Appendix

.1. Detailed Proofs

In this section, we provide the detailed proofs of the manuscript.

Theorem .1. *Let $H = \{\mathbf{h}_i \mid \forall i \in [N], \mathbf{h}_i \in \mathcal{R}^{D_E}\}$ be set of embedding features of N points, and the corresponding label set is given as $Y = \{y_i \mid \forall i \in [N], y \in [K]\}$. For a fixed batch size $|B|$, we define a set of sub-sampling index sets of size $|B|$ as $\mathcal{B}$ such that*

$$\mathcal{B} = \{\{n_1, n_2, \ldots, n_B\} | n_i \in [N], \forall i \in [B]\}.$$

We have

$$\mathcal{L}_{OCL}(H;Y) \geq \sum_{l=2}^{|B|} l M_l \log\left(l - 1 + \frac{|B|-1}{e}\right) \tag{6}$$

where $M_l = \sum_{y\in[K]} |\{B \in \mathcal{B} || B_y| = l\}|$, and the set B_y consists of all samples with label y in the batch B. Equality is attained if and only if there are K orthonormal vectors $\xi_1, \xi_2, \ldots, \xi_K \in R^{D_E}$ with a large D_E, s.t. $K < D_E$ can be obtained under the condition that $\forall n \in [N] : \mathbf{h}_n = \xi_{y_n}$.

Several steps are presented in order to prove Theorem .1 as follows.

Step 1: First let us define B_y^C to be the complementary set of B_y such that $B_y + B_y^C = B$. For any class y and any batch $B \in \mathcal{B}$, the class-specific loss $\mathcal{L}_{OCL}(H;Y,B,y)$ can be bounded by

$$\begin{aligned} &\mathcal{L}_{OCL}(H;Y,B,y) \\ \geq &|B_y| \log(|B_y| - 1 + |B_y^C| \exp(S(H;Y,B,y))) \end{aligned} \tag{7}$$

where function S can be defined as

$$S(H;Y,B,y) = S_{att}(H;Y,B,y) + S_{rep}(H;Y,B,y) \tag{8}$$

In Eq. (9), we further introduce the two functions $S_{att}()$ and $S_{rep}()$ respectively below

$$\begin{aligned} &S_{att}(H;Y,B,y) = -\frac{1}{|B_y|(|B_y|-1)} \sum_{i\in B_y} \sum_{j\in B_y \setminus \{\{i\}\}} \langle \mathbf{h}_i, \mathbf{h}_j \rangle \\ &S_{rep}(H;Y,B,y) \\ &= \begin{cases} \frac{1}{|B_y||B_y^C|} \sum_{i\in B_y} \sum_{j\in B_y^C} |\langle \mathbf{h}_i, \mathbf{h}_j \rangle|, & if |B_y| \neq |B| \\ 0, & if\ |B_y| = |B| \end{cases} \end{aligned} \tag{9}$$

Lemma .2. *For any class y and any batch $B \in \mathcal{B}$, the class-specific loss $\mathcal{L}_{OCL}(H;Y,B,y)$ can be bounded by*

$$\begin{aligned} &\mathcal{L}_{OCL}(H;Y,B,y) \\ \geq &|B_y| \log(|B_y| - 1 + |B_y^C| \exp(S(H;Y,B,y))) \end{aligned} \tag{10}$$

where equality holds iff all of the following hold:

(A1) $\forall i \in B$ there is a $C_i(B,y)$ such that $\forall j \in B_y \setminus \{\{i\}\}$, $\langle \mathbf{h}_i, \mathbf{h}_j \rangle = C_i(B,y)$.

(A2) $\forall i \in B$ there is a $D_i(B,y)$ such that $\forall j \in B_y^C$, $|\langle \mathbf{h}_i, \mathbf{h}_j \rangle| = D_i(B,y)$.

Proof.

$$\begin{aligned} &\mathcal{L}_{OCL}(H;Y,B,y) \\ &= -\sum_{i\in B_y} \frac{1}{|B_{y_i}|-1} \sum_{j\in B_{y_i}\setminus\{\{i\}\}} \log\left(\frac{\exp(\langle \mathbf{h}_i, \mathbf{h}_j \rangle)}{\sum_{k\in B\setminus\{\{i\}\}} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k \rangle|)}\right) \\ &= \sum_{i\in B_y} \log\left(\frac{\sum_{k\in B\setminus\{\{i\}\}} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k \rangle|)}{\Pi_{j\in B_{y_i}\setminus\{\{i\}\}} \exp(|\langle \mathbf{h}_i, \mathbf{h}_j \rangle|)^{1/(|B_{y_i}|-1)}}\right) \\ &= \sum_{i\in B_y} \log\left(\frac{\sum_{k\in B\setminus\{\{i\}\}} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k \rangle|)}{\exp((|B_{y_i}|-1)^{-1} \sum_{j\in B_{y_i}\setminus\{\{i\}\}} |\langle \mathbf{h}_i, \mathbf{h}_j \rangle|)}\right) \end{aligned} \tag{11}$$

In Eq. (11), we can further reorganize the numerator below.

$$\sum_{k\in B\setminus\{\{i\}\}} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|) = \sum_{k\in B_y\setminus\{\{i\}\}} \exp(\langle \mathbf{h}_i, \mathbf{h}_k\rangle) + \sum_{k\in B_y^C} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|) \tag{12}$$

Using Jensen's inequality on both sums, one can attain In Eq. (11), we can further reorganize the numerator below.

$$\begin{aligned}\sum_{k\in B_y\setminus\{\{i\}\}} \exp(\langle \mathbf{h}_i, \mathbf{h}_k\rangle) &\overset{(A1)}{\geq} |B_y\setminus\{\{i\}\}| \exp\left(\frac{\sum_{k\in B_y\setminus\{\{i\}\}}\langle \mathbf{h}_i, \mathbf{h}_k\rangle}{|B_y\setminus\{\{i\}\}|}\right)\\ \sum_{k\in B_y^C} \exp(|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|) &\overset{(A2)}{\geq} |B_y^C| \exp\left(\frac{\sum_{k\in B_y\setminus\{\{i\}\}}|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|}{|B_y^C|}\right)\end{aligned} \tag{13}$$

where the the equality holds if and only if

(A1) $\exists C_i(B,y)$ such that $\forall j \in B_y\setminus\{\{i\}\}, |\langle \mathbf{h}_i, \mathbf{h}_j\rangle| = C_i(B,y)$.

(A2) $\exists D_i(B,y)$ such that $\forall j \in B_y^C, |\langle \mathbf{h}_i, \mathbf{h}_j\rangle| = D_i(B,y)$.

Plugging Eq. (14) in Eq. (12), we obtain the bound of each addend as

$$\begin{aligned}&\frac{\sum_{k\in B\setminus\{\{i\}\}}\exp(|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|)}{\exp((|B_{y_i}|-1)^{-1}\sum_{j\in B_{y_i}\setminus\{\{i\}\}}|\langle \mathbf{h}_i, \mathbf{h}_j\rangle|)}\\ &\geq |B\setminus\{\{i\}\}| + |B_y^C|\exp\left(\frac{\sum_{k\in B_y^C}|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|}{|B_y^C|} - \frac{\sum_{k\in B\setminus\{\{i\}\}}|\langle \mathbf{h}_i, \mathbf{h}_k\rangle|}{|B\setminus\{\{i\}\}|}\right)\end{aligned} \tag{14}$$

So with the definition of $S(H;Y,B,y)$, we can obtain the claimed bound

$$\begin{aligned}&\mathcal{L}_{OCL}(H;Y,B,y)\\ \geq&|B_y|\log(|B_y|-1+|B_y^C|\exp(S(H;Y,B,y)))\end{aligned} \tag{15}$$

□

Lemma .3. *Let $l \in \{2,\dots,|B|\}$. For $Y \in [K]$ and H, we have $L_{LOCL}(H,Y) = \sum_{B\in\mathcal{B}}\sum_{y\in[K]}\mathcal{L}_{OCL}(H;Y,B,y)$, we have*

$$\begin{aligned}&\frac{1}{M_l}\sum_{B\in\mathcal{B}}\sum_{y\in[K]}\log(l-1+(|B|-l)\exp(S(H;Y,B,y)))\\ \geq&\log\left(l-1+(|B|-l)\exp\left(\frac{1}{M_l}S(H;Y,B,y)\right)\right)\end{aligned} \tag{16}$$

where $M_l = \sum_{y\in[K]}|\mathcal{B}_{y,l}|$ and $\mathcal{B}_{y,l}$ is an auxiliary partition of $\mathcal{B}$ such that $\mathcal{B}_{y,l} = \{B_{y_i} | |B_{y_i}| = l, \forall i \in [K]\}$. The equality holds if and only if

(A3) $l = |B|$ or there exists $D(l)$ such that for every $y \in [K]$ and $B \in \mathcal{B}_{y,l}$ the values of $S(H;Y,B,y) = D(l)$ agree.

Proof. Since $f(x) = \log(l-1+(|B|-l)\exp(|x|))$ is a convex function, using Jensen's inequality, for every $y \in [K]$ and $B \in \mathcal{B}_{y,l}$, we have

$$\frac{1}{|\mathcal{B}_{y,l}|}\sum_{B\in\mathcal{B}}\sum_{y\in[K]} f(S(H;Y,B,y)) \overset{(A3)}{\geq} f\left(\frac{1}{|\mathcal{B}_{y,l}|}\sum_{B\in\mathcal{B}}\sum_{y\in[K]} S(H;Y,B,y)\right) \tag{17}$$

where the equality can be obtained if and only if A3 holds. □

Step 2: Next, we use the bound of $\mathcal{L}_{OCL}(H;Y,B,y)$ derived from Lemma .2 and Lemma .3 to get the bound for $\mathcal{L}_{OCL}(H,Y)$.

Lemma .4. *For every Y and H the orthonormal contrastive loss $\mathcal{L}_{OCL}$ is bounded by*

$$\mathcal{L}_{OCL} \geq \sum_{l=2}^{|B|} l M_l \log\left(l-1+(|B|-l)\exp\left(\frac{1}{M_l}S(H;Y,B,y)\right)\right) \tag{18}$$

where the equality holds if and only if

(B1) $\forall n,m \in [N]$, if $y_n = y_m$, it implies $\langle h_n, h_m\rangle \equiv \eta$.

(B2) $\forall n,m \in [N]$, if $y_n \neq y_m$, it implies $|\langle h_n, h_m\rangle| \equiv \gamma$.

Proof.

$$\begin{aligned}\mathcal{L}_{OCL}(H,Y)=&\sum_{B\in\mathcal{B}}\sum_{y\in[K]}\mathcal{L}_{OCL}(H;Y,B,y)\\=&\sum_{l=2}^{|B|}\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}}\mathcal{L}_{OCL}(H;Y,B,y)\\\geq&\sum_{l=2}^{|B|}\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}}l\log(l-1+(|B|-l)\exp(S(H;Y,B,y)))\\\geq&\sum_{l=2}^{|B|}lM_l\log\left(l-1+(|B|-l)\exp\left(\frac{1}{M_l}\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}}S(H;Y,B,y)\right)\right)\end{aligned}\tag{19}$$

The first and second inequality can be attained via Lemma .2 and Lemma .3. The equality can be achieved if and only if (A1), (A2), and (A3) are true. It can be further proved that $(A1)\&(A2)\&(A3)\Leftrightarrow(B1)\&(B2)$.

We first prove " $\Leftarrow$".

(A1) For an arbitrary $l\in\{2,\ldots,|B|\}$, $y\in Y$, $B\in\mathcal{B}_{y,l}$ and $i\in B$, we let $j\in B_y\setminus\{\{i\}\}$, i.e., $y_j=y_i=y$. Then we have $\langle h_i,h_j\rangle=\eta=C_i(B,y)$.

(A2) For an arbitrary $l\in\{2,\ldots,|B|\}$, $y\in Y$, $B\in\mathcal{B}_{y,l}$ and $i\in B$, we let $j\in B_y^C$, i.e., $y_j=y_i=y$. Then we have $|\langle h_i,h_j\rangle|=\gamma=D_i(B,y)$.

(A3) For an arbitrary $l\in\{2,\ldots,|B|-1\}$, $y\in Y$, and $B\in\mathcal{B}_{y,l}$, with condition (B1), $S_{att}(H;Y,B,y)=-\eta$, and by condition (A2), $S_{rep}(H;Y,B,y)=-\gamma$. So we have $S(H;Y,B,y)=S_{att}(H;Y,B,y)+S_{rep}(H;Y,B,y)=\gamma-\eta=D(l)$.

Next, we prove " $\Rightarrow$".

(B1) We aim to prove that given y,y' and $m,n,m',n'\in[N]$ with $y_m=y_n=y$ and $y_{m'}=y_{n'}=y'$, we can induce that $|\langle h_n,h_m\rangle|=|\langle h_{n'},h_{m'}\rangle|$.

Case I:

If $y\neq y'$, we choose $l=2$ and we specify the batch $B'=\{\{n,m,n',\ldots,n'\}\}$ with the size b. We can get

$$\begin{aligned}&S(H,Y,B',y)\\=&S_{att}(H;Y,B,y)+S_{rep}(H;Y,B,y)\\=&-\langle h_n,h_m\rangle+\frac{|\langle h_n,h_{n'}\rangle|}{2}+\frac{|\langle h_{n'},h_m\rangle|}{2}\end{aligned}\tag{20}$$

With (A2), we can further get $S(H;Y,B,y)=-|\langle h_n,h_m\rangle|+|\langle h_{n'},h_n\rangle|$. Similarly, we can specify the batch $B''=\{\{m',n',n,\ldots,n\}\}$ with the size b and we can get $S(H,Y,B'',y=-|\langle h_{n'},h_{m'}\rangle|+|\langle h_{n'},h_n\rangle|)$. Combining these two equations with condition (A3), one can deduce that $|\langle h_n,h_m\rangle|=|\langle h_{n'},h_{m'}\rangle|$.

Case II: If $y=y'$, we choose $l=2$ and we specify the batch $B'=\{\{m,n,p,\ldots,p\}\}$ with the size b. Following the similar procedure in Case I, with (A2), we can further get $S(H,Y,B',y)=-|\langle h_m,h_n\rangle|+|\langle h_n,h_p\rangle|$. Similarly, we can specify the batch $B''=\{\{m',n',p,\ldots,p\}\}$ with the size b and we can get $S(H,Y,B',y=-\langle h_{n'},h_{m'}\rangle+\langle h_{n'},h_p\rangle)$. Combining these two equations with condition (A3), one can deduce that $-|\langle h_n,h_m\rangle|+|\langle h_n,h_p\rangle|=|\langle h_{n'},h_{m'}\rangle|+|\langle h_{n'},h_p\rangle|$.

Now, pick the batch $B_3=\{\{h_m,h_m,p,\ldots,p\}\}$. With condition (A2), we have $|\langle h_n,p\rangle|=|\langle h_m,p\rangle|$ and thus $|\langle h_{n'},h_{m'}\rangle|=|\langle h_n,h_m\rangle|$.

(B2) We aim to prove that given $y\neq y'$, $|\langle h_n,h_{n'}\rangle|=|\langle h_m,h_{m'}\rangle|$.

We still choose $l=2$ and we specify two batches as $B'=\{\{n,n,n',\ldots,n'\}\}$ with the size $|B|$ and $B''=\{\{m,m,m',\ldots,m'\}\}$ with the size $|B|$. Assuming $S_{att}(H;Y,B,y)=-\eta$ and thus

$$\begin{aligned}&S(H,Y,B',y)\\=&-\eta+S_{rep}(H,Y,B',y)\\=&-\eta+\frac{1}{2(|B|-2)}\sum_{i\in B'_y}\sum_{j\in B'^C_y}|\langle h_i,h_j\rangle|\\=&-\eta+|\langle h_n,h_{n'}\rangle|\end{aligned}\tag{21}$$

Similar to Eq. 21, we have $S(H, Y, B'', y) = -\eta + |\langle h_m, h_{m'} \rangle|$. With (A3), we have $S(H, Y, B'', y) = S(H, Y, B', y)$ so that $|\langle h_n, h_{n'} \rangle| = |\langle h_m, h_{m'} \rangle|$.

With (A2), we can further get $S(H; Y, B, y) = -|\langle h_n, h_m \rangle| + |\langle h_{n'}, h_n \rangle|$. Similarly, we can specify the batch $B'' = \{\{m', n', n, \ldots, n\}\}$ with the size b and we can get $S(H, Y, B'', y = -|\langle h_{n'}, h_{m'} \rangle| + |\langle h_{n'}, h_n \rangle|)$. Combining these two equations with condition (A3), one can deduce that $|\langle h_n, h_m \rangle| = |\langle h_{n'}, h_{m'} \rangle|$. □

Step 3:

Now we will partition the bounding problem into two components which characterize the intra-class bound and the inter-class bound respectively. Mathematically, a decomposition can be written as

$$\begin{aligned} &\sum_{y \in Y} \sum_{B \in \mathcal{B}_{y,l}} S(H; Y, B, y) \\ =&\sum_{y \in Y} \sum_{B \in \mathcal{B}_{y,l}} S_{att}(H; Y, B, y) + \sum_{y \in Y} \sum_{B \in \mathcal{B}_{y,l}} S_{rep}(H; Y, B, y) \end{aligned} \tag{22}$$

We first address the first addend in Eq. 23 in the following lemma. And the rest of the lemmas focus on the second addend.

Lemma .5. *Let $l \in \{2, \ldots, |B|\}$ and let H to be the unit vector on a unit sphere. For every Y and H, it holds that*

$$\sum_{y \in Y} \sum_{B \in \mathcal{B}_{y,l}} S_{att}(H; Y, B, y) \geq -\left(\sum_{y \in Y} |B_{y,l}| \right) \tag{23}$$

where the equality is attained if and only if: (A4) $\forall m, n \in [N]$, $y_m = y_n$ implies $h_m = h_n$.

Proof.

$$\begin{aligned} S_{att}(H; Y, B, y) &= -\frac{1}{|B_y||B_y \setminus \{\{i\}\}|} \sum_{i \in B_y} \sum_{j \in \mathcal{B}_y \setminus \{\{i\}\}} \langle h_i, h_j \rangle \\ &\geq -\frac{1}{|B_y||B_y \setminus \{\{i\}\}|} \sum_{i \in B_y} \sum_{j \in \mathcal{B}_y \setminus \{\{i\}\}} h_i h_j \\ &= -1 \end{aligned} \tag{24}$$

which can be obtained by using Cauchy-Schwarz inequality. The equality holds if and only if h_i and h_j are identical since the h_i and h_j are unit vectors. So the equality condition can be written as (A4) $\forall m, n \in [N]$, $y_m = y_n$ implies $h_m = h_n$. □

Now, we use Lemma 3 and Lemma 4 to prove the bound for our orthonormal supervised contrastive loss.

Lemma .6. *The orthonormal contrastive loss $\mathcal{L}_{OCL}(H, Y)$ is bounded from below by*

$$\mathcal{L}_{OCL}(H, Y) \geq \sum_{l=2}^{|B|} l M_l \log\left(l - 1 + \frac{|B| - 1}{e} \right) \tag{25}$$

where equality is achieved if and only if there exists $\{\xi_1, \ldots, \xi_Y\}$ such that the following conditions hold:

(C1) $\forall n \in [N]$, $h_n = \xi_{y_n}$.

(C2) $\{\xi_1, \ldots, \xi_Y\}$ are pairwise orthonormal.

Proof. Utilizing the lower bound of S_{att} in Lemma .5, we can bound the exponential term in Lemma .4 first below

$$\begin{aligned} &\sum_{y \in [K]} \sum_{B \in \mathcal{B}_{y,l}} S(H; Y, B, y) \\ \geq &\sum_{y \in [K]} \sum_{B \in \mathcal{B}_{y,l}} S_{att}(H; Y, B, y) + \sum_{y \in [K]} \sum_{B \in \mathcal{B}_{y,l}} S_{rep}(H; Y, B, y) \\ \geq &\sum_{y \in Y} |B_{y,l}| \times (-1) + 0 \\ = &-|Y| \sum_{y \in Y} |B_{y,l}| \end{aligned} \tag{26}$$

where the second term $\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}} S_{rep}(H;Y,B,y) \geq 0$ and $\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}} S_{rep}(H;Y,B,y) = 0$ if and only if $\{\xi_1,\ldots,\xi_Y\}$ are pairwise orthonormal and $\forall n \in [N]$, $h_n = \xi_{y_n}$. So we can further derive the bound for $\mathcal{L}_{OCL}$ as follows.

$$
\begin{aligned}
&\mathcal{L}_{OCL}(H,Y) \\
\geq &\sum_{y\in[K]}\sum_{B\in\mathcal{B}_{y,l}} S(H;Y,B,y) \\
\geq &\sum_{l=2}^{|B|} lM_l \log\left(l-1+(|B|-l)\exp\left(\frac{1}{M_l}S(H;Y,B,y)\right)\right) \\
\geq &\sum_{l=2}^{|B|} lM_l \log\left(l-1+(|B|-l)\exp\left(-\frac{\sum_{y\in Y}|B_{y,l}|}{M_l}\right)\right) \\
\geq &\sum_{l=2}^{|B|} lM_l \log\left(l-1+(|B|-l)\exp\left(-\frac{\sum_{y\in Y}|B_{y,l}|}{\sum_{y\in Y}|B_{y,l}|}\right)\right) \\
\geq &\sum_{l=2}^{|B|} lM_l \log\left(l-1+\frac{|B|-l}{e}\right)
\end{aligned}
\tag{27}
$$

□

With Lemma 5, Theorem 1 is readily attained.

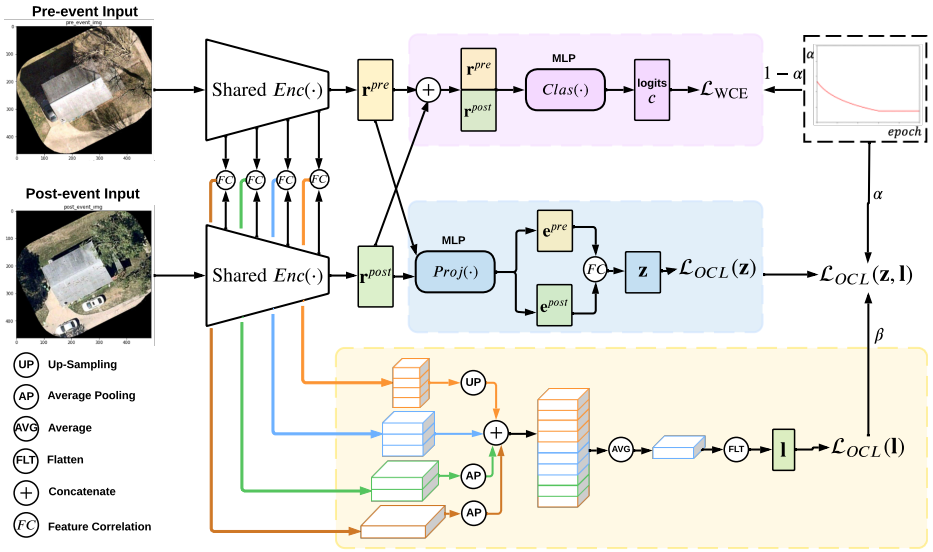

Figure 8. The end-to-end learning of the ***LOCAL*** model.

.2. End-to-End Learning of Paired Inputs

Fig. 8 shows the overview framework for learning both representation and classification based on paired inputs. It includes two parts: the first part learns a feature mapping with the property of intra-class compactness and inter-class separability; whereas the second part is expected to learn a less biased classifier based on the orthonormal representations produced by the first part. We take the damage detection task as an example where pre- and post-disaster image pair is denote by $(\mathbf{x}^{pre}, \mathbf{x}^{post})$.
Encoder Network, $Enc(\cdot)$, can employ any suitable backbone network, e.g., ResNet[13], and maps either image $\mathbf{x}^{pre}$ and $\mathbf{x}^{post}$ in the pair to a vector representation, $\mathbf{r}^{pre} = Enc(\mathbf{x}^{pre}) \in \mathbb{R}^{D_R}$ and $\mathbf{r}^{post} = Enc(\mathbf{x}^{post}) \in \mathbb{R}^{D_R}$, whereas $\mathbf{r}^{pre}$ and $\mathbf{r}^{post}$ are normalized to be on the unit hypersphere in $\mathbb{R}^{D_R}$.
Latent Hierarchical Feature Correlation Module, $Lat(\cdot)$, is a module injected into the backbone network to learn latent hierarchical joint representation between $\mathbf{x}^{pre}$ and $\mathbf{x}^{post}$. Specifically, from the backbone network, we firstly extract the

multi-scale outputs of each block (e.g., four blocks for ResNet) and then the feature correlation between the pre- and post-event outputs from each block. Denote the output from each block as $Enc_i'(\cdot)$, $i \in [1, ..., 4]$. Our feature correlation module can be computed as $Enc_i'(\mathbf{x}^{pre}), Enc_i'(\mathbf{x}^{post}))) = \mathbf{W}_i([Enc_i'(\mathbf{x}^{pre}), Enc_i'(\mathbf{x}^{post})])$ where the matrix $\mathbf{W}_i \in \mathbb{R}^{d_i \times 2d_i}$ denotes the correlation parameters, and $[Enc_i'(\mathbf{x}^{pre}), Enc_i'(\mathbf{x}^{post})] \in \mathbb{R}^{2d_i}$ is the concatenated vector of $Enc_i'(\mathbf{x}^{pre}), Enc_i'(\mathbf{x}^{post}) \in \mathbb{R}^{d_i}$. Practically, $\mathbf{W}_i$ can be set to compute the difference between pre- and post-event outputs, yielding satisfactory results. Then the feature correlation maps (lower left part of Fig. 6) illustrate the variation in dual images. These maps are resized to the same size via average pooling and up-sampling, and concatenated to form a hierarchical latent feature maps. Then, averaging over all channels produces a single channel feature map, which is then flattened and normalized to give a latent feature embedding $\mathbf{l} = Lat(\mathbf{x}^{pre}, \mathbf{x}^{post}) \in \mathbb{R}^{D_L}$. This embedding incorporates latent supervision from the backbone network, and then contributes to the computation of latent orthonormal contrastive loss $\mathcal{L}_{OCL}(\mathbf{l})$.

Projection Network with Feature Correlation Module, $Proj(\cdot)$ maps $\mathbf{r}^{pre}$ and $\mathbf{r}^{post}$ to the corresponding embedding vectors $\mathbf{e}^{pre} = Proj(\mathbf{r}^{pre}) \in \mathbb{R}^{D_E}$ and $\mathbf{e}^{post} = Proj(\mathbf{r}^{post}) \in \mathbb{R}^{D_E}$. This network is a multi-layer perceptron (MLP) with a hidden layer and an output layer of size D_E. It has been shown that such a projection module improves the quality of the embeddings of the layers preceding it [4, 17]. We apply an ℓ_2 normalization to $\mathbf{e}^{pre}$ and $\mathbf{e}^{post}$ to ensure that the inner product can be used as the $cosine$ similarity measure. A similar feature correlation module is incorporated to learn the variation between the pre- and post- outputs of the projection network. $\mathbf{z} = \mathbf{W}([\mathbf{e}^{pre}, \mathbf{e}^{post}])$ is used to compute the OCL loss $\mathcal{L}_{OCL}(\mathbf{z})$.

In our framework, the proposed OCL takes effect on both the latent hierarchical feature embedding $\mathbf{l}$ and the joint embbedding of the paired inputs $\mathbf{z}$, leading to the latent OCL loss:

$$\mathcal{L}_{OCL}(\mathbf{z}, \mathbf{l}) = \mathcal{L}_{OCL}(\mathbf{z}) + \beta \mathcal{L}_{OCL}(\mathbf{l}) \tag{28}$$

where $\beta \geq 0$ is a hyperparameter for tuning and $\mathcal{L}_{OCL}(\mathbf{l})$ can be regarded as a regularizer which regularizes the orthonormality class representations of lower-level features (with high contrast) that are extracted by the deep neural network. Imposing this regularizer helps learn the final inter-class orthonormal embeddings.

Classification Network, $Clas(\cdot)$, takes in the concatenated representation, $concat(\mathbf{r}^{pre}, \mathbf{r}^{post})$, from the *Encoder Network*. A non-linear MLP with a hidden layer and an output layer of the class size is employed to predict the class-wise logit values $\mathbf{c} \in \mathbb{R}^{D_C}$ of the input image pair, which are used to compute the weighted cross-entropy (WCE) loss $\mathcal{L}_{WCE}$. Combining with the WCE loss for classifier learning where the weight is the reciprocal of the appearance frequency of each class, we arrive at our final loss function of our proposed LOCAL: Latent Orthonormal Contrastive Learning Framework:

$$Total\ loss = \alpha \mathcal{L}_{OCL}(\mathbf{z}, \mathbf{l}) + (1 - \alpha)\mathcal{L}_{WCE} \tag{29}$$

where $0 \leq \alpha \leq 1$ is a weighting coefficient inversely proportional to the number of epochs.

.3. Implementation Details of *SCL* and *L-SCL*

Fig. 9 shows the detailed architecture of ***SCL*** model, where $\mathcal{L}_{\textbf{\textit{SCL}}} = \alpha \mathcal{L}_{SCL}(\mathbf{z}) + (1 - \alpha)\mathcal{L}_{WCE}$. Fig. 10 shows the detailed architecture of ***L-SCL*** model, where $\mathcal{L}_{\textbf{\textit{L-SCL}}} = \alpha \mathcal{L}_{SCL}(\mathbf{z}, \mathbf{l}) + (1 - \alpha)\mathcal{L}_{WCE}$.

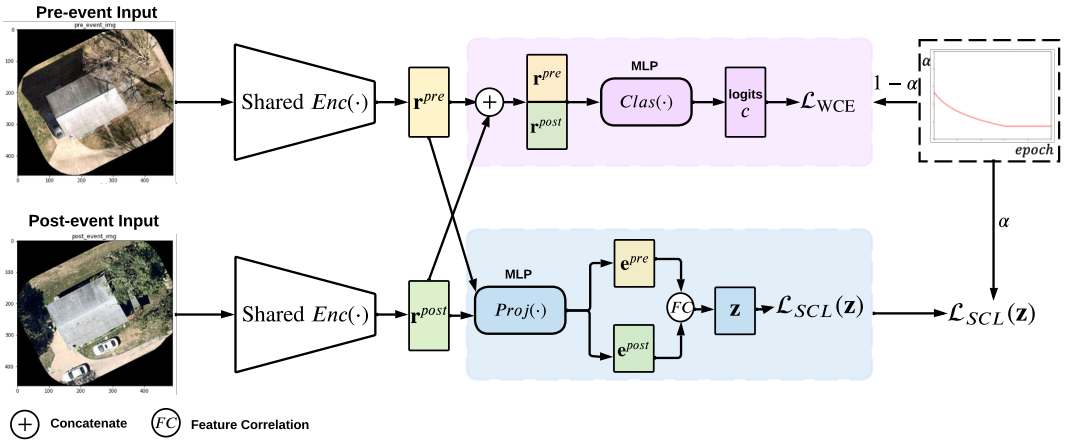

Figure 9. The end-to-end learning of the ***SCL*** model, $\mathcal{L}_{\textbf{SCL}} = \alpha\mathcal{L}_{SCL}(\mathbf{z}) + (1-\alpha)\mathcal{L}_{WCE}$.

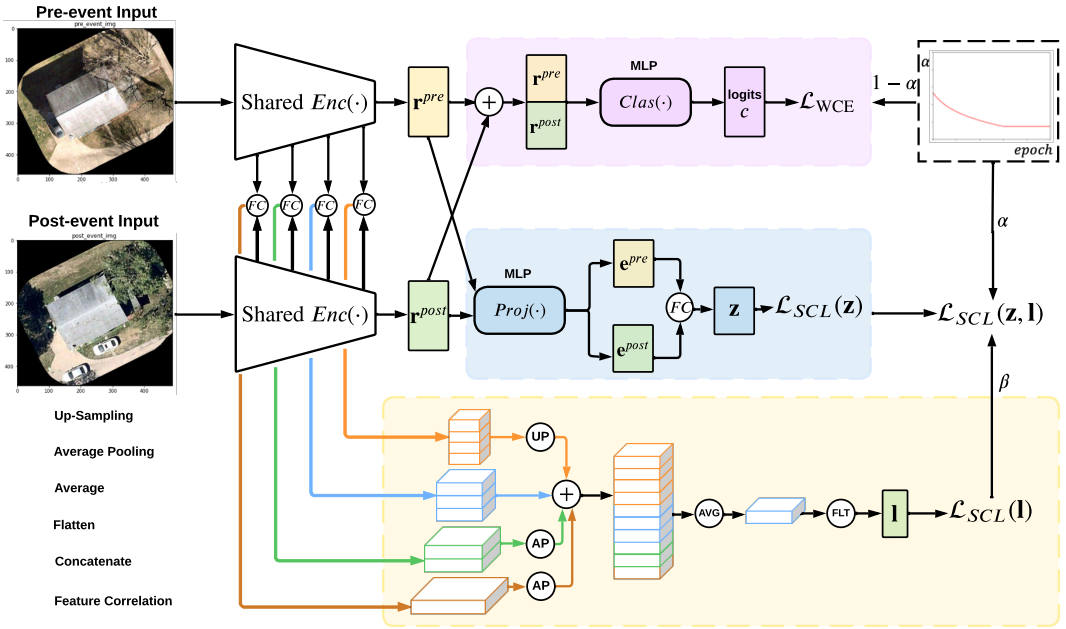

Figure 10. The end-to-end learning of the ***L-SCL*** model, $\mathcal{L}_{\textbf{L-SCL}} = \alpha\mathcal{L}_{SCL}(\mathbf{z},\mathbf{l}) + (1-\alpha)\mathcal{L}_{WCE}$.

.4. Generalizability on Benchmark Datasets with Single Image as Input

Additionally, we have performed experiments on benchmark datasets with single image as sample such as CIFAR-10-LT and CIFAR-10-LT and iNatualist-LT, to verify the effectiveness of OCL over SCL, as shown in Table 6. The obtained results suggest that encouraging orthonormality leads to improved performance, especially with small batch sizes: employing relatively small batch sizes for training: (4, 8, 12), (32, 64, 80) and (4, 8, 16) for each of the corresponding datasets.

We also conduct experiments on ImageNet-LT with limited epochs (thus not fully trained). We test different small batch sizes such as 4, 8, and 16 to simulate memory constraints and evaluate how OCL performs under such conditions. OCL is expected to show better performance over SCL, particularly in small batch sizes and under short training durations, as OCL optimizes representation more efficiently by encouraging orthonormality, which can help even under limited epochs.

Dataset	BS	SCL		OCL	
		Accuracy	F1-macro	Accuracy	F1-macro
CIFAR-10-LT	4	88.25 ± 1.47	67.32 ± 1.23	**88.79** ± 1.32	**71.25** ± 1.55
	8	92.29 ± 1.23	80.48 ± 1.35	**92.58** ± 1.42	**80.70** ± 1.21
	12	92.67 ± 1.10	80.90 ± 1.20	**93.25** ± 1.31	**81.42** ± 1.41
CIFAR-100-LT	32	74.90 ± 1.42	49.21 ± 1.33	**75.30** ± 1.35	**50.85** ± 1.25
	64	77.95 ± 1.20	53.43 ± 1.14	**78.20** ± 1.30	**53.55** ± 1.40
	80	78.35 ± 1.10	54.20 ± 1.23	**78.65** ± 1.17	53.77 ± 1.31
iNaturalist-LT	4	87.51 ± 1.38	65.09 ± 1.21	**87.73** ± 1.32	**70.42** ± 1.42
	8	93.10 ± 1.15	86.77 ± 1.09	**93.51** ± 1.24	**87.07** ± 1.16
	16	93.93 ± 1.08	88.99 ± 1.12	**94.25** ± 1.15	**90.02** ± 1.19

Table 6. Performance on benchmark datasets.

Dataset	BS	SCL		OCL	
		Accuracy	F1-macro	Accuracy	F1-macro
ImageNet-LT	4	5.74 ± 0.69	4.35 ± 0.52	**7.12** ± 0.88	**5.66** ± 0.74
	8	15.58 ± 1.23	13.55 ± 1.07	**17.55** ± 1.40	**15.29** ± 1.32
	16	26.71 ± 1.53	23.65 ± 1.48	**29.32** ± 1.62	**26.06** ± 1.55

Table 7. Performance on ImageNet-LT.

.5. Generalizability on Benchmark Datasets of Natural Language Inference with Paired Sentences as Input

We evaluate the applicability of OCL for NLP on the Stanford Natural Language Inference (SNLI) dataset (Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326 (2015).). This corpus consists of sentence pairs written and annotated by humans. Each sentence pair is comprised of a "premise" and "hypothesis" sentence. The relationship between the premise and hypothesis is assigned to one of three labels: "entailment", "contradiction", or "neutral". To craft a challenging, imbalanced subset of this collection to demonstrate OCL's robust performance, we randomly select only 50%, 20%, and 5% of the pairs in the categories, respectively, to be retained from the SNLI training dataset. This creates a 1:10 imbalance ratio in training dataset; we leave the SNLI validation and testing sets unchanged.

Batch Size	PairSCL-Baseline	PairSCL-OCL
4	86.86 (86.82-86.89)	**87.33** (87.29-87.37)
8	86.92 (86.83-86.87)	**87.3** (87.27-87.33)
16	87.37 (87.35-87.4)	87.19 (87.16-87.23)
32	87.1 (87.08-87.13)	**87.33** (87.28-87.36)

Table 8. Performance on imbalanced SNLI dataset for NLP

We implement OCL on the existing framework of PairSCL [19], and replace the SCL term with our OCL to learn orthonormal embeddings and evaluate the subsequent performance on the crafted imbalanced SNLI dataset. In Table 8, we report the baseline performance of PairSCL, i.e., the unchanged implementation of PairSCL, against PairSCL-OCL, i.e, PairSCL modified with OCL.

Our proposed method consistently outperforms, or at least matches, the baseline of SCL loss.Meanwhile, our OCL loss exhibits greater stability across various batch sizes, ranging from 4 to 32. It may be confusing that SCL excel with a batch size of 16. SCL has been demonstrated to be effective on imbalanced datasets with a moderately sized batch, as overly large batches can introduce an excessive number of negative examples for samples from minority classes, thereby reducing the effectiveness of learning from these minority samples (Mitrovic, Jovana, Brian McWilliams, and Melanie Rey. "Less can be more in contrastive learning." (2020): 70-75.). Conversely, overly small batch sizes may lead to unstable learning outcomes for SCL. Determining the optimal batch size for SCL is challenging, as evidenced by our empirical findings that SCL reaches its peak performance with a batch size of 16. In contrast, our proposed OCL loss shows relative stability across a range of batch sizes.