A. Proof of Eq 7.

$$
\begin{aligned}
Z_k^{(l+1)} &= \sum_{i=1}^{N_l} X_i^{(l)} \circledast W_{k,i}^{(l+1)} \\
&= \sum_{i=1,i\neq j}^{N_l} X_i^{(l)} \circledast W_{k,i}^{(l+1)} + X_j^{(l)} \circledast W_{k,j}^{(l+1)} \\
&\approx \sum_{i=1,i\neq j}^{N_l} B(Z_i^{(l)}) \circledast W_{k,i}^{(l+1)} + \\
&\quad B(\sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times Z_i^{(l)}) \circledast W_{k,j}^{(l+1)} \\
&= \sum_{i=1,i\neq j}^{N_l} B(Z_i^{(l)}) \circledast (W_{k,i}^{(l+1)} + \hat{s}_i \times W_{k,j}^{(l+1)}),
\end{aligned} \tag{1}
$$

B. Proof of Eq 10.

$$
\begin{aligned}
e_p &= {Z_k}^{(l+1)} - \hat{Z}_k^{(l+1)} \\
&= (X_j^{(l)} - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times X_i^{(l)}) \circledast W_{k,j}^{(l+1)} \\
&= (B(Z_j^{(l)}) - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)})) \circledast W_{k,j}^{(l+1)} \\
&= \{\{\frac{\gamma_j(Z_j^{(l)} - \mu_j)}{\sigma_j} + \beta_j\} - \\
&\quad \sum_{i=1,i\neq j}^{N_l} \hat{s}_i\{\frac{\gamma_i(Z_i^{(l)} - \mu_i)}{\sigma_i} + \beta_i\}\} \circledast W_{k,j}^{(l+1)} \\
&= \{\frac{\gamma_j}{\sigma_j}(X^{(l-1)} \circledast W_j^{(l)}) - \\
&\quad \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \frac{\gamma_i}{\sigma_i}(X^{(l-1)} \circledast W_i^{(l)}) + \beta_j - \frac{\gamma_j\mu_j}{\sigma_j} + \\
&\quad \sum_{i=1,i\neq j}^{N_l} \hat{s}_i(\frac{\gamma_i\mu_i}{\sigma_i} - \beta_i)\} \circledast W_{k,j}^{(l+1)} \\
&= \{\frac{\gamma_j}{\sigma_j}\{X^{(l-1)} \circledast (W_j^{(l)} - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \frac{\gamma_i\sigma_j}{\sigma_i\gamma_j} W_i^{(l)})\} + \\
&\quad (\beta_j - \frac{\gamma_j\mu_j}{\sigma_j}) - (\sum_{i=1,i\neq j}^{N_l} \hat{s}_i(\beta_i - \frac{\gamma_i\mu_i}{\sigma_i}))\} \circledast W_{k,j}^{(l+1)}
\end{aligned} \tag{2}
$$

C. Proof of Eq 11.

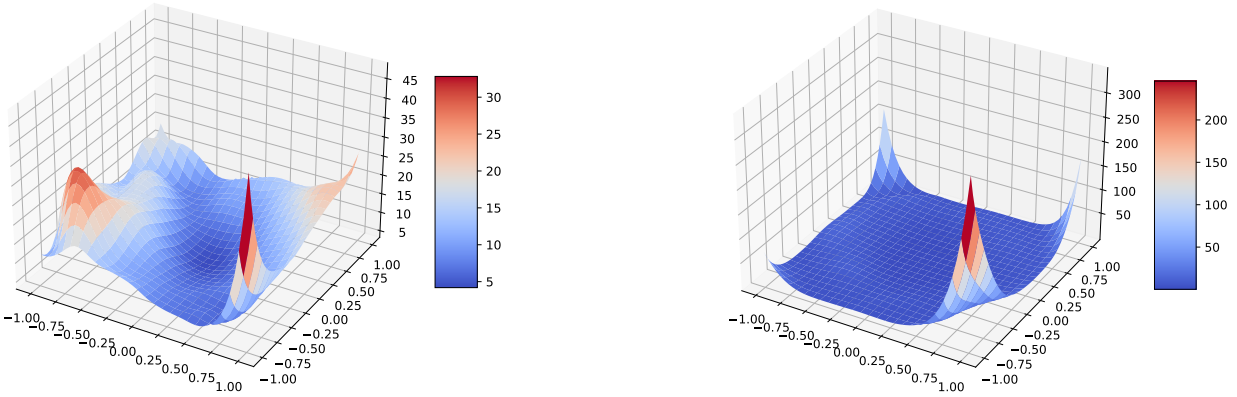
$$
\begin{aligned}
e_p &= \Theta(B(Z_j^{(l)})) - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times \Theta(B(Z_i^{(l)})) \\
&= Max(B(Z_j^{(l)}), 0) - Max(\sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)}), 0) \\
&= \frac{1}{2}(B(Z_j^{(l)}) + |B(Z_j^{(l)}| - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)}) - \\
&\quad |\sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)})|) \\
&\leqslant \frac{1}{2}(B(Z_j^{(l)}) - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)}) + \\
&\quad |B(Z_j^{(l)}) - \sum_{i=1,i\neq j}^{N_l} \hat{s}_i \times B(Z_i^{(l)})|) \\
&= \frac{1}{2}(A + |A|),
\end{aligned} \tag{3}
$$

D. Proof of Eq 13.

$$
\begin{aligned}
e_q &= {Z_k}^{(l+1)} - \tilde{Z}_k^{(l+1)} \\
&= (X_m^{(l)} - \tilde{s}_m \tilde{X}_m^{(l)}) \circledast W_{k,j}^{(l+1)} \\
&= (B(Z_m^{(l)}) - \tilde{s}_m B(\tilde{Z}_m^{(l)})) \circledast W_{k,m}^{(l+1)} \\
&= \{\{\frac{\gamma_m(Z_m^{(l)} - \mu_m)}{\sigma_m} + \beta_m\} - \\
&\quad \tilde{s}_m\{\frac{\gamma_m(\tilde{Z}_m^{(l)} - \mu_m)}{\sigma_m} + \beta_m\}\} \circledast W_{k,m}^{(l+1)} \\
&= (\frac{\gamma_m Z_m^{(l)}}{\sigma_m} - \tilde{s}_m \frac{\gamma_m \tilde{Z}_m^{(l)}}{\sigma_m} + \tilde{s}_m \frac{\gamma_m\mu_m}{\sigma_m} - \frac{\gamma_m\mu_m}{\sigma_m} \\
&\quad + \beta_m - \tilde{s}_m\beta_m) \circledast W_{k,m}^{(l+1)} \\
&= \{(\frac{\gamma_m W_m^{(l)}}{\sigma_m} - \tilde{s}_m \frac{\gamma_m \tilde{W}_m^{(l)}}{\sigma_m}) \circledast X^{(l-1)} + \tilde{s}_m \frac{\gamma_m\mu_m}{\sigma_m} \\
&\quad - \frac{\gamma_m\mu_m}{\sigma_m} + \beta_m - \tilde{s}_m\beta_m\} \circledast W_{k,m}^{(l+1)}
\end{aligned} \tag{4}
$$

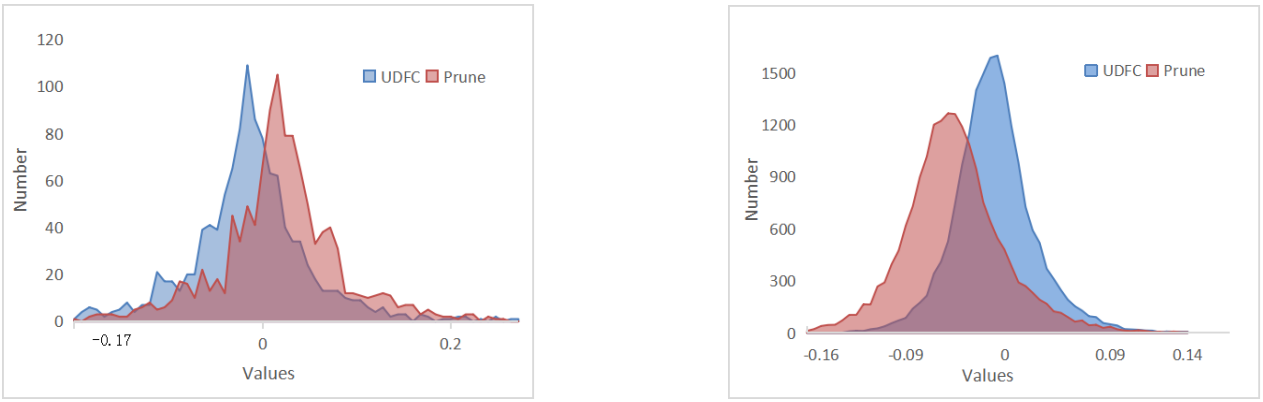
E. Effectiveness of UDFC.

To better understand the behavior of UDFC in the CNN architecture, we visualize the loss landscape and weights offset of ResNet-56 on CIFAR-10.



(a) Prune. (b) UDFC.

Figure 1: The loss landscape of ResNet-50 before and after compensation.



(a) ResNet-56.3.2. (b) ResNet-56.2.2.

Figure 2: The loss weights offset of Prune and UDFC on ResNet-56.

Figure 1a shows the landscape of Prune and Figure 1b shows the landscape of UDFC. It is clear that the Prune method has a coarser plane of loss landscape than our method. This is due to the fact that pruning without compensation destroys the integrity of the original model. The loss landscape of UDFC becomes smoothed out, proving the effectiveness of our method. As shown in Figure 2, the compressed model produces noisy and divergent weights distribution. In contrast, our method reverts the weight average to zero, which is closer to original model.