Image

ResNet101

I

Visual Feature

Encoder

IE

$\mathcal{T}_\mu$

$\mathcal{T}_\sigma$

Information Mixer

Ψ

$\mathcal{L}_{VAE}$

z_{img}

Decoder

I'

$\mathcal{L}_{DA}$

Vision-Semantic Alignment

Classifier

Classifier

$\mathcal{L}_{ACMR}$

VSA

Semantic Feature

Attribute

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

A

Encoder

IE

$\mathcal{T}_\mu$

$\mathcal{T}_\sigma$

Information Mixer

Ψ

z_{att}

$\mathcal{L}_{VAE}$

Decoder

A'