

I_u	user domain images taken by users
I_s	shop domain images from e-commerce websites
$o \in I_u$	anchor image or query image
$p \in I_s$	positive image (matched image to o)
$q \in I_s$	negative image (unmatched image to o)
\mathbf{o}^p (resp. \mathbf{o}^q)	feature of o using the attention w.r.t to p (resp. q)
$\mathbf{x}_l \in R^C$	feature at location l , $x \in \{o, p, q\}$
$\mathbf{x}^t \in \{0, 1\}^T$	tags vector, $x \in \{p, q\}$
T	total number of tags
$a_l \in R$	attention weight at location l
$\mathbf{W} \in R^{T \times C}$	tag embedding matrix
$\mathbf{v} \in R^C$	weight vector in Equation ??
$\mathbf{U} \in R^{L \times C}$	weight matrix in Equation ??
L	total number of locations, i.e. height*width
$g_u()$, $g_s()$	feature alignment function in Equation ?? and ??