A The Difficulty of Directly Applying Batch Normalization to Mixed Operation

Before Node Normalization is adapted to j-th node x_j, the feature map of node is as follows:

$$x_j = \sum_{i<j} \bar{o}^{(i,j)}(\hat{x}_i) = \sum_{i<j} \sum_{o \in O} \beta_o^{(i,j)} o(\hat{x}_i). \tag{A1}$$

Simply introducing Batch Normalization to the above expression yields:

$$x_j = \sum_{i<j} BN(\bar{o}^{(i,j)}(\hat{x}_i)) = \sum_{i<j} \frac{\bar{o}^{(i,j)}(\hat{x}_i) - \mu_{\bar{o}^{(i,j)}(\hat{x}_i)}}{\sigma_{\bar{o}^{(i,j)}(\hat{x}_i)}} = \sum_{i<j} \sum_{o \in O} \frac{\beta_o^{(i,j)} o(\hat{x}_i)}{\sigma_{\bar{o}^{(i,j)}(\hat{x}_i)}}, \tag{A2}$$

where BN means Batch Normalization and $\mu_{\bar{o}^{(i,j)}(x_i)}$ is negligible because all operations are normalized. According to Eq. (A2), introducing batch normalization to mixed operation rescales $\beta o(x)$. Therefore, comparing the values of *beta* once again leads to the lack of reliability of β problem because the variances of edges are almost always different. To avoid this problem, we use the proposed $\sqrt{\beta}$-continuous relaxation instead of batch normalization.

B Empirical results of $\sqrt{\beta}$-Continuous Relaxation

As shown in Figure A1, the ratio of the variance of the output of the mixed operation to that of the input of the mixed operation is close to 1 after 10 epochs when using node normalization and $\sqrt{\beta}$-Continuous Relaxation, which implies C in Eq.(6) remains relatively unchanged. In DARTS, this ratio monotonically decreases, and in DARTS-L2, this ratio is visibly unstable; but in VS-DARTS, we observe that the variances of the input and the output feature stay close to 1, making the optimization landscape smoother.

C Settings for CIFAR-10

Settings for Search: Our search algorithm is evaluated under the cell-based micro search space [12], where the cell consists of two input nodes from two previous cells and four intermediate nodes, and its output is a concatenation of the intermediate nodes. For a fair comparison, following usual settings of DARTS [12], the super-net consists of eight cells, *i.e.*, six normal cells and two reduction cells, and its first stem layer is based on 3x3 convolutional layer with 16 initial channels. The search process is executed for 50 epochs, with a batch size of 64. Weights of operations in the super-net are optimized by momentum SGD, with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of 5E-3, where the learning rate is annealed down to zero according to a cosine schedule. Following Zela *et al.* [21], we use a larger weight decay factor than default DARTS to account for the serious performance drop due to the increase in the number of skip connections. Architecture parameters α are optimized by the Adam optimizer [9], with a fixed learning rate of 0.0003, a momentum of (0.5, 0.999), and a weight decay of 1E-3.

Settings for Evaluation: For the evaluation of the searched architectures, we follow DARTS settings [12] to train the final architectures on CIFAR-10, where 50,000 training images and 10,000 test images are included. The network with 36 initial channels consists of 20 cells, *i.e.*, 18 normal cells and two reduction cells; each type of the cells share the

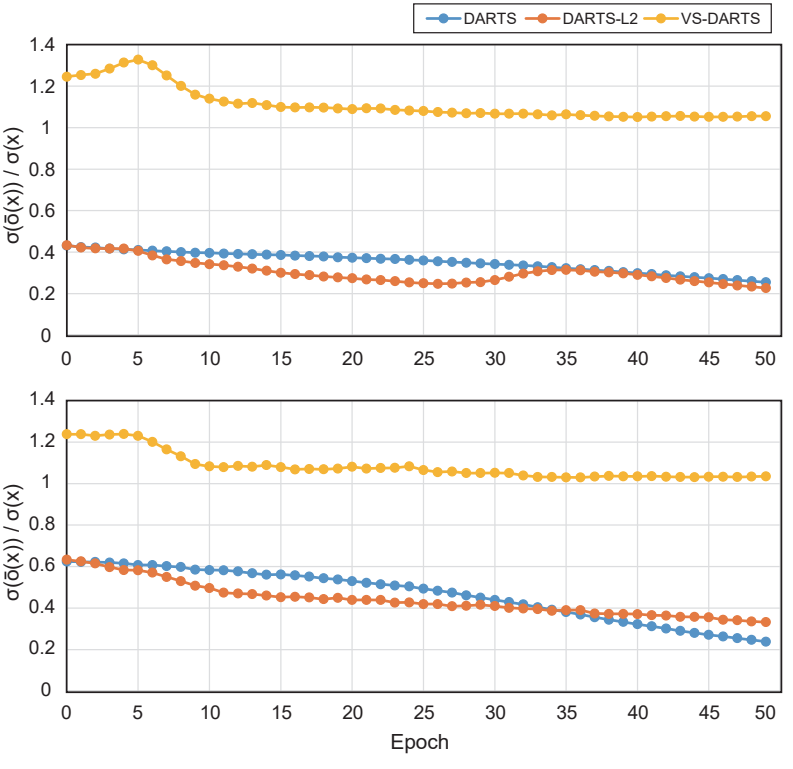

Figure A1: The ratio of the variance of the output of the mixed operation to that of the input of the mixed operation in the other 2 randomly selected edges related with Figure **??**

cell architectures obtained in the search process. In the evaluation, the network including an auxiliary loss is trained for 600 epochs, with a batch size of 96. The momentum SGD optimizer is used, with an initial learning rate of 0.025 following cosine scheduled annealing, a momentum of 0.9, a weight decay of 3E-4, and a norm of gradient clipping at 5. With aforementioned data augmentation techniques, cutout[5] is additionally used, and drop-path with a rate of 0.2 is used for regularization.

D Experimental Settings for Retraining (ImageNet)

We evaluate two architectures on ImageNet, where 1.28M training images and 5K test images are included: the best architecture searched from CIFAR-10 and the architecture directly searched on ImageNet using pxoxy dataset proposed by Na *et al*. [13]. We follow PC-DARTS [18] settings, while using four V100 GPUs for ImageNet training with batch size of 1536. The network consists of 14 cells: 12 normal cells and two reduction cells. Before stacking the cells, the network with 48 initial channels is equipped with stem layers which are composed of three 3x3 convolution layers of stride 2 to reduce 224×224 input images to 28×28. In the training process, the network is trained from scratch for 250 epochs, and the warm-up process is applied in the first 5 epochs. The momentum SGD optimizer is used, with an initial learning rate of 0.5 which is decayed down to zero linearly, a momentum of 0.9, a weight decay of 3E-5, and a norm gradient clipping at 5. Label smoothing with a rate of 0.1 and an auxiliary loss are used to enhance the training.

E Derived Architectures on CIFAR-10(S0)

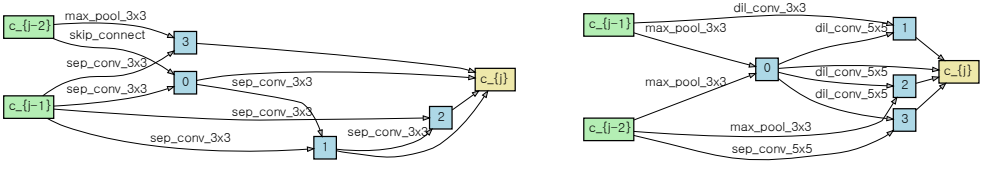

Figure A2: Normal(left) and reduction(right) cells found by VS-DARTS in CIFAR-10 and DARTS search space. Validation results are shown in Table 1(best) and Table 2.

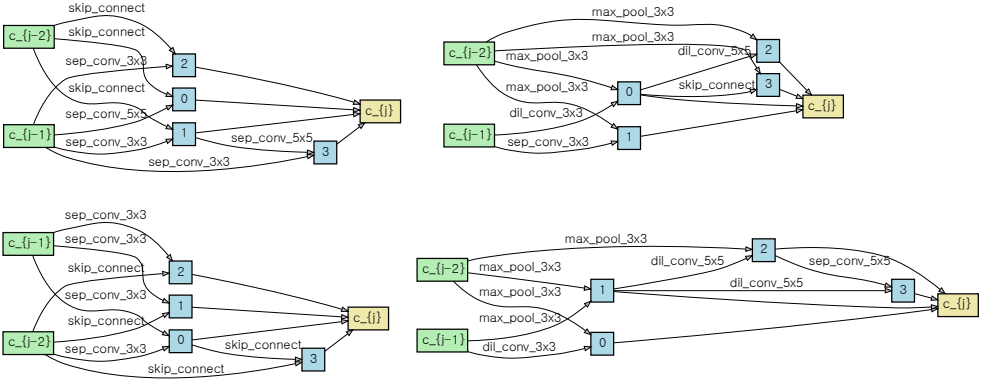

Figure A3: Two additional normal(left) and reduction(right) cells found by VS-DARTS in CIFAR-10 and DARTS search space. These architectures are used for VS-DARTS results(avg) in Table 1.