A Appendix

A.1 Theoretical Proofs

Notations of Convolutional Operations. In our paper, we express convolution operation as $\mathbf{Z}_u = \boldsymbol{\alpha}\mathbf{X}_p\mathbf{D}_u$. More explicitly, the formulation writes as $\mathbf{Z}_u = \boldsymbol{\alpha}\mathbf{X}_p \star \mathbf{D}_u$, where $\star$ is the convolutional operation. By converting convolutional kernel $\mathbf{D}_u$ into a Toeplitz matrix, we can replace the convolution operation $\mathbf{X}_p \star \mathbf{D}_u$ with matrix multiplication $\mathbf{X}_p\mathbf{D}_u$. We also modify α by $I_{hw} \bigotimes \alpha$, where $\bigotimes$ is Kronecker product, to enable the matrix multiplication $\boldsymbol{\alpha}\mathbf{X}_p\mathbf{D}_u$.

Proposition A.1. *Suppose $\mathbf{D}_u$ and $\mathbf{D}_v$ are two different sets of filter atoms for a convolutional layer with the common atom coefficients $\boldsymbol{\alpha}$, we can upper bound the changes in the corresponding features $\mathbf{Z}_u, \mathbf{Z}_v$ with atom changes,*

$$||\mathbf{Z}_u - \mathbf{Z}_v||_F \leq (||\boldsymbol{\alpha}||_F \lambda)\sqrt{|\mathcal{B}|} \cdot ||(\mathbf{D}_u - \mathbf{D}_v)||_F, \quad \textit{with } \lambda = \sup_{b \in \mathcal{B}} ||\mathbf{X}||_{F,N_b}, \tag{8}$$

Proof. Recall the decomposed convolution can be expressed as,

$$\mathbf{Z} = \sum_{i=1}^{m} \boldsymbol{\alpha}_i \langle \mathbf{X}, \mathbf{D}[i] \rangle_{N_b} \tag{9}$$

$\forall b$ we have,

$$\begin{aligned} |\mathbf{Z}_u(b) - \mathbf{Z}_v(b)| &= |\sum_{i=1}^{m} \boldsymbol{\alpha}_i \langle \mathbf{X}, \mathbf{D}_u[i] \rangle_{N_b} - \sum_{i=1}^{m} \boldsymbol{\alpha}_i \langle \mathbf{X}, \mathbf{D}_v[i] \rangle_{N_b}| \\ &\leq ||\boldsymbol{\alpha}||_F (\sum_{i=1}^{m} |\langle \mathbf{X}, (\mathbf{D}_u[i] - \mathbf{D}_v[i]) \rangle_{N_b}|^2)^{1/2}. \end{aligned} \tag{10}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned} |\langle \mathbf{X}, (\mathbf{D}_u[i] - \mathbf{D}_v[i]) \rangle_{N_b}| &\leq ||\mathbf{X}||_{F,N_b} \cdot ||\mathbf{D}_u[i] - \mathbf{D}_v[i]||_{F,N_b} \\ &\leq \lambda \cdot ||\mathbf{D}_u[i] - \mathbf{D}_v[i]||_{F,N_b} \end{aligned} \tag{11}$$

we have that

$$\begin{aligned} \sum_{b \in \mathcal{B}} |\mathbf{Z}_u(b) - \mathbf{Z}_v(b)|^2 &\leq ||\boldsymbol{\alpha}||_F^2 \sum_{b} \sum_{i=1}^{m} |\langle \mathbf{X}, (\mathbf{D}_u[i] - \mathbf{D}_v[i]) \rangle_{N_b}|^2 \\ &\leq ||\boldsymbol{\alpha}||_F^2 \sum_{b} \sum_{i=1}^{m} ||\mathbf{X}||_{F,N_b}^2 \cdot ||(\mathbf{D}_u[i] - \mathbf{D}_v[i])||_{F,N_b}^2 \\ &\leq (||\boldsymbol{\alpha}||_F \lambda)^2 \sum_{b,i} ||(\mathbf{D}_u[i] - \mathbf{D}_v[i])||_{F,N_b}^2 \end{aligned} \tag{12}$$

and observe that

$$\sum_{b,i} ||(\mathbf{D}_u[i] - \mathbf{D}_v[i])||_{F,N_b}^2 = \sum_{b \in \mathcal{B}} \sum_{i=1}^{m} ||(\mathbf{D}_u[i] - \mathbf{D}_v[i])||_{F,N_b}^2 = |\mathcal{B}| \cdot ||(\mathbf{D}_u - \mathbf{D}_v)||_F^2, \tag{13}$$

where $|\mathcal{B}|$ is the area of the domain of $\mathbf{X}$. Then Eq. 12 becomes

$$\sum_{b \in \mathcal{B}} |\mathbf{Z}_u(b) - \mathbf{Z}_v(b)|^2 \leq (||\boldsymbol{\alpha}||_F \lambda)^2 |\mathcal{B}| \cdot ||(\mathbf{D}_u - \mathbf{D}_v)||_F^2, \tag{14}$$

which proves that $||\mathbf{Z}_u - \mathbf{Z}_v||_F \leq (||\boldsymbol{\alpha}||_F \lambda)\sqrt{|\mathcal{B}|} \cdot ||(\mathbf{D}_u - \mathbf{D}_v)||_F$ as claimed.

□

Proposition A.2. *Assume filter atoms $\mathbf{D}_u, \mathbf{D}_v$ are orthogonal matrices, then $\mathcal{S}_{Gras} = \mathcal{S}_{Atom}$.*

Proof. Since $\mathbf{D}_u, \mathbf{D}_v \in \mathbb{R}^{k^2 \times m}$ are orthogonal matrices, i.e., $\mathbf{D}_u^T \mathbf{D}_u = \mathbf{D}_v^T \mathbf{D}_v = I$, the Grassmann similarity can be represented as,

$$\mathcal{S}_{Gras}(\mathcal{F}_u, \mathcal{F}_v) = \frac{1}{m} \sum_i^m \mathbf{cos}\theta_i = \frac{1}{m} \sum_i^m \sigma_i, \tag{15}$$

where $\sigma_i = \Sigma_{ii}, U\Sigma V = \mathbf{D}_u^T \mathbf{D}_v$.

$\mathcal{S}_{Atom}$ is defined as,

$$\mathcal{S}_{Atom}(\mathcal{F}_u, \mathcal{F}_v) = \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v) = \frac{< vec(\mathbf{D}_u), vec(\mathbf{D}_v) >}{||vec(\mathbf{D}_u)||_F \cdot ||vec(\mathbf{D}_v)||_F}. \tag{16}$$

Analyze each part separately, we have $< vec(\mathbf{D}_u), vec(\mathbf{D}_v) >= \mathbf{Tr}(\mathbf{D}_u^T \mathbf{D}_v) = \sum_i^m \sigma_i$, $||vec(\mathbf{D}_u)||_F = \sqrt{\mathbf{Tr}(\mathbf{D}_u^T \mathbf{D}_u)} = \sqrt{\mathbf{Tr}(I)} = \sqrt{m}$, and also $||vec(\mathbf{D}_v)||_F = \sqrt{m}$. In total, the filter subspace similarity becomes,

$$\mathcal{S}_{Atom}(\mathcal{F}_u, \mathcal{F}_v) = \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v) = \frac{\sum_i^m \sigma_i}{m}, \tag{17}$$

which equals $\mathcal{S}_{Gras}$. The claimed theorem is proved.

$\square$

Lemma A.3. *For two positive semidefinite matrices* $\mathbf{A}, \mathbf{B}$*,*

$$\mathbf{Tr}(\mathbf{AB}) \geq \sigma_{min}(\mathbf{A})\mathbf{Tr}(\mathbf{B}), \tag{18}$$

where σ_{min} *denotes the minimum eigenvalue of* A*.*

Proof. It is equivalent to prove that,

$$\mathbf{Tr}((\mathbf{A} - \sigma_{min}(\mathbf{A})\mathbf{I})\mathbf{B}) \geq 0. \tag{19}$$

Let $\mathbf{C}, \mathbf{D}$ be matrices such that $\mathbf{A} - \sigma_{min}(\mathbf{A})\mathbf{I} = \mathbf{C}^\intercal\mathbf{C}$, $\mathbf{B} = \mathbf{D}^\intercal\mathbf{D}$, then

$$\begin{aligned} \mathbf{Tr}((\mathbf{A} - \sigma_{min}(\mathbf{A})\mathbf{I})\mathbf{B}) &= \mathbf{Tr}(\mathbf{C}^\intercal\mathbf{C}\mathbf{D}^\intercal\mathbf{D}) \\ &= \mathbf{Tr}(\mathbf{C}\mathbf{D}^\intercal\mathbf{D}\mathbf{C}^\intercal) \\ &= \mathbf{Tr}((\mathbf{D}\mathbf{C}^\intercal)^\intercal(\mathbf{D}\mathbf{C}^\intercal)) \geq 0. \end{aligned} \tag{20}$$

$\square$

Theorem A.4. *Suppose the forward of decomposed convolution layer for the* u*-th model is* $\mathbf{Z}_u = \alpha\mathbf{X}\mathbf{D}_u$*.* $\mathbf{Z}_u, \mathbf{Z}_v$ *nearly have zero-mean since* $\mathbf{X}_p$ *is preprocessed to be normalized. CCA coefficient is defined as* $S(\mathbf{Z}_u, \mathbf{Z}_v) = \sqrt{\frac{1}{c}\sum_{i=1}^c \sigma_i^2}$*, where* σ_i^2 *denotes the* i*-th eigenvalue of* $\Lambda_{u,v} = {Q_u}^\intercal Q_v$*,* $Q_u = \mathbf{Z}_u(\mathbf{Z}_u^\intercal\mathbf{Z}_u)^{-\frac{1}{2}}$*. Then* $\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v)$ *is upper bounded,*

$$\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v) \leq \frac{c^{\frac{3}{2}}\mathcal{T}}{\mathcal{C}} \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v), \tag{21}$$

where $\mathcal{T} = \mathbf{Tr}(\mathbf{X}^\intercal\boldsymbol{\alpha}^\intercal\boldsymbol{\alpha}\mathbf{X}), \mathcal{C} = \sigma_{min}(\mathbf{X}^\intercal\boldsymbol{\alpha}^\intercal\boldsymbol{\alpha}\mathbf{X})$.

Proof. Consider $\mathcal{S}^2 = \frac{1}{c}\sum_{i=1}^c \sigma_i^2$.

$$\mathcal{S}^2 = \frac{1}{c}\sum_{i=1}^c \sigma_i^2 = \frac{1}{c}\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^\intercal). \tag{22}$$

where

$$\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^\intercal) = \mathbf{Tr}(Q_u^\intercal Q_v Q_v^\intercal Q_u) = \mathbf{Tr}(Q_v Q_v^\intercal Q_u Q_u^\intercal). \tag{23}$$

As defined above, we have

$$Q_u Q_u^\intercal = \mathbf{Z}_u(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-\frac{1}{2}}(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-\frac{1}{2}}\mathbf{Z}_u^\intercal = \mathbf{Z}_u(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}\mathbf{Z}_u^\intercal$$
$$Q_v Q_v^\intercal = \mathbf{Z}_v(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-\frac{1}{2}}(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-\frac{1}{2}}\mathbf{Z}_v^\intercal = \mathbf{Z}_v(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1}\mathbf{Z}_v^\intercal. \tag{24}$$

Then Equation 23 becomes,

$$\begin{aligned}\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^\intercal) &= \mathbf{Tr}(\mathbf{Z}_u(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}\mathbf{Z}_u^\intercal \mathbf{Z}_v(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1}\mathbf{Z}_v^\intercal) \\ &= \mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}\mathbf{Z}_u^\intercal \mathbf{Z}_v(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1}\mathbf{Z}_v^\intercal \mathbf{Z}_u).\end{aligned} \tag{25}$$

By Cauchy-Schwartz Inequality,

$$\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^\intercal) \leq \mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1})\mathbf{Tr}((\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1})\mathbf{Tr}(\mathbf{Z}_u^\intercal \mathbf{Z}_v)^2. \tag{26}$$

Then we analyze these terms individually,

$$\begin{aligned}\mathbf{Tr}(\mathbf{Z}_u^\intercal \mathbf{Z}_v) &= \mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} \mathbf{X} \mathbf{D}_v) = \mathbf{Tr}(\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} \mathbf{X} \mathbf{D}_v \mathbf{D}_u^\intercal) \\ &\leq \mathbf{Tr}(\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} \mathbf{X})\mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{D}_v) \leq \mathcal{T} \cdot \mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{D}_v)\end{aligned} \tag{27}$$

As for $\mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1})$, let $\lambda_1, \lambda_2, ..., \lambda_c$ be eigenvalues for $\mathbf{Z}_u^\intercal \mathbf{Z}_u$ listed in descending order ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_c$), and assume the condition number of $\mathbf{Z}_u^\intercal \mathbf{Z}_u$ and $\mathbf{Z}_v^\intercal \mathbf{Z}_v$ satisfy $\lambda_{max}/\lambda_{min} \leq \gamma$, then,

$$\mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}) = \sum_{i=1}^{c} \frac{1}{\lambda_i} \leq c \cdot \frac{1}{\lambda_c} \leq \frac{\gamma c}{\lambda_1}, \tag{28}$$

where $\lambda_1 = ||\mathbf{Z}_u^\intercal \mathbf{Z}_u||_2$, $||\cdot||_2$ denotes the operator norm induced by the vector L_2-norm. With the norm inequalities of any positive semidefinite matrix A,

$$||A||_2 \geq \frac{1}{\sqrt{c}}||A||_F \geq \frac{1}{c}||A||_* \geq \frac{1}{c}\mathbf{Tr}(A), \tag{29}$$

where $||\cdot||_F, ||\cdot||_*$ denote the Frobenius norm and the nuclear norm, respectively.

Equation (30) then becomes,

$$\mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}) \leq c \cdot \frac{1}{||\mathbf{Z}_u^\intercal \mathbf{Z}_u||_2} \leq \frac{\gamma c^2}{\mathbf{Tr}(\mathbf{Z}_u^\intercal \mathbf{Z}_u)}. \tag{30}$$

By Lemma A.3,

$$\begin{aligned}\mathbf{Tr}(\mathbf{Z}_u^\intercal \mathbf{Z}_u) &= \mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} \mathbf{X} \mathbf{D}_u) \\ &= \mathbf{Tr}(\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}^\intercal \mathbf{X} \mathbf{D}_u \mathbf{D}_u^\intercal) \\ &\geq \sigma_{min}(\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}^\intercal \mathbf{X})\mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{D}_u) \\ &\geq \mathcal{C} \cdot \mathbf{Tr}(\mathbf{D}_u^\intercal \mathbf{D}_u) \\ &\geq \mathcal{C} \cdot ||vec(\mathbf{D}_u)||_2^2,\end{aligned} \tag{31}$$

where $vec(\cdot)$ denotes vectorization of a matrix.

Then Equation 30 is further derived as,

$$\mathbf{Tr}((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1}) \leq \frac{\gamma c^2}{\mathcal{C} \cdot ||vec(\mathbf{D}_u)||_2^2}. \tag{32}$$

Similarly, we have

$$\mathbf{Tr}((\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1}) \leq \frac{\gamma c^2}{\mathcal{C} \cdot ||vec(\mathbf{D}_v)||_2^2}. \tag{33}$$

Finally, with $\mathbf{Tr}(\mathbf{D}_u^{\intercal}\mathbf{D}_v) =< vec(\mathbf{D}_u), vec(\mathbf{D}_v) >$, we have

$$\begin{aligned}\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^{\intercal}) &\leq \frac{\gamma^2\mathcal{T}^2c^4(< vec(\mathbf{D}_u), vec(\mathbf{D}_v) >)^2}{\mathcal{C}^2||vec(\mathbf{D}_u)||_2^2 \cdot ||vec(\mathbf{D}_v)||_2^2} \\ &\leq \frac{\gamma^2\mathcal{T}^2c^4}{\mathcal{C}^2} \cdot \mathbf{cos}^2(\mathbf{D}_u, \mathbf{D}_v),\end{aligned} \tag{34}$$

and thus,

$$\begin{aligned}\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v) &= \sqrt{\frac{1}{c}\mathbf{Tr}(\Lambda_{u,v}\Lambda_{u,v}^{\intercal})} \\ &\leq \frac{\gamma\mathcal{T}c^{\frac{3}{2}}}{\mathcal{C}} \cdot \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v).\end{aligned} \tag{35}$$

Then the claimed theorem is proved.

□

Lemma A.5. *For two matrices* $\mathbf{A}$, $\mathbf{B}$, *their frobenius norm satisfies,*

$$\|\mathbf{AB}\|_F = \|\mathbf{A}\|_F\|\mathbf{B}\|_F\sqrt{1 - \frac{\Delta_1}{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}}, \tag{36}$$

where $\Delta_1 = \sum_{ij}(\sum_k A_{ik}^2)(\sum_k B_{kj}^2) \cdot \sin^2(\langle A_{i:}, B_{:j}\rangle)$.

Proof. According to the definition of frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{ij}|A_{ij}|^2}$ we have,

$$\|\mathbf{AB}\|_F = \sqrt{\sum_{ij}(\sum_k A_{ik}B_{kj})^2}. \tag{37}$$

Note that $(\sum_i x_iy_i)^2 = (\sum_i x_i^2)(\sum_i y_i^2) \cdot \cos^2(\langle x, y\rangle) = (\sum_i x_i^2)(\sum_i y_i^2) - (\sum_i x_i^2)(\sum_i y_i^2) \cdot \sin^2(\langle x, y\rangle)$, where $\langle x, y\rangle$ is the angle of two vectors x and y. We have,

$$\begin{aligned}&\sqrt{\sum_{ij}(\sum_k A_{ik}B_{kj})^2} \\ =&\sqrt{\sum_{ij}\left[(\sum_k A_{ik}^2)(\sum_k B_{kj}^2) - (\sum_k A_{ik}^2)(\sum_k B_{kj}^2) \cdot \sin^2(\langle A_{i:}, B_{:j}\rangle)\right]} \\ =&\sqrt{\sum_{ik}A_{ik}^2}\sqrt{\sum_{kj}B_{kj}^2}\sqrt{1 - \frac{\sum_{ij}(\sum_k A_{ik}^2)(\sum_k B_{kj}^2) \cdot \sin^2(\langle A_{i:}, B_{:j}\rangle)}{\sum_{ik}A_{ik}^2\sum_{kj}B_{kj}^2}} \\ =&\|\mathbf{A}\|_F\|\mathbf{B}\|_F\sqrt{1 - \frac{\sum_{ij}(\sum_k A_{ik}^2)(\sum_k B_{kj}^2) \cdot \sin^2(\langle A_{i:}, B_{:j}\rangle)}{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}} \\ =&\|\mathbf{A}\|_F\|\mathbf{B}\|_F\sqrt{1 - \frac{\Delta_1}{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2}},\end{aligned} \tag{38}$$

where $A_{i:}$ is the i-th row of $\mathbf{A}$ and $B_{:j}$ is the j-th column of $\mathbf{B}$, $\Delta_1 = \sum_{ij}(\sum_k A_{ik}^2)(\sum_k B_{kj}^2) \cdot \sin^2(\langle A_{i:}, B_{:j}\rangle)$. As $A_{i:}$ and $B_{:j}$ are more correlated, $\langle A_{i:}, B_{:j}\rangle \to 0$, thus, $\Delta_1 \ll \|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$.

□

Lemma A.6.

$$\|\mathbf{A}^{1/2}\|_F = \|\mathbf{A}\|_F^{1/2}(1 + \frac{\Delta_{1\mathbf{A}^{1/2}}}{\|\mathbf{A}\|_F^2})^{1/4}. \tag{39}$$

Proof. According to Lemma A.5, we have,

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}^{1/2}\|_F^4 - \Delta_1. \tag{40}$$

Thus,

$$\|\mathbf{A}^{1/2}\|_F = \|\mathbf{A}\|_F^{1/2}(1 + \frac{\Delta_{1A^{1/2}}}{\|\mathbf{A}\|_F^2})^{1/4}, \tag{41}$$

where $\Delta_{1\mathbf{A}^{1/2}} = \sum_{ij}(\sum_k (A^{1/2})_{ik}^2)(\sum_k (A^{1/2})_{kj}^2) \cdot \sin^2(\langle (A^{1/2})_{i:}, (A^{1/2})_{:j}\rangle)$. As $(A^{1/2})_{i:}$ and $(A^{1/2})_{:j}$ are more correlated, $\langle (A^{1/2})_{i:}, (A^{1/2})_{:j}\rangle \to 0$, thus, $\Delta_{1A^{1/2}} \ll \|\mathbf{A}\|_F^2$.

□

Lemma A.7. *For three matrices* $\mathbf{A}$*,* $\mathbf{B}$*, and* $\mathbf{C}$*, their frobenius norm satisfies,*

$$\|\mathbf{A}\|_F = \|\mathbf{A}\|_F\|\mathbf{B}\|_F\|\mathbf{C}\|_F\sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2}}, \tag{42}$$

where $\Delta_2 = \frac{1}{2}[\|\mathbf{A}\|_F^2 \sum_{kj}(\sum_l B_{kl}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle B_{k:}, C_{:j}\rangle) + \|\mathbf{C}\|_F^2 \sum_{il}(\sum_k A_{ik}^2)(\sum_k B_{kl}^2) \cdot \sin^2(\langle A_{i:}, B_{:l}\rangle)]$ *and* $\Delta_3 = \frac{1}{2}[\sum_{ij}(\sum_k A_{ik}^2)(\sum_k (BC)_{kj}^2) \cdot \sin^2(\langle A_{i:}, (BC)_{:j}\rangle) + \sum_{ij}(\sum_l (AB)_{il}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle (AB)_{i:}, C_{:j}\rangle)]$.

Proof. Based on Lemma A.5, we have,

$$\begin{aligned}
&\|\mathbf{ABC}\|_F^2 \\
=&\|\mathbf{AB}\|_F^2\|\mathbf{C}\|_F^2 - \sum_{ij}(\sum_l (AB)_{il}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle (AB)_{i:}, C_{:j}\rangle) \\
=&\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 - \|\mathbf{C}\|_F^2 \sum_{il}(\sum_k A_{ik}^2)(\sum_k B_{kl}^2) \cdot \sin^2(\langle A_{i:}, B_{:l}\rangle) \\
&- \sum_{ij}(\sum_l (AB)_{il}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle (AB)_{i:}, C_{:j}\rangle)
\end{aligned} \tag{43}$$

Symmetrically, we also have,

$$\begin{aligned}
&\|\mathbf{ABC}\|_F^2 \\
=&\|\mathbf{A}\|_F^2\|\mathbf{BC}\|_F^2 - \sum_{ij}(\sum_k A_{ik}^2)(\sum_k (BC)_{kj}^2) \cdot \sin^2(\langle A_{i:}, (BC)_{:j}\rangle) \\
=&\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 - \|\mathbf{A}\|_F^2 \sum_{kj}(\sum_l B_{kl}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle B_{k:}, C_{:j}\rangle) \\
&- \sum_{ij}(\sum_k A_{ik}^2)(\sum_k (BC)_{kj}^2) \cdot \sin^2(\langle A_{i:}, (BC)_{:j}\rangle)
\end{aligned} \tag{44}$$

Thus,

$$\begin{aligned}
&\|\mathbf{ABC}\|_F^2 \\
=&\frac{1}{2}[\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 - \|\mathbf{A}\|_F^2 \sum_{kj}(\sum_l B_{kl}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle B_{k:}, C_{:j}\rangle) \\
&- \sum_{ij}(\sum_k A_{ik}^2)(\sum_k (BC)_{kj}^2) \cdot \sin^2(\langle A_{i:}, (BC)_{:j}\rangle) \\
&+ \|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 - \|\mathbf{C}\|_F^2 \sum_{il}(\sum_k A_{ik}^2)(\sum_k B_{kl}^2) \cdot \sin^2(\langle A_{i:}, B_{:l}\rangle) \\
&- \sum_{ij}(\sum_l (AB)_{il}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle (AB)_{i:}, C_{:j}\rangle)] \\
=&\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 - \Delta_2 - \Delta_3,
\end{aligned} \tag{45}$$

where $\Delta_2 = \frac{1}{2}[\|A\|_F^2 \sum_{kj}(\sum_l B_{kl}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle B_{k:}, C_{:j}\rangle) + \|C\|_F^2 \sum_{il}(\sum_k A_{ik}^2)(\sum_k B_{kl}^2) \cdot \sin^2(\langle A_{i:}, B_{:l}\rangle)]$ and $\Delta_3 = \frac{1}{2}[\sum_{ij}(\sum_k A_{ik}^2)(\sum_k (BC)_{kj}^2) \cdot \sin^2(\langle A_{i:}, (BC)_{:j}\rangle) + \sum_{ij}(\sum_l (AB)_{il}^2)(\sum_l C_{lj}^2) \cdot \sin^2(\langle (AB)_{i:}, C_{:j}\rangle)]$. Therefore,

$$\|\mathbf{ABC}\|_F = \|\mathbf{A}\|_F \|\mathbf{B}\|_F \|\mathbf{C}\|_F \sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}\|_F^2}}. \tag{46}$$

As $A_{i:}$ and $B_{:l}$, $B_{k:}$ and $C_{:j}$ are more correlated, $\langle A_{i:}, B_{:l}\rangle, \langle B_{k:}, C_{:j}\rangle, \langle A_{i:}, (BC)_{:j}\rangle, \langle (AB)_{i:}, C_{:j}\rangle \to 0$, thus, $\Delta_2 \ll \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}\|_F^2$ and $\Delta_3 \ll \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}\|_F^2$.

□

Lemma A.8.

$$\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\|_F = \kappa_F(\mathbf{A}^{1/2})\kappa_F(\mathbf{C}^{1/2}) \frac{\|\mathbf{B}\|_F}{\|\mathbf{A}^{1/2}\|_F \|\mathbf{C}^{1/2}\|_F} \sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|\mathbf{A}^{-1/2}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}^{-1/2}\|_F^2}}, \tag{47}$$

where $\kappa_F(\mathbf{A}^{1/2})$ and $\kappa_F(\mathbf{C}^{1/2})$ are the condition number of $\mathbf{A}^{1/2}$ and $\mathbf{C}^{1/2}$, $\kappa_F(\mathbf{A}^{1/2}) = \sqrt{(\sum \sigma_i^2(\mathbf{A}^{1/2}))(\sum \frac{1}{\sigma_i^2(\mathbf{A}^{1/2})})}$ and $\kappa_F(\mathbf{C}^{1/2}) = \sqrt{(\sum \sigma_i^2(\mathbf{C}^{1/2}))(\sum \frac{1}{\sigma_i^2(\mathbf{C}^{1/2})})}$; $\sigma_i^2(\mathbf{A}^{1/2})$ are singular value of $\mathbf{A}^{1/2}$ and $\sigma_i^2(\mathbf{C}^{1/2})$ are singular value of $\mathbf{C}^{1/2}$.

Proof. Based on Lemma A.7, we have,

$$\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\|_F = \|\mathbf{A}^{-1/2}\|_F \|\mathbf{B}\|_F \|\mathbf{C}^{-1/2}\|_F \sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|\mathbf{A}^{-1/2}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}^{-1/2}\|_F^2}}. \tag{48}$$

By the definition of condition number $\kappa_F(\mathbf{X}) = \|\mathbf{X}\|_F \|\mathbf{X}^{-1}\|_F = \sqrt{(\sum \sigma_i^2(\mathbf{X}))(\sum \frac{1}{\sigma_i^2(\mathbf{X})})}$,

$$\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\|_F = \kappa_F(\mathbf{A}^{1/2})\kappa_F(\mathbf{C}^{1/2}) \frac{\|\mathbf{B}\|_F}{\|\mathbf{A}^{1/2}\|_F \|\mathbf{C}^{1/2}\|_F} \sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|\mathbf{A}^{-1/2}\|_F^2 \|\mathbf{B}\|_F^2 \|\mathbf{C}^{-1/2}\|_F^2}}. \tag{49}$$

□

Theorem A.9. *Suppose the forward of decomposed convolution layer for the u-th model is $\mathbf{Z}_u = \alpha \mathbf{X}\mathbf{D}_u$, CCA coefficient be $S(\mathbf{Z}_u, \mathbf{Z}_v) = \sqrt{\frac{1}{c}\sum_{i=1}^{c} \sigma_i^2}$, where σ_i^2 denotes the i-th eigenvalue of $\Lambda_{u,v} = {Q_u}^\intercal Q_v$, $Q_u = \mathbf{Z}_u(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-\frac{1}{2}}$. Then $\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v)$ is approximately linear to filter subspace similarity,*

$$\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v) = \frac{\gamma_1 \gamma_2 \gamma_3}{\sqrt{c}} \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v), \tag{50}$$

Proof. Based on $S(\mathbf{Z}_u, \mathbf{Z}_v) = \sqrt{\frac{1}{c}\sum_{i=1}^{c} \sigma_i^2}$ and $\|\Lambda_{u,v}\|_F = \sqrt{\sum_{i=1}^{c} \sigma_i^2}$, where σ_i are the singular value of $\Lambda_{u,v}$,

$$\mathcal{S} = \sqrt{\frac{1}{c}\sum_{i=1}^{c} \sigma_i^2} = \frac{1}{\sqrt{c}} \|\Lambda_{u,v}\|_F = \frac{1}{\sqrt{c}} \|(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-\frac{1}{2}} \mathbf{Z}_u^\intercal \mathbf{Z}_v (\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-\frac{1}{2}}\|_F. \tag{51}$$

According to Lemma. A.8, we have

$$\frac{1}{\sqrt{c}} \|(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-\frac{1}{2}} \mathbf{Z}_u^\intercal \mathbf{Z}_v (\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-\frac{1}{2}}\|_F = \frac{\gamma_1 \gamma_2}{\sqrt{c}} \frac{\|\mathbf{Z}_u^\intercal \mathbf{Z}_v\|_F}{\|(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{\frac{1}{2}}\|_F \|(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{\frac{1}{2}}\|_F}, \tag{52}$$

where $\gamma_1 = \kappa_F((\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{\frac{1}{2}}) \cdot \kappa_F((\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{\frac{1}{2}})$ and $\gamma_2 = \sqrt{1 - \frac{\Delta_2 + \Delta_3}{\|(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{-1/2}\|_F^2 \|\mathbf{Z}_u^\intercal \mathbf{Z}_v\|_F^2 \|(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{-1/2}\|_F^2}}$.
As $\mathbf{Z}_u = \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u$ and $\mathbf{Z}_v = \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v$, we have

$$\begin{aligned}
&\frac{\gamma_1\gamma_2}{\sqrt{c}} \frac{\|\mathbf{Z}_u^\intercal \mathbf{Z}_v\|_F}{\|(\mathbf{Z}_u^\intercal \mathbf{Z}_u)^{\frac{1}{2}}\|_F \|(\mathbf{Z}_v^\intercal \mathbf{Z}_v)^{\frac{1}{2}}\|_F} \\
=&\frac{\gamma_1\gamma_2}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v\|_F}{\|(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u)^{\frac{1}{2}}\|_F \|(\mathbf{D}_v^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v)^{\frac{1}{2}}\|_F}.
\end{aligned} \tag{53}$$

According to Lemma A.6,

$$\begin{aligned}
&\frac{\gamma_1\gamma_2}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v\|_F}{\|(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u)^{\frac{1}{2}}\|_F \|(\mathbf{D}_v^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v)^{\frac{1}{2}}\|_F} \\
=&\frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v\|_F}{\|(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u)\|_F^{\frac{1}{2}} \|(\mathbf{D}_v^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v)\|_F^{\frac{1}{2}}},
\end{aligned} \tag{54}$$

where $\gamma_3 = (1 + \frac{\Delta_1}{\|(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u)\|_F^2})^{-\frac{1}{4}} (1 + \frac{\Delta_1}{\|(\mathbf{D}_v^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v)\|_F^2})^{-\frac{1}{4}}$.
As Assumption 2.6 holds, it becomes

$$\begin{aligned}
&\frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v\|_F}{\|(\mathbf{D}_u^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_u)\|_F^{\frac{1}{2}} \|(\mathbf{D}_v^\intercal \mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\mathbf{D}_v)\|_F^{\frac{1}{2}}} \\
=&\frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{D}_v\|_F \|\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\|_F}{\|\mathbf{D}_u^\intercal\|_F^{\frac{1}{2}} \|\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\|_F^{\frac{1}{2}} \|\mathbf{D}_u\|_F^{\frac{1}{2}} \|\mathbf{D}_v^\intercal\|_F^{\frac{1}{2}} \|\mathbf{X}^\intercal \boldsymbol{\alpha}^\intercal \boldsymbol{\alpha}\mathbf{X}\|_F^{\frac{1}{2}} \|\mathbf{D}_v\|_F^{\frac{1}{2}}} \\
=&\frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \frac{\|\mathbf{D}_u^\intercal \mathbf{D}_v\|_F}{\|\mathbf{D}_u\|_F \|\mathbf{D}_v\|_F} \\
=&\frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v).
\end{aligned} \tag{55}$$

Thus, we have

$$\mathcal{S}(\mathbf{Z}_u, \mathbf{Z}_v) = \frac{\gamma_1\gamma_2\gamma_3}{\sqrt{c}} \mathbf{cos}(\mathbf{D}_u, \mathbf{D}_v). \tag{56}$$

Specifically, we have $\gamma_2 = \sqrt{1 - \frac{\Delta}{\gamma_1^2\gamma_3^2} \frac{1}{cos^2(\mathbf{D}_u, \mathbf{D}_v)}}$, and since Δ are small, with Taylor expansion, $\gamma_2 \approx 1 - \frac{1}{2} \frac{\Delta}{\gamma_1^2\gamma_3^2} \frac{1}{cos^2(\mathbf{D}_u, \mathbf{D}_v)}$. The term $\frac{1}{cos^2(\mathbf{D}_u, \mathbf{D}_v)}$ causes non-linearity in the relation between CCA and filter subspace similarity. $\square$

A.2 Experiment Settings

Model training of Federated Learning. In each experiment we have 100 clients in total and sample a ratio $r = 0.1$ of all the clients on every round. All models are randomly initialized and trained for $T = 100$ communication rounds for the CIFAR datasets. At each round, the client executes 15 epochs of SGD with momentum to train the local model, the learning rate is 0.01 and momentum is 0.9. Accuracies are computed by taking the average local accuracies for all users at the final communication round. As shown in the Table 3, we have different settings for CIFAR-10 and CIFAR-100. For example, $(100, 2)$ means 100 clients with 2 classes on each client. For each method, the training takes about 12 hours on Nvidia RTX A5000.

Comparison with other FL approaches. We compare our approach by evolving shared atom coefficients with various personalized federated learning methods and federated learning methods with local finetuning. Among these methods, FedPer [2] and FedRep[6] have the similar ideas by learning shared global representation and personalized local heads. Ditto [25] and FedProx [27] induce global regularization to improve the model performance. We also compare our method with FedAvg [32]. FedRep [6] approaches the common knowledge with shared representation. The codes

Table 3: Compare accuracy with different approaches

	CIFAR-100		CIFAR-10		
(# client, # classes per client)	(100, 5)	(100, 20)	(100, 2)	(100, 5)	(1000, 2)
FedAvg	82.39	62.92	86.37	70.63	86.12
FedProx	80.77	59.7	85.90	69.94	84.83
FedPer	81.46	62.52	81.74	68.24	81.74
FedRep	72.98	37.71	80.55	67.3	82.98
Local	81.21	49.25	90.24	72.05	97.80
Ours	81.03	52.13	83.37	65.63	82.54

are adapted from [1]. We evaluate the test accuracy on CIFAR-10 and CIFAR-100 with different FL setting. As shown in Table 3, our method achieves comparable performance among different methods.

Fine-tuning models for ensemble. We select 3 models with different similarity measures for ensemble. For feature-based similarity methods, we randomly select 1000 examples from CIFAR-100 dataset. The fully-connected layer of each model is fine-tuned on the user's local data with 100 epochs. The fine-tuning takes about 12 hours on Nvidia RTX A5000. After fine-tuning, the accuracy is measured on local test data, with the predictions of current model and 3 selected models.

A.3 Extra Experiments

Representation dependency on filter atoms. We first validate the dependency of deep features on filter atoms in Proposition 2.1 with a simple experiment. The model $\mathcal{F}$ here is a 2-layer CNN with coefficient $\boldsymbol{\alpha}$ and atom $\mathbf{D}$ generated from normal distribution $\mathcal{N}(0,1)$. The input sample $\mathbf{X}$ is also generated from normal distribution $\mathcal{N}(0,1)$. Figure 8(a) shows the relation between $\|\mathbf{Z}_u - \mathbf{Z}_v\|_F$ and $\|\mathbf{D}_u - \mathbf{D}_v\|_F$ by fixing coefficient $\boldsymbol{\alpha}$ and input sample $\mathbf{X}$ and randomly varying filter atoms $\mathbf{D}$. All the points are below the line which is the bound provided by Proposition 2.1, reflecting that the representation variations are dominated by filter atoms.

Correlation between probing-based and filter subspace-based methods. In addition, we empirically verify that CCA and filter subspace similarity have a strong correlation with AlexNet. In this experiment, 10 tasks are generated from CIFAR100 [21] with 10 classes in each task. Only the filter atoms of each task are trained while the atom coefficients are fixed. We calculate CCA and filter subspace similarity among 45 pairs of models. The correlation between CCA and filter subspace similarity is *0.8638* which is shown in Figure 9(b). Similarly, the correlation between CKA and filter subspace similarity is also reported in Figure 9 (Table). These results clearly show that the proposed filter subspace similarity has high linear relationship with popular probing-based similarities, which agrees with Theorem 2.5 and Theorem 2.7.

[1] `https://github.com/lgcollins/FedRep`

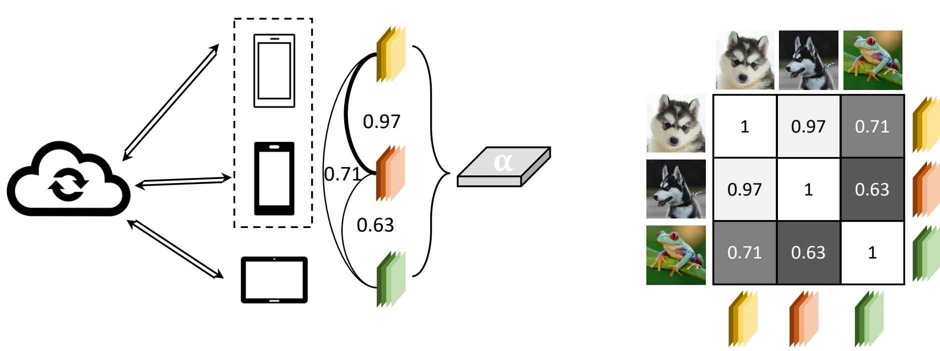
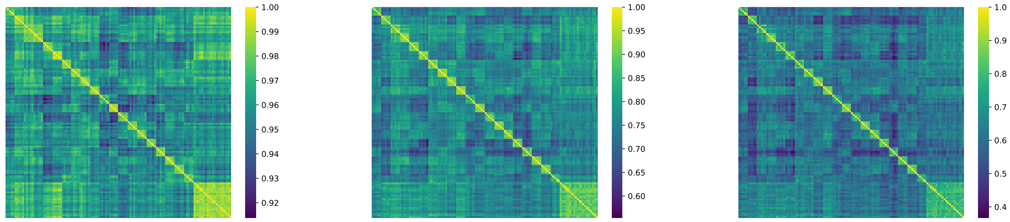
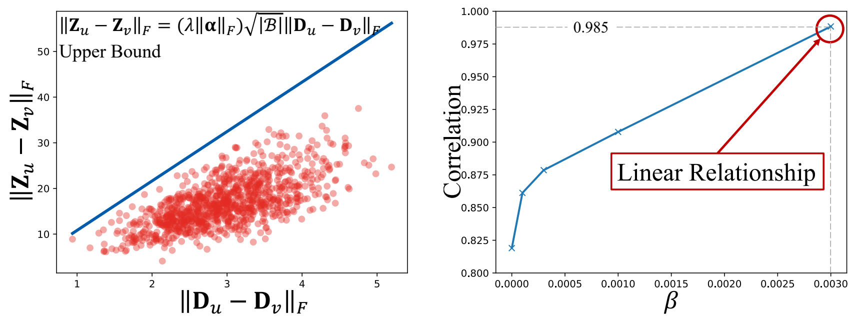


Figure 6: The shared coefficients and user-specific atoms represent common knowledge and personalized information. The filter subspace similarity is used to calculate the relations among users. Users with heterogeneous data result in lower similarity, as illustrated in a similarity matrix.



(a) 10th epoch (b) 50th epoch (c) 90th epoch

Figure 7: Similarity matrices that show relations among 120 users in FL with our filter subspace similarity through the training process.



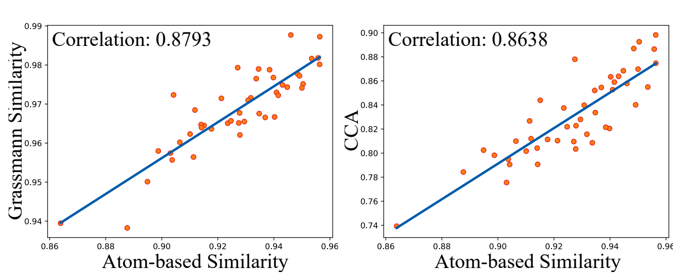
(a) (b)

Figure 8: (a) The change of features $\|\mathbf{Z}_u - \mathbf{Z}_v\|_F$ is bounded by the change of atoms $\|\mathbf{D}_u - \mathbf{D}_v\|_F$. (b) The channel decorrelation leads to a higher correlation between CCA and filter subspace similarity. And the correlation can reach 0.985 with $\beta = 3 \times 10^{-3}$, which means a near linear relation between CCA and filter subspace similarity.

Effect of channel decorrelation. We further design a regularization term $\beta \sum_{i \neq j} (\mathbf{Z}_u^\intercal \mathbf{Z}_u)_{ij}^2$ to approach $(\mathbf{Z}_u^\intercal \mathbf{Z}_u)_{ii} \gg (\mathbf{Z}_u^\intercal \mathbf{Z}_u)_{ij}$ in Assumption. 2.6. As shown in Figure 8(b), the correlation between CCA and filter subspace similarity keeps increasing as β increases. The correlation reaches 0.985 when $\beta = 3 \times 10^{-3}$, indicating a near-linear relationship, which is aligned with Theorem. 2.7.

Similar representations across datasets. Similar to [19], we can use filter subspace similarity to compare networks trained on different datasets. In Figure 10(a), we show that pairs of models that are both trained on CIFAR-10 and CIFAR-100 have high atom-based similarities. Models learned on two datasets respectively still show high similarity. In contrast, similarities between trained and untrained models are significantly lower.

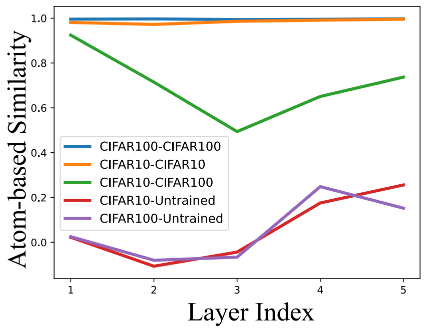
Limitation of probing-based methods. As shown in Figure 10(b), to illustrate sensitivity of probing-based similarities to probing data, we perform a simple regression task with data, $\{(x_i = 0, y_i, z_i)\}_{i=1}^n$, where $z_i = f(x_i, y_i) + \epsilon_i$ and $y_i, \epsilon_i \sim \mathcal{N}(0.5, 0.1)$. Two NN models $\mathcal{F}_1$ and $\mathcal{F}_2$



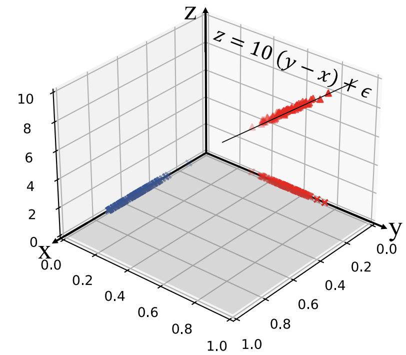
(a) (b)

	Correlation
CCA [40]	0.8638
CKA [19]	0.7358
Grassmann Distance [33]	0.8793

Figure 9: (a) Correlation between Grassmann similarity and filter subspace similarity; (b) Correlation between CCA and filter subspace similarity. (Table) Correlation between filter subspace similarity and other approaches.



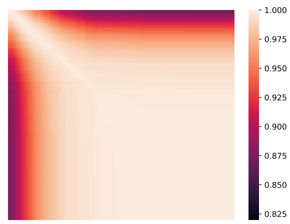
(a)



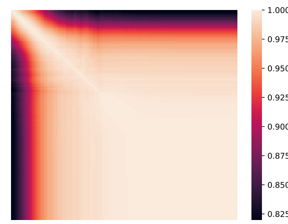
(b)

Figure 10: (a) Using filter subspace similarity, models trained on different datasets (CIFAR-10 and CIFAR-100) are similar among themselves, but they differ from untrained models. (b) Illustration of limitations of probing-based similarities. Input data from "red" ($\{(x_i = 0, y_i)\}$) and "blue" ($\{(x'_i = y_i, y'_i = 0)\}$) are orthogonal. Since two models are learned on "red" data, their similarity should be 1, which can be faithfully indicated by our atom similarity. However, probing-based similarities will become 0 with the "blue" probing data.

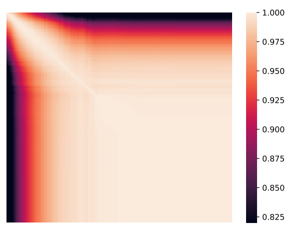
(a) 1st conv



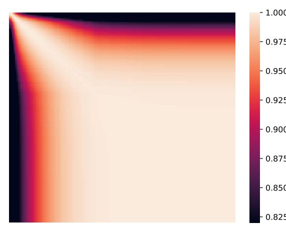
(b) 2nd conv



(c) 3rd conv



(d) 4th conv



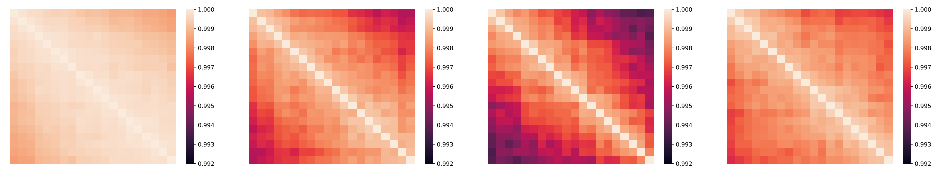
(e) 5th conv

Figure 11: Similarity of AlexNet with atoms from different time point during the training.

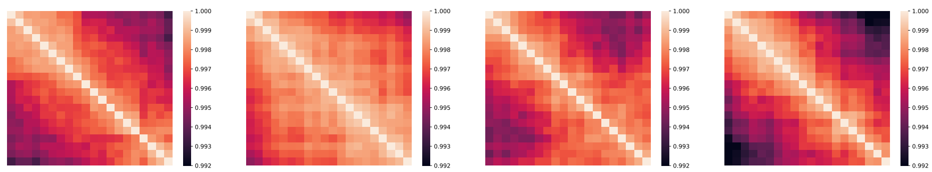
with the same initialization and atom coefficients are trained for their different atoms to learn $\mathcal{F} : (X, Y) \to Z$. It is can be simply found that the filter subspace similarity of $\mathcal{F}_1$ and $\mathcal{F}_2$ is 1 and the probing-based similarity is also 1 with the same $\{(x_i = 0, y_i)\}$ as the probing data. However, if we choose $\{(x'_i = y_i, y'_i = 0)\}$ as the probing data, then the probing-based similarities directly become **0** as the data are now orthogonal to model parameters.

A.4 Training dynamics.

We investigate the training dynamics of AlexNet [22] and VGG [47] separately on CIFAR-100 [21] and ImageNet [44]. The details of training dynamics of models with atoms from different time point during the training are shown in Figure 11 and Figure 12. Moreover, we examine the similarity between the two participated models shared the same initialization trained only with atoms on two different tasks. The results is shown in Figure 13 and Figure 14. The difference is less on the first few layers, but more on the middle layers. It reflects the middle layer is more critical than other layers, which is aligned with previous work [36].

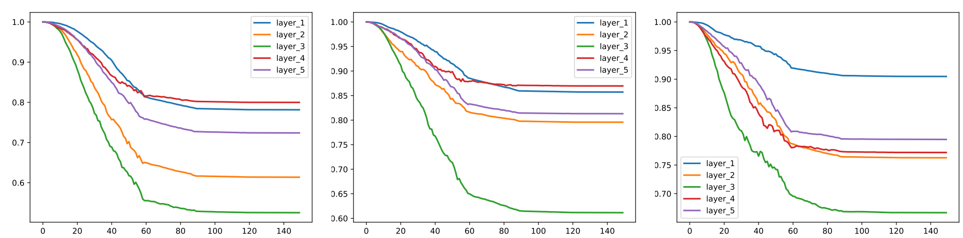


(a) 1st conv (b) 2nd conv (c) 3rd conv (d) 4th conv

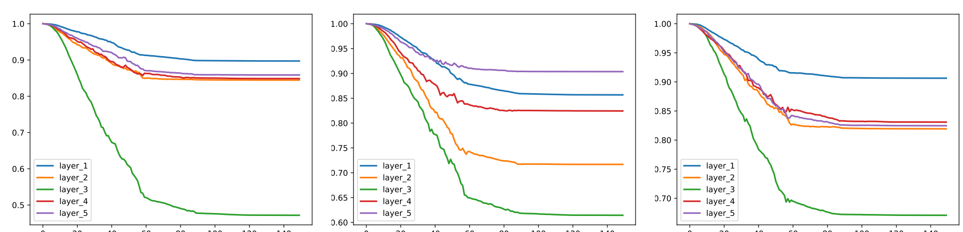


(e) 5th conv (f) 6th conv (g) 7th conv (h) 8th conv

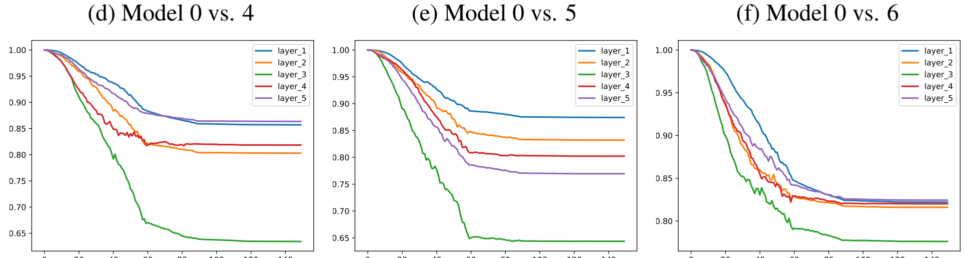
Figure 12: Similarity of VGG with atoms from different time point during the training.



(a) Model 0 vs. 1 (b) Model 0 vs. 2 (c) Model 0 vs. 3

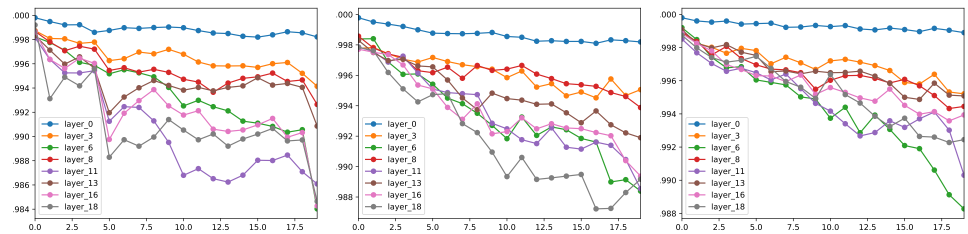


(d) Model 0 vs. 4 (e) Model 0 vs. 5 (f) Model 0 vs. 6

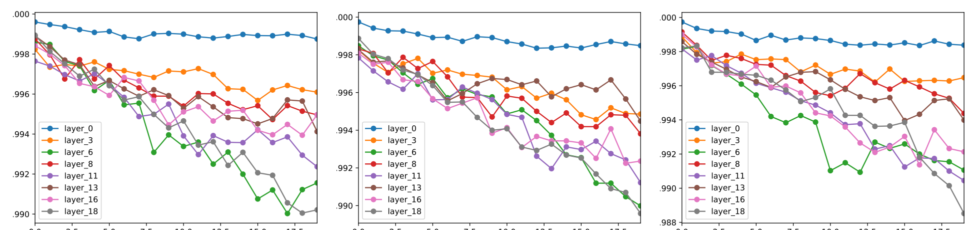


(g) Model 0 vs. 7 (h) Model 0 vs. 8 (i) Model 0 vs. 9

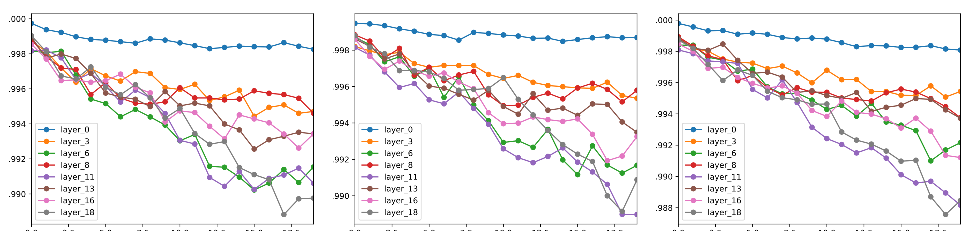
Figure 13: Similarity of AlexNet trained on different tasks during the training.



(a) Model 0 vs. 1 (b) Model 0 vs. 2 (c) Model 0 vs. 3



(d) Model 0 vs. 4 (e) Model 0 vs. 5 (f) Model 0 vs. 6



(g) Model 0 vs. 7 (h) Model 0 vs. 8 (i) Model 0 vs. 9

Figure 14: Similarity of VGG trained on different tasks during the training.