Optimizing Neural Network Effectiveness via Non-monotonicity Refinement

Koushik Biswas[1], Amit Reza[2*], Meghana Karri[1*], Debesh Jha[1], Hongyi Pan[1],
Nikhil Tomar[1], Aliza Subedi[1], Smriti Regmi[1], Ulas Bagci[1]
[1]Machine and Hybrid Intelligence Lab, Northwestern University, Chicago, IL, USA.
[2]Space Research Institute, Austrian Academy of Science, Graz, Austria
koushik.biswas@northwestern.edu

Supplementary Material

Proposition: The following two equations (1 & 2) are equivalent.

$$max(x_1, x_2) = \begin{cases} x_1 & \text{if } x_1 \geq x_2 \\ x_2 & \text{otherwise} \end{cases} \tag{1}$$

We can rewrite the equation (1) as follows:

$$max(x_1, x_2) = x_1 + max(0, x_2 - x_1) \tag{2}$$

Proof:
Case-1: $x_1 = x_2$, then,

$$max(x_1, x_2) = x_1 = x_2$$

Case-2: $x_1 > x_2$, then, $x_2 - x_1 < 0$ and $max(0, x_2 - x_1) = 0$ so,

$$max(x_1, x_2) = x_1$$

Case-3: $x_1 < x_2$, then, $x_2 - x_1 > 0$ and $max(0, x_2 - x_1) = x_2 - x_1$, so,

$$max(x_1, x_2) = x_2$$

The SoftSign activation is defined as follows:

$$f(x) = \frac{x}{1 + |x|} \tag{3}$$

The SoftSign is not a differentiable function in the real line due to the $|x|$ function. $|x|$ can be approximated by $x\text{erf}(\alpha x)$, $x\tanh(\alpha x)$, $\sqrt{x^2 + \alpha^2}$. These three functions are infinite time differentiable, where α is a non-negative hyper-parameter. Thus, replacing $|x|$ with three functions, we have three approximations as follows:

$$g_1(x; \alpha) = \frac{x}{1 + x\text{erf}(\alpha x)} \tag{4}$$

$$g_2(x; \alpha) = \frac{x}{1 + x\tanh(\alpha x)} \tag{5}$$

$$g_3(x; \alpha) = \frac{x}{1 + \sqrt{(x^2 + \alpha^2)}} \tag{6}$$

These functions have a range in $[-1, 1]$. We need to transfer it to the whole real line, by using the transformation $x \to e^x$, we have replace x by e^x in equation (4), equation (5), and equation (6).

$$g_1(x; \alpha) = \frac{e^x}{1 + e^x\text{erf}(\alpha e^x)} = \frac{1}{e^{-x} + \text{erf}(\alpha e^x)} \tag{7}$$

$$g_2(x; \alpha) = \frac{e^x}{1 + e^x\tanh(\alpha e^x)} = \frac{1}{e^{-x} + \tanh(\alpha e^x)} \tag{8}$$

$$g_3(x; \alpha) = \frac{e^x}{1 + \sqrt{(e^{2x} + \alpha^2)}} = \frac{1}{e^{-x} + \sqrt{(1 + e^{-2x}\alpha^2)}} \tag{9}$$

Activation functions should be non-monotonic. So, to make these functions non-monotonic, multiply by x. Also, We add another parameter, β, to control the smoothness of these functions.

$$g_1(x; \alpha, \beta) = \frac{x}{e^{-\beta x} + \text{erf}(\alpha e^x)} \tag{10}$$

$$g_2(x; \alpha, \beta) = \frac{x}{e^{-\beta x} + \tanh(\alpha e^x)} \tag{11}$$

$$g_3(x; \alpha, \beta) = \frac{x}{e^{-\beta x} + \sqrt{(1 + \alpha^2 e^{-2x})}} \tag{12}$$

Now, if we consider $\beta \to \infty$, for large positive x, $g_i(x; \alpha, \beta), i = 1, 2, 3$, will converge to x. Also, note that, for $\beta \to \infty$, and for large negative x, $g_i(x; \alpha, \beta), i = 1, 2, 3$, will converge 0. So for large β, $g_i(x; \alpha, \beta), i = 1, 2, 3$ converges to ReLU activation function.

*Amit Reza and Meghana Karri contributed equally to this work.

1. Image Classification

1.1. MNIST, Fashion MNIST, and SVHN

We report more experiments on MNIST, Fashion MNIST, and SVHN datasets with LeNet [10] and AlexNet [9] architectures in Table 1 and Table 2 respectively. The experimental setup is the same as reported in the main document.

Table 1. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on MNIST, Fashion MNIST, and SVHN datasets for image classification problem on LeNet architecture. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	MNIST	Fashion MNIST	SVHN
ReLU	99.15 ± 0.11	91.34 ± 0.19	92.01 ± 0.19
Leaky ReLU	99.09 ± 0.11	91.32 ± 0.21	92.20 ± 0.21
ReLU6	99.24 ± 0.12	91.40 ± 0.21	92.32 ± 0.17
PReLU	99.20 ± 0.11	91.43 ± 0.21	92.12 ± 0.21
ELU	99.26 ± 0.10	91.32 ± 0.22	92.20 ± 0.19
Softplus	99.00 ± 0.21	91.03 ± 0.24	91.86 ± 0.25
GELU	99.30 ± 0.07	91.66 ± 0.15	92.50 ± 0.14
Swish	99.25 ± 0.10	91.75 ± 0.14	92.32 ± 0.16
PAU	99.36 ± 0.11	91.72 ± 0.13	92.32 ± 0.17
Mish	99.34 ± 0.06	**91.72** ± 0.13	92.36 ± 0.16
AMSU-1	**99.50** ± 0.04	91.61 ± 0.14	**92.70** ± 0.16
AMSU-2	**99.45** ± 0.06	91.55 ± 0.14	**92.77** ± 0.14
AMSU-3	**99.40** ± 0.07	91.54 ± 0.15	**92.60** ± 0.14

Table 2. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on MNIST, Fashion MNIST, and SVHN datasets for image classification problem on AlexNet architecture. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	MNIST	Fashion MNIST	SVHN
ReLU	99.45 ± 0.08	92.77 ± 0.21	95.12 ± 0.17
Leaky ReLU	99.55 ± 0.08	92.80 ± 0.16	95.17 ± 0.15
ReLU6	99.50 ± 0.04	92.88 ± 0.13	95.13 ± 0.12
PReLU	99.40 ± 0.08	92.70 ± 0.18	95.18 ± 0.14
ELU	99.50 ± 0.06	92.91 ± 0.14	95.10 ± 0.14
Softplus	99.20 ± 0.10	92.36 ± 0.26	94.57 ± 0.21
GELU	99.50 ± 0.06	93.17 ± 0.11	95.24 ± 0.12
Swish	99.55 ± 0.06	92.95 ± 0.15	95.34 ± 0.17
PAU	99.54 ± 0.11	93.04 ± 0.15	95.37 ± 0.13
Mish	99.60 ± 0.06	93.19 ± 0.16	95.38 ± 0.14
AMSU-1	**99.70** ± 0.04	**93.38** ± 0.10	**95.63** ± 0.14
AMSU-2	**99.72** ± 0.05	**93.34** ± 0.10	**95.68** ± 0.11
AMSU-3	**99.63** ± 0.06	**93.26** ± 0.12	**95.52** ± 0.12

1.2. CIFAR

We report more experiments on the CIFAR100 and CIFAR10 datasets with AlexNet, Inception V3, DenseNet-121, WideResNet 28-10, ShuffleNet V1, Googl-eNet, SqueezeNet, VGG 16, and LeNet architectures in Table 5 and Table 9 respectively. An extension to the Table 2 (from main paper) on CIFAR100 dataset with baseline activations are reported in Table 6 & Table 7 and an extension to the Table 3 (from main paper) on CIFAR10 dataset with baseline activations are reported in Table 8 & Table 10.

1.3. Semantic Segmentation

We report more experiments on semantic segmentation with the CityScapes [4] dataset with the UNet model and the Pascal Voc dataset with PspNet [19]. We consider the same experimental setup as mentioned in the main document. The results are reported in Table 3 and Table 4.

Table 3. Comparison between baseline activations and AMSU-1, AMSU-2, & AMSU-3 on the CityScapes dataset for semantic segmentation problem on UNet. We report mIOU for the mean of 10 different runs. mean±std is reported in the table.

Activation Function	mIOU
ReLU	69.5 ± 0.07
Leaky ReLU	69.7 ± 0.07
PReLU	69.7 ± 0.08
ReLU6	69.6 ± 0.10
ELU	69.5 ± 0.06
SoftPlus	69.2 ± 0.10
Swish	70.0 ± 0.08
Mish	70.1 ± 0.09
GELU	69.9 ± 0.10
PAU	70.1 ± 0.06
AMSU-1	**70.5** ± 0.09
AMSU-2	**70.7** ± 0.09
AMSU-3	**70.3** ± 0.10

Table 4. Comparison between baseline activations and AMSU-1, AMSU-2, & AMSU-3 on the Pascal VOC dataset for semantic segmentation problem on PSPNet. We report mIOU for the mean of 10 different runs. mean±std is reported in the table.

Activation Function	mIOU
ReLU	78.9 ± 0.07
Leaky ReLU	79.0 ± 0.06
PReLU	79.1 ± 0.04
ReLU6	79.1 ± 0.04
ELU	79.2 ± 0.05
SoftPlus	78.4 ± 0.10
Swish	79.6 ± 0.04
Mish	79.5 ± 0.06
GELU	79.4 ± 0.04
PAU	79.4 ± 0.04
AMSU-1	**79.9** ± 0.04
AMSU-2	**80.0** ± 0.04
AMSU-3	**79.5** ± 0.06

2. Machine Translation

This section presents a detailed experimental evaluation of the proposed activation functions and baseline functions on the machine translation problem. This problem deals with translating text or speech data from one language to another without the help of any human being. The WMT 2014 English→German dataset is used for our experiment. The database contains a 4.5 million training dataset (i.e., sentences). We use an attention-based [18] 8-head transformer network with Adam optimizer [8], 0.1 dropout rate [16], and trained up to 10^5 steps without changing the Other hyperparameters [18]. We evaluate the network performance on the newstest2014 dataset using the BLEU score metric. We report the mean of 10 different runs in Table 11 on the test dataset(newstest2014). It shows that the results are stable on different runs (mean±std), and we got $0.6\%, 0.6\%, 0.5\%$ boost in BLEU score for the proposed AMSU-1, AMSU-2, & AMSU-3 activations respectively compared to ReLU.

Table 5. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on CIFAR100 dataset for image classification problem. We report Top-1 test accuracy (in %) for the mean of 20 different runs. mean±std is reported in the table.

Activation Function	Alex Net	Shuffle Net V1	Google Net	Inception V3	Dense Net 121	WideRes Net 28-10	Squeeze Net	VGG 16	LeNet
ReLU	54.78 ±0.27	65.70 ±0.27	72.50 ±0.33	74.23 ±0.25	75.76 ±0.26	76.44 ±0.27	66.11 ±0.28	71.79 ±0.28	45.42 ±0.26
Leaky ReLU	55.50 ±0.29	65.90 ±0.27	72.49 ±0.29	74.40 ±0.27	75.80 ±0.26	76.50 ±0.28	66.23 ±0.30	71.99 ±0.28	45.69 ±0.27
ReLU6	55.77 ±0.25	66.23 ±0.27	72.57 ±0.25	74.53 ±0.25	75.89 ±0.27	76.63 ±0.26	66. ±0.26	71.55 ±0.25	45.80 ±0.27
PReLU	55.50 ±0.26	65.82 ±0.31	72.51 ±0.26	74.22 ±0.27	76.22 ±0.26	76.67 ±0.26	66.30 ±0.26	71.86 ±0.29	45.50 ±0.30
ELU	55.88 ±0.27	65.79 ±0.26	72.81 ±0.26	74.55 ±0.26	75.77 ±0.22	76.17 ±0.27	66.30 ±0.27	71.70 ±0.31	46.15 ±0.27
Softplus	54.90 ±0.37	65.01 ±0.35	71.68 ±0.37	74.11 ±0.35	75.33 ±0.34	75.52 ±0.36	65.60 ±0.34	70.90 ±0.30	44.01 ±0.37
GELU	57.21 ±0.28	67.17 ±0.24	73.10 ±0.24	75.60 ±0.26	76.77 ±0.24	77.20 ±0.23	66.88 ±0.28	71.80 ±0.28	47.50 ±0.24
Swish	57.47 ±0.29	67.14 ±0.28	73.39 ±0.28	75.59 ±0.26	76.63 ±0.30	77.47 ±0.26	66.48 ±0.25	71.82 ±0.27	47.42 ±0.25
PAU	57.20 ±0.28	67.39 ±0.27	73.59 ±0.26	75.77 ±0.28	76.81 ±0.28	77.17 ±0.24	66.70 ±0.22	71.68 ±0.24	47.42 ±0.30
Mish	58.10 ±0.25	67.77 ±0.26	74.06 ±0.25	76.20 ±0.23	77.32 ±0.24	77.36 ±0.21	67.45 ±0.25	72.44 ±0.21	**47.55** ±0.30
AMSU-1	**61.35** ±0.20	**69.30** ±0.25	**74.50** ±0.22	**77.69** ±0.23	**78.41** ±0.20	**78.69** ±0.21	**68.66** ±0.23	**73.49** ±0.23	47.41 ±0.23
AMSU-2	**61.27** ±0.22	**69.38** ±0.23	**74.51** ±0.24	**77.58** ±0.20	**78.30** ±0.23	**78.61** ±0.24	**68.51** ±0.20	**73.33** ±0.20	47.30 ±0.22
AMSU-3	**61.10** ±0.22	**68.86** ±0.20	**74.20** ±0.22	**77.06** ±0.20	**78.23** ±0.22	**78.27** ±0.22	**68.29** ±0.24	72.41 ±0.22	47.18 ±0.20

Table 6. This is an extension to Table 2. We report Top-1 test accuracy (in %) on the CIFAR100 dataset for baseline functions for the mean of 20 different runs. mean±std is reported in the table. SF V2 stands for ShuffleNet v2.

Activation Function	SF V2 0.5x	SF V2 1.0x	SF V2 1.5x	SF V2 2.0x	SeNet 18	SeNet 34	SeNet 50	Res-Next	Xcep-tion	EffitientNet B0
Leaky ReLU	62.18 ±0.30	65.28 ±0.32	67.41 ±0.30	67.70 ±0.30	74.55 ±0.22	75.01 ±0.22	76.15 ±0.20	74.50 ±0.22	71.15 ±0.24	76.70 ±0.26
ReLU6	62.27 ±0.32	65.84 ±0.30	67.49 ±0.27	68.19 ±0.23	74.62 ±0.23	75.22 ±0.24	76.60 ±0.23	74.52 ±0.23	71.28 ±0.22	76.55 ±0.23
PReLU	62.14 ±0.33	65.19 ±0.30	67.29 ±0.32	67.80 ±0.27	74.28 ±0.26	75.30 ±0.25	76.26 ±0.26	74.39 ±0.28	71.30 ±0.26	76.55 ±0.28
ELU	62.55 ±0.32	65.48 ±0.29	67.70 ±0.24	67.77 ±0.30	74.67 ±0.20	75.01 ±0.20	76.30 ±0.20	74.20 ±0.20	71.40 ±0.22	76.61 ±0.27
Softplus	61.70 ±0.35	64.59 ±0.33	67.29 ±0.30	68.60 ±0.34	74.20 ±0.32	74.67 ±0.30	75.16 ±0.35	74.20 ±0.36	71.17 ±0.36	76.44 ±0.37
GELU	64.48 ±0.27	66.67 ±0.23	69.80 ±0.28	70.19 ±0.28	74.98 ±0.21	76.12 ±0.20	77.22 ±0.20	75.29 ±0.22	72.01 ±0.21	77.42 ±0.21
Swish	63.66 ±0.24	66.90 ±0.24	69.66 ±0.29	70.32 ±0.26	74.49 ±0.20	75.82 ±0.22	76.77 ±0.21	75.28 ±0.26	72.30 ±0.20	77.31 ±0.19
PAU	64.17 ±0.25	66.70 ±0.24	69.41 ±0.26	70.60 ±0.27	74.77 ±0.22	75.82 ±0.26	77.01 ±0.21	75.57 ±0.24	72.50 ±0.25	77.35 ±0.22
Mish	64.82 ±0.22	67.92 ±0.24	70.56 ±0.23	71.40 ±0.20	75.21 ±0.20	76.59 ±0.23	77.79 ±0.21	76.29 ±0.22	73.61 ±0.24	78.26 ±0.24

Table 7. This is an extension to Table 2. We report Top-1 test accuracy (in %) on the CIFAR100 dataset for baseline functions for the mean of 20 different runs. mean±std is reported in the table.

Activation Function	ResNet 18	ResNet 34	ResNet 50	PreAct ResNet 18	PreAct ResNet 34	PreAct ResNet 50	MobileNet V1	MobileNet V2
Leaky ReLU	73.01 ±0.23	73.30 ±0.27	74.30 ±0.23	73.20 ±0.21	73.30 ±0.22	74.20 ±0.21	71.20 ±0.24	74.10 ±0.22
ReLU6	73.30 ±0.22	73.50 ±0.24	74.20 ±0.22	73.40 ±0.21	73.50 ±0.20	74.35 ±0.21	71.47 ±0.22	74.40 ±0.21
PReLU	73.10 ±0.25	73.45 ±0.26	74.20 ±0.26	73.10 ±0.23	73.55 ±0.28	74.16 ±0.24	71.51 ±0.31	74.59 ±0.29
ELU	73.30 ±0.22	73.60 ±0.24	74.33 ±0.25	73.30 ±0.20	73.41 ±0.22	74.59 ±0.24	71.30 ±0.21	74.32 ±0.22
Softplus	72.67 ±0.34	73.29 ±0.35	74.00 ±0.37	72.83 ±0.37	73.22 ±0.33	73.82 ±0.34	71.00 ±0.37	74.41 ±0.34
GELU	73.81 ±0.21	74.12 ±0.22	75.70 ±0.21	74.91 ±0.23	74.30 ±0.23	74.80 ±0.24	71.70 ±0.24	75.19 ±0.24
Swish	73.55 ±0.22	74.04 ±0.22	75.23 ±0.22	75.20 ±0.22	74.70 ±0.19	75.19 ±0.20	71.99 ±0.20	75.26 ±0.21
PAU	74.19 ±0.21	74.40 ±0.23	75.91 ±0.22	74.80 ±0.23	74.80 ±0.22	75.81 ±0.20	71.71 ±0.23	75.01 ±0.20
Mish	74.49 ±0.21	74.68 ±0.20	76.10 ±0.24	75.01 ±0.20	75.24 ±0.20	76.89 ±0.17	72.20 ±0.22	75.33 ±0.21

3. 3D Medical Imaging

The following subsections present detailed experimental results on 3D medical image classification, 3D medical image segmentation, and 2D medical image classification problems.

3.1. 3D Medical Image Classification

In this section, we report experimental results for the 3D image classification problem on MosMed dataset [14]. The dataset has CT scans with COVID-19-related findings (CT1-CT4) and without any findings (CT0). The dataset

Table 8. This is an extension to Table 3. We report Top-1 test accuracy (in %) on the CIFAR10 dataset for baseline functions for the mean of 20 different runs. mean±std is reported in the table.

Activation Function	ResNet 18	ResNet 34	ResNet 50	PreAct ResNet 18	PreAct ResNet 34	PreAct ResNet 50	MobileNet V1	MobileNet V2
Leaky ReLU	94.09 ±0.22	94.14 ±0.23	94.32 ±0.25	93.43 ±0.18	94.20 ±0.20	94.25 ±0.19	92.47 ±0.19	94.17 ±0.17
ReLU6	94.22 ±0.23	94.27 ±0.26	94.30 ±0.38	93.58 ±0.22	94.12 ±0.22	94.44 ±0.21	92.58 ±0.18	94.15 ±0.18
PReLU	94.15 ±0.25	94.20 ±0.26	94.20 ±0.25	93.47 ±0.21	94.20 ±0.27	94.38 ±0.29	92.44 ±0.17	94.30 ±0.21
ELU	94.20 ±0.21	94.14 ±0.21	94.27 ±0.24	93.50 ±0.20	94.30 ±0.20	94.30 ±0.21	92.60 ±0.18	94.110 ±0.17
Softplus	93.71 ±0.28	93.81 ±0.28	93.68 ±0.28	93.00 ±0.27	93.95 ±0.31	94.17 ±0.29	91.89 ±0.29	93.81 ±0.27
GELU	94.27 ±0.19	94.44 ±0.21	94.69 ±0.21	93.56 ±0.24	94.29 ±0.22	94.57 ±0.20	92.80 ±0.23	94.26 ±0.14
Swish	94.30 ±0.22	94.35 ±0.23	94.76 ±0.20	93.64 ±0.19	94.23 ±0.23	94.52 ±0.21	92.57 ±0.20	94.38 ±0.16
PAU	94.26 ±0.22	94.50 ±0.24	94.55 ±0.21	93.71 ±0.21	94.17 ±0.20	94.66 ±0.22	93.12 ±0.12	94.39 ±0.12
Mish	94.43 ±0.21	94.29 ±0.20	94.68 ±0.21	93.81 ±0.20	94.41 ±0.21	94.71 ±0.22	92.72 ±0.22	94.71 ±0.16

Table 9. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on CIFAR10 dataset for image classification problem. We report Top-1 test accuracy (in %) for the mean of 20 different runs. mean±std is reported in the table.

Activation Function	Alex Net	Shuffle Net V1	Google Net	Inception V3	Dense Net 121	WideRes Net 28-10	Squeeze Net	VGG 16	LeNet
ReLU	84.17 ±0.21	91.28 ±0.21	92.82 ±0.20	94.10 ±0.20	94.82 ±0.17	95.11 ±0.18	90.50 ±0.21	93.53 ±0.17	75.71 ±0.20
Leaky ReLU	84.15 ±0.20	91.48 ±0.21	92.90 ±0.16	94.18 ±0.20	94.55 ±0.20	94.91 ±0.18	90.60 ±0.18	93.63 ±0.18	75.87 ±0.21
ReLU6	84.68 ±0.18	91.60 ±0.17	92.83 ±0.14	94.10 ±0.17	94.61 ±0.21	95.40 ±0.21	90.90 ±0.20	93.78 ±0.20	75.70 ±0.20
PReLU	84.21 ±0.22	91.70 ±0.22	92.90 ±0.21	94.31 ±0.23	94.44 ±0.20	95.13 ±0.23	90.88 ±0.24	93.49 ±0.20	75.81 ±0.19
ELU	84.77 ±0.20	91.80 ±0.20	92.87 ±0.18	94.52 ±0.18	94.82 ±0.20	95.30 ±0.20	90.90 ±0.18	93.90 ±0.17	75.95 ±0.20
Softplus	84.12 ±0.27	91.17 ±0.27	92.61 ±0.30	94.20 ±0.29	94.39 ±0.27	94.71 ±0.25	90.54 ±0.29	93.29 ±0.26	75.50 ±0.32
GELU	85.10 ±0.20	91.89 ±0.21	93.40 ±0.20	94.38 ±0.15	94.80 ±0.22	95.27 ±0.20	90.77 ±0.14	93.52 ±0.17	77.77 ±0.20
Swish	85.27 ±0.20	91.36 ±0.21	93.19 ±0.17	94.29 ±0.17	94.71 ±0.19	95.39 ±0.20	91.21 ±0.19	93.73 ±0.19	77.61 ±0.21
PAU	84.96 ±0.22	91.81 ±0.20	93.07 ±0.17	94.20 ±0.22	94.42 ±0.21	94.93 ±0.21	90.59 ±0.20	93.49 ±0.22	77.77 ±0.22
Mish	85.80 ±0.18	91.88 ±0.15	93.37 ±0.19	94.38 ±0.14	95.10 ±0.15	95.50 ±0.17	91.20 ±0.18	93.83 ±0.20	**77.93** ±0.18
AMSU-1	**87.37** ±0.12	**92.49** ±0.16	**94.21** ±0.18	**95.70** ±0.12	**96.03** ±0.13	**96.12** ±0.12	**91.86** ±0.12	**94.52** ±0.11	77.75 ±0.18
AMSU-2	**87.30** ±0.14	**92.50** ±0.17	**94.10** ±0.15	**95.79** ±0.14	**96.14** ±0.12	**96.03** ±0.11	**91.70** ±0.14	**94.60** ±0.12	77.61 ±0.16
AMSU-3	**86.98** ±0.15	**92.20** ±0.12	**93.88** ±0.18	**95.39** ±0.16	**95.60** ±0.14	**95.91** ±0.12	**91.49** ±0.14	**94.40** ±0.16	77.45 ±0.14

Table 10. This is an extension to Table 3. We report Top-1 test accuracy (in %) on the CIFAR10 dataset for baseline functions for the mean of 20 different runs. mean±std is reported in the table. SF V2 stands for ShuffleNet v2.

Activation Function	SF V2 0.5x	SF V2 1.0x	SF V2 1.5x	SF V2 2.0x	SeNet 18	SeNet 34	SeNet 50	Res-Next	Xcep-tion	EffitientNet B0
Leaky ReLU	88.20 ±0.23	91.28 ±0.26	91.11 ±0.23	91.79 ±0.26	94.23 ±0.25	94.61 ±0.24	94.39 ±0.19	93.33 ±0.20	90.89 ±0.23	95.49 ±0.14
ReLU6	88.59 ±0.20	91.27 ±0.22	91.21 ±0.21	91.76 ±0.19	94.25 ±0.19	94.61 ±0.22	94.60 ±0.20	93.39 ±0.20	91.11 ±0.19	95.21 ±0.14
PReLU	88.24 ±0.21	91.11 ±0.20	91.33 ±0.21	91.68 ±0.22	94.21 ±0.22	94.47 ±0.22	94.51 ±0.21	93.22 ±0.21	91.18 ±0.22	95.31 ±0.20
ELU	88.09 ±0.20	91.02 ±0.23	91.40 ±0.23	91.77 ±0.21	94.30 ±0.23	94.30 ±0.24	94.60 ±0.23	93.41 ±0.22	91.51 ±0.20	95.30 ±0.20
Softplus	87.86 ±0.29	90.31 ±0.28	91.11 ±0.30	90.91 ±0.28	93.91 ±0.27	94.11 ±0.28	94.31 ±0.28	93.17 ±0.26	90.32 ±0.24	94.92 ±0.22
GELU	88.81 ±0.21	91.55 ±0.22	91.85 ±0.18	92.17 ±0.16	94.32 ±0.18	94.72 ±0.18	94.67 ±0.14	93.55 ±0.21	92.08 ±0.21	95.40 ±0.19
Swish	89.15 ±0.18	91.77 ±0.20	91.72 ±0.23	92.05 ±0.21	94.20 ±0.20	94.60 ±0.20	94.40 ±0.18	93.50 ±0.18	91.55 ±0.18	95.41 ±0.14
PAU	89.30 ±0.20	91.59 ±0.22	92.11 ±0.18	92.18 ±0.18	94.21 ±0.20	94.65 ±0.21	94.61 ±0.19	93.57 ±0.18	91.99 ±0.23	95.40 ±0.16
Mish	89.30 ±0.18	91.85 ±0.19	92.30 ±0.18	92.40 ±0.17	94.37 ±0.17	94.68 ±0.17	94.88 ±0.14	93.90 ±0.16	92.20 ±0.22	95.77 ±0.14

has 1110 studies. We consider 70% of the data as training data and 30% of the data used for testing. We use 0.0001 initial learning rate and decay the learning rate with cosine annealing learning rate scheduler [11]. We use Adam optimizer [8], $5e^{-4}$ weight decay, a batch size of 8, and trained up to 200 epochs with 3D ResNet-18 model. Experimental results are reported in Table 12 for the mean of 10 different runs.

3.2. 3D Medical Image Segmentation

In this section, we report experimental results for 3D brain tumor segmentation using 3D-UNet [20] on the BraTS

Table 11. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on WMT2014 dataset for machine translation problem. We report the BLEU score for the mean of 10 different runs. mean±std is reported in the table.

Activation Function	BLEU Score
ReLU	26.2 ± 0.13
Leaky ReLU	26.3 ± 0.17
PReLU	26.2 ± 0.15
ReLU6	26.1 ± 0.16
ELU	25.2 ± 0.14
SoftPlus	23.5 ± 0.18
Swish	26.4 ± 0.13
Mish	26.3 ± 0.14
GELU	26.4 ± 0.14
PAU	26.3 ± 0.17
AMSU-1	**26.8** ± 0.11
AMSU-2	**26.8** ± 0.12
AMSU-3	**26.5** ± 0.14

Table 12. Comparison between AMSU-1, AMSU-2, & AMSU-3 and other baseline activations on MosMedData dataset for 3D image classification with ResNet-18 model. We report Top-1 accuracy for the mean of 10 different runs.

Activation Function	Accuracy
ReLU	79.57
Leaky ReLU	79.74
PReLU	79.88
ReLU6	79.77
ELU	79.83
SoftPlus	79.62
Swish	80.03
Mish	80.18
GELU	79.99
PAU	80.11
AMSU-1	**80.70**
AMSU-2	**80.55**
AMSU-3	**80.22**

Table 13. Comparison between baseline activations and AMSU-1, AMSU-2, & AMSU-3 on the CIFAR10 dataset for FGSM Adversarial attack ($\epsilon = 0.03$). We report Top-1 accuracy for the mean of 5 different runs. mean±std is reported in the table.

Activation Function	ShuffleNet V2 (2.0x)	MobileNet V2
ReLU	87.45 ± 0.18	90.61 ± 0.16
Leaky ReLU	87.57 ± 0.13	90.68 ± 0.11
PReLU	87.81 ± 0.12	90.58 ± 0.11
ReLU6	87.88 ± 0.12	90.69 ± 0.12
ELU	87.99 ± 0.14	90.82 ± 0.14
SoftPlus	87.54 ± 0.20	90.78 ± 0.17
Swish	89.67 ± 0.11	91.50 ± 0.12
Mish	89.37 ± 0.14	91.69 ± 0.14
GELU	89.18 ± 0.16	91.51 ± 0.12
PAU	89.97 ± 0.11	91.70 ± 0.13
AMSU-1	**91.29** ± 0.10	**92.38** ± 0.12
AMSU-2	**91.12** ± 0.12	**92.47** ± 0.11
AMSU-3	**90.67** ± 0.14	**92.14** ± 0.11

2020 dataset [1, 2, 13]. This data set has 369 samples for training and 125 samples for validation. For this experiment, we consider a batch size of 2, and Adam optimizer [8] with $5e^{-4}$ weight decay, the initial learning rate is 0.001, and cosine annealing learning rate scheduler, and we train this 3D model for 150 epochs. We presented the network performance analysis for AMSU-1, AMSU-2, & AMSU-3 and the baseline activation functions with this network in Table 14 in terms of Accuracy and Dice Score to measure the performance.

Table 14. Comparison between AMSU-1, AMSU-2, & AMSU-3 and other baseline activations on BraTS 2020 Dataset for 3D image segmentation with 3D-Unet Model.

Activation Function	Accuracy	Dice Score
ReLU	95.03	97.40
Leaky ReLU	94.99	97.36
PReLU	94.98	97.35
ReLU6	94.91	97.36
ELU	95.08	97.39
SoftPlus	94.88	97.37
Swish	95.15	97.48
Mish	95.10	97.50
GELU	95.15	97.31
PAU	95.22	97.52
AMSU-1	**95.47**	**97.70**
AMSU-2	**95.49**	**97.67**
AMSU-3	**95.38**	**97.59**

3.3. 2D Medical Image Classification

We use RadImageNet [12] database for 2D medical image classification tasks, which is an open-access medical imaging database designed to improve transfer learning performance on downstream medical imaging applications and perhaps it is the largest ever medical imaging dataset so far. We run experiment on CT abdominal/pelvis and Lung data from the whole dataset, consisting of 28 and 6 disease classes, respectively. We also run experiments on MRI abdominal/pelvis and Brain data, consisting of 26 and 10 disease classes. No pre-trained weight is used for our experiments. We consider a batch size 32, 0.00001 initial learning rate, Adam [8] optimizer, and $1e^{-4}$ weight decay. The results are reported with ResNet-50 in Table 17. We got almost 6%, 3%, 3%, and 2% improvement compared to ReLU on Abdomen (CT), Lung (CT), Abdomen (MRI), and Brain (MRI), respectively. We also report results with ShuffleNet and MobileNet V2 in Table 15, and Table 18 on Abdomen (CT) and Abdomen (MRI).

Table 15. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on the RadImageNet dataset for 2D medical image classification problem on **ShuffleNet** architecture. We report Top-1 test accuracy (in %) for the mean of 3 different runs. mean±std is reported in the table.

Activation Function	Abdomen (CT)	Abdomen (MRI)
ReLU	62.61 ± 0.22	84.41 ± 0.15
Leaky ReLU	62.75 ± 0.21	84.56 ± 0.18
ReLU6	62.66 ± 0.21	84.76 ± 0.20
PReLU	62.89 ± 0.19	84.60 ± 0.22
ELU	63.20 ± 0.20	84.89 ± 0.20
Softplus	62.55 ± 0.20	84.50 ± 0.16
GELU	63.91 ± 0.19	85.06 ± 0.17
Swish	64.26 ± 0.20	85.02 ± 0.19
PAU	64.59 ± 0.18	85.17 ± 0.20
Mish	64.97 ± 0.18	85.65 ± 0.18
AMSU-1	**65.78** ± 0.17	**86.10** ± 0.17
AMSU-2	**65.39** ± 0.19	**85.89** ± 0.18
AMSU-3	**65.19** ± 0.19	**85.90** ± 0.17

3.4. 2D Medical Image Segmentation

We consider the Liver Tumor Segmentation Benchmark (LiTS) [3] dataset for the segmentation tasks. LiTS is a multi-center dataset collected from seven clinical centers.

Table 16. Baseline table for AMSU-3. These numbers represent the total number of models in which AMSU-3 outperforms, equals, or underperforms compared to the baseline activation functions

Baselines	ReLU	Leaky ReLU	ELU	SoftPlus	PReLU	ReLU6	Swish	Mish	GELU	PAU
AMSU-3 $>$ Baseline	101	101	101	102	101	101	97	96	97	96
AMSU-3 $=$ Baseline	0	0	0	0	0	0	0	0	0	0
AMSU-3 $<$ Baseline	1	1	1	0	1	1	5	6	5	6

Table 17. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on the RadImageNet dataset for 2D medical image classification problem on **ResNet-50** architecture. We report Top-1 test accuracy (in %) for the mean of 3 different runs. mean$\pm$std is reported in the table.

Activation Function	Abdomen (CT)	Lung (CT)	Abdomen (MRI)	Brain (MRI)
ReLU	67.50 $\pm$ 0.20	84.38 $\pm$ 0.20	86.67 $\pm$ 0.17	87.10 $\pm$ 0.18
Leaky ReLU	68.03 $\pm$ 0.18	84.56 $\pm$ 0.15	86.89 $\pm$ 0.16	87.37 $\pm$ 0.20
ReLU6	68.18 $\pm$ 0.20	84.58 $\pm$ 0.18	87.10 $\pm$ 0.17	87.31 $\pm$ 0.18
PReLU	68.62 $\pm$ 0.17	84.72 $\pm$ 0.17	87.22 $\pm$ 0.22	87.50 $\pm$ 0.20
ELU	68.45 $\pm$ 0.20	84.70 $\pm$ 0.19	87.45 $\pm$ 0.20	87.57 $\pm$ 0.18
Softplus	68.12 $\pm$ 0.20	84.30 $\pm$ 0.21	87.10 $\pm$ 0.24	87.41 $\pm$ 0.21
GELU	70.60 $\pm$ 0.15	85.85 $\pm$ 0.18	88.12 $\pm$ 0.20	88.02 $\pm$ 0.18
Swish	70.88 $\pm$ 0.16	86.00 $\pm$ 0.17	88.60 $\pm$ 0.21	87.91 $\pm$ 0.20
PAU	71.10 $\pm$ 0.15	86.27 $\pm$ 0.17	88.51 $\pm$ 0.20	88.25 $\pm$ 0.18
Mish	71.45 $\pm$ 0.17	86.10 $\pm$ 0.18	88.31 $\pm$ 0.20	87.87 $\pm$ 0.18
AMSU-1	**73.37** $\pm$ 0.15	**87.50** $\pm$ 0.16	**89.68** $\pm$ 0.14	**89.30** $\pm$ 0.16
AMSU-2	**72.92** $\pm$ 0.16	**87.20** $\pm$ 0.14	**89.78** $\pm$ 0.17	**89.12** $\pm$ 0.17
AMSU-3	**72.47** $\pm$ 0.17	**86.91** $\pm$ 0.16	**89.30** $\pm$ 0.16	**88.80** $\pm$ 0.17

Table 18. Comparison between AMSU-1, AMSU-2, & AMSU-3 activations and other baseline activations on the RadImageNet dataset for 2D medical image classification problem on **MobileNet V2** architecture. We report Top-1 test accuracy (in %) for the mean of 3 different runs. mean$\pm$std is reported in the table.

Activation Function	Abdomen (CT)	Abdomen (MRI)
ReLU	58.51 $\pm$ 0.18	82.77 $\pm$ 0.17
Leaky ReLU	58.73 $\pm$ 0.17	82.90 $\pm$ 0.19
ReLU6	59.02 $\pm$ 0.17	83.01 $\pm$ 0.17
PReLU	58.81 $\pm$ 0.19	82.81 $\pm$ 0.20
ELU	59.22 $\pm$ 0.20	83.21 $\pm$ 0.18
Softplus	58.70 $\pm$ 0.22	82.87 $\pm$ 0.20
GELU	59.74 $\pm$ 0.18	82.90 $\pm$ 0.18
Swish	59.97 $\pm$ 0.16	82.87 $\pm$ 0.17
PAU	60.45 $\pm$ 0.19	83.26 $\pm$ 0.18
Mish	60.78 $\pm$ 0.18	83.40 $\pm$ 0.17
AMSU-1	**61.23** $\pm$ 0.19	**83.89** $\pm$ 0.18
AMSU-2	**61.04** $\pm$ 0.18	**83.98** $\pm$ 0.16
AMSU-3	**60.86** $\pm$ 0.18	**83.68** $\pm$ 0.17

The dataset contains 201 CT images of the abdomen. The whole dataset is distributed into a training dataset with 130 CT scans, and the test dataset has 71 CT scans. We consider a batch size 16 and learning rate $1e^{-4}$ with Adam optimizer. We consider the input image size of 256×256. The results are reported in Table 19 and Table 20 with Unet [15], NanoNet-A [7], Unext [17], and ResUnet++ [6].

4. Adversarial Attack

Top-1 accuracy is presented in Table 21 ($\epsilon = 0.04$) on the CIFAR10 dataset for the FGSM Adversarial attack [5] for a mean of 10 different runs.

Table 19. Comparison of different activation functions on liver segmentation benchmark (LiTS) dataset.

Activation Function	UNet				UNext			
	mDSC	**mIoU**	**Rec.**	**Prec.**	**mDSC**	**mIoU**	**Rec.**	**Prec.**
ReLU	82.11	73.51	77.96	91.18	80.29	71.29	**78.90**	87.80
LReLU	82.47	73.98	78.46	91.17	80.92	72.15	76.84	90.85
ReLU6	82.34	73.50	78.04	91.31	80.20	71.17	78.55	87.49
PReLU	82.71	73.91	78.77	91.21	79.08	69.95	77.38	88.55
ELU	82.89	73.70	78.97	91.39	79.20	69.77	77.17	88.27
SoftPlus	82.00	73.77	77.55	91.01	79.10	70.80	78.55	87.98
GELU	83.20	73.88	79.01	91.22	81.01	72.61	78.79	88.09
Swish	83.01	73.99	78.57	90.45	81.16	72.19	78.87	88.29
PAU	82.68	73.60	78.78	90.90	80.67	72.01	**78.90**	88.01
Mish	82.48	73.88	78.44	90.46	80.39	72.55	78.80	87.55
AMSU-1	**83.62**	**75.08**	**80.59**	**93.29**	**82.10**	**73.10**	78.60	**91.17**
AMSU-2	**83.71**	**74.98**	**80.44**	**93.40**	**82.10**	**73.09**	78.77	**91.34**
AMSU-3	**83.22**	**74.56**	**80.12**	**93.20**	**81.67**	**72.91**	78.11	**90.89**

Table 20. Comparison of different activation functions on liver segmentation benchmark (LiTS) dataset.

Activation Function	NanoNet-A				ResUnet++			
	mDSC	**mIoU**	**Rec.**	**Prec.**	**mDSC**	**mIoU**	**Rec.**	**Prec.**
ReLU	75.17	66.62	73.40	83.11	76.90	68.03	79.81	83.43
LReLU	74.10	65.10	73.29	84.12	74.98	66.47	74.55	83.92
ReLU6	75.28	66.60	73.49	83.30	76.99	67.80	79.67	83.33
PReLU	74.88	66.21	72.50	85.39	74.30	65.87	74.21	81.80
ELU	74.90	66.40	72.40	85.10	74.67	65.60	74.20	81.60
SoftPlus	75.20	66.20	73.67	83.20	76.67	68.10	80.10	81.55
GELU	74.99	66.56	72.58	85.45	74.43	65.34	74.45	84.61
Swish	74.67	66.50	72.20	85.18	74.21	65.22	74.23	84.40
PAU	74.45	66.42	72.01	84.89	74.20	64.91	74.01	84.17
Mish	74.50	66.20	71.67	84.55	74.31	64.90	74.17	84.55
AMSU-1	**76.81**	**67.52**	**76.10**	**84.90**	**78.37**	**69.76**	**81.85**	**84.70**
AMSU-2	**76.55**	**67.80**	**76.11**	**84.97**	**78.56**	**69.70**	**81.55**	84.39
AMSU-3	**76.70**	**67.41**	**75.78**	**84.70**	**78.01**	**69.54**	**81.51**	84.01

Table 21. Comparison between baseline activations and AMSU-1, AMSU-2, & AMSU-3 on the CIFAR10 dataset for FGSM Adversarial attack ($\epsilon = 0.04$). We report Top-1 accuracy for the mean of 5 different runs. mean$\pm$std is reported in the table.

Activation Function	ShuffleNet V2 (2.0x)	MobileNet V2
ReLU	87.36 $\pm$ 0.15	90.57 $\pm$ 0.17
Leaky ReLU	87.50 $\pm$ 0.12	90.60 $\pm$ 0.11
PReLU	87.78 $\pm$ 0.12	90.49 $\pm$ 0.13
ReLU6	87.86 $\pm$ 0.14	90.59 $\pm$ 0.14
ELU	87.87 $\pm$ 0.15	90.74 $\pm$ 0.12
SoftPlus	87.49 $\pm$ 0.19	90.62 $\pm$ 0.16
Swish	89.60 $\pm$ 0.12	91.43 $\pm$ 0.11
Mish	89.29 $\pm$ 0.13	91.66 $\pm$ 0.15
GELU	89.09 $\pm$ 0.13	91.40 $\pm$ 0.12
PAU	89.87 $\pm$ 0.13	91.74 $\pm$ 0.12
AMSU-1	**91.22** $\pm$ 0.11	**92.31** $\pm$ 0.10
AMSU-2	**91.02** $\pm$ 0.10	**92.39** $\pm$ 0.12
AMSU-3	**90.56** $\pm$ 0.14	**92.02** $\pm$ 0.09

References

[1] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*,

4(1):1–13, 2017. 5

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 5

[3] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 5

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016. 2

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 6

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *Proc. of int. symposium on multimedia*, pages 225–2255, 2019. 6

[7] Debesh Jha, Nikhil Kumar Tomar, Sharib Ali, Michael A Riegler, Håvard D Johansen, Dag Johansen, Thomas de Lange, and Pål Halvorsen. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *Proceedings of the International Symposium on Computer-Based Medical Systems (CBMS)*, pages 37–43, 2021. 6

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2, 4, 5

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 2

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[11] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 4

[12] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. RadImageNet: RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022. 5

[13] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 5

[14] S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset, 2020. 3

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 6

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014. 2

[17] Jeya Maria Jose Valanarasu and Vishal M. Patel. Unext: Mlp-based rapid medical image segmentation network, 2022. 6

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network, 2017. 2

[20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016. 4