Supplementary Material:
Architecture Disentanglement for Deep Neural Networks

A. Formula Derivations

A.1. Derivation of Eq. 3

According to the definition of mutual information, we expand the first term of Eq. 2 under the joint distribution of input x^c and representation r^c_{n-1}:

$$\begin{aligned}\mathcal{I}(x^c; r^c_{n-1}) &= \int\int P(r^c_{n-1}, x^c) \log \frac{P(r^c_{n-1}|x^c)}{P(r^c_{n-1})} dx^c dr^c_{n-1} \\ &= \int\int P(r^c_{n-1}, x^c) \log P(r^c_{n-1}|x^c) dx^c dr^c_{n-1} - \int\int P(x^c|r^c_{n-1}) P(r^c_{n-1}) \log P(r^c_{n-1}) dx^c dr^c_{n-1} \\ &= \int\int P(r^c_{n-1}, x^c) \log P(r^c_{n-1}|x^c) dx^c dr^c_{n-1} - \int P(r^c_{n-1}) \log P(r^c_{n-1}) dr^c_{n-1}.\end{aligned} \tag{1}$$

Let $Q(r^c_{n-1})$ be a variational approximation of $P(r^c_{n-1})$, we have:

$$KL\big[P(r^c_{n-1})||Q(r^c_{n-1})\big] \geq 0 \Rightarrow \int P(r^c_{n-1}) \log P(r^c_{n-1}) dr^c_{n-1} \geq \int P(r^c_{n-1}) \log Q(r^c_{n-1}) dr^c_{n-1}. \tag{2}$$

Therefore, the trackable upper bound after applying the variational approximation is:

$$\begin{aligned}\mathcal{I}(r^c_{n-1}; x^c) &\leq \int\int P(r^c_{n-1}|x^c) P(x^c) \log \frac{P(r^c_{n-1}|x^c)}{Q(x^c)} dx^c dr^c_{n-1} \\ &= \mathbb{E}_{x^c \sim P(x^c)} \Big[KL\big[P(r^c_{n-1}|x^c)||Q(r^c_{n-1})\big] \Big].\end{aligned} \tag{3}$$

For the second term of Eq. 2, we expand it as the joint distribution of representation r^c_{n-1} and the target y^c:

$$\begin{aligned}\mathcal{I}(r^c_{n-1}; y^c) &= \int\int P(r^c_{n-1}, y^c) \log \frac{P(y^c|r^c_{n-1})}{P(y^c)} dr^c_{n-1} dy^c \\ &= \int\int P(r^c_{n-1}, y^c) \log P(y^c|r^c_{n-1}) dy^c dr^c_{n-1} - \int P(y^c) \log P(y^c) dy^c \\ &= \int\int P(r^c_{n-1}, y^c) \log P(y^c|r^c_{n-1}) dy^c dr^c_{n-1} + \mathcal{H}(y^c) \\ &\geq \int\int P(y^c|r^c_{n-1}) P(r^c_{n-1}) \log P(y^c|r^c_{n-1}) dy^c dr^c_{n-1},\end{aligned} \tag{4}$$

where $\mathcal{H}(y^c) \geq 0$ is the information entropy of y^c. Let $Q(y^c|r^c_{n-1})$ be a variational approximation of $P(y^c|r^c_{n-1})$, we have:

$$KL\big[P(y^c|r^c_{n-1})||Q(y^c|r^c_{n-1})\big] \geq 0 \Rightarrow \int P(y^c|r^c_{n-1}) \log P(y^c|r^c_{n-1}) dy^c \geq \int P(y^c|r^c_{n-1}) \log Q(y^c|r^c_{n-1}) dy^c. \tag{5}$$

By applying the variational approximation, the trackable lower bound of the mutual information between r^c_{n-1} and y^c is:

$$\mathcal{I}(r^c_{n-1}; y^c) \geq \int\int P(r^c_{n-1}, y^c) \log Q(y^c|r^c_{n-1}) dy^c dr^c_{n-1}. \tag{6}$$

Assuming that the representation r_{n-1}^c is independent of the label y^c, *i.e.*, $P(r_{n-1}^c|x^c, y^c) = P(r_{n-1}^c|x^c)$, we have:

$$P(x^c, r_{n-1}^c, y^c) = P(r_{n-1}^c|x^c, y^c)P(y^c|x^c)P(x^c) = P(r_{n-1}^c|x^c)P(y^c|x^c)P(x^c). \tag{7}$$

Then, the joint distribution of r_{n-1}^c and y^c can be written as:

$$P(r_{n-1}^c, y^c) = \int P(x^c, r_{n-1}^c, y^c)dx^c = \int P(r_{n-1}^c|x^c)P(y^c|x^c)P(x^c)dx^c. \tag{8}$$

Combining Eq. 22 with Eq. 24, we get the lower bound:

$$\begin{aligned}\mathcal{I}(r_{n-1}^c; y^c) &\geq \int\int\int P(x^c)P(r_{n-1}^c|x^c)P(y^c|x^c)\log Q(y^c|r_{n-1}^c)dy^c dr_{n-1}^c dx^c \\ &= \mathbb{E}_{x^c\sim P(x^c)}\Big[\mathbb{E}_{r_{n-1}^c\sim P(r_{n-1}^c|x^c)}\big[\int P(y^c|x^c)\log Q(y^c|r_{n-1}^c)dy^c\big]\Big] \\ &= \mathbb{E}_{x^c\sim P(x^c)}\Big[\mathbb{E}_{r_{n-1}^c\sim P(r_{n-1}^c|x^c)}\big[\log Q(y^c|r_{n-1}^c)\big]\Big].\end{aligned} \tag{9}$$

With the Eq. 19 and Eq. 25, the variational upper bound of Eq. 2 is derived to Eq. 3 as:

$$\widetilde{\mathcal{L}}_{IB} = \mathbb{E}_{x^c\sim P(x^c)}\Big[\beta KL\big[P(r_{n-1}^c|x^c)||Q(r_{n-1}^c)\big] - \mathbb{E}_{r_{n-1}^c\sim P(r_{n-1}^c|x^c)}\big[\log Q(y^c|r_{n-1}^c)\big]\Big]. \tag{10}$$

The derivation of Eq. 11 is the same as the derivation of the second term of Eq. 3.

A.2. Derivation of Eq. 9 and Eq. 13

For the first term of Eq. 6, the KL divergence can be expanded with Eq. 8. Taking the univariate Gaussian distribution as example, the KL divergence is:

$$\begin{aligned}KL\Big[\mathcal{N}(0,1)||\mathcal{N}\big(r_i^c\cdot\mu_i^c, (r_i^c\cdot\sigma_i^c)^2\big)\Big] =& \int \frac{1}{\sqrt{2\pi}r_i^c\sigma_i^c}e^{-(x-r_i^c\mu_i^c)^2/2(r_i^c\sigma_i^c)^2}\big(\log\frac{e^{-(x-r_i^c\mu_i^c)^2/2(r_i^c\sigma_i^c)^2}}{r_i^c\sigma_i^c e^{-x^2/2}}\big)dx \\ =& \frac{1}{2}\int \frac{1}{\sqrt{2\pi}r_i^c\sigma_i^c}e^{-(x-r_i^c\mu_i^c)^2/2(r_i^c\sigma_i^c)^2}\big(x^2 - \log(r_i^c\sigma_i^c)^2 - (x-r_i^c\mu_i^c)^2/(r_i^c\sigma_i^c)^2\big)dx \\ =& \frac{1}{2}\big((r_i^c\cdot\sigma_i^c)^2 - \log(r_i^c\cdot\sigma_i^c)^2 + (r_i^c\cdot\mu_i^c)^2 - 1\big).\end{aligned} \tag{11}$$

For the second term of Eq. 6, the log-likelihood function of the univariate Gaussian distribution is:

$$\begin{aligned}-\log Q(\widetilde{r}_i^c|r_i^c) =& -\log\frac{1}{\sqrt{2\pi}r_i^c\sigma_i^c}e^{-(r_i^c - r_i^c\mu_i^c)^2/2(r_i^c\sigma_i^c)^2} \\ =& \frac{1}{2}\big(||r_i^c - \mu_i^c\cdot r_i^c||_2^2 + \log 2\pi + \log(r_i^c\cdot\sigma_i^c)^2\big).\end{aligned} \tag{12}$$

Combining Eq. 27 and Eq. 28, the constraint in Eq. 6 for the i-th hidden layer can be derived to Eq. 9 as:

$$\begin{aligned}\widetilde{\mathcal{L}}_i^c =& \frac{\beta}{2}\big((r_i^c\cdot\sigma_i^c)^2 - \log(r_i^c\cdot\sigma_i^c)^2 + (r_i^c\cdot\mu_i^c)^2 - 1\big) \\ &+ \frac{1}{2}\big(||r_i^c - \mu_i^c\cdot r_i^c||_2^2 + \log 2\pi + \log(r_i^c\cdot\sigma_i^c)^2\big).\end{aligned} \tag{13}$$

After reducing the noisy level by fixing ϵ_i in Eq. 7 to its mean value, *i.e.*, 0, σ_i^c is freed to be any value that does not affect the optimization process. Therefore, Eq. 29 can be simplified to Eq. 13 as:

$$\widetilde{\mathcal{L}}_i^c = \beta(r_i^c\cdot\mu_i^c)^2 + ||r_i^c - \mu_i^c\cdot r_i^c||_2^2. \tag{14}$$

A.3. Derivation of Eq. 12

For Eq. 11, the log-likelihood function of the univariate Bernoulli distribution can be directly written as:

$$\begin{aligned} -\log Q(y^c|\widetilde{y}^c) &= -\log \left(f_n(r^c_{n-1})\right)^{y^c} \left(1-f_n(r^c_{n-1})\right)^{1-y^c} \\ &= -y^c \log f_n(r^c_{n-1}) - (1-y^c)\log\left(1-f_n(r^c_{n-1})\right) \\ &= -y^c \log \widetilde{y}^c - (1-y^c)\log(1-\widetilde{y}^c). \end{aligned} \tag{15}$$

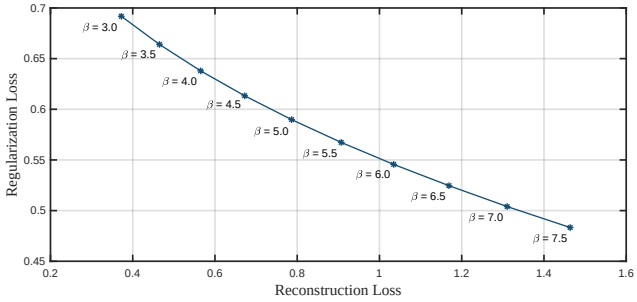
B. Hyper-Parameter β

We use the first 5% labels for determining the hyper-parameter β according the Elbow method, with which the reconstruction loss and regularization loss are balanced. For VGG16 in Figs. 1a and 1e, the hyper-parameter $\beta = 4.5$ and $\beta = 5.0$ balance the losses well on ImageNet and Place365, respectively. For ResNet50 in Figs. 1b and 1f, the hyper-parameter $\beta = 0.02$ and $\beta = 0.02$ balance the losses well on ImageNet and Place365, respectively. For DenseNet121 in Figs. 1c and 1g, the hyper-parameter $\beta = 0.02$ and $\beta = 0.02$ balance the losses well on ImageNet and Place365, respectively. For DARTS-Net in Figs. 1d and 1h, the hyper-parameter $\beta = 0.45$ and $\beta = 0.20$ balance the losses well on ImageNet and Place365, respectively.

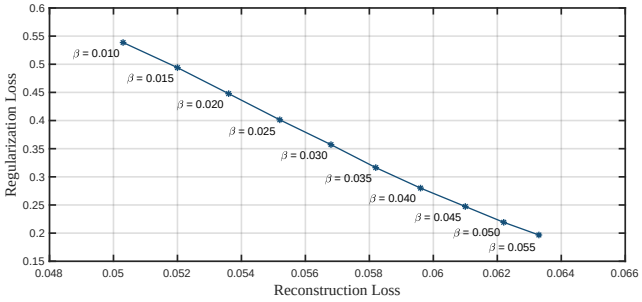
C. More Visualization Examples

More results of the visualization with the activated feature maps are shown from Fig. 7 to Fig. 16.

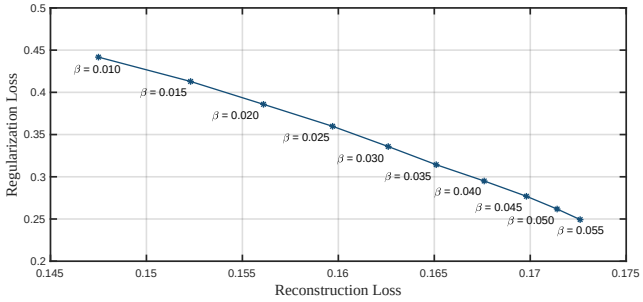
(a) Results of VGG16 on ImageNet



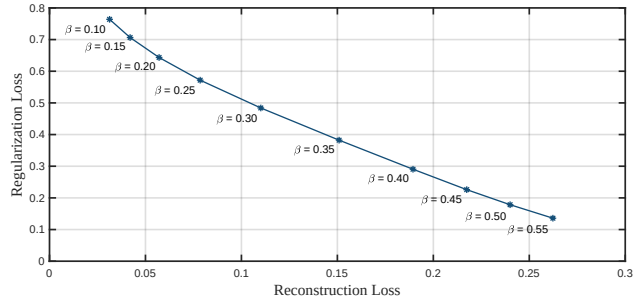
(b) Results of ResNet50 on ImageNet



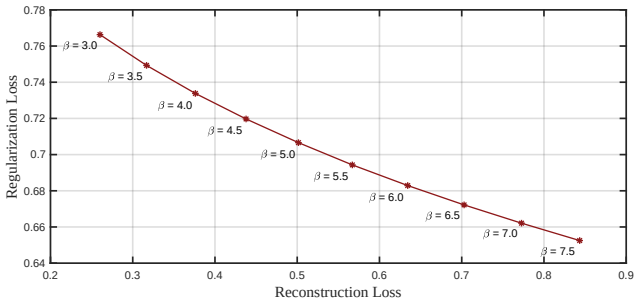
(c) Results of DenseNet121 on ImageNet



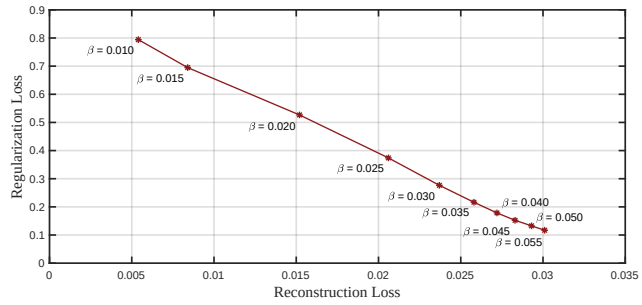
(d) Results of DARTS-Net on ImageNet



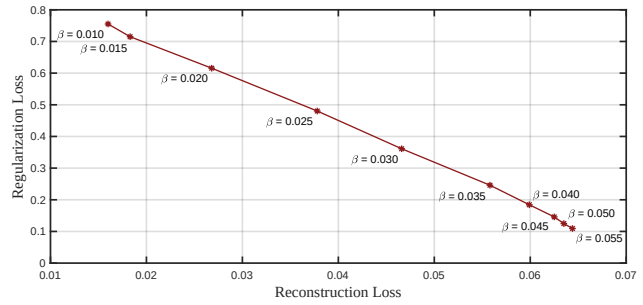
(e) Results of VGG16 on Place365



(f) Results of ResNet50 on Place365



(g) Results of DenseNet121 on Place365



(h) Results of DARTS-Net on Place365

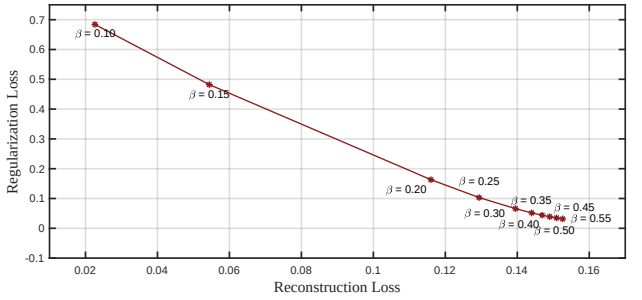


Figure 1: Reconstruction loss vs. regularization loss with different hyper-parameter β.

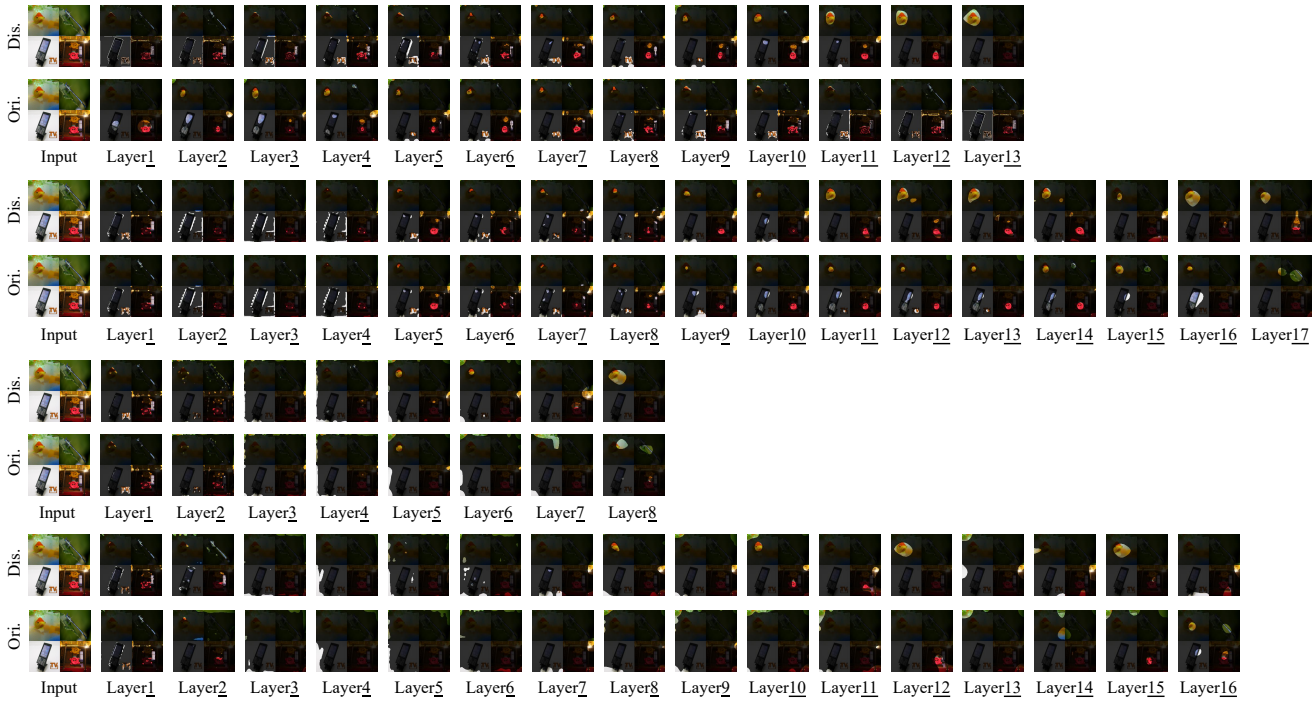


Figure 2: Example combining the images with 'label-concept': '1-Goldfish', '320-Damselfly', '487-MobilePhone', and '489-ChainLinkFence' from the validation set of ImageNet. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

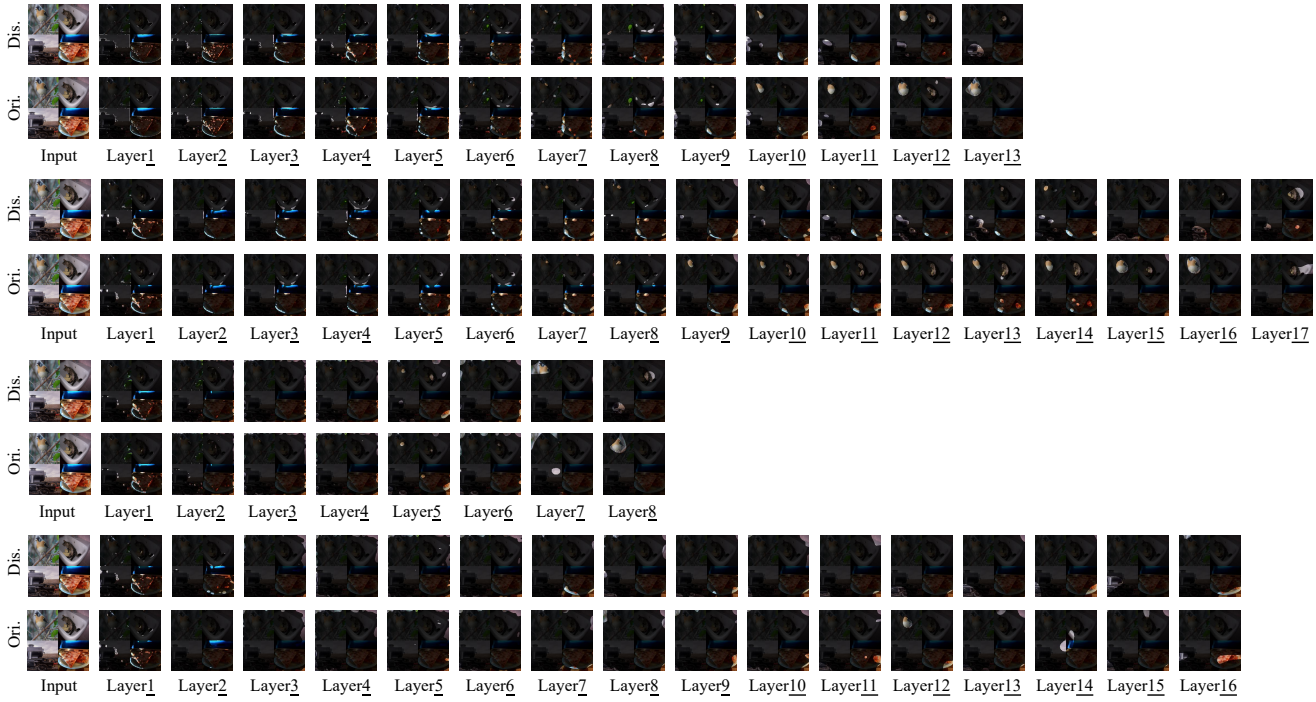


Figure 3: Example combining the images with 'label-concept': '10-Goldfinch', '896-Washer', '820-SteelArchBridge', and '963-Potpie' from the validation set of ImageNet. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

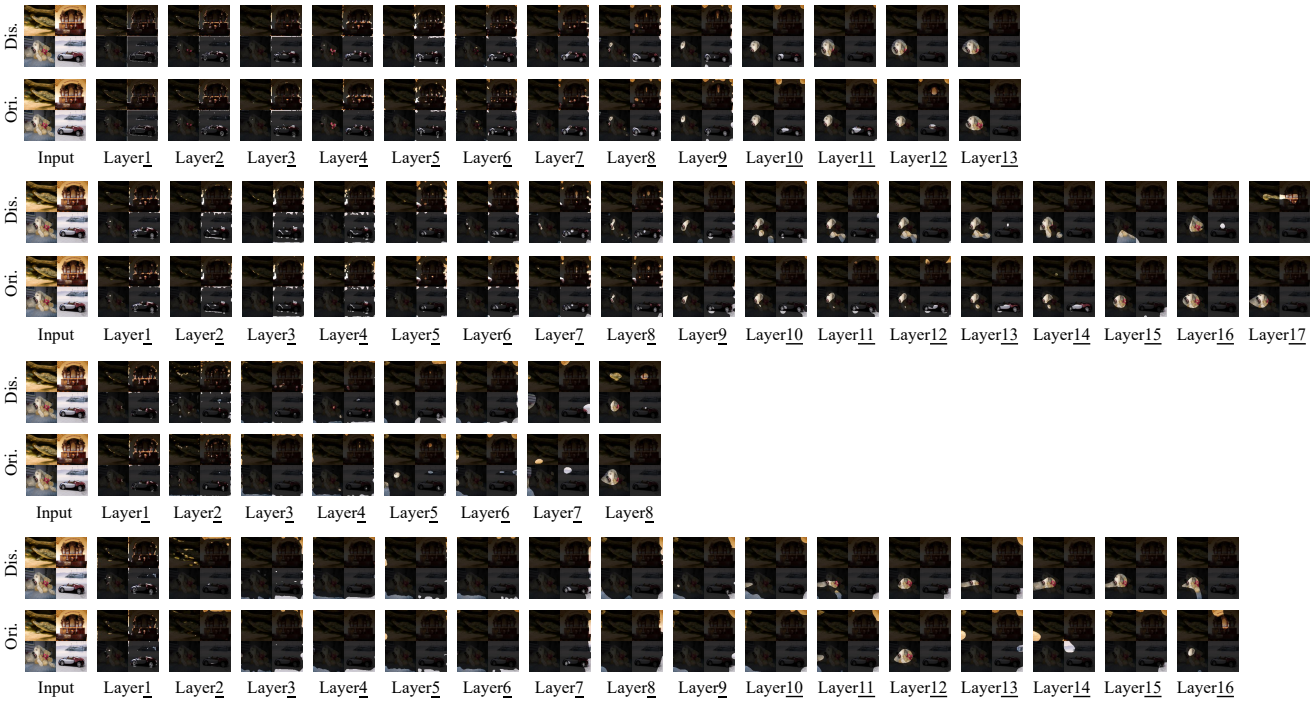


Figure 4: Example combining the images with 'label-concept': '35-Terrapin', '687-Oscilloscope', '257-Samoyed', and '511-Corkscrew' from the validation set of ImageNet. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

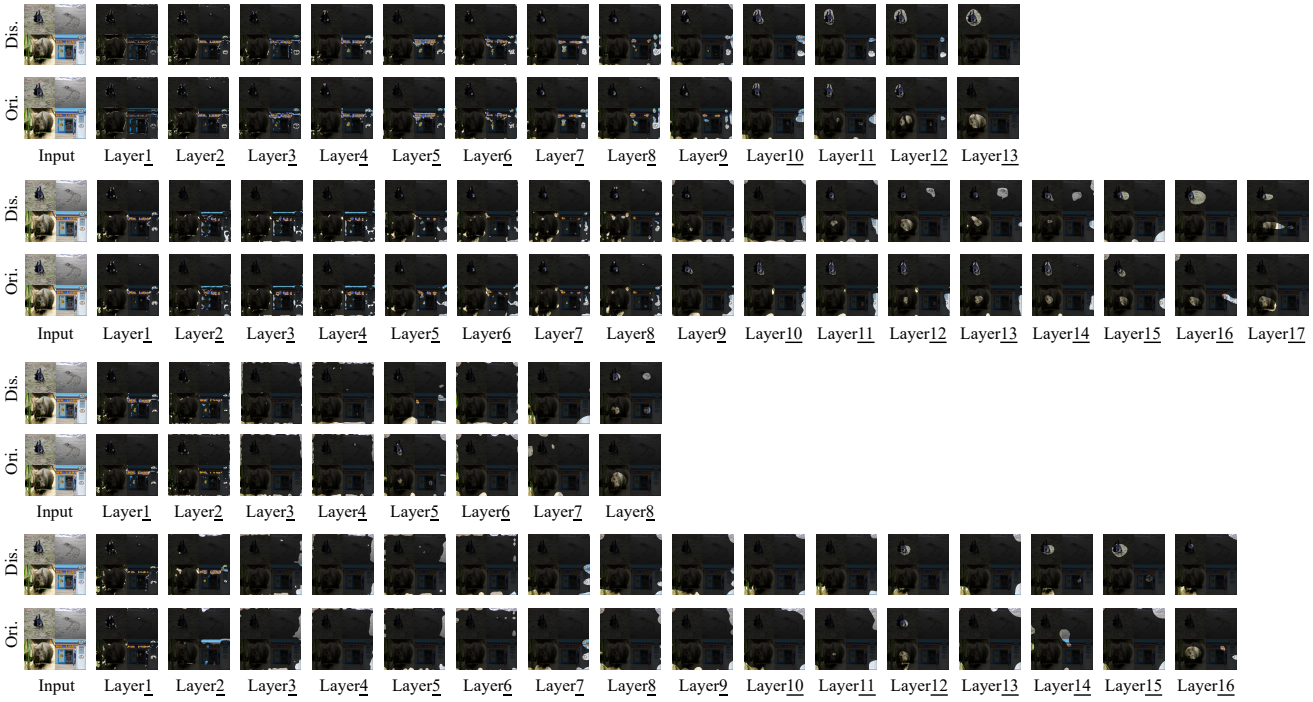


Figure 5: Example combining the images with 'label-concept': '80-Ptarmigan', '145-Albatross', '106-jellyfish', and '571-Goblet' from the validation set of ImageNet. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

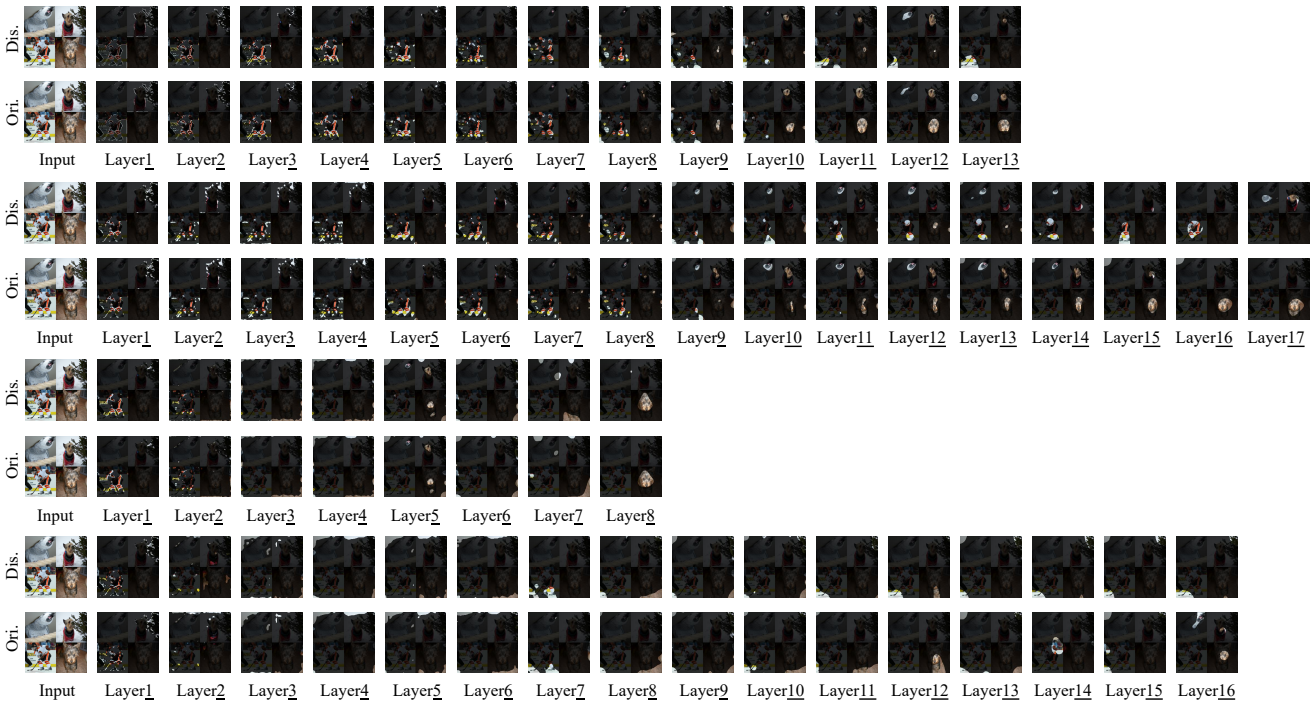


Figure 6: Example combining the images with 'label-concept': '87-Macaw', '193-DandieDinmont', '746-PunchingBag', and '187-WireHairedFoxTerrier' from the validation set of ImageNet. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

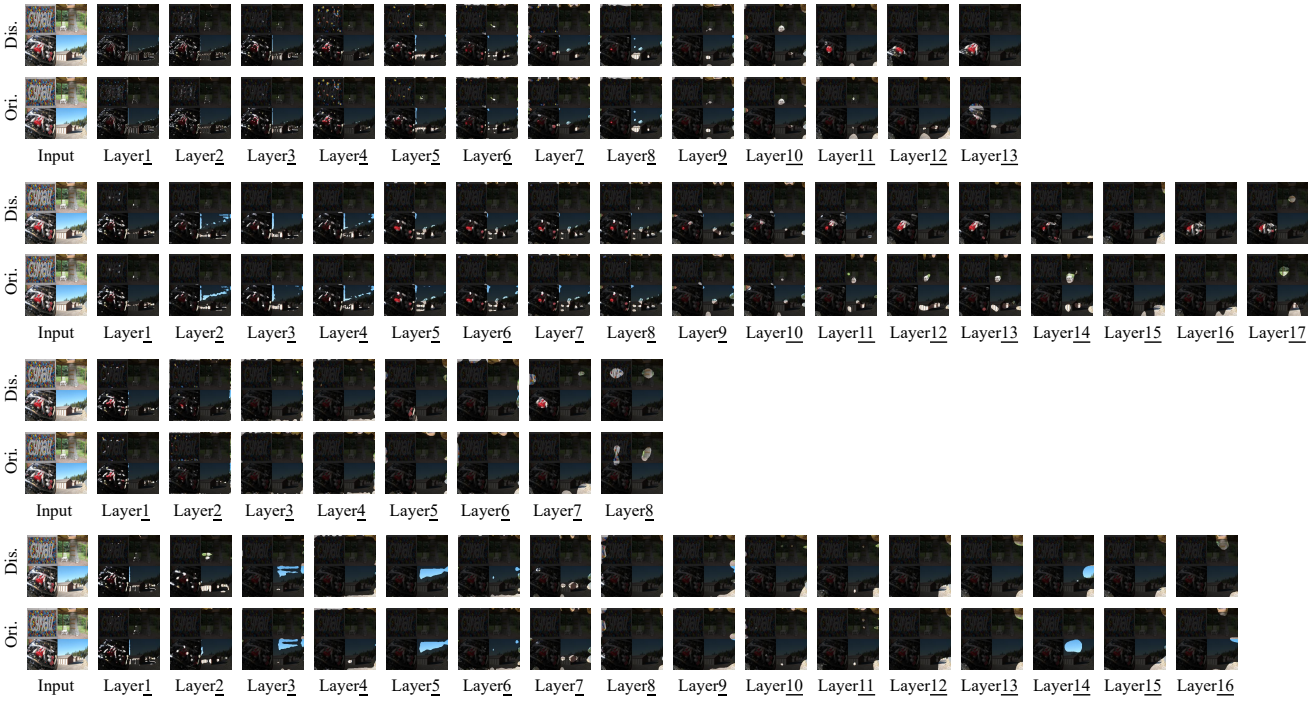


Figure 7: Example combining the images with 'label-concept': '34-BallPit', '339-TreeHouse', '28-AutoFactory', and '221-ManufacturedHome' from the validation set of Place365. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

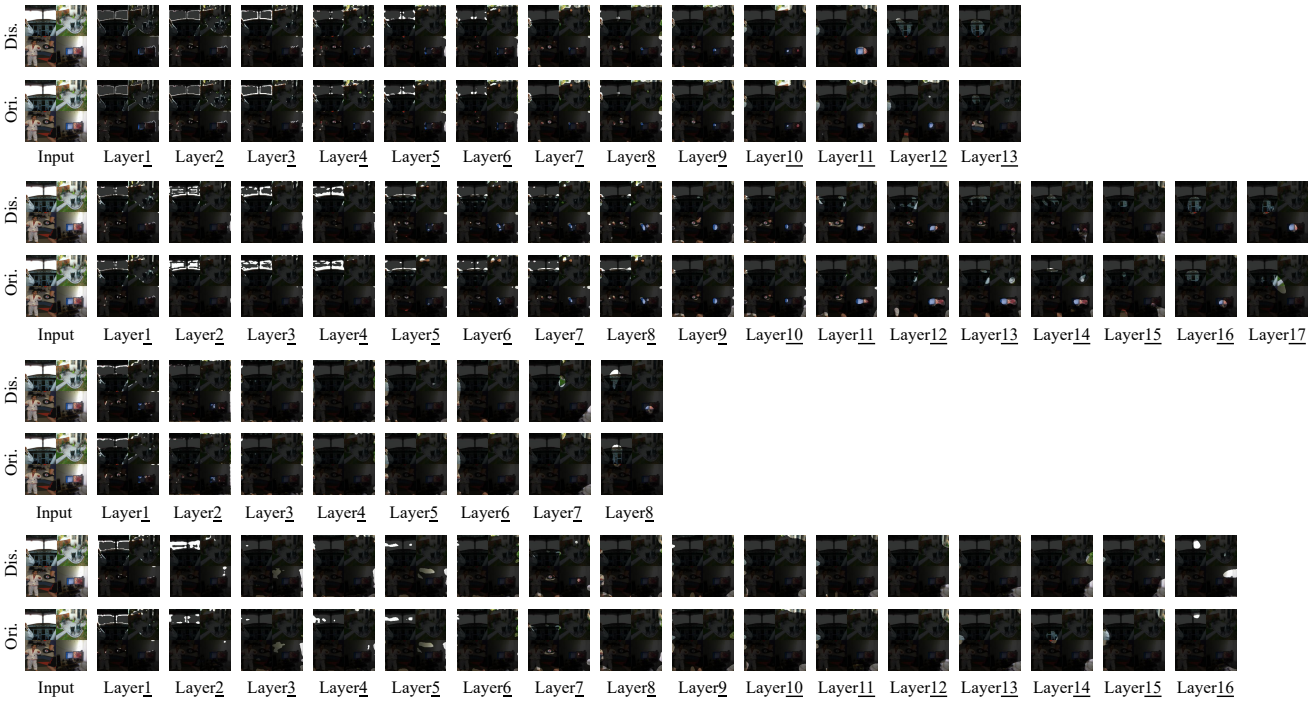


Figure 8: Example combining the images with 'label-concept': '98-Cockpit', '259-Patio', '225-MartialArtsGym', and '100-ComputerRoom' from the validation set of Place365. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

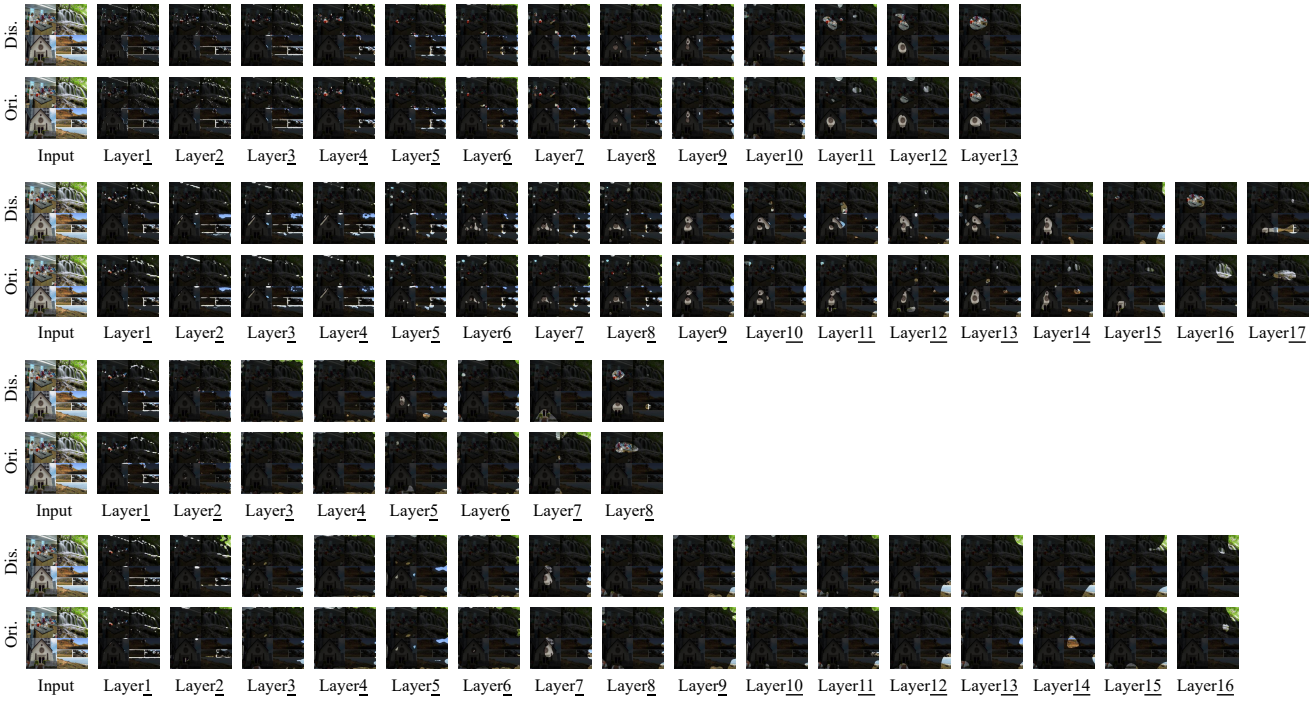


Figure 9: Example combining the images with 'label-concept': '56-BiologyLaboratory', '355-Waterfall', '327-Synagogue', and '233-MountainPath' from the validation set of Place365. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

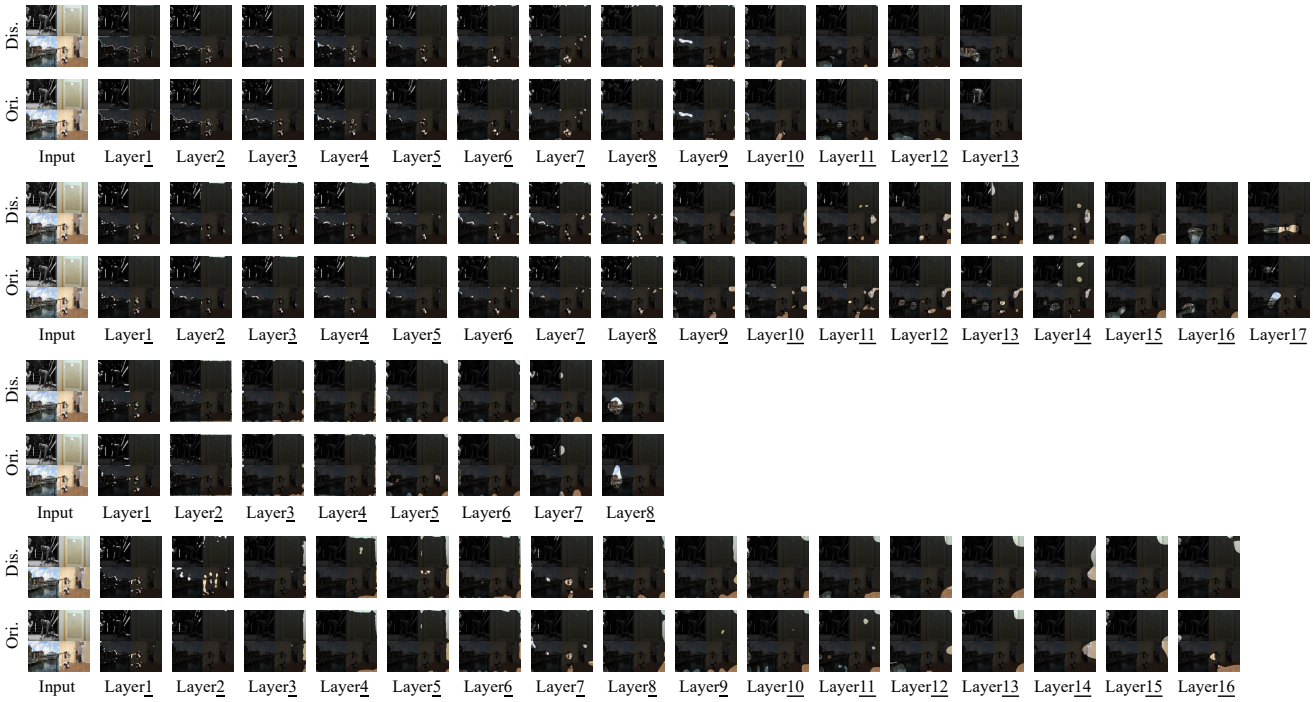


Figure 10: Example combining the images with 'label-concept': '65-BoxingRing', '19-ArtGallery', '57-Boardwalk', and '168-Gymnasium' from the validation set of Place365. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.

Figure 11: Example combining the images with 'label-concept': '70-BusInterior', '8-ApartmentBuilding', '345-VegetableGarden', and '206-Landfill' from the validation set of Place365. The results from top to bottom are from VGG16, ResNet50, DenseNet121, and DARTS-Net, respectively.