



Image



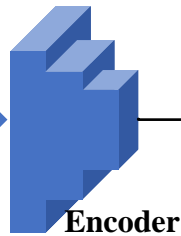
ResNet101

$I$

Visual Feature



$I$



Encoder



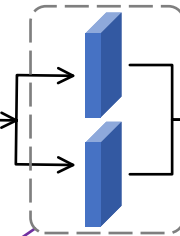
IE

$\tau_\mu$

$\tau_\sigma$

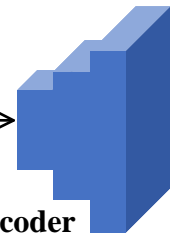
Information Mixer

$\Psi$



$Z_{img}$

Decoder



$I'$

$\mathcal{L}_{DA}$



Vision-Semantic Alignment

Classifier

Classifier

VSA

$\mathcal{L}_{ACMR}$

Attribute

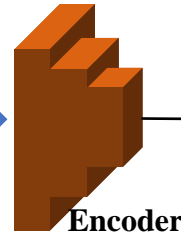
black white brown gray  
furry toughskin tail hooves  
longleg longneck  
fast strong muscle  
active agility quadrapedal  
vegetation grazer

$A$

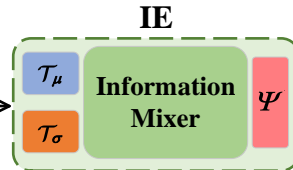
Semantic Feature



$A$



Encoder



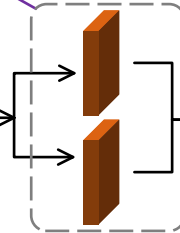
IE

$\tau_\mu$

$\tau_\sigma$

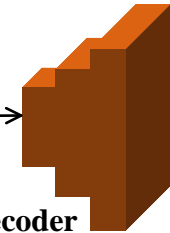
Information Mixer

$\Psi$



$Z_{att}$

Decoder



$A'$

$\mathcal{L}_{VAE}$