point $(\varrho_{z_0}), 0$ and we have:

$$\varsigma(\varrho_x) \geqslant |\varsigma(\varrho_{z_0})| = 0 \geqslant |\varsigma(\varrho_y)| = -\varsigma(\varrho_y) \Longrightarrow |\varsigma(\varrho_y)| = 0 \, . \tag{16}$$

But this deduced conclusion leads to a paradox to the assumption: $\varsigma(\varrho_y) < 0$, so we cannot find such a ϱ_y and $\varsigma(\varrho_y)$.

Besides, it can be deduced that both the cases $\exists\varsigma(\varrho_x) > 0, \varsigma(\varrho_y) = 0$ and $\exists\varsigma(\varrho_x) = 0, \varsigma(\varrho_y) < 0$ does not lead to paradoxes. That is, with the above deductions, we have $\forall\varrho_x, \varsigma(\varrho_x) \geqslant 0 \vee \varsigma(\varrho_x) \leqslant 0$.

Next, we first consider the condition: $\varsigma(\varrho_x) \geqslant 0$. Then, Property 1 can be specified to: $\forall\varrho_x, \varrho_y$ in the domain, $|\varsigma(\varrho_x)| = \varsigma(\varrho_x) \geqslant \varsigma(\varrho_y) = |\varsigma(\varrho_y)|$ if $\varrho_x \geqslant \varrho_y$. Therefore, Property 1 is monotonically increasing about ϱ_x, $\forall\varrho_x$.

Similarly, with the condition: $\varsigma(\varrho_x) \leqslant 0$, Property 1 can be specified to: $\forall\varrho_x, \varrho_y$ in the domain, $|\varsigma(\varrho_x)| = -\varsigma(\varrho_x) \geqslant -\varsigma(\varrho_y) = |\varsigma(\varrho_y)|$ if $\varrho_x \geqslant \varrho_y$, *i.e.*, $\varsigma(\varrho_x) \leqslant \varsigma(\varrho_y)$. Therefore, Property 1 is monotonically decreasing about ϱ_x, $\forall\varrho_x$.

This completes the proof. ■

Proposition $\Longleftarrow$. Property 1 $\Longleftarrow$ (1) $\varsigma(\varrho_x)$ is monotonically increasing (*i.e.*, non-decreasing) about $\varrho_x \wedge \varsigma(\varrho_x) \geqslant 0 \vee$ (2) $\varsigma(\varrho_x)$ is monotonically decreasing (*i.e.*, non-increasing) about $\varrho_x \wedge \varsigma(\varrho_x) \leqslant 0$.

Proof. With the condition (1): $\forall\varrho_x, |\varsigma(\varrho_x)| = \varsigma(\varrho_x)$ and $\varsigma(\varrho_x)$ is monotonically increasing about ϱ_x, we have: $\forall\varrho_x, \varrho_y$ in the domain, $|\varsigma(\varrho_x)| = \varsigma(\varrho_x) \geqslant \varsigma(\varrho_y) = |\varsigma(\varrho_y)|$ if $\varrho_x \geqslant \varrho_y$. This ensures the Property 1.

Similarly, with the condition (2), we have: $|\varsigma(\varrho_x)| = -\varsigma(\varrho_x) \geqslant -\varsigma(\varrho_y) = |\varsigma(\varrho_y)|$ if $\varrho_x \geqslant \varrho_y$. This ensures the Property 1.

This completes the proof. ■

Summary. We complete the proofs for both the partial propositions (*i.e.*, directions "$\Longrightarrow$" and "$\Longleftarrow$") of Proposition 2, which ensures Proposition 2.

B.3. Properties 2, 3, and 4

In the main paper, we introduce the Intuition 4 (***C**onstraint on **N**egative **I**nfluence (CNI)*), 5 (***P**reservation on **P**ositive **I**nfluence (PPI)*), and 6 (***O**riented **D**iscriminativeness (OD)*), which inspire the Properties 2 (***(CNI)***), 3 (***(PPI)***), 4 (***(OD)***), respectively. In the following, we introduce Properties 2, 3, and 4 and the Propositions 3, 4 corresponded to Properties 2, 3 respectively.

A retrospect of the Intuitions 4, 5, and 6. In the main paper, we aim to find further properties to embody the term-*B* ν and adjuster ς by specifying the relationships of the ideal similarity ϱ and its adjuster ς, with the preceding deductions and new intuitive assumptions, termed as **CNI**, **PPI**, and **OD**:

- ***CNI***: We suppose any non-important candidate have constrained influence.
- ***PPI***: We suppose any important candidates $\boldsymbol{x}, \boldsymbol{y}$ with close importance scores $\varrho(\tilde{x})$ and $\varrho(\tilde{y})$ will have comparable influence, *i.e.*, the influence of the one with lower weight will not be covered by the higher one.
- ***OD***: We suppose the core of the Act model, *i.e.*, the reweighting function ρ, has a sufficient capability to differentiate between important/non-important candidates.

They 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 a set of simple constraints: (1) $\phi(-\infty) = 0$ (*i.e.*, we adopt the boundedness constraint for self-gated Act functions [22] to ensure the stability and convergence of training, with the pre-condition that $\rho(\tilde{x})$ is lower-bounded); (2) $\nabla_{\tilde{x}}\varrho(\tilde{x})$ is bounded; (3) $\varrho(\tilde{x})^{-1}\tilde{x}$ is bounded at $\forall\varrho(\tilde{x}) \neq 0$. We formalize the corresponding Properties as follows:

Property 2. ***(CNI)*** $\exists\eta \in \mathbb{R}, \mathcal{M}_{x^-} \geqslant 0$ *s.t.* $\forall\varrho(\tilde{x}) < \eta$ we have $|\rho(\tilde{x})\tilde{x}|\;|_{\varrho(\tilde{x})<\eta} \leqslant \mathcal{M}_{x^-}$.

Property 3. ***(PPI)*** $\exists\eta \in \mathbb{R}, \mathcal{M}_{x^+} \geqslant 0$, *s.t.* $\forall\varrho(\tilde{x}) > \eta$ we have $|\nabla_{\tilde{x}}\rho(\tilde{x})\tilde{x}|\;|_{\varrho(\tilde{x})>\eta} \leqslant \mathcal{M}_{x^+}$ at any $\tilde{x}$ where $\phi(\tilde{x})$ is differentiable.

Property 4. ***(OD)*** $\exists\eta \in \mathbb{R}$ and $\exists\epsilon_\rho, \delta_\rho > 0$ *s.t.* if $\varrho(\tilde{x}) > \eta > \varrho(\tilde{y})$, then, $\forall\varrho(\tilde{x}) - \varrho(\tilde{y}) > \epsilon_\rho$ we have $\varsigma(\varrho(\tilde{x})) - \varsigma(\varrho(\tilde{y})) > \delta_\rho$. (Note that δ_ρ is big enough to prevent gradient vanishing)

Then, with all the preceding assumed and deduced conditions, we formalize the corresponding Propositions as follows:

Proposition 3. $\phi(-\infty) = 0 \Longrightarrow$ *Property 2.*

Proposition 4. *(1) $\rho(\tilde{x})$ and $\nabla_{\tilde{x}}\varrho(\tilde{x})$ are bounded $\wedge$ (2) $\varrho(\tilde{x})^{-1}\tilde{x}$ is bounded at $\forall\varrho(\tilde{x}) \neq 0 \Longrightarrow$ Property 3.*

The suggested Intuitions and their inspired Properties lay a basis for us to present IIEU.

B.4. Proof of Proposition 3

Proposition. With the assumed/deduced pre-conditions, we have: $\phi(-\infty) = 0 \Longrightarrow \exists \eta \in \mathbb{R}, \mathcal{M}_{x^-} \geqslant 0$ *s.t.* $\forall \varrho(\tilde{x}) < \eta$ we have $|\rho(\tilde{x})\tilde{x}|\,|_{\varrho(\tilde{x})<\eta} \leqslant \mathcal{M}_{x^-}$. Note that $\phi(\tilde{x}) = \rho(\tilde{x})\tilde{x} = \varsigma(\varrho(\tilde{x}))\tilde{x}$.

Proof.

a. Basic case. First, we consider the basic case where $\varsigma(\varrho(\tilde{x}))$ is fully continuous and differentiable about $\varrho(\tilde{x})$.

Then, as we have the pre-condition: $\varrho(\tilde{x})$ is continuous and differentiable about $\tilde{x}$ on $\mathbb{R}$ (mentioned in Property 1), for $\forall \tilde{x} \in [a, b]$, where $a, b \in \mathbb{R}$ and $[a, b]$ an arbitrary finite interval, then, $\varrho(\tilde{x})$ and $\varsigma(\varrho(\tilde{x}))$ are bounded, simultaneously.

Then, because $\varsigma(\varrho(\tilde{x}))$ and $\tilde{x}$ are both bounded on $\tilde{x} \in [a, b]$, we have $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ bounded. As the upper-bound of $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ exists, then, without loss of generality, let $\mathbb{M}_{x^-}$ denote the set of the upper-bound, such that we have $\mathcal{M}_{x^-} \in \mathbb{M}_{x^-}$. That is, $\exists \mathcal{M}_{x^-}$ *s.t.* $|\phi(\tilde{x})| \leqslant \mathcal{M}_{x^-}$. As such, the conclusion: $|\rho(\tilde{x})\tilde{x}|\,|_{\varrho(\tilde{x})<\eta} \leqslant \mathcal{M}_{x^-}$ holds as long as $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ is upper-bounded when $\varrho(\tilde{x}) < \eta$.

With the above deduction, that $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ is unbounded is only possible when $\varrho(\tilde{x})$ approaches $-\infty$ where $\tilde{x}$ approaches to $-\infty$ or $+\infty$. Note that now the direction is unknown. But, as the given condition $\phi(-\infty) = 0$ constraints that:

$$\lim_{\varrho(\tilde{x}) \to -\infty} |\varsigma(\varrho(\tilde{x}))\tilde{x}| = 0\,, \tag{17}$$

then, no matter $\tilde{x}$ approaches to $-\infty$ or $+\infty$, we have $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ bounded, *i.e.*, $|\rho(\tilde{x})\tilde{x}|\,|_{\varrho(\tilde{x})<\eta} \leqslant \mathcal{M}_{x^-}$.

This completes the proof. ■

b. Extended case. Here, we discuss the extended case: $\varsigma(\varrho(\tilde{x}))$ is fully continuous about $\varrho(\tilde{x})$ while has a finite number of non-differentiable points where the corresponding left-hand and right-hand limits exist but are unequal.

As the left-hand and right-hand limits always exist for any point on $\varsigma(\varrho(\tilde{x}))$, the boundedness of $\varsigma(\varrho(\tilde{x}))$ is ensured at any finite interval. That is, like in the fully continuous case, the conclusion is only possible to be violated when $\varrho(\tilde{x})$ approaches $-\infty$.

But, because the number of the non-differentiable points is finite and the continuity always holds, we can still find such a η which is smaller than all the $\varrho(\tilde{x})$ where $\varsigma(\varrho(\tilde{x}))$ are non-differentiable but both the corresponding left-hand and right-hand limits exist. Therefore, the proof of the case **a** can be directly generalized to the case **b**.

This completes the proof. ■

Further discussion. As a corollary to Proposition 3, we identify a more specific condition to ensure Proposition 3. This specific condition is easier to apply to help the design of activation models, which we suggest as: (1) $\varsigma(-\infty) = 0$; (2) $\exists \eta_\varrho \in \mathbb{R}$ and $\forall k \in \mathbb{R}$, if $|\varsigma(\varrho(\tilde{x}))| \leqslant \left|\frac{k}{\tilde{x}}\right|$ holds for $\forall \varrho(\tilde{x}) < \eta_\varrho$.

That is, in intuition, we suppose as long as the (absolute) reweighting function $|\varsigma(\varrho(\tilde{x}))|$ changes slower than the reference function $\left|\frac{k}{\tilde{x}}\right|$ when the ideal similarity $\varrho(\tilde{x})$ gradually approaches to $-\infty$, the Proposition 3 holds.

Proof. As discussed in the proofs of the cases **a** and **b**, the boundedness of $|\phi(\tilde{x})|$ is only possible to be violated when $\varrho(\tilde{x})$ approaches $-\infty$ where $\tilde{x}$ approaches to $-\infty$ or $+\infty$. Then, as we have the condition: $|\varsigma(\varrho(\tilde{x}))| \leqslant \left|\frac{k}{\tilde{x}}\right|$ for $\forall \varrho(\tilde{x}) < \eta_\varrho$, we have:

$$\begin{aligned} |\varsigma(\varrho(\tilde{x}))\tilde{x}|\,|_{\varrho(\tilde{x})<\eta_\varrho} &= |\varsigma(\varrho(\tilde{x}))|\,|\tilde{x}|\,|_{\varrho(\tilde{x})<\eta_\varrho} \\ &\leqslant \left|\frac{k}{\tilde{x}}\right| |\tilde{x}|\,|_{\varrho(\tilde{x})<\eta_\varrho} = |k|\,|_{\varrho(\tilde{x})<\eta_\varrho}\,, \end{aligned} \tag{18}$$

where $\lim_{\varrho(\tilde{x}) \to -\infty} |k| = |k|$. That is, $|\varsigma(\varrho(\tilde{x}))\tilde{x}|$ is bounded.

Therefore, we complete the proof. ■

Summary. We complete the proofs for the cases **a** and **b** of Proposition 3, which ensures Proposition 3. We further the discussion to a more specific condition that we find easier to apply to help the design of neural feature activation models.

B.5. Proof of Proposition 4

Proposition. (1) $\rho(\tilde{x})$ and $\nabla_{\tilde{x}}\varrho(\tilde{x})$ are bounded $\wedge$ (2) $\varrho(\tilde{x})^{-1}\tilde{x}$ is bounded at $\forall \varrho(\tilde{x}) \neq 0 \Longrightarrow \exists \eta \in \mathbb{R}, \mathcal{M}_{x^+} \geqslant 0$, *s.t.* $\forall \varrho(\tilde{x}) > \eta$ we have $|\nabla_{\tilde{x}}\rho(\tilde{x})\tilde{x}|\,|_{\varrho(\tilde{x})>\eta} \leqslant \mathcal{M}_{x^+}$ at any $\tilde{x}$ where $\phi(\tilde{x})$ is differentiable.

A weaker condition: $\varrho(\tilde{x}) = \tilde{x}$**.** We begin by considering this weaker condition and then extend the corresponding proof to the general case.

Proof.

a. Basic case. First, we discuss the basic case where $\varsigma(\varrho(\tilde{x}))$ is fully continuous and differentiable about $\varrho(\tilde{x})$.

Then, combining the pre-condition: $\varrho(\tilde{x})$ is continuous and differentiable about $\tilde{x}$ on $\mathbb{R}$ (mentioned in Property 1), we have $\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\,\tilde{x}$ bounded at $\tilde{x}$ for all $\tilde{x} \in [a, b]$, where $a, b \in \mathbb{R}$ and $[a, b]$ an arbitrary finite interval, such that $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\,\tilde{x}|$ is also bounded on the finite interval of $\tilde{x}$. That is, $\exists \mathcal{M}_{x^+} \geqslant 0$ *s.t.* $|\nabla_{\tilde{x}}\rho(\tilde{x})\,\tilde{x}|\;|_{\tilde{x}\in[a,b]} \leqslant \mathcal{M}_{x^+}$. Therefore, the only case that is possible to violate the boundedness of $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\,\tilde{x}|$ is when $\varrho(\tilde{x})$ approaching $+\infty$. Note that the relevant condition $\varrho(\tilde{x})$ approaching $-\infty$ is excluded, since we have the condition: $\varrho(\tilde{x}) > \eta, \eta \in \mathbb{R}$.

Then, as in this case we discuss $\varrho(\tilde{x}) = \tilde{x}$, we have:

$$\begin{aligned}
|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\,\tilde{x}| = |\nabla_{\tilde{x}}\varsigma(\tilde{x})\,\tilde{x}| &= \left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x} + \varsigma(\tilde{x})\right| \\
&\leqslant \left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right| + |\varsigma(\tilde{x})|\;,
\end{aligned} \tag{19}$$

where $\rho(\tilde{x}) = \varsigma(\tilde{x})$ is bounded (*i.e.*, the given condition (1)). As such, the upper-boundedness of $|\nabla_{\tilde{x}}\varsigma(\tilde{x})\,\tilde{x}|$ can be ensured by $\left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right|$ if $\left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right|$ is upper-bounded.

In order to deduce the upper-boundedness of $\left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right|$, we introduce a reference function $\ln(\tilde{x})$ which does not have an upper-bound on $\tilde{x} \in \mathbb{R}$. Moreover, as noted in the main paper, we discuss the case where $\varsigma(\varrho(\tilde{x})) \geqslant 0$ without loss of generality. This brings a deduced condition: $\varsigma(\varrho(\tilde{x}))$ is monotonically increasing about $\varrho(\tilde{x})$ (*i.e.*, ϱ_x denoted in Appendix B.2), which we have proved in Appendix B.2. As such, for the assumed case $\varrho(\tilde{x}) = \tilde{x}$, we have $\varsigma(\tilde{x})$ is monotonically increasing about $\tilde{x}$. Then, as $\varsigma(\tilde{x})$ is upper-bounded, we suppose that $\varsigma(\tilde{x}) < \mathcal{M}_\rho$ for all $\tilde{x} \in \mathbb{R}$. Further, as we have the conditions (1) $\ln(\tilde{x})$ and $\varsigma(\tilde{x})$ both are continuous, differentiable, and monotonically increasing about $\tilde{x}$; (2) $\varsigma(\tilde{x}) < \mathcal{M}_\rho$ for all $\tilde{x}$; and (3) $\ln(\tilde{x})$ does not have an upper-bound, we have the conclusion: $\exists \eta \in \mathbb{R}^+$ *s.t.* (1) $\ln(\tilde{x})\;|_{\tilde{x}>\eta} > \mathcal{M}_\rho$ and (2) $\nabla_{\tilde{x}}\ln(\tilde{x}) > \nabla_{\tilde{x}}\varsigma(\tilde{x}) \geqslant 0$ for all $\tilde{x} > \eta$.

Combining the above-given conditions and the deduced conclusions, for any $\tilde{x} > \eta$, we have:

$$\begin{aligned}
\left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right| &= \left|\frac{\partial\varsigma}{\partial\tilde{x}}\right||\tilde{x}| = |\nabla_{\tilde{x}}\varsigma(\tilde{x})|\,|\tilde{x}| \\
&< |\nabla_{\tilde{x}}\ln(\tilde{x})|\,|\tilde{x}| = \left|\frac{1}{\tilde{x}}\right||\tilde{x}| = 1\,.
\end{aligned} \tag{20}$$

That is,

$$\lim_{\tilde{x}\to+\infty}\left|\frac{\partial\varsigma}{\partial\tilde{x}}\tilde{x}\right| < \left|\frac{1}{\tilde{x}}\right||\tilde{x}| = 1\,. \tag{21}$$

So it ensures the conclusion: $\exists \eta \in \mathbb{R}$ *s.t.* $|\nabla_{\tilde{x}}\varsigma(\tilde{x})\,\tilde{x}|\;|_{\tilde{x}>\eta} \leqslant \mathcal{M}_{x^+}$.

Therefore, we complete the proof. ■

b. Extended case. Here, we discuss the extended case: $\varsigma(\varrho(\tilde{x}))$ is fully continuous about $\varrho(\tilde{x})$ while has a finite number of non-differentiable points where the corresponding left-hand and right-hand limits exist but are unequal.

As the left-hand and right-hand limits always exist for any point on $\varsigma(\varrho(\tilde{x}))$, the boundedness of $\varsigma(\varrho(\tilde{x}))$ is ensured at any finite interval. That is, like in the case **a**, the conclusion is only possible to be violated when $\varrho(\tilde{x})$ approaches $+\infty$. But, because the number of the non-differentiable points is finite and the continuity always holds, we can still find such a η which is larger than any $\varrho(\tilde{x})$ where $\varsigma(\varrho(\tilde{x}))$ are non-differentiable but both the corresponding left-hand and right-hand limits exist. Therefore, the proof of the case **a** can be directly generalized to the case **b**.

Therefore, we complete the proof. ■

The general condition. Here, we extend the above proof of the weaker condition to the general condition.

Proof.

a. Basic case. First, we discuss the basic case where $\varsigma(\varrho(\tilde{x}))$ is fully continuous and differentiable about $\varrho(\tilde{x})$.

As we deduced in the weaker condition, the conclusion to be proved can be ensured by: $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\tilde{x}|$ is upper-bounded, which holds on any finite interval and only possible to be violated when $\varrho(\tilde{x})$ is approaching $+\infty$.

Then, since we have:

$$\begin{aligned} |\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\tilde{x}| &= \left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x} + \varsigma(\varrho(\tilde{x}))\right| \\ &\leqslant \left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right| + |\varsigma(\varrho(\tilde{x}))| \,, \end{aligned} \tag{22}$$

where $|\varsigma(\varrho(\tilde{x}))|$ is bounded (*i.e.*, the given condition (1)) such that the upper-boundedness of $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\tilde{x}|$ can be ensured by $\left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right|$ if $\left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right|$ is upper-bounded.

In order to deduce the upper-boundedness of $\left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right|$, we introduce a reference function $\ln(\varrho(\tilde{x}))$ which does not have an upper-bound on $\varrho(\tilde{x}) \in \mathbb{R}$. Moreover, like in the weaker condition, without loss of generality, we discuss the case where $\varsigma(\varrho(\tilde{x})) \geqslant 0$ such that we can adopt the proved conclusion as a new condition: $\varsigma(\varrho(\tilde{x}))$ is monotonically increasing about $\varrho(\tilde{x})$, which we have proved in Appendix B.2. Then, as $\varsigma(\varrho(\tilde{x}))$ is upper-bounded, we suppose that $\varsigma(\varrho(\tilde{x})) < \mathcal{M}_\rho$ for all $\varrho(\tilde{x})$. Further, as we have the conditions: (1) $\ln(\varrho(\tilde{x}))$ and $\varsigma(\varrho(\tilde{x}))$ both are continuous, differentiable, and monotonically increasing about $\varrho(\tilde{x})$; (2) $\varsigma(\varrho(\tilde{x})) < \mathcal{M}_\rho$ for all $\varrho(\tilde{x})$; and (3) $\ln(\varrho(\tilde{x}))$ does not have an upper-bound, we have the conclusion: $\exists\eta \in \mathbb{R}^+$ *s.t.* (1) $\ln(\varrho(\tilde{x}))\,|_{\varrho(\tilde{x})>\eta} > \mathcal{M}_\rho$, and (2) $\nabla_{\varrho(\tilde{x})}\ln(\varrho(\tilde{x})) > \nabla_{\varrho(\tilde{x})}\varsigma(\varrho(\tilde{x})) \geqslant 0$ for all $\varrho(\tilde{x}) > \eta$.

Combining the above-given conditions and the deduced conclusions, for any $\varrho(\tilde{x}) > \eta$, we have:

$$\begin{aligned} \left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right| &= \left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\right||\tilde{x}| = \left|\frac{\partial\varsigma}{\partial\varrho}\right|\left|\frac{\partial\varrho}{\partial\tilde{x}}\right||\tilde{x}| \\ &< \left|\frac{1}{\varrho(\tilde{x})}\right|\left|\frac{\partial\varrho}{\partial\tilde{x}}\right||\tilde{x}| = \left|\varrho(\tilde{x})^{-1}\tilde{x}\right|\left|\frac{\partial\varrho}{\partial\tilde{x}}\right| \,, \end{aligned} \tag{23}$$

where $\left|\varrho(\tilde{x})^{-1}\tilde{x}\right|$ and $\left|\frac{\partial\varrho}{\partial\tilde{x}}\right|$ are bounded (at $\forall\varrho(\tilde{x}) \neq 0$) since both the boundednesses of $\varrho(\tilde{x})^{-1}\tilde{x}$ (at $\forall\varrho(\tilde{x}) \neq 0$) and $\nabla_{\tilde{x}}\varrho(\tilde{x})$ are given conditions. Then, without loss of generality, suppose that $\left|\varrho(\tilde{x})^{-1}\tilde{x}\right| < \mathcal{M}_{u_1}$ for all $\varrho(\tilde{x}) \neq 0$ and $\left|\frac{\partial\varrho}{\partial\tilde{x}}\right| < \mathcal{M}_{u_2}$, we have:

$$\lim_{\varrho(\tilde{x})\to+\infty}\left|\frac{\partial\varsigma}{\partial\varrho}\frac{\partial\varrho}{\partial\tilde{x}}\tilde{x}\right| < \left|\varrho(\tilde{x})^{-1}\tilde{x}\right|\left|\frac{\partial\varrho}{\partial\tilde{x}}\right| < \mathcal{M}_{u_1}\mathcal{M}_{u_2} \,. \tag{24}$$

Note in particular that $\varrho(\tilde{x}) = 0$ does not violate the boundedness of $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\tilde{x}|$ because it can be included in a given finite interval which we proved to preserve the conclusion. That is, let $\mathcal{M}_{x^+} = \mathcal{M}_{u_1}\mathcal{M}_{u_2}$, we have the conclusion: $\exists\eta \in \mathbb{R}$, *s.t.* $|\nabla_{\tilde{x}}\varsigma(\varrho(\tilde{x}))\tilde{x}|\,|_{\varrho(\tilde{x})>\eta} \leqslant \mathcal{M}_{x^+}$.

Therefore, we complete the proof. ■

b. Extended case. Here, we consider the extended case: $\varsigma(\varrho(\tilde{x}))$ is fully continuous about $\varrho(\tilde{x})$ while has a finite number of non-differentiable points where the corresponding left-hand and right-hand limits exist but are unequal.

As the left-hand and right-hand limits always exist for any point on $\varsigma(\varrho(\tilde{x}))$, the boundedness of $\varsigma(\varrho(\tilde{x}))$ is ensured at any finite interval. That is, like in the case **a**, the conclusion is only possible to be violated when $\varrho(\tilde{x})$ is approaching $+\infty$. But, because the number of the non-differentiable points is finite and the continuity always holds, we can still find such a η which is larger than any $\varrho(\tilde{x})$ where $\varsigma(\varrho(\tilde{x}))$ are non-differentiable but both the corresponding left-hand and right-hand limits exist. Therefore, the proof of the case **a** can be directly generalized to the case **b**.

Therefore, we complete the proof. ■

Summary. We complete the proofs for the cases **a** and **b** of Proposition 4, which ensures Proposition 4.

C. Calculations for Section 2.3

C.1. The Range of Term-*S*

In the following, we show the derivations for Equation (4) (*i.e.*, the range of the term-*S* of IIEU-B) of the main paper.

We discuss the common case with BN [11] applied (denoted by ψ), *i.e.*, now we have:

$$\tilde{x} := \psi\left(\langle \boldsymbol{w}, \boldsymbol{x} \rangle\right) = \gamma \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \mu}{\sigma} + \beta\,, \tag{25}$$

where $\gamma, \beta \in \mathbb{R}$ denote the channel scaling and shift factors of BN; $\sigma \in \mathbb{R} \neq 0$ and $\mu \in \mathbb{R}$ denote the standard deviation and mean of $\tilde{x}$ for the channel-c (*i.e.*, the current channel).

Let $E = \|\boldsymbol{x}\|\,\|\boldsymbol{w}\| \neq 0$. As the vanilla cosine similarity $\frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle}{\|\boldsymbol{w}\|\|\boldsymbol{x}\|} \in [0, 1]$, the codomain of term-*S* , *i.e.*, $\frac{\tilde{x}}{\|\boldsymbol{w}\|\|\boldsymbol{x}\|}$ can be calculated as:

$$\begin{cases} -\frac{|\gamma|}{\sigma} - \frac{|\gamma|\mu}{E\sigma} + \frac{\beta}{E} \leqslant \frac{\tilde{x}}{E} \leqslant \frac{|\gamma|}{\sigma} - \frac{|\gamma|\mu}{E\sigma} + \frac{\beta}{E}\,, & \gamma \geqslant 0\,, \\ -\frac{|\gamma|}{\sigma} + \frac{|\gamma|\mu}{E\sigma} + \frac{\beta}{E} \leqslant \frac{\tilde{x}}{E} \leqslant \frac{|\gamma|}{\sigma} + \frac{|\gamma|\mu}{E\sigma} + \frac{\beta}{E}\,. & \gamma < 0\,. \end{cases} \tag{26}$$

Then, let $r = \frac{\gamma}{\sigma}$, we have:

$$-|r| + \frac{\beta - r\mu}{E} \leqslant \frac{\tilde{x}}{E} \leqslant |r| + \frac{\beta - r\mu}{E}\,, \tag{27}$$

i.e., the Equation (4) we present in the main paper.

C.2. The Derivative of Term-S about w

We show the calculation of Equation (5) of the main paper, *i.e.*, the (partial) derivative of the term-*S* $s(\boldsymbol{w})$ about $\boldsymbol{w}$ ($\nabla_{\boldsymbol{w}} s(\boldsymbol{w})$) as follows:

$$\begin{aligned} \nabla_{\boldsymbol{w}} s(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle}{\|\boldsymbol{w}\|\,\|\boldsymbol{x}\|} &= \|\boldsymbol{x}\|^{-1} \left(\frac{\partial \|\boldsymbol{w}\|^{-1}}{\partial \boldsymbol{w}} \cdot \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{x} \cdot \|\boldsymbol{w}\|^{-1} \right) \\ &= \|\boldsymbol{x}\|^{-1} \left(-\|\boldsymbol{w}\|^{-2} \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + \frac{\boldsymbol{x}}{\|\boldsymbol{w}\|} \right) \\ &= \|\boldsymbol{x}\|^{-1} \left(\frac{\|\boldsymbol{w}\|^2 \boldsymbol{x} - \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}}{\|\boldsymbol{w}\|^3} \right) \\ &= \frac{\|\boldsymbol{w}\|^2 \boldsymbol{x} - \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}}{\|\boldsymbol{x}\|\,\|\boldsymbol{w}\|^3}\,. \end{aligned} \tag{28}$$

C.3. The Derivative of Term-B about w

We show the calculation of Equation (6) of the main paper, *i.e.*, the (partial) derivative of the term-*B* $\nu(\boldsymbol{w})$ about $\boldsymbol{w}$ ($\nabla_{\boldsymbol{w}} \nu(\boldsymbol{w})$) as follows:

$$\begin{aligned} \nabla_{\boldsymbol{w}} \nu(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \delta\left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right) &= \frac{\partial \delta\left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right)}{\partial \left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right)} \cdot \frac{\partial \left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right)}{\partial \boldsymbol{w}} \\ &= \delta\left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right) \left(1 - \delta\left(\dot{\gamma}\overline{\langle \boldsymbol{w}, \boldsymbol{x} \rangle} + \dot{\beta}\right)\right) \cdot \dot{\gamma} \cdot \frac{1}{N} \cdot \sum\nolimits_{n=1}^{N} \boldsymbol{x}(n) \\ &= \delta\left(\frac{\dot{\gamma}}{N} \boldsymbol{w}^{\mathrm{T}} \sum\nolimits_{n=1}^{N} \boldsymbol{x}(n)\right) \left(1 - \delta\left(\frac{\dot{\gamma}}{N} \boldsymbol{w}^{\mathrm{T}} \sum\nolimits_{n=1}^{N} \boldsymbol{x}(n)\right)\right) \cdot \frac{\dot{\gamma}}{N} \sum\nolimits_{n=1}^{N} \boldsymbol{x}(n) \\ &= \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}} \overline{\boldsymbol{x}} + \dot{\beta}\right) \left(1 - \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}} \overline{\boldsymbol{x}} + \dot{\beta}\right)\right) \dot{\gamma} \overline{\boldsymbol{x}}\,, \end{aligned} \tag{29}$$

where $N = H \times L$ denotes the number of feature vectors in the current feature map (a tensor) of the layer-τ (*i.e.*, $\mathbf{X}$ with a spatial resolution of $H \times L$, as assumed in Section 2.1 of the main paper). Note that δ denotes the Sigmoid function and we adopt the known derivation rule of the Sigmoid function, *i.e.*, $\forall x \in \mathbb{R}$, $\delta(x) = \delta(x)\left(1 - \delta(x)\right)$. This derivation rule can be directly generalized to the case of vector-valued inputs.

C.4. Calculation of Equation (7)

From Equation (7) of the main paper, 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 the following, we show the calculation of Equation (7):

$$\begin{aligned}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} &= \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_C \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_C \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_C \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{w}_1\boldsymbol{w}_1 & \boldsymbol{w}_1\boldsymbol{w}_2 & \dots & \boldsymbol{w}_1\boldsymbol{w}_C \\ \boldsymbol{w}_2\boldsymbol{w}_1 & \boldsymbol{w}_2\boldsymbol{w}_2 & \dots & \boldsymbol{w}_2\boldsymbol{w}_C \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_C\boldsymbol{w}_1 & \boldsymbol{w}_C\boldsymbol{w}_2 & \dots & \boldsymbol{w}_C\boldsymbol{w}_C \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_C \end{bmatrix} \\ &= \boldsymbol{w}\left(\sum\nolimits_{c=1}^{C} \boldsymbol{w}_c\boldsymbol{x}_c\right) = \left(\sum\nolimits_{c=1}^{C} \boldsymbol{w}_c\boldsymbol{x}_c\right)\boldsymbol{w}\,. \end{aligned} \tag{30}$$

C.5. Proof of The Inequality: $|\nabla_{\boldsymbol{w}}\nu(\boldsymbol{w})| \leqslant \frac{1}{4}|\dot{\gamma}|\,|\overline{\boldsymbol{x}}|$

First, we adopt the conclusion for $\nabla_{\boldsymbol{w}}\nu(\boldsymbol{w})$ in Appendix C.3: $\nabla_{\boldsymbol{w}}\nu(\boldsymbol{w}) = \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\overline{\boldsymbol{x}} + \dot{\beta}\right)\left(1 - \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\overline{\boldsymbol{x}} + \dot{\beta}\right)\right)\dot{\gamma}\overline{\boldsymbol{x}}$. As $\dot{\gamma}, \dot{\beta} \in \mathbb{R}$, $\boldsymbol{w}, \overline{\boldsymbol{x}} \in \mathbb{R}^C$, and $\boldsymbol{w}^{\mathrm{T}}\overline{\boldsymbol{x}} \in \mathbb{R}$, let $z = \dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\overline{\boldsymbol{x}} + \dot{\beta} \in \mathbb{R}$ without loss of generality. Then, we have:

$$\begin{aligned}|\nabla_{\boldsymbol{w}}\nu(\boldsymbol{w})| &= \left|\delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\overline{x} + \dot{\beta}\right)\left(1 - \delta\left(\dot{\gamma}\boldsymbol{w}^{\mathrm{T}}\overline{\boldsymbol{x}} + \dot{\beta}\right)\right)\dot{\gamma}\overline{\boldsymbol{x}}\right| = \left|\delta(z)\left(1 - \delta(z)\right)\dot{\gamma}\overline{\boldsymbol{x}}\right| \\ &\leqslant \sup\left(\left|\delta(z)\left(1 - \delta(z)\right)\right|\right)\cdot|\dot{\gamma}|\cdot|\overline{\boldsymbol{x}}| = \frac{1}{2}\left(1 - \frac{1}{2}\right)|\dot{\gamma}|\,|\overline{\boldsymbol{x}}| = \frac{1}{4}|\dot{\gamma}|\,|\overline{\boldsymbol{x}}|\,. \end{aligned} \tag{31}$$

That is: $|\nabla_{\boldsymbol{w}}\nu(\boldsymbol{w})| \leqslant \frac{1}{4}|\dot{\gamma}|\,|\overline{\boldsymbol{x}}|$. Therefore, we complete the proof. ■

D. Training Configures

In the following, we detail the training configures we adopt in the experiments on ImageNet [6] Classification.

D.1. ImageNet Classification

Training configures for ResNet To make fair comparisons with existing activation models trained with various configures, we adopt the three different basic configures applied in [25], [15], and [26] to train ResNets equipped with our IIEU-B/-DC and compare with the baseline and popular/SoTA activation functions using the corresponding configures, respectively, in Section 4 of the main paper (*i.e.*, Experiment), where we denote these three configures by **cfg**-1, -2, and -3, respectively. This allows us to investigate the stability of activation models with different training conditions. We detail the **cfg**-1, -2, and -3 as follows:

1. **cfg-1**. This training configure applies 120 epochs using the basic SGD optimizer with the weight decay of 1^{-4} and momentum of 0.9, where the first 5 epochs are the linear warm-up epochs. The learning rate starts from 0.1 with a batch size of 256 by default and decays to 1^{-5} following the cosine schedule. After the main training schedule, it applies an extra 10 cool-down epochs with the minimum learning rate 1^{-5} to stabilize the model weights. It follows the common practice to first randomly resize the input images and then crop the input images to the size of 224×224. In the test phase, each input images are center cropped to 224×224. It adopts the standard data augmentation strategy used in [13, 25, 15, 10].

2. **cfg-2**. **cfg-2** has two differences compared to **cfg-1**: (1) it applies the linear learning rate schedule which starts from 0.1 and decays to 1^{-5} (*i.e.*, the minimum learning rate); (2) it removes the extra 10 cool-down epochs.

3. **cfg-3**. **cfg-3** has one difference compared to **cfg-1**: (1) it applies a cosine learning rate with only 100 epochs.

Training configures for MobileNetV2 and ShuffleNetV2 We train MobileNetV2(s) and ShuffleNetV2(s) with two different configures, where the former is a standard configure used in [9, 20, 17, 3, 16, 15] and the later replaces the linear learning rate scheduler in the former with the cosine learning rate scheduler (denoted by **cfg**-*l* and –*c*, respectively). We detail the **cfg**-*l* and **cfg**-*c* as follows:

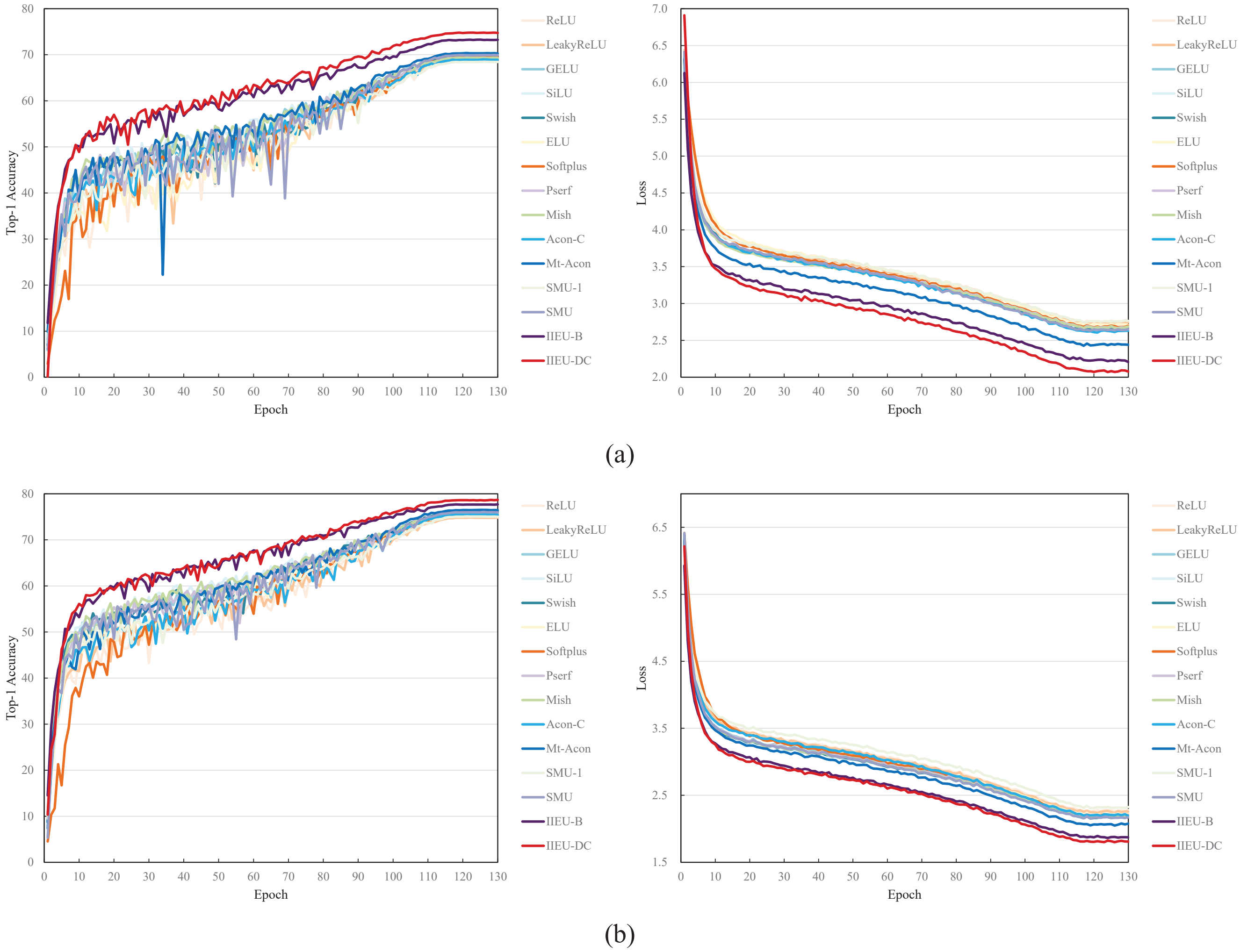
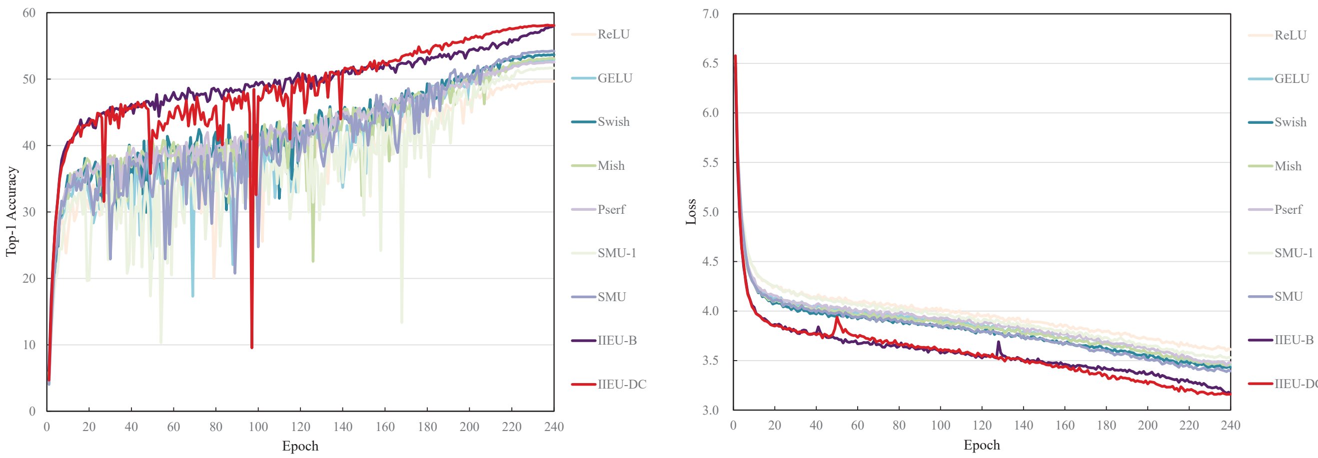


Figure 1. The accuracy and loss curves of the (a) ResNet-14 and (b) ResNet-26 with different activation models.

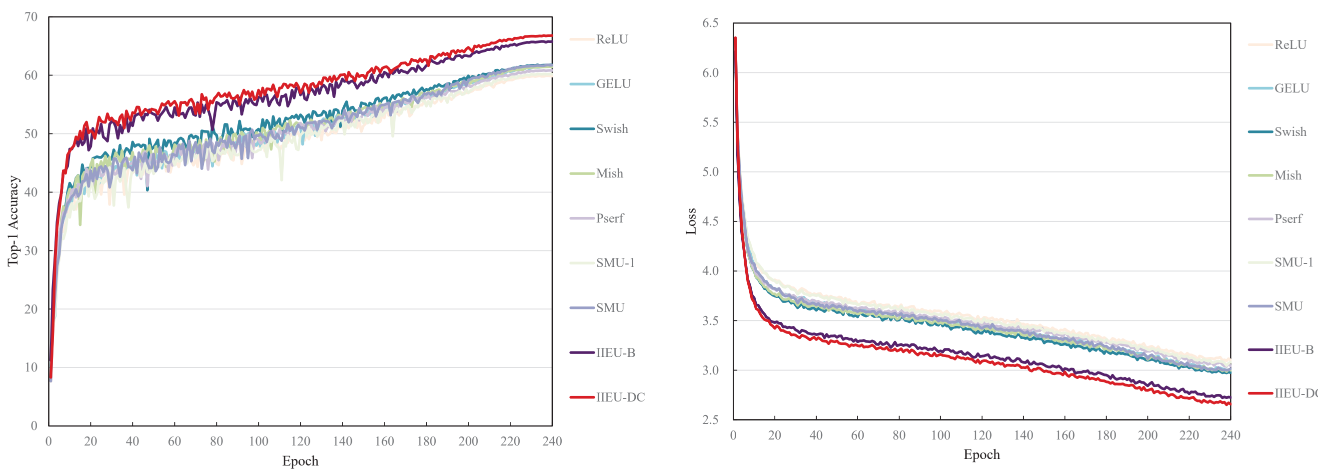
1. **cfg**-*l*. This training configure applies the basic SGD optimizer with the weight decay of 4×10^{-5} and momentum of 0.9. Each network is trained with a batch size of 1024 for 300k iterations (*i.e.*, 240 epochs as for the number of images in the training set of ImageNet). The learning rate starts from 0.5 and decreases to 1^{-5} (*i.e.*, the minimum learning rate) following the linear schedule. It follows the common practice to first randomly resize the input images and then crop each input images to the size of 224×224. In the test phase, each input images are center cropped to 224×224. It adopts the standard data augmentation strategy used in [13, 25, 15, 10].
2. **cfg**-*c*. **cfg**-*c* has one difference compared to **cfg**-*l*: it

D.2. CIFAR-100 Classification

In the experiment on CIFAR-100, we apply the same training and evaluation configure for the CIFAR-ResNet, CIFAR-MobileNetV2, and CIFAR-ShuffleNetV2. For fair comparisons, we adopt the standard data augmentations used in [13] to train all the networks with our and compared activation models by a basic SGD optimizer with the weight decay of 5×10^{-4} and momentum of 0.9. Each model is trained for 350 epochs with a batch size of 256. The learning rate starts from 0.1 and decreases to 1^{-6} following the cosine schedule. All the input images are fixed to the size of 32×32.



(c)



(d)

Figure 2. The accuracy and loss curves of the (c) MobileNetV2 $0.17\times$ and (d) ShuffleNetV2 $0.5\times$ with different activation models.

E. Supplementary Results on ImageNet Classification

E.1. Convergence Analysis

We show the convergence curves of ResNet-14/-26 [8], MobileNetV2 $0.17\times$ [20], and ShuffleNetV2 $0.5\times$ [17] with our IIEU-B/-DC and the compared baseline/popular/SoTA activation models. Each of the models is trained by the **cfg-1** [25] from scratch to convergence, respectively.

Figures 1 and 2 depict the convergence trends in Top-1 accuracy (the higher the better) and training loss (the lower the better) of the ResNet-14, ResNet-26, MobileNetV2 $0.17\times$, and ShuffleNetV2 $0.5\times$ equipped with our IIEUs and the compared activation models, respectively. ReLU networks are the baselines and Pserf (AAAI'22) [2], ACON-C/Mt-ACON (*i.e.*, Meta-ACON, CVPR'21) [15], and SMU-1/SMU (CVPR'22) [3] are current SoTAs. It is worth noting that our IIEU-B and IIEU-DC consistently achieve the relatively highest Top-1 accuracies and lowest loss values on different networks over the varying of epochs.

Tables 1 and 2 reports the number of training epochs to convergence for the networks (*i.e.*, ResNet-14 and ResNet-26, respectively) of different activation models, where we select the epoch that *each corresponding network reaches its lowest training loss value* as the criterion of *convergence*. Moreover, for detailed comparisons of convergence speed, we also show the specific epochs that the loss of each network first drops below the specific values (*i.e.*, $\mathrm{epoch}_{\mathcal{L}<3.0}$, $\mathrm{epoch}_{\mathcal{L}<2.5}$, and $\mathrm{epoch}_{\mathcal{L}<2.0}$ are selected, where $\mathcal{L}$ denotes "loss value"). Our two major observations are: (1) IIEU-B and IIEU-DC demonstrate improved convergence properties. That is, IIEU-B and IIEU-DC reach each of the corresponding loss thresholds

Table 1. Convergence analysis with *ResNet-14* backbone. We show the results for different activation models with three valid digits. $\mathcal{L}$ denotes "loss value." Note that each model is trained for 130 epochs using **cfg-1** [25] (*i.e.*, 120 main epochs with 10 cool-down epochs).

Metric	ReLU	LkReLU	GELU	SiLU	Swish	ELU	Softplus	Pserf	Mish	ACON-C	Mt-ACON	SMU-1	SMU	IIEU-B	IIEU-DC
$\mathcal{L}_{min} \downarrow$	2.74	2.72	2.66	2.66	2.65	2.71	2.67	2.66	2.66	2.61	2.43	2.75	2.63	**2.21**	**2.07**
$\text{epoch}_{\mathcal{L}_{\min}}$	119	122	122	120	117	119	122	117	122	121	117	117	119	130	124
$\text{epoch}_{\mathcal{L}_{<3.0}}$	98	98	93	93	91	98	94	93	93	91	79	99	91	**57**	**44**
$\text{epoch}_{\mathcal{L}_{<2.5}}$	–	–	–	–	–	–	–	–	–	–	112	–	–	**97**	**90**
$\text{epoch}_{\mathcal{L}_{<2.0}}$	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Top-1(%)$\uparrow$	68.7	68.8	69.6	69.6	69.9	69.1	69.5	69.4	69.4	69.0	70.4	68.5	70.0	**73.2**	**74.8**

Table 2. Convergence analysis with *ResNet-26* backbone. We show the results for different activation models with three valid digits. $\mathcal{L}$ denotes "loss value." Note that each model is trained for 130 epochs using **cfg-1** [25] (*i.e.*, 120 main epochs with 10 cool-down epochs).

Metric	ReLU	LkReLU	GELU	SiLU	Swish	ELU	Softplus	Pserf	Mish	ACON-C	Mt-ACON	SMU-1	SMU	IIEU-B	IIEU-DC
$\mathcal{L}_{min} \downarrow$	2.26	2.24	2.18	2.16	2.15	2.20	2.19	2.19	2.17	2.19	2.05	2.31	2.15	**1.86**	**1.80**
$\text{epoch}_{\mathcal{L}_{\min}}$	119	121	119	119	119	119	117	119	120	122	126	119	119	124	126
$\text{epoch}_{\mathcal{L}_{<3.0}}$	68	65	57	53	53	60	60	57	56	62	46	75	53	**23**	**20**
$\text{epoch}_{\mathcal{L}_{<2.5}}$	102	100	96	96	96	99	99	98	96	99	90	106	95	**74**	**71**
$\text{epoch}_{\mathcal{L}_{<2.0}}$	–	–	–	–	–	–	–	–	–	–	–	–	–	**107**	**104**
Top-1(%)$\uparrow$	74.9	74.9	75.7	75.8	76.1	75.5	75.7	75.7	75.8	75.6	76.5	75.1	76.1	**77.7**	**78.7**

Table 3. Comparisons of FLOPs and parameters of IIEUs with the ReLU baselines on ResNet backbones. We show the official Top-1 of the ReLU ResNet-50 adopted from [25]. All the models are trained by the **cfg-1** [25] (including the ReLU ResNet-50).

Method	Metric	ResNet-14 [8]	ResNet-26 [8]	ResNet-50 [8]
ReLU [18]	Params.	10.1M	16.0M	25.6M
IIEU-B (ours)		10.1M	16.0M	25.6M
IIEU-DC (ours)		10.8M	17.5M	28.3M
ReLU [18]	FLOPs	1.5G	2.4G	4.1G
IIEU-B (ours)		1.5G	2.4G	4.2G
IIEU-DC (ours)		1.5G	2.4G	4.2G
ReLU [18]	Top-1(%)$\uparrow$	68.7	74.9	77.2
IIEU-B (ours)		**73.2**	**77.7**	**79.7**
IIEU-DC (ours)		**74.8**	**78.7**	**80.3**

with relatively fewer training epochs. (2) IIEU-B and IIEU-DC reach clearly lower minimum training loss values (*i.e.*, $\mathcal{L}_{min}$) than other compared SoTA/popular/baseline activation models. This validates the convergence property of IIEU.

E.2. FLOPs & Parameters Added to The ReLU Baselines

Table 3 shows the additional FLOPs and parameters of our IIEU-B and IIEU-DC to the ReLU baselines on ResNet-14, ResNet-26, and ResNet-50 [8], respectively. Our IIEU-B adds approximately 0.3% parameters and 1.3% FLOPs to the ReLU counterparts. IIEU-DC shows closed FLOPs to IIEU-B with a relatively slight increase in parameters. Both IIEU-B and IIEU-DC introduce significant gains in accuracy with relatively low computational overhead.

F. Ablation Study on Normalization Operations of Term-*B*

We consider Layer Normalization (LayerNorm) [1] as an effective operation for the term-*B* (*i.e.*, ν) of IIEU-B to perform flexible channel-dependent scaling and shift to channel statistics with negligible cost (introduced in Formulation, Section 2.3). Herein, we further investigate the effectiveness (*i.e.*, suitability) of LayerNorm for the learning of term-*B* by comparing it to different relevant parametric normalization operations that are commonly applied in neural networks. Specifically, a targeted ablation study of applying alternative parametric normalization layers in term-*B* of IIEU-B is conducted on CIFAR-100 [12] dataset with CIFAR-ResNet-56 backbone [8, 21], where five **c**ontrol **g**roups (**cg**) are set up: **(1) LayerNorm [1] (*i.e.*, the original setting);** (2) Group Normalization (GroupNorm) [23] with groups (denoted by G) 2, 4, and C; (3) Batch Normalization (BatchNorm) [11]; (4) the blank group which applies updatable element-wise affine but removing the normalization operation (*i.e.*, Z-Scoring); (5) the ReLU [18] baseline.

We report $\mathbf{mean} \pm \mathbf{std}$ of the Top-1 accuracy in Table 4, where our five major observations are: **(1) the LayerNorm group (cg-1) achieves the highest Top-1 accuracy of all the compared groups;** (2) the GroupNorm group (**cg-2**) demonstrates inferior Top-1 accuracy with $G = C$ while yields close accuracies with $G = 2$ and $G = 4$; (3) the BatchNorm group (**cg-3**) shows relatively low Top-1 accuracy; (4) the blank group (**cg-4**) improves the ReLU baseline (**cg-5**) by a large margin and also clearly outperforms the BatchNorm group; (5) **cg-1** to **cg-4** all enjoy clear accuracy improvements to the ReLU baseline. Note that for single vector input (*i.e.*, the case of the term-*B* in IIEU-B), (1) "GroupNorm of $G = 1$" equals to "LayerNorm;" (2) "GroupNorm of $G = C$" equals to "using biases only;" (3) Instance Normalization (InstNorm) is non-applicable. This validates LayerNorm for the learning of the adaptive shift (*i.e.* term-*B*) in IIEU-B.

Table 4. Ablation study on normalization operations of the term-*B* in IIEU-B.

(1) LayerNorm	(2) GroupNorm			(3) BatchNorm	(4) Blank Group	(5) ReLU Baseline
	$G = C$	$G = 4$	$G = 2$			
$\mathbf{77.2 \pm 0.3}$	76.4 ± 0.2	76.9 ± 0.2	77.0 ± 0.3	*75.4 ± 0.3*	76.6 ± 0.3	74.4 ± 0.3

G. Limitation

Despite the marginal additional parameters and theoretical computational overhead, we find that IIEU-B introduces relatively more throughput decrease than the FLOPs it adds to ReLU [18] baseline. To investigate this phenomenon, we conduct a comparative evaluation of FLOPs and throughput by comparing IIEU-B to ReLU [18] baseline and popular/SoTA activation models, including SiLU [7], Meta-ACON [15], Pserf [2], and SMU [3] on ImageNet [6] with ResNet-26 [8], implemented with $1 \times \mathrm{RTX\ A100}$ GPU. Note that we follow the common practice to fix the input images to the size of 224×224.

Table 5 reports the comparative results of *Parameters*, *FLOPs*, *Throughput (image / s)*, and *Top-1 accuracy* for the corresponding ResNet-26s with IIEU-B and other activation models. Our major observations are: (a) Compared to the marginal additional FLOPs, IIEU-B shows relatively heavier decreases in throughput to the ReLU baseline and SiLU. (b) IIEU-B has close throughput to other SoTA activation models (*i.e.*, Meta-ACON, Pserf, and SMU). (c) IIEU-B enjoys significant improvements in Top-1 accuracy to the baseline and popular/SoTA activation models.

Table 5. Comparisons of FLOPs/throughput of different activation models.

Method	FLOPs	Throughput	Params.	Top-1(%)↑
ReLU [18]	2.4G	4688.4	16.0M	74.9
SiLU [7]	2.4G	4559.5	16.0M	75.8
Mt-ACON [15]	2.4G	3449.1	16.1M	76.5
Pserf [2]	2.4G	3417.8	16.0M	75.7
SMU [3]	2.4G	3444.6	16.0M	76.1
IIEU-B (ours)	2.4G	3630.3	16.0M	77.7

G.1. MS COCO Object Detection

Implementation details. As generic activation models, our IIEUs can be easily extended to other vision tasks. We evaluate our IIEU-B and IIEU-DC on MS COCO [14] object detection using the popular efficient detector RetinaNet. We compare our IIEUs to the baseline ReLU [18], the popular Swish [19], and the current SoTAs Meta-ACON [15] and SMU [3]. For fair comparisons, we adopt the default implementation configures ($1\times$ schedule) defined by the MMDetection toolbox [5] and report the standard evaluation metrics, *i.e.*, mAP (the primary metric of averaged precisions), AP_{50}, AP_{75}, AP_S, AP_M, AP_L (specific APs at different scales). We employ the ResNet-50 backbones equipped with different activation functions, each applied with their corresponding ImageNet pre-trained weights. Note that we keep using the deterministic mode for each of the implementations to ensure reproducibility.

Experimental results. We show the experimental results in Table 6, where our IIEUs enjoy clear gains in accuracy compared to different baseline/popular/SoTA activation models. This validates the scalability and versatility of IIEU. Note that we report the official results for Meta-ACON (*i.e.*, Mt-Acon) as our re-implemented results are lower (which is possibly caused by the different implementation environments).

Table 6. Comparison of different activation models on the COCO object detection [14].

Method	Backbone	Params.	FLOPs	mAP (%)↑	AP_{50}(%)↑	AP_{75}(%)↑	AP_S(%)↑	AP_M(%)↑	AP_L(%)↑
ReLU [18]	ResNet-50 [8]	37.7M	238.9G	36.7	56.0	39.3	21.0	40.2	48.2
Swish [19]		37.7M	238.9G	37.2	56.3	39.9	21.0	41.1	47.8
Mt-ACON [15]		37.9M	238.9G	36.5	55.9	38.9	19.9	40.7	50.6
SMU [3]		37.7M	238.9G	37.5	56.6	40.2	21.5	41.5	48.4
IIEU-B (ours)		37.7M	239.0G	**38.2**	**58.2**	**40.6**	**23.2**	**42.1**	**49.2**
IIEU-DC (ours)		40.4M	239.0G	**38.6**	**59.0**	**40.8**	**22.2**	**42.6**	**50.7**

H. KITTI-Materials Road Scene Material Segmentation

Implementation details. We evaluate our and compared activation models on an emerging task, *i.e.*KITTI-Materials [4] RGB road scene material segmentation, using ResNet-50 backbone. To ensure fair comparisons, we adopt the official implementation protocols applied in [4].

Experimental results. We report the results of our IIEU-B and the compared activation models in Appendix H, where IIEU-B achieves significant accuracy gains to ReLU baseline and also shows clear improvements on SoTAs Swish, SMU, and Meta-ACON. It is worth noting that our IIEU shows consistent significant accuracy improvements on various vision benchmarks to the baselines and SoTAs.

Table 7. Comparison of different activation models on KITTI-Materials [4] RGB road scene material segmentation.

Method	Encoder	Decoder	Params.	mIoU(%)↑
ReLU [18]	ResNet-50 [8]	All-MLP [24]	31.7M	40.2
Swish [19]			31.7M	41.2
Mt-ACON [15]			31.9M	41.7
SMU [3]			31.7M	40.6
IIEU-B (ours)			31.7M	**42.4**

References

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

[2] 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.

[3] 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.

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

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

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

[23] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[25] 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.

[26] 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.