RSLab

Remote Sensing Laboratory
National Technical University of Athens
√ Sensing √ Analytics √ Monitoring

Learning
Visual and Multimodal Representations

Summary

1. Introduction

2. Visual Representations

3. Multimodal Representations

4. Conclusion

5. Future Work

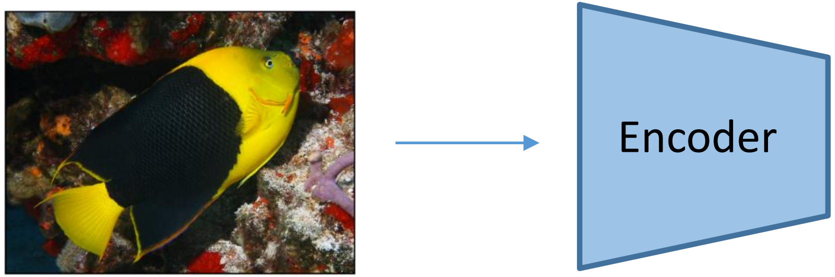
6. Publications

1. Introduction

Representation Learning

Input
Image

Representation Learning

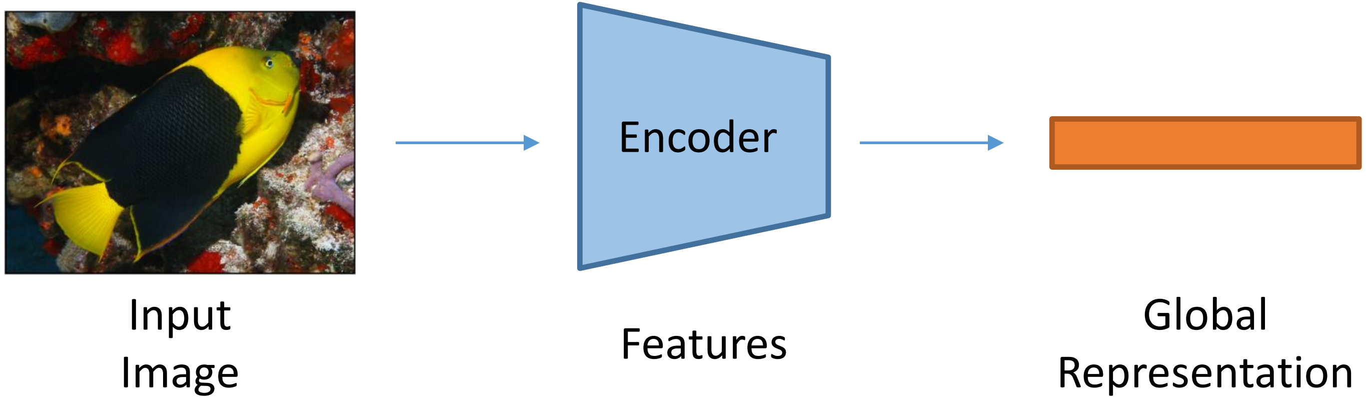


Encoder

Input
Image

Features

Representation Learning: Graphical Abstract



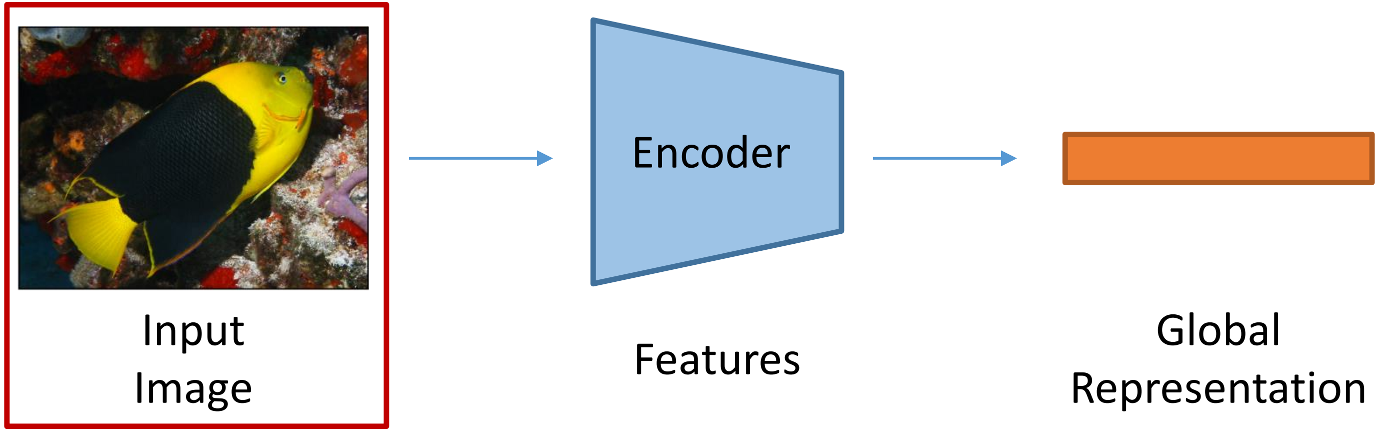
Encoder

Input
Image

Features

Global
Representation

Representation Learning: Focus on Data



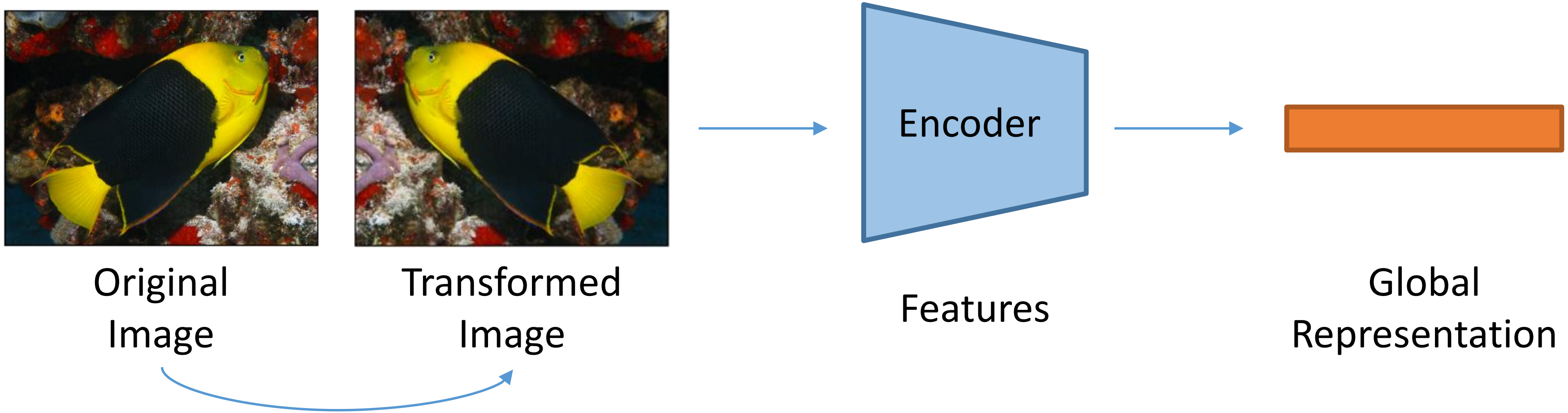
Input Image

Encoder

Features

Global Representation

Data Augmentation



Original Image

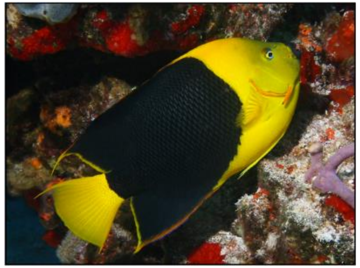
Transformed Image

Encoder

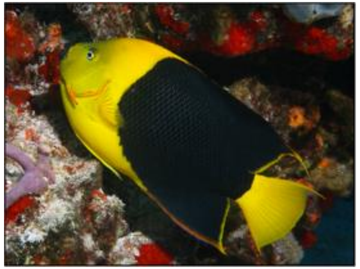
Features

Global Representation

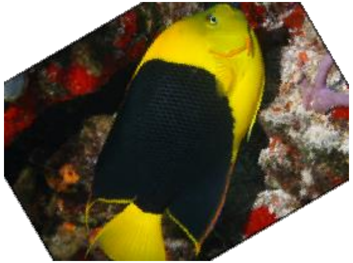
Data Augmentation



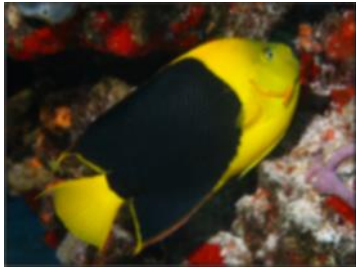
Original Image



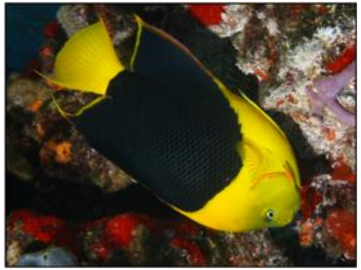
Horizontal Flip



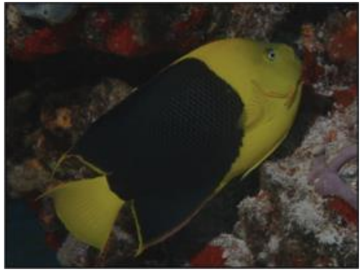
Rotation



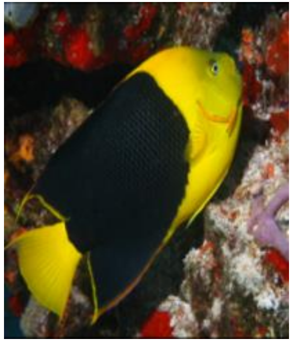
Gaussian Blur



Vertical Flip



Color Jitter



Resized Crop

Mixup: Advanced Data Augmentation

fish: 1.0

shark: 0.0

fish: 0.0

shark: 1.0

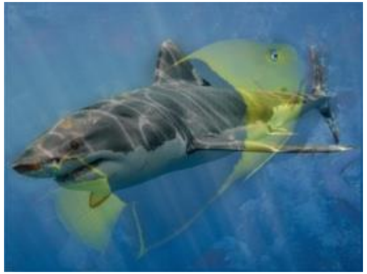
Zhang et al., mixup: Beyond Empirical Risk Minimization, ICLR 2018

Mixup: Advanced Data Augmentation

fish: 1.0
shark: 0.0

fish: 0.0
shark: 1.0

interpolating pairs of images



Zhang et al., mixup: Beyond Empirical Risk Minimization, ICLR 2018

Mixup: Advanced Data Augmentation

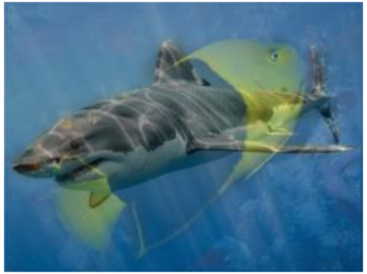
fish: 1.0
shark: 0.0

fish: 0.0
shark: 1.0

interpolating pairs of images

→

and their target labels



fish: 0.4
shark: 0.6

Zhang et al., mixup: Beyond Empirical Risk Minimization, ICLR 2018

Mixup: Advanced Data Augmentation

x_1

y_1

fish: 1.0
shark: 0.0

x_2

y_2

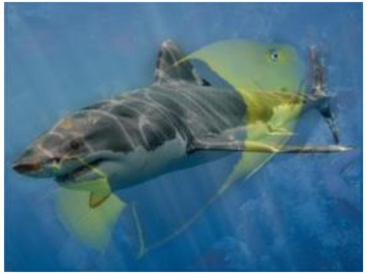
fish: 0.0
shark: 1.0

interpolating pairs of images

and their target labels

Mixed image: $x_{mix} = \lambda x_1 + (1-\lambda)x_2$
Mixed label: $y_{mix} = \lambda y_1 + (1-\lambda)y_2$

λ is called interpolation factor



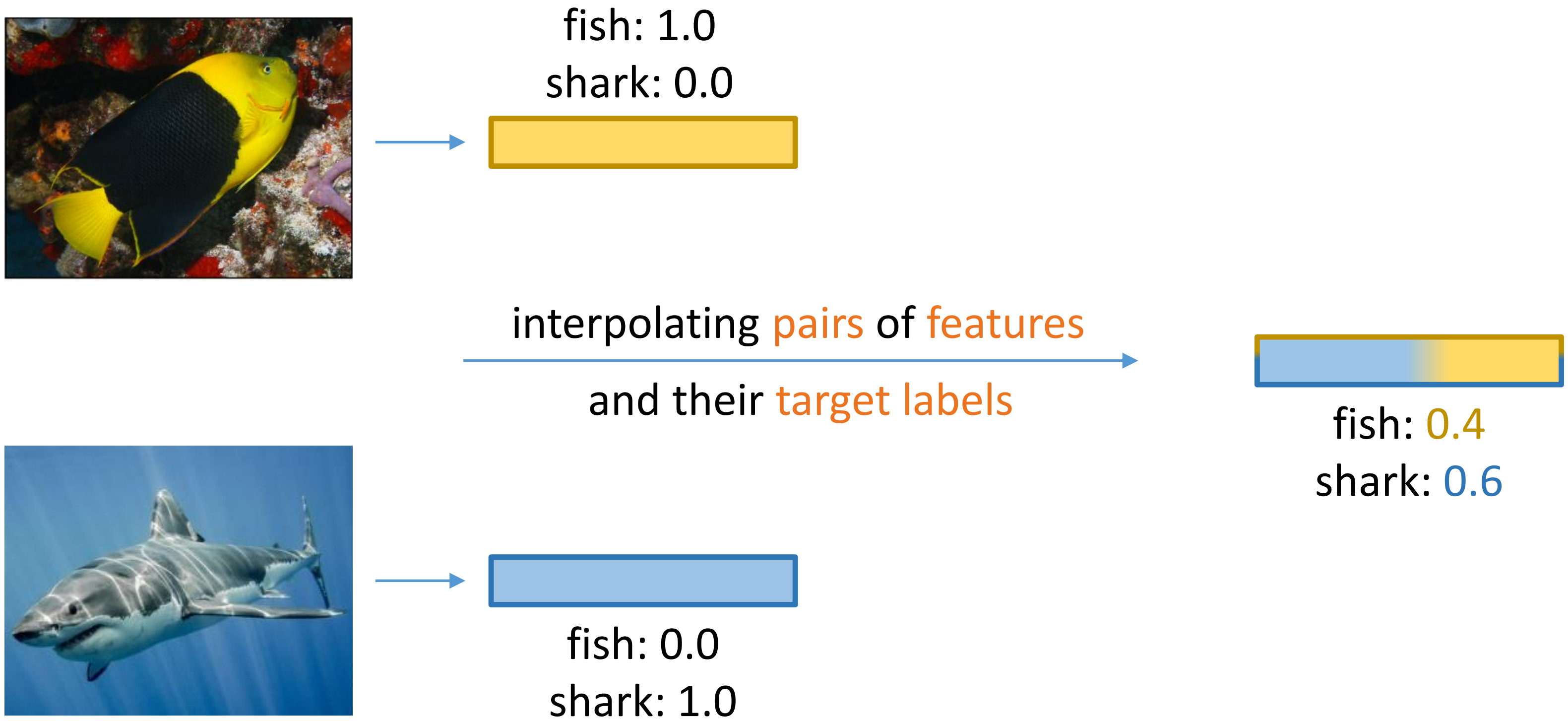
x_{mix}

fish: 0.4
shark: 0.6

y_{mix}

Zhang et al., mixup: Beyond Empirical Risk Minimization, ICLR 2018

Manifold Mixup: Advanced Data Augmentation



fish: 1.0
shark: 0.0

interpolating pairs of features

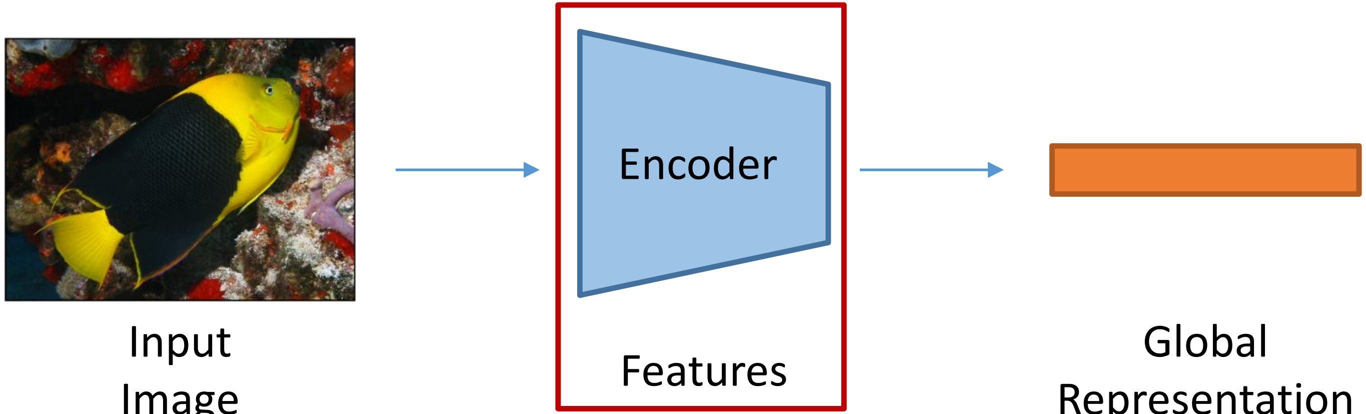
and their target labels

fish: 0.4
shark: 0.6

fish: 0.0
shark: 1.0

Verma et al., Manifold Mixup: Better Representations by Interpolating Hidden States, ICML 2018

Representation Learning: Focus on the Encoder



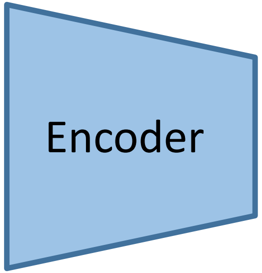
Input Image

Encoder

Features

Global Representation

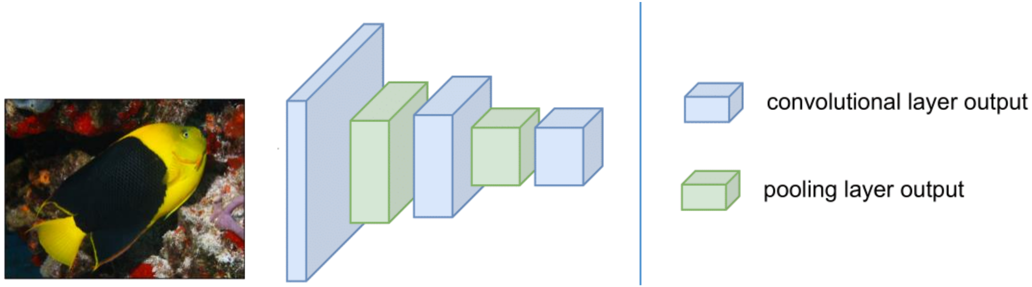
Encoder: Convolutional Neural Network



Encoder

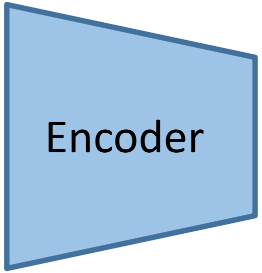
Features

Convolutional Neural Network



Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks, NeurIPS 2012

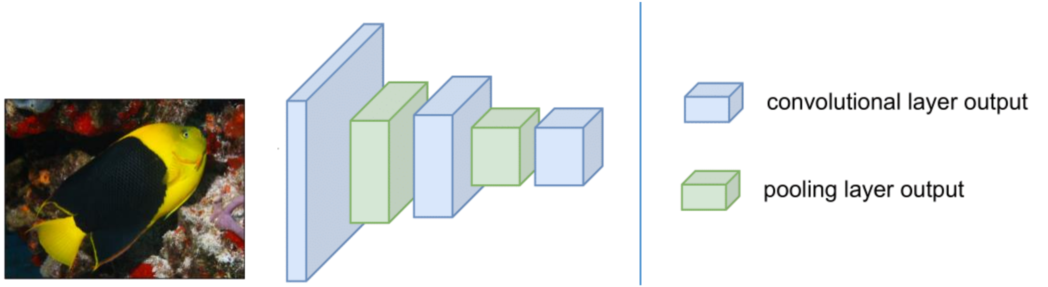
Encoder: Vision Transformer



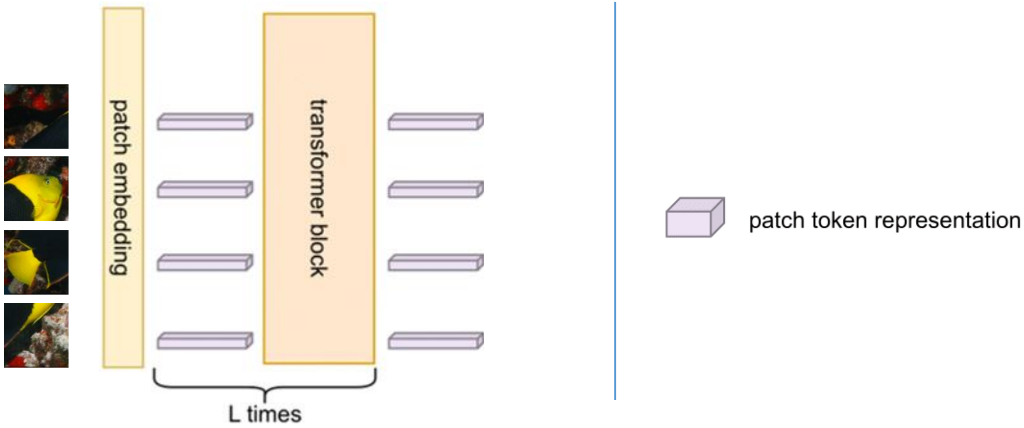
Encoder

Features

Convolutional Neural Network

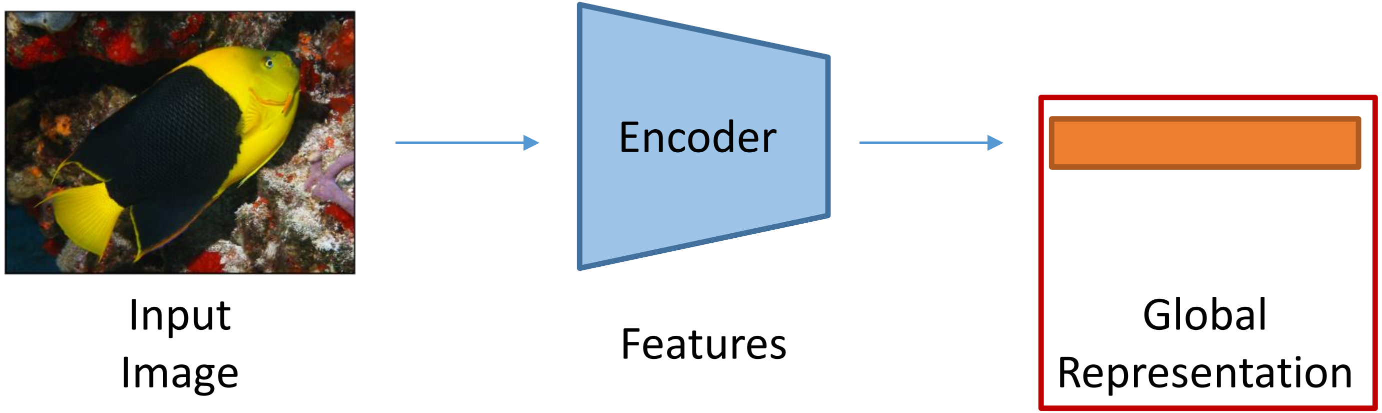


Vision Transformer



Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Representation Learning: Focus on the Global Representation



Input Image

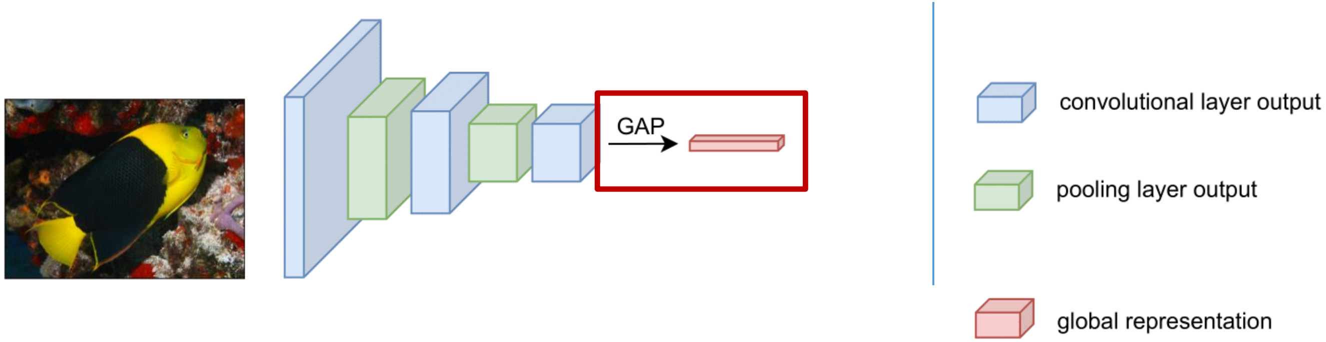
Encoder

Features

Global Representation

Global Representation in CNNs

Convolutional
Neural Network

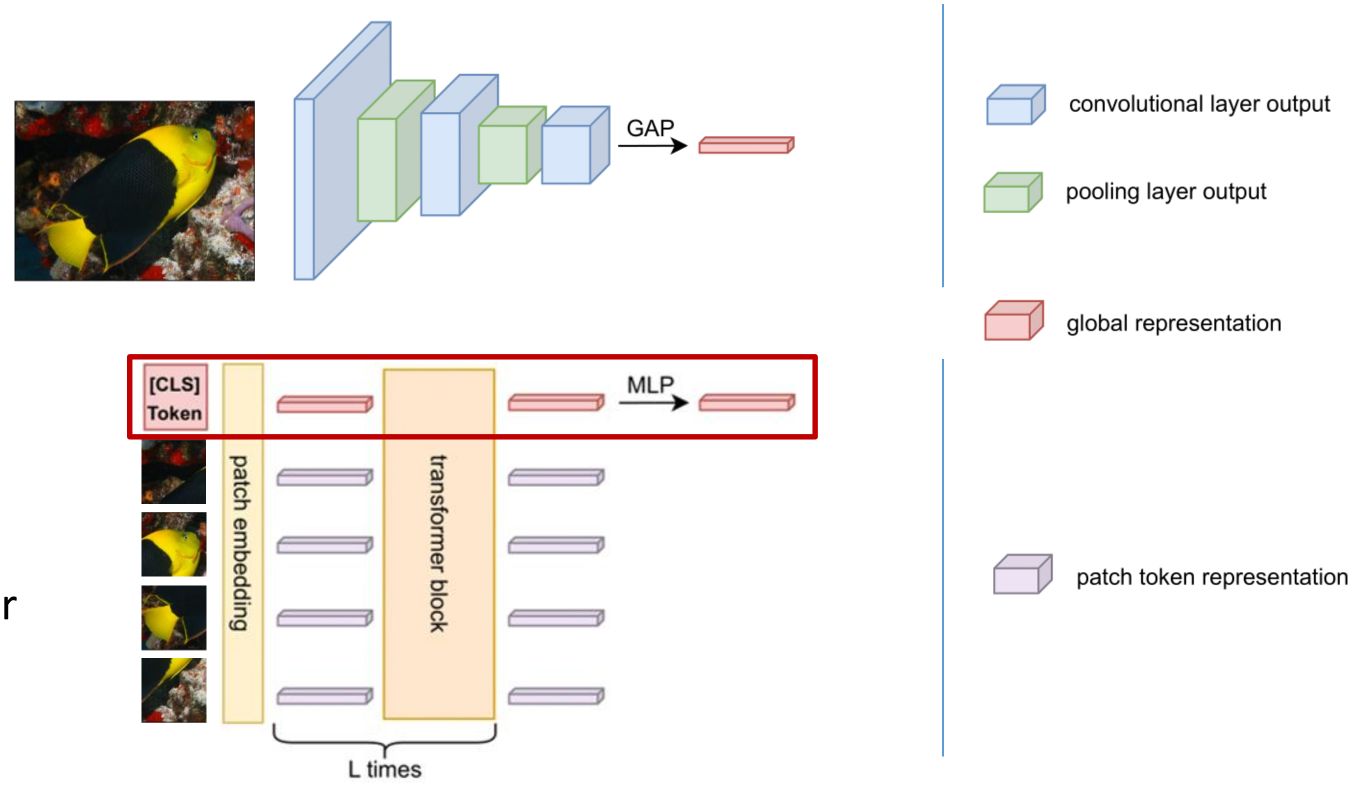


He et al., Deep residual learning for image recognition, CVPR 2016

Global Representation in ViTs

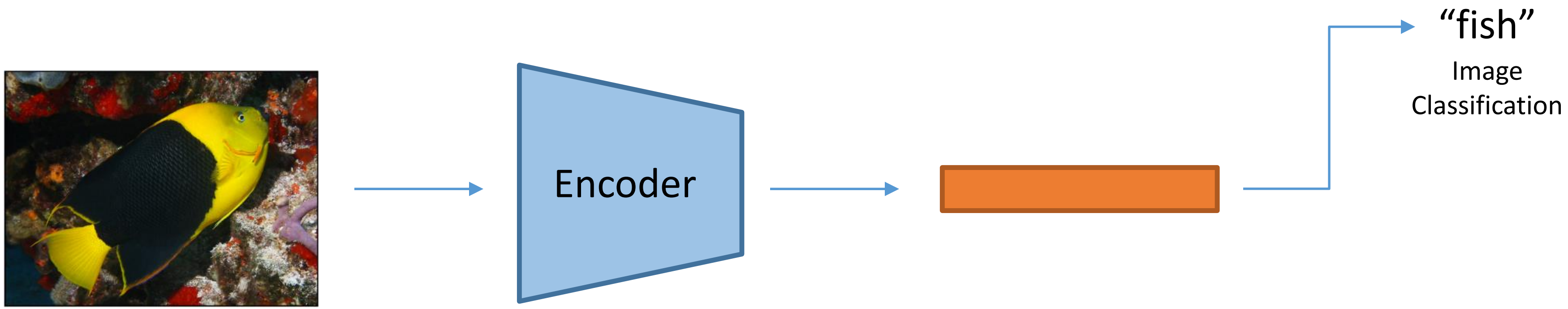
Convolutional Neural Network

Vision Transformer



Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Representation Learning: Tasks



Encoder

“fish”

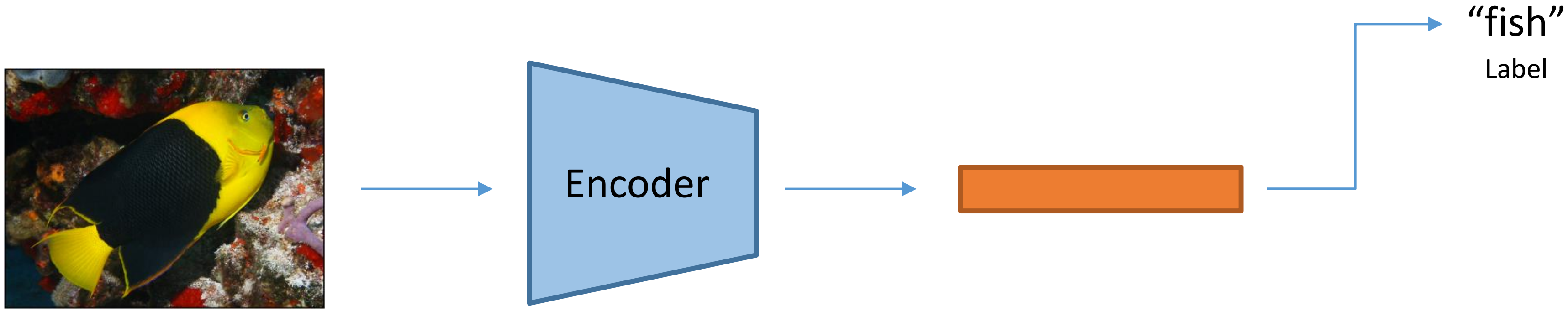
Image Classification

Input Image

Features

Global Representation

Representation Learning: Tasks



Encoder

“fish”

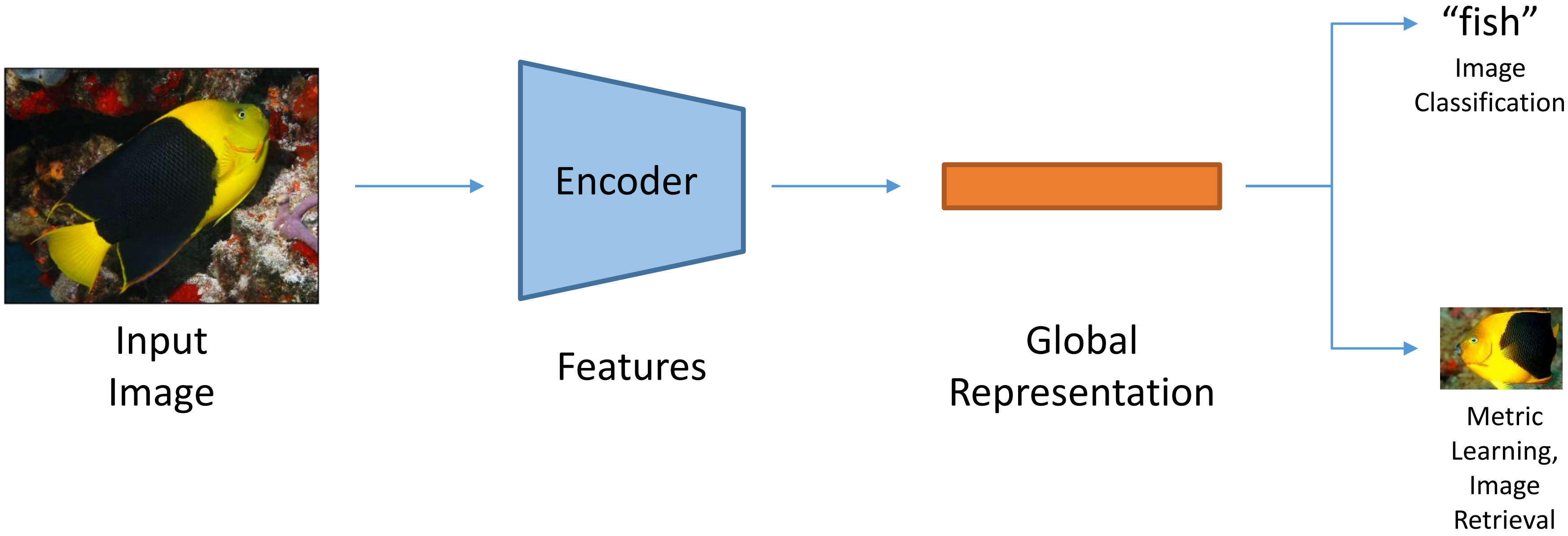
Label

Input Image

Features

Global Representation

Representation Learning: Tasks



Input Image

Encoder

Features

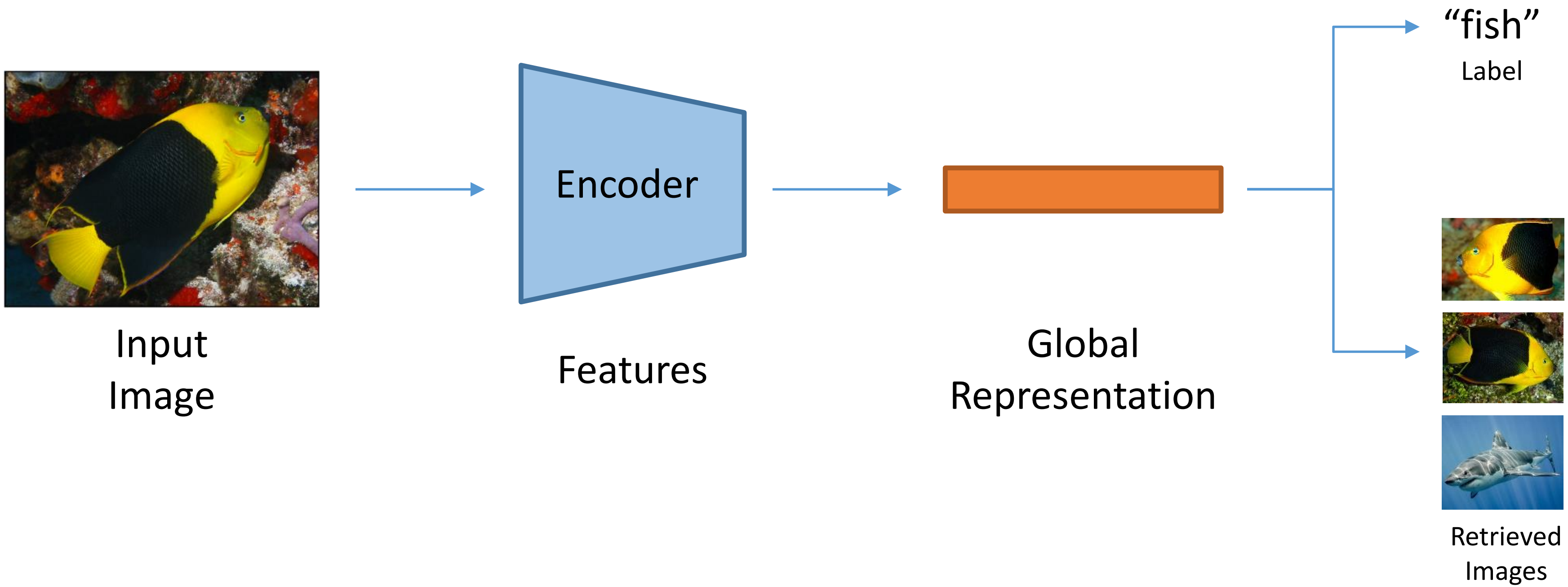
Global Representation

“fish”

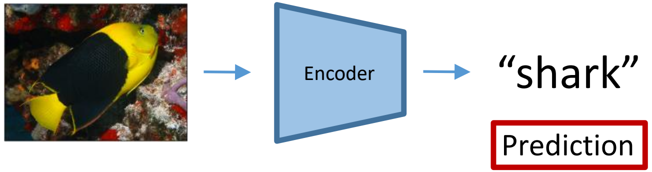
Image Classification

Metric Learning, Image Retrieval

Representation Learning: Tasks



Tasks: Image Classification vs. Metric Learning



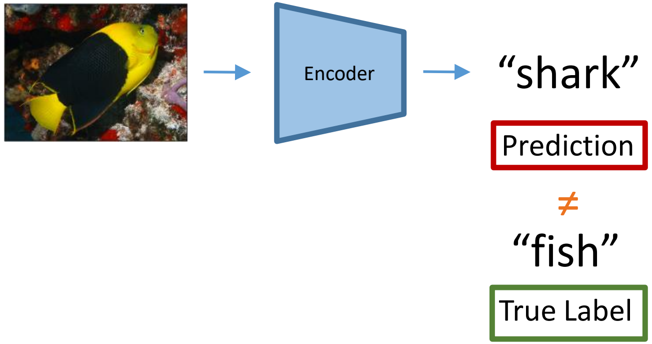
"shark"

Prediction

Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\left(y \cdot \log(\boxed{\hat{y}}) + (1-y) \cdot \log(1-\boxed{\hat{y}})\right)$$

Tasks: Image Classification vs. Metric Learning



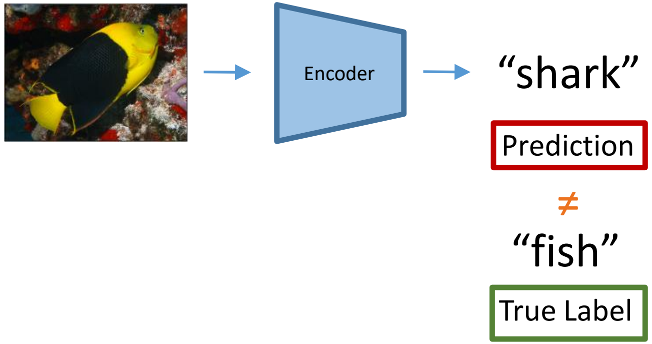
Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\boxed{y}\cdot\log\boxed{(\hat{y})} + (1-\boxed{y})\cdot\log(1-\boxed{\hat{y})})$$

It penalizes the model if the predicted probability ŷ does not match the true label y

Cross Entropy is thus additive over examples

Tasks: Image Classification vs. Metric Learning

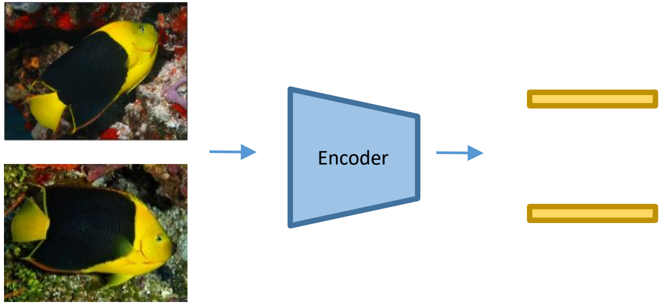


Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\boxed{y}\cdot\log\boxed{(\hat{y})} + (1-\boxed{y)}\cdot\log(1-\boxed{\hat{y})})$$

It penalizes the model if the predicted probability ŷ does not match the true label y

Cross Entropy is thus additive over examples

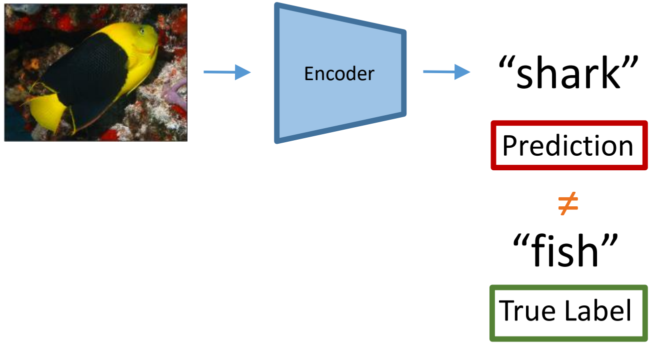


Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

$$\mathcal{L}_{\text{contrastive}} = y\cdot(1-s) + (1-y)\cdot\max(0, s-m)$$

Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR 2006

Tasks: Image Classification vs. Metric Learning

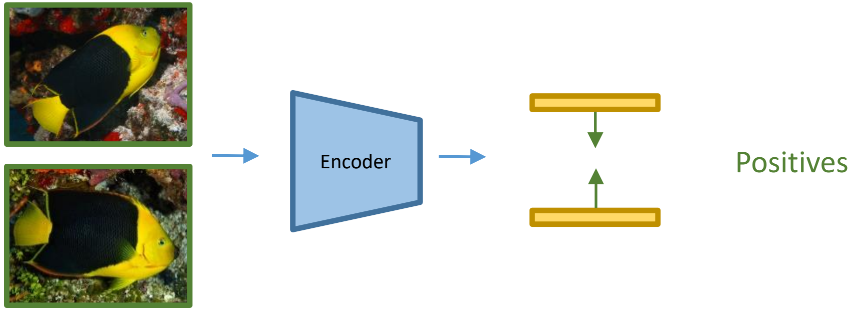


Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\boxed{y}\cdot\log\boxed{(\hat{y})} + (1-\boxed{y})\cdot\log(1-\boxed{\hat{y}})$$

It penalizes the model if the predicted probability ŷ does not match the true label y (here, binary)

Cross Entropy is thus additive over examples



Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

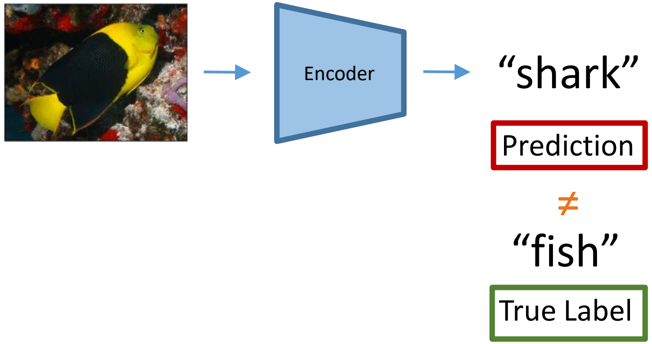
$$\mathcal{L}_{\text{contrastive}} = \boxed{y\cdot(1-s)} + (1-y)\cdot\max(0, s-m)$$

It penalizes the model when the similarity s between two examples does not align with their label y, pulling positives closer together

Contrastive is thus non-additive over examples

Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR 2006

Tasks: Image Classification vs. Metric Learning

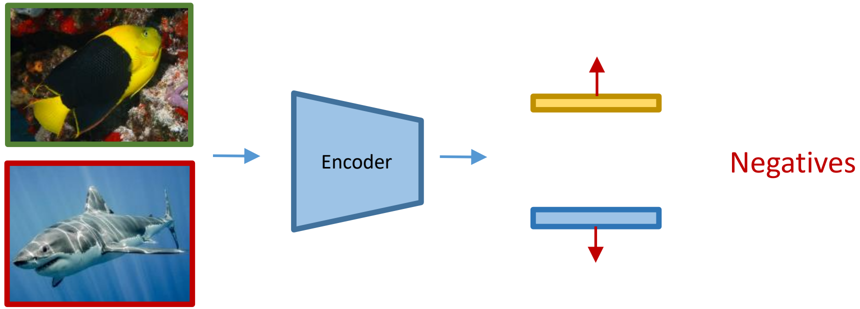


Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\boxed{y}\cdot\log\boxed{(\hat{y})} + (1-\boxed{y})\cdot\log(1-\boxed{\hat{y}})$$

It penalizes the model if the predicted probability ŷ does not match the true label y (here, binary)

Cross Entropy is thus additive over examples



Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

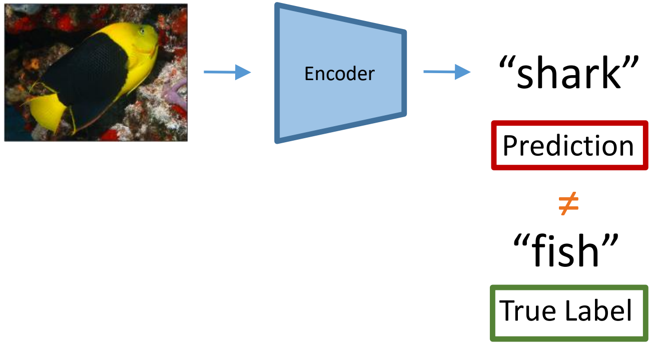
$$\mathcal{L}_{\text{contrastive}} = y\cdot(1-s) + \boxed{(1-y)\cdot\max(0, s-m)}$$

It penalizes the model when the similarity s between two examples does not align with their label y, pushing negatives further apart

Contrastive is thus non-additive over examples

Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR 2006

Tasks: Image Classification vs. Metric Learning

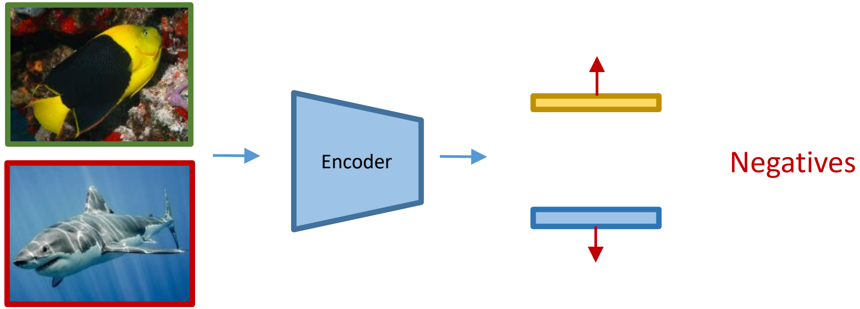


Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\boxed{y}\cdot\log\boxed{(\hat{y})} + (1-\boxed{y})\cdot\log(1-\boxed{\hat{y}})$$

It penalizes the model if the predicted probability ŷ does not match the true label y (here, binary)

Classes are exactly the same in training and test



Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

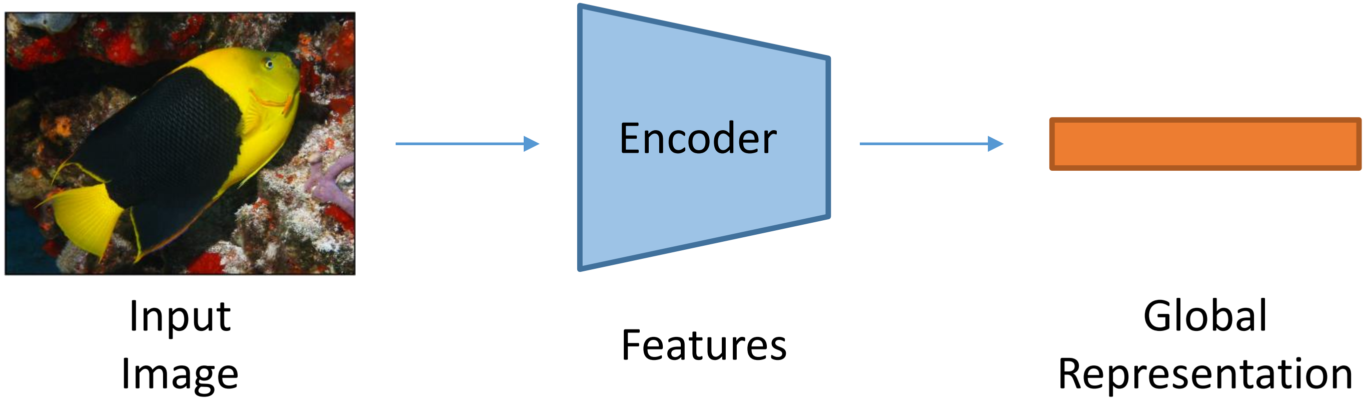
$$\mathcal{L}_{\text{contrastive}} = y\cdot(1-s) + \boxed{(1-y)\cdot\max(0, s-m)}$$

It penalizes the model when the similarity s between two examples does not align with their label y, pushing negatives further apart

Classes are different in training and test (zero-shot recognition)

Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR 2006

Visual Representation Learning: Graphical Abstract



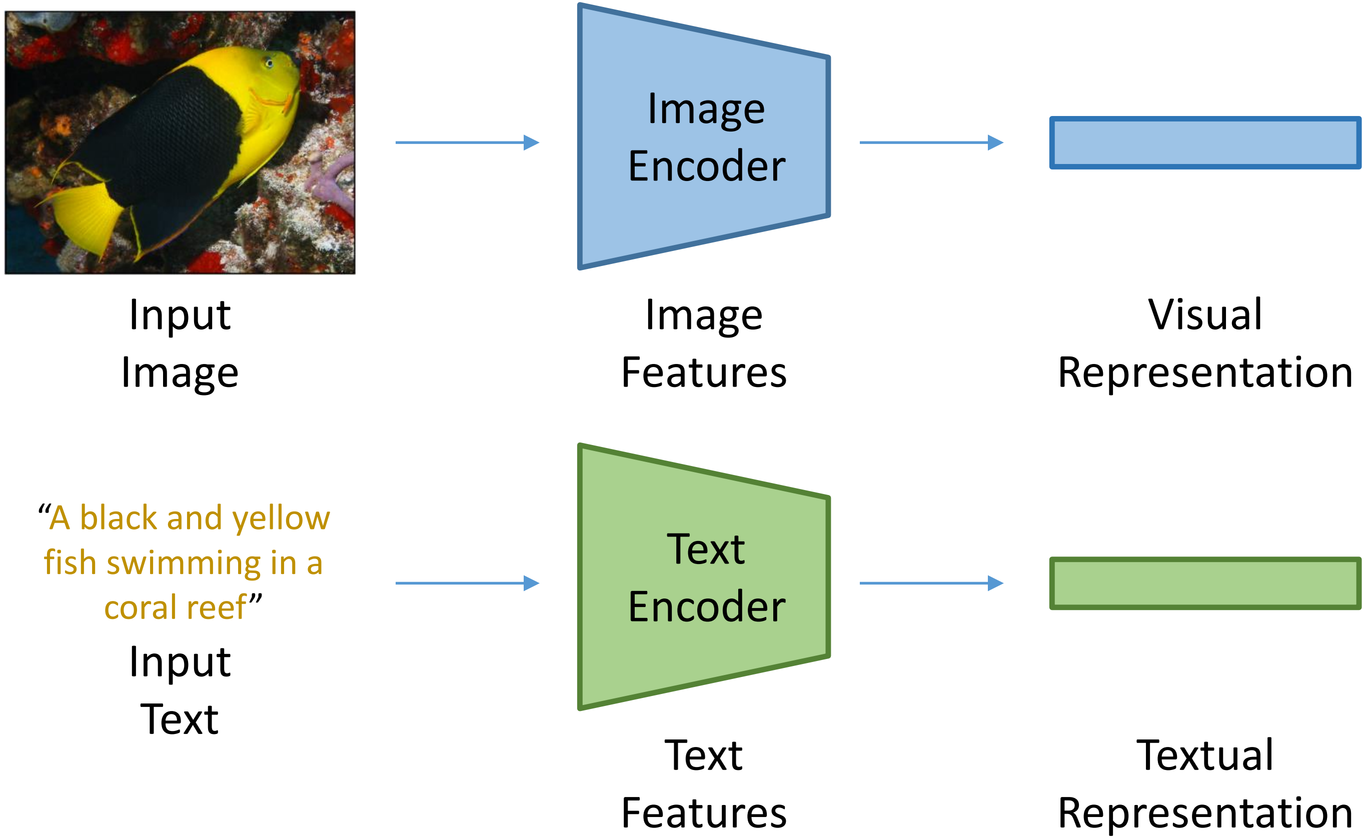
Encoder

Input
Image

Features

Global
Representation

Multimodal Representation Learning: Graphical Abstract



Input Image

Image Encoder

Image Features

Visual Representation

“A black and yellow fish swimming in a coral reef”

Input Text

Text Encoder

Text Features

Textual Representation

Multimodal Representation Learning: Full Graphical Abstract

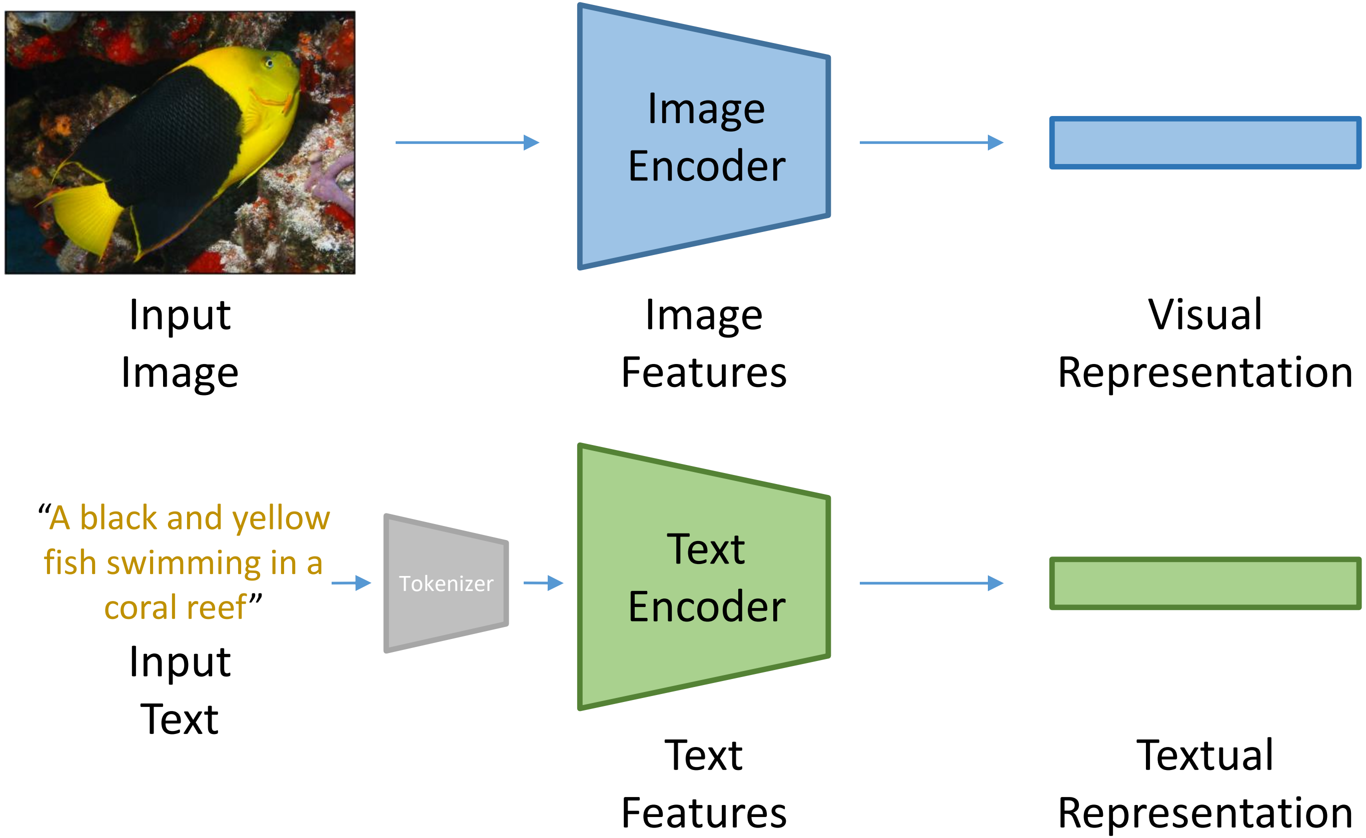


Image Encoder

Input Image

Image Features

Visual Representation

“A black and yellow fish swimming in a coral reef”

Input Text

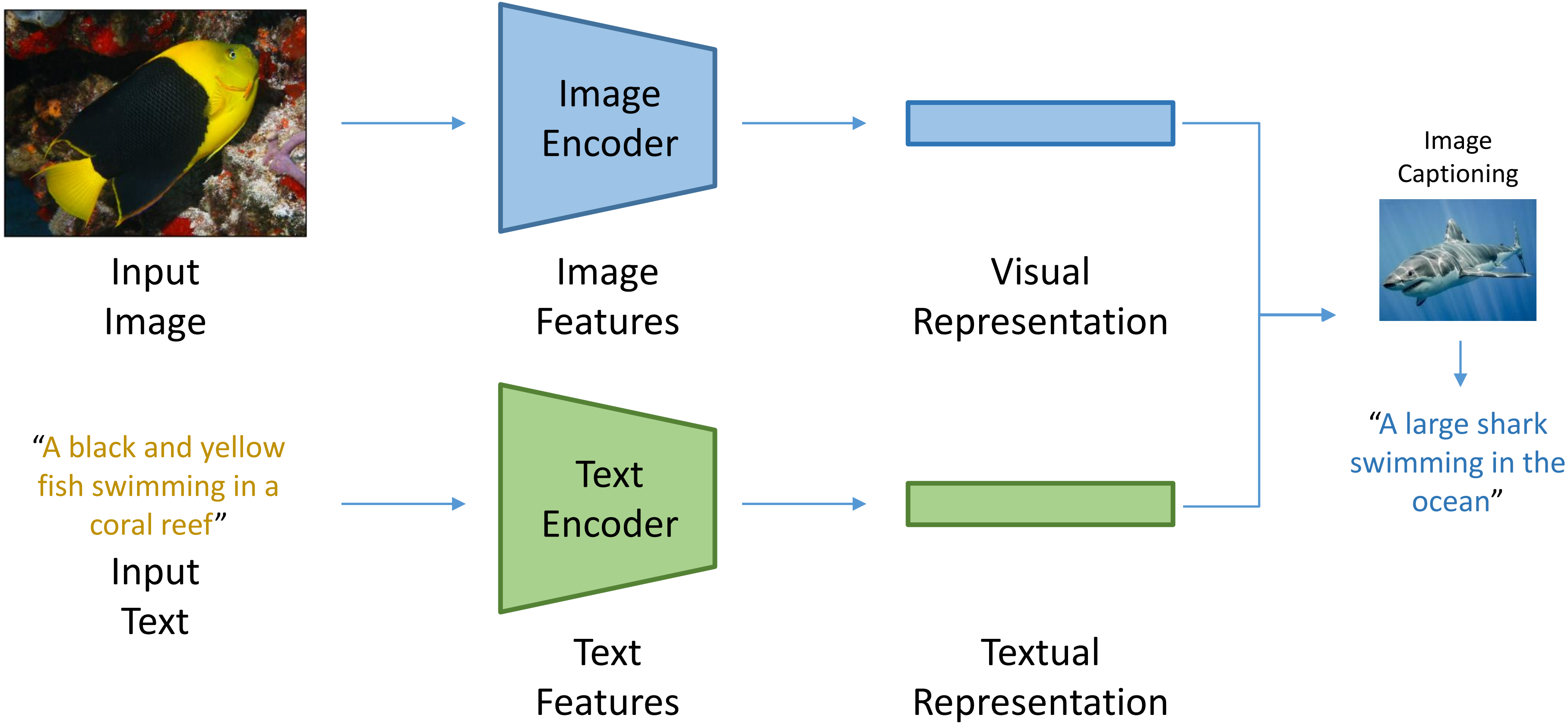
Tokenizer

Text Encoder

Text Features

Textual Representation

Multimodal Representation Learning: Tasks



Input Image

Image Encoder

Image Features

Visual Representation

“A black and yellow fish swimming in a coral reef”

Input Text

Text Encoder

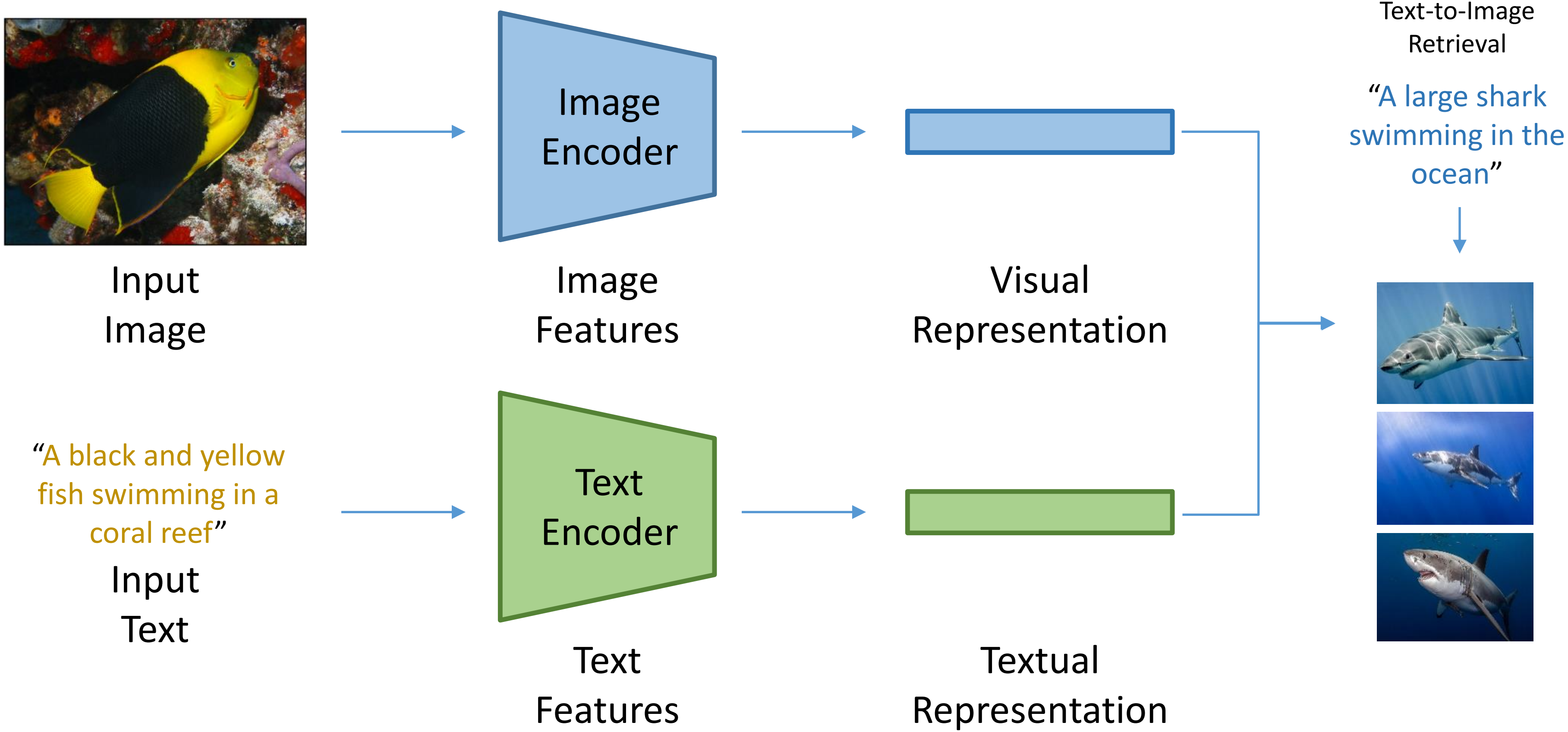
Text Features

Textual Representation

Image Captioning

“A large shark swimming in the ocean”

Multimodal Representation Learning: Tasks



Input Image

Image Encoder

Image Features

Visual Representation

“A black and yellow fish swimming in a coral reef”

Input Text

Text Encoder

Text Features

Textual Representation

Text-to-Image Retrieval

“A large shark swimming in the ocean”

CLIP: Contrastive Language-Image Pretraining

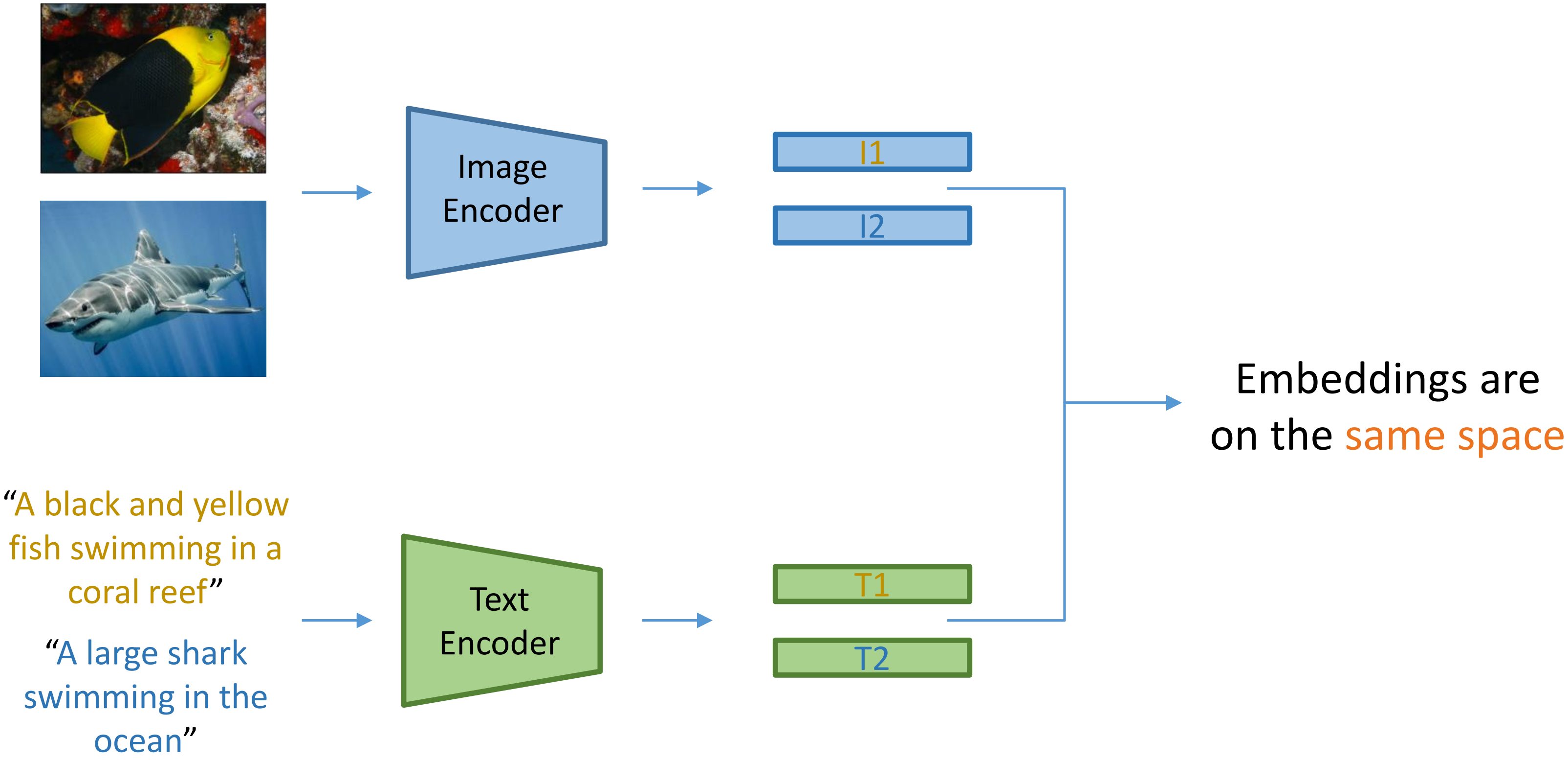


Image Encoder

I1

I2

“A black and yellow fish swimming in a coral reef”

“A large shark swimming in the ocean”

Text Encoder

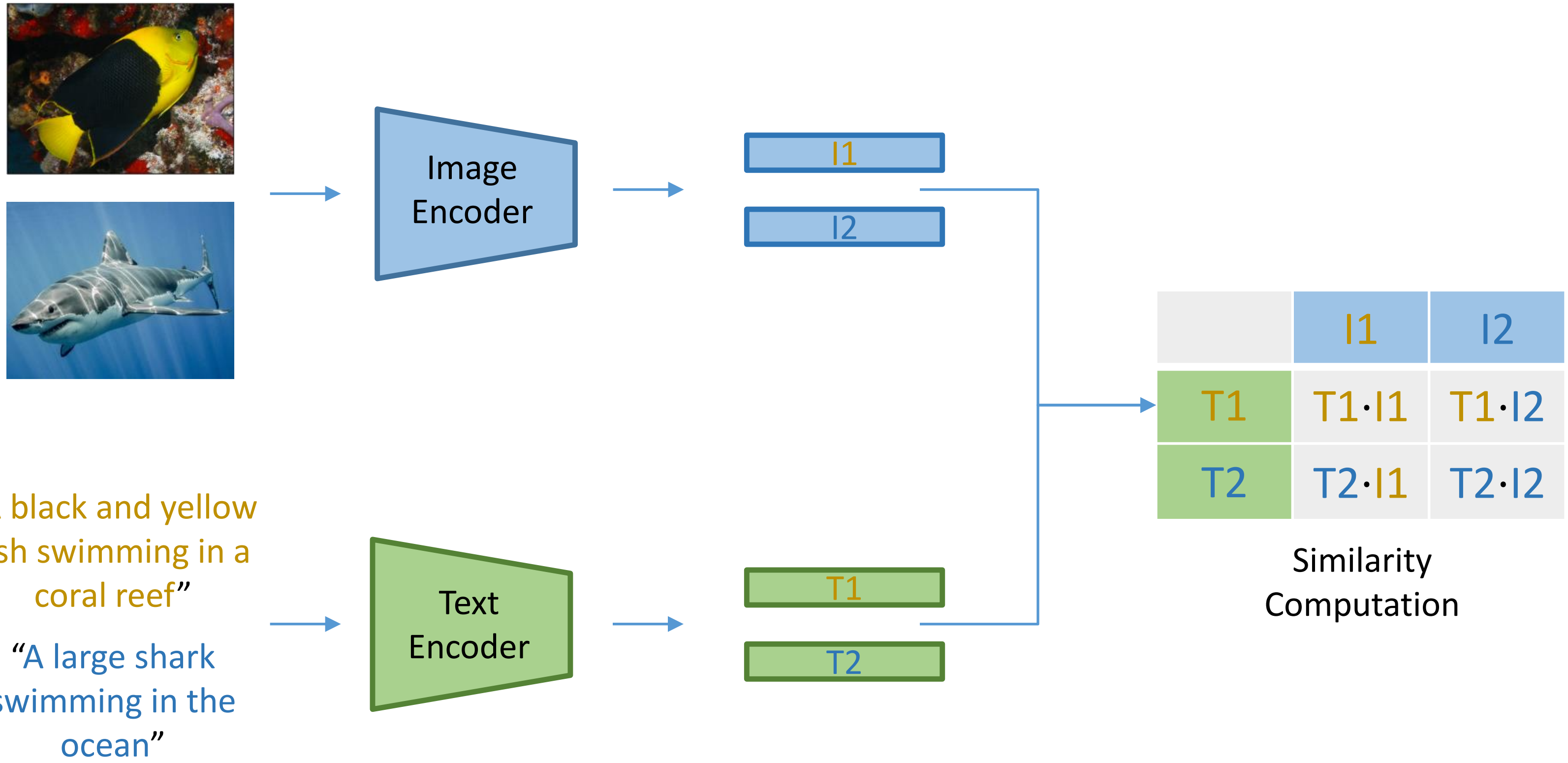
T1

T2

Embeddings are on the same space

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

CLIP: Contrastive Language-Image Pretraining



“A black and yellow fish swimming in a coral reef”

“A large shark swimming in the ocean”

	I1	I2
T1	T1·I1	T1·I2
T2	T2·I1	T2·I2

Similarity Computation

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

CLIP: Contrastive Language-Image Pretraining

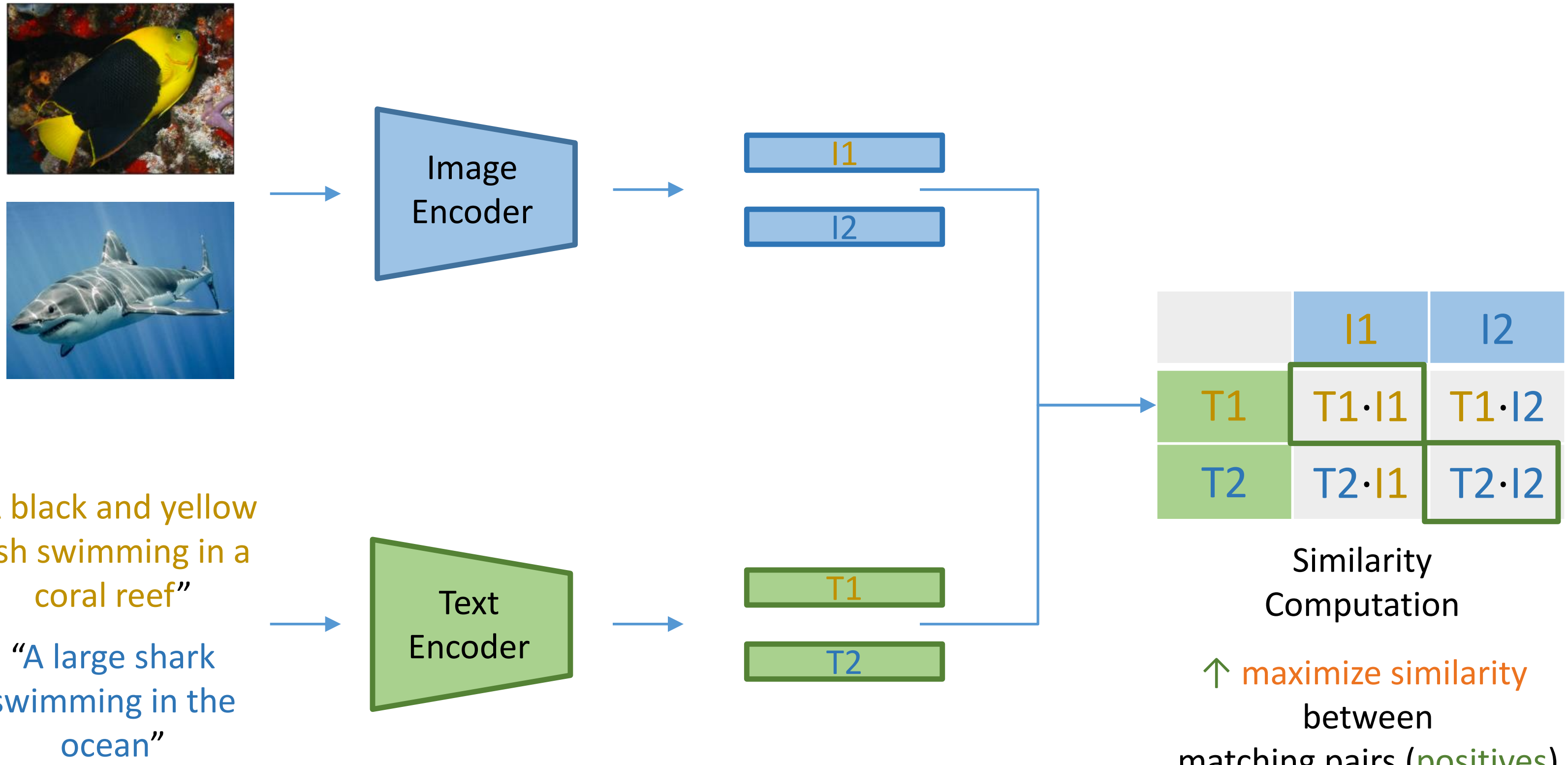


Image Encoder
I1
I2
“A black and yellow fish swimming in a coral reef”
“A large shark swimming in the ocean”
Text Encoder
T1
T2

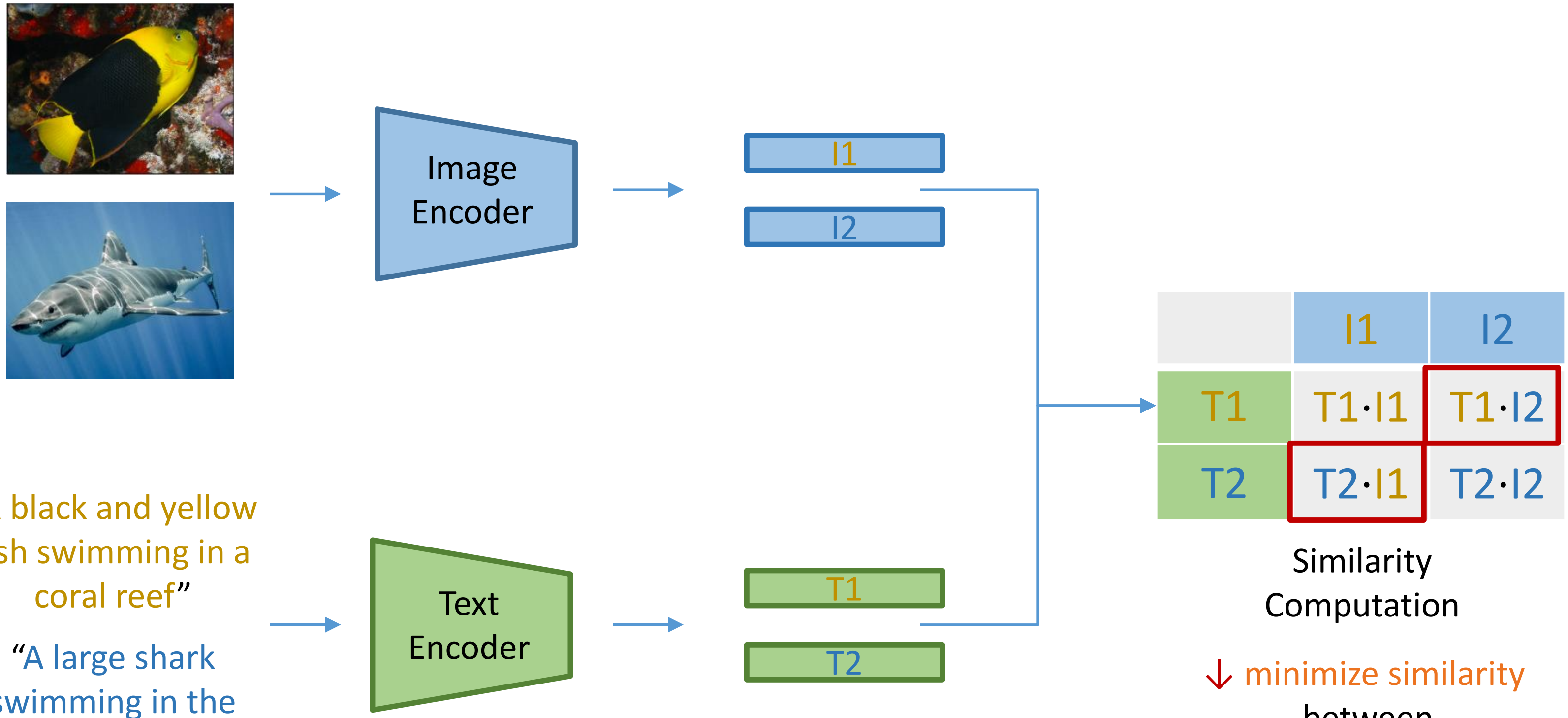
	I1	I2
T1	T1·I1	T1·I2
T2	T2·I1	T2·I2

Similarity Computation

↑ maximize similarity between matching pairs (positives)

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

CLIP: Contrastive Language-Image Pretraining



"A black and yellow fish swimming in a coral reef"

"A large shark swimming in the ocean"

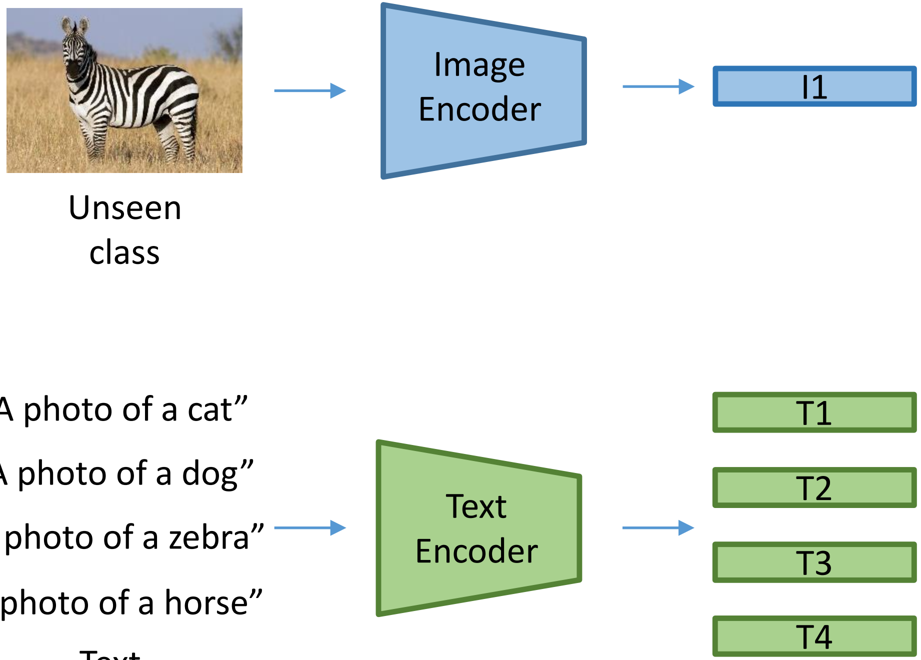
	I1	I2
T1	T1·I1	T1·I2
T2	T2·I1	T2·I2

Similarity Computation

↓ minimize similarity between non-matching pairs (negatives)

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Zero-Shot Recognition with CLIP



Unseen class

Image Encoder

I1

"A photo of a cat"

"A photo of a dog"

"A photo of a zebra"

"A photo of a horse"

Text descriptions

Text Encoder

T1

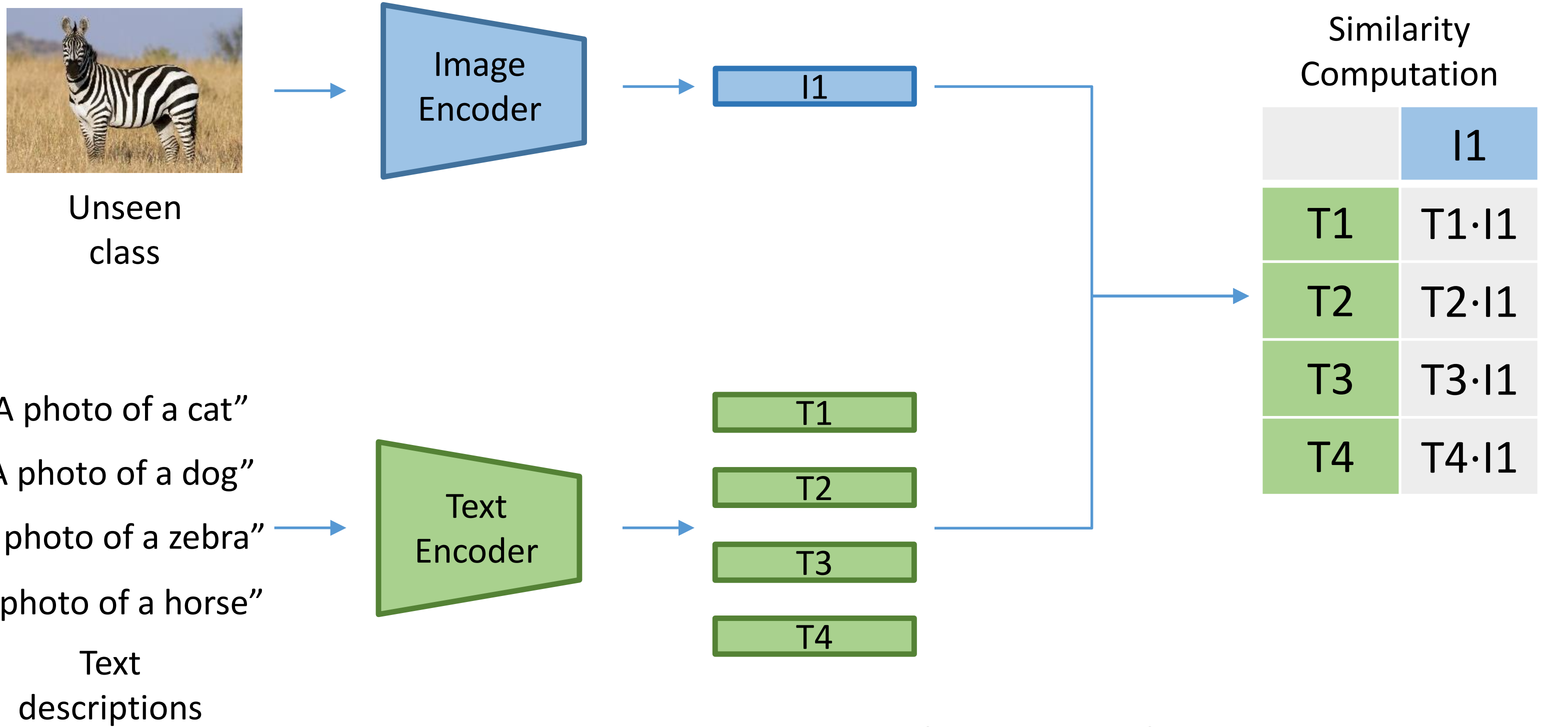
T2

T3

T4

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Zero-Shot Recognition with CLIP



Unseen class

Image Encoder

I1

“A photo of a cat”

“A photo of a dog”

“A photo of a zebra”

“A photo of a horse”

Text descriptions

Text Encoder

T1

T2

T3

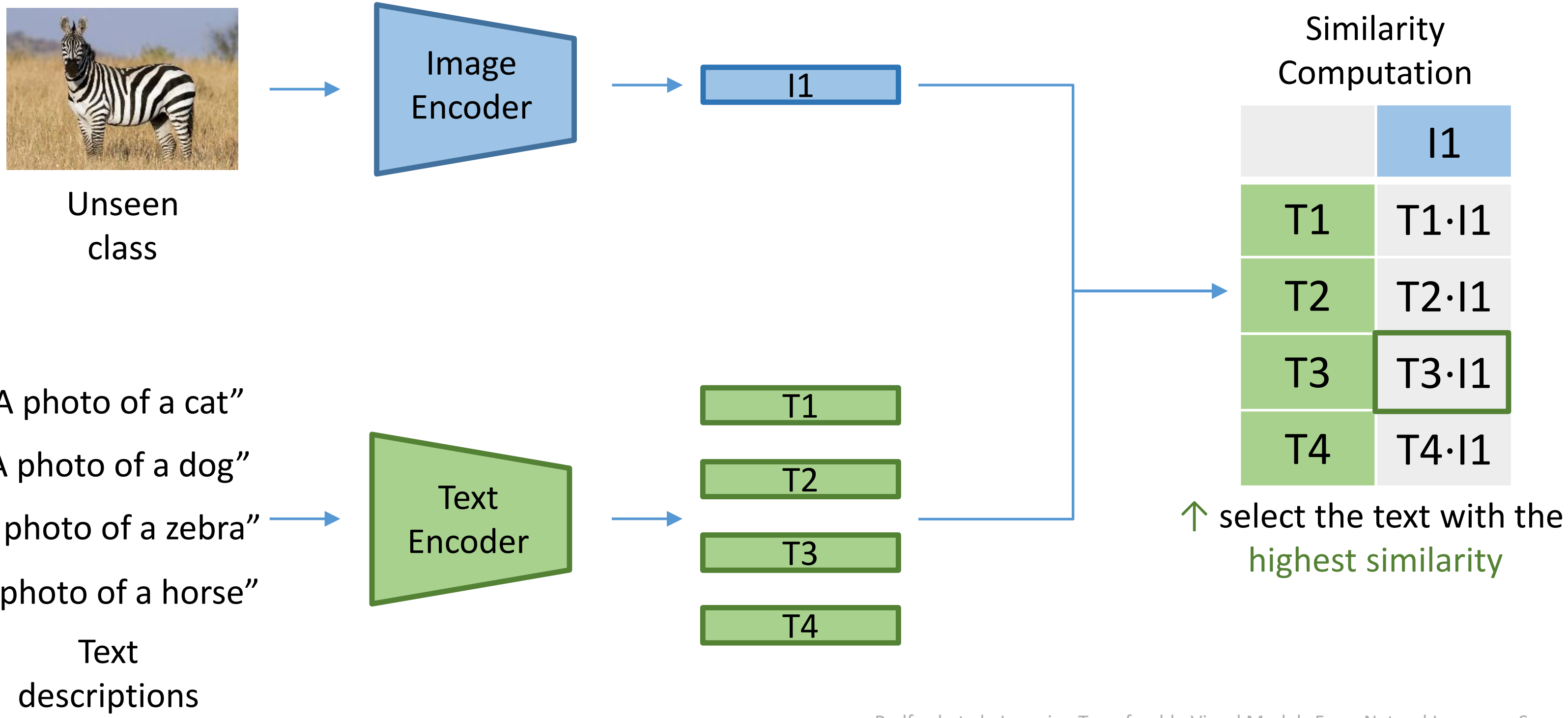
T4

Similarity Computation

	I1
T1	T1·I1
T2	T2·I1
T3	T3·I1
T4	T4·I1

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Zero-Shot Recognition with CLIP

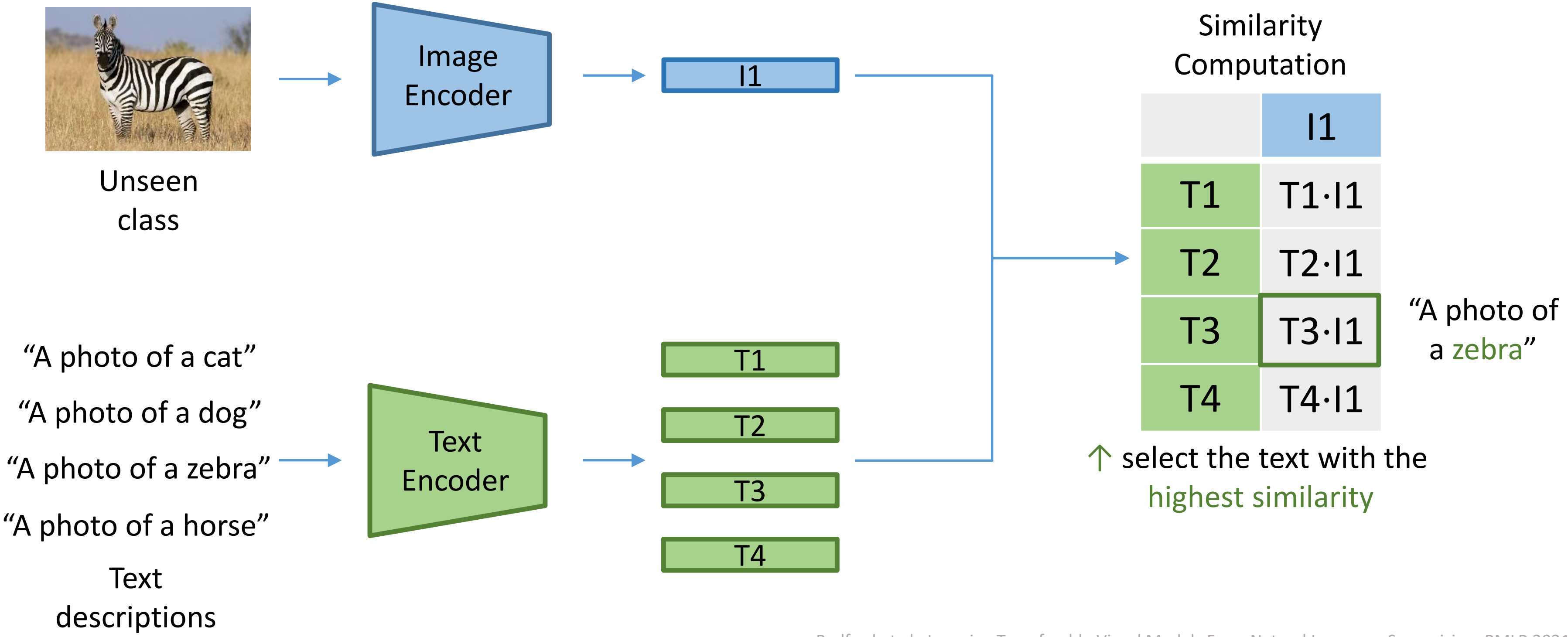


	I1
T1	T1·I1
T2	T2·I1
T3	T3·I1
T4	T4·I1

↑ select the text with the highest similarity

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Zero-Shot Recognition with CLIP



Unseen class

Image Encoder

I1

"A photo of a cat"

"A photo of a dog"

"A photo of a zebra"

"A photo of a horse"

Text descriptions

Text Encoder

T1

T2

T3

T4

Similarity Computation

	I1
T1	T1·I1
T2	T2·I1
T3	T3·I1
T4	T4·I1

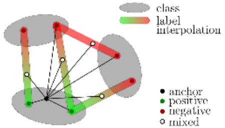
"A photo of a zebra"

↑ select the text with the highest similarity

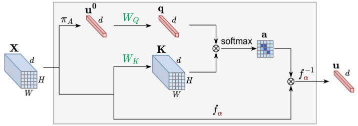
Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Main Objectives

1. Address the challenges of learning and improving visual representations from a:

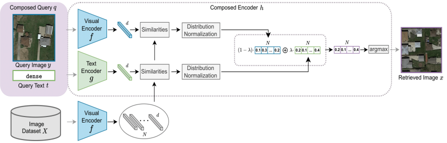


Data-centric perspective via advanced data augmentation (mixup)

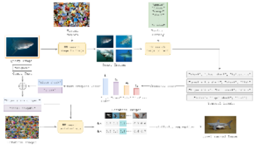


Model-centric perspective via model architecture component (pooling)

2. Leverage the multimodal capabilities of a pre-trained, frozen VLM to:

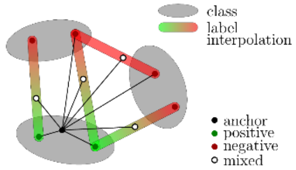


Introduce a new task into Remote Sensing and a new flexible method



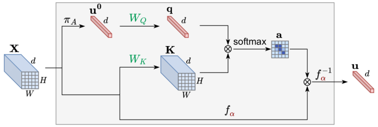
Expand the task of domain conversion in composed image retrieval
and introduce a discrete-space memory-based textual inversion method

Main Contributions



✓ Develop a generic way of representing and interpolating labels, allowing the straightforward extension of any kind on mixup to metric learning.

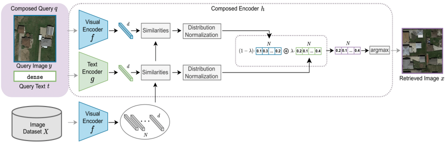
✓ Introduce and systematically evaluate a novel mixup method (Metrix).



✓ Formulate a generic pooling framework, allowing easy inspection of a wide range of methods. Utilize it to derive an attention-based pooling (SimPool)

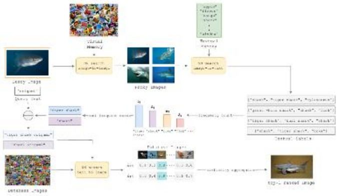
✓ Quantitatively evaluate the attention maps of SimPool through experiments using them for object localization and object discovery.

Main Contributions



✓ Introduce a new task, Remote Sensing Composed Image Retrieval, accompanied by a new benchmark dataset, PatternCom, to facilitate evaluation.

✓ Develop a training-free method, WeiCom, leveraging a pre-trained, frozen VLM, utilizing a control parameter for more image- or text-oriented search results.



✓ Expand the task of domain conversion in composed image retrieval, by introducing new benchmark datasets.

✓ Develop a training-free, discrete-space memory-based textual inversion method, FreeDom, leveraging a pre-trained, frozen VLM.

2. Visual Representations

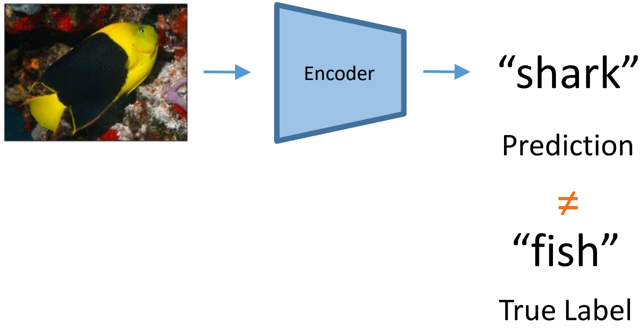
Learning Visual Representations via Data Augmentation [Metrix]

Venkataramanan*, **Psomas*** et al. *It Takes Two to Tango: Mixup for Deep Metric Learning*, **ICLR 2022**

Source code: https://github.com/billpsomas/metrix

*equal contribution

Recap: Image Classification vs. Metric Learning


Encoder

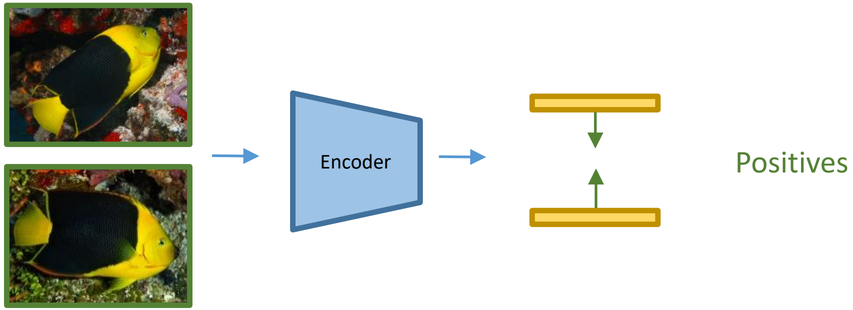
“shark”

Prediction

≠

“fish”

True Label


Encoder

Positives

Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -\left(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right)$$

Cross Entropy is thus additive over examples

Classes are exactly the same in training and test

Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

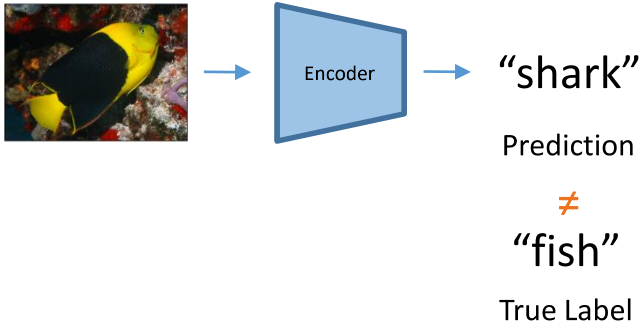
$$\mathcal{L}_{\text{contrastive}} = y \cdot (1-s) + (1-y) \cdot \max(0, s-m)$$

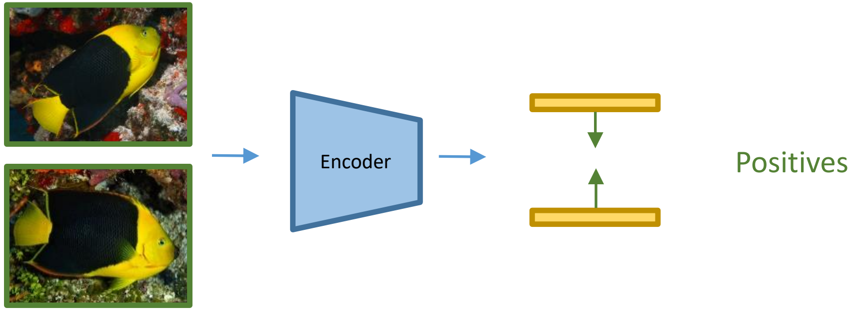
Contrastive is thus non-additive over examples

Classes are different in training and test

Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR 2006

Motivation: Mixup for Metric Learning?





Loss function (Cross Entropy) considers one example at a time, independently of others

$$\mathcal{L}_{\text{CE}} = -(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y}))$$

Cross Entropy is thus additive over examples

Classes are exactly the same in training and test

Loss functions (e.g. Contrastive) do not consider single examples (i.e. pairs)

$$\mathcal{L}_{\text{contrastive}} = y \cdot (1-s) + (1-y) \cdot \max(0, s-m)$$

Contrastive is thus non-additive over examples

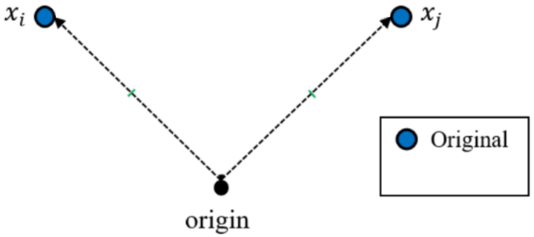
Classes are different in training and test

Motivation: Mixup should play a significant role!

Related Work: Mixup for Metric Learning

Motivation: Mixup should play a significant role!

Embedding Expansion

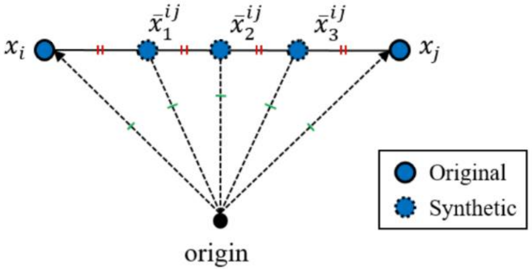


Ko & Gu, Augmentation in embedding space for deep metric learning, CVPR 2020
Gu et al., Proxy synthesis: Learning with synthetic classes for deep metric learning, AAAI 2021
Kalantidis et al., Hard negative mixing for contrastive learning, NeurIPS 2020

Related Work: Mixup for Metric Learning

Motivation: Mixup should play a significant role!

Embedding Expansion



Interpolates pairs of embeddings in a deterministic way within the same class (positives).

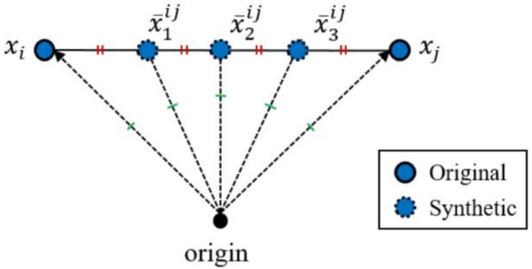
Interpolates only positives; does not interpolate labels.

Ko & Gu, Augmentation in embedding space for deep metric learning, CVPR 2020
Gu et al., Proxy synthesis: Learning with synthetic classes for deep metric learning, AAAI 2021
Kalantidis et al., Hard negative mixing for contrastive learning, NeurIPS 2020

Related Work: Mixup for Metric Learning

Motivation: Mixup should play a significant role!

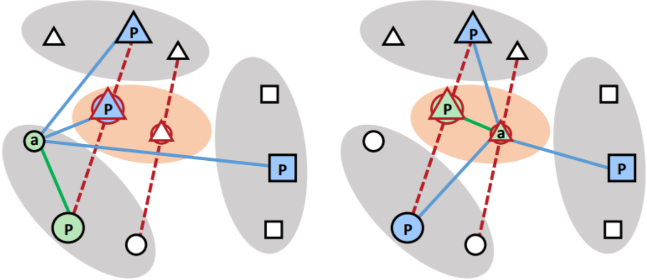
Embedding Expansion



Interpolates pairs of embeddings in a deterministic way within the same class (positives).

Interpolates only positives; does not interpolate labels.

Proxy Synthesis



Interpolates between classes, applying to proxy-based losses only.

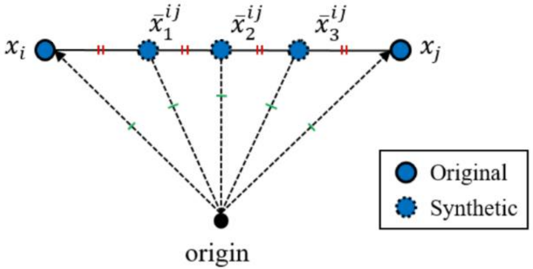
Risks synthesizing false negatives when the interpolation factor λ is close to 0 or 1.

Ko & Gu, Augmentation in embedding space for deep metric learning, CVPR 2020
Gu et al., Proxy synthesis: Learning with synthetic classes for deep metric learning, AAAI 2021
Kalantidis et al., Hard negative mixing for contrastive learning, NeurIPS 2020

Related Work: Mixup for Metric Learning

Motivation: Mixup should play a significant role!

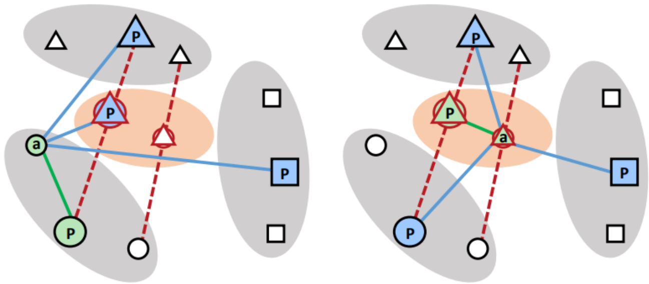
Embedding Expansion



Interpolates pairs of embeddings in a deterministic way within the same class (positives).

Interpolates only positives; does not interpolate labels.

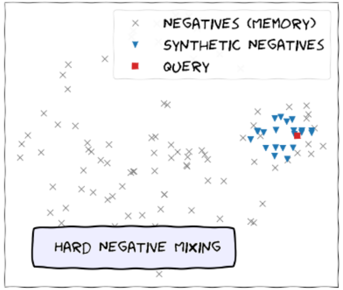
Proxy Synthesis



Interpolates between classes, applying to proxy-based losses only.

Risks synthesizing false negatives when the interpolation factor λ is close to 0 or 1.

MoCHi



Interpolates anchor with negative embeddings.

Does not interpolate labels, chooses λ in [0, 0.5] to avoid false negatives.

Ko & Gu, Augmentation in embedding space for deep metric learning, CVPR 2020
Gu et al., Proxy synthesis: Learning with synthetic classes for deep metric learning, AAAI 2021
Kalantidis et al., Hard negative mixing for contrastive learning, NeurIPS 2020

Why Mixup is difficult in Metric Learning?

Why Mixup is difficult in Metric Learning?

Recap!

Cross Entropy is additive over examples

Contrastive is non-additive over examples

Why Mixup is difficult in Metric Learning?

Recap!

Cross Entropy is additive over examples

Contrastive is non-additive over examples

Mixup involves linear interpolation between examples and their labels

Why Mixup is difficult in Metric Learning?

Recap!

Cross Entropy is additive over examples

Contrastive is non-additive over examples

Mixup involves linear interpolation between examples and their labels

In Classification, we can easily interpolate between two labels, e.g., for y=1 and y'=0, we can mix them as $\lambda y + (1-\lambda) y'$.

Why Mixup is difficult in Metric Learning?

Recap!

Cross Entropy is additive over examples

Contrastive is non-additive over examples

Mixup involves linear interpolation between examples and their labels

In Classification, we can easily interpolate between two labels, e.g., for $y=1$ and $y'=0$, we can mix them as $\lambda y+(1-\lambda)y'$.

In Metric Learning, we are dealing with binary labels for pairs (positives = 1, negatives = 0).

If we mix a positive and a negative example, how we define the label for the mixed example?

Why Mixup is difficult in Metric Learning?

Recap!

Cross Entropy is additive over examples

Contrastive is non-additive over examples

Mixup involves linear interpolation between examples and their labels

In Classification, we can easily interpolate between two labels, e.g., for $y=1$ and $y'=0$, we can mix them as $\lambda y + (1-\lambda) y'$.

In Metric Learning, we are dealing with binary labels for pairs (positives = 1, negatives = 0).

If we mix a positive and a negative example, how we define the label for the mixed example?

Problem with label interpolation!

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(\boxed{a,}x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(\boxed{a,}x))\right)\right)$$

Anchor point from the training set

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{\boxed{(x,y)}\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{\boxed{(x,y)}\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

Pair of an example and its binary label
(y=1 for positives, y=0 for negatives)

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in \boxed{U(a)}} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in \boxed{U(a)}} (1-y)\rho^{-}(s(a,x))\right)\right)$$

The union of positives and negatives of anchor

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^+\left(\sum_{(x,y)\in U(a)} y\boxed{\rho^+(s(a,x))}\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)\right)$$

A decreasing function of similarity between anchor and positive example

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\boxed{\rho^{-}(s(a,x))}\right)\right)$$

An increasing function of similarity between anchor and negative example

Towards Solution: The Generic Loss Formulation

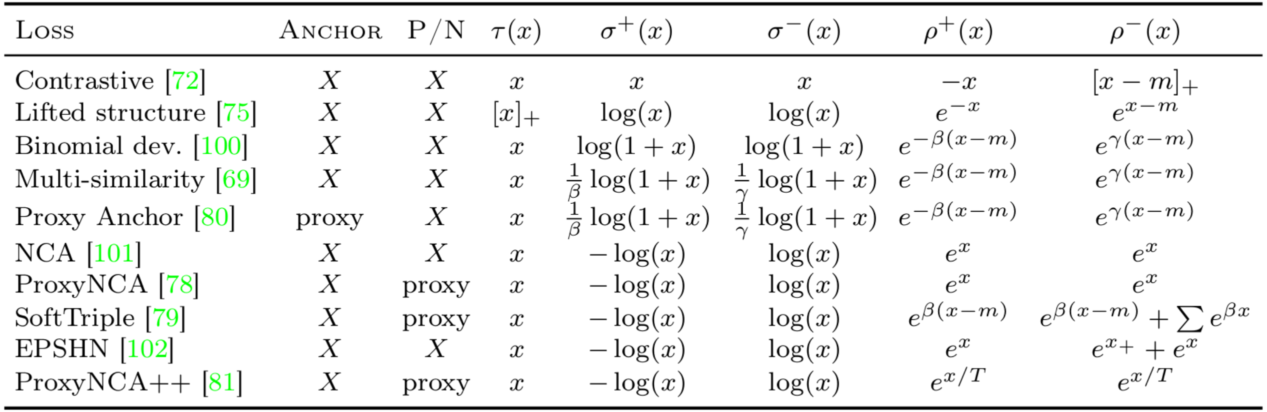
$$\ell(a;\theta) := \boxed{\tau}\left(\boxed{\sigma^{+}}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \boxed{\sigma^{-}}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

Non-linear functions

Towards Solution: The Generic Loss Formulation

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

This formulation serves as a unifying framework for various existing loss functions



Loss	Anchor	P/N	$\tau(x)$	$\sigma^{+}(x)$	$\sigma^{-}(x)$	$\rho^{+}(x)$	$\rho^{-}(x)$
Contrastive [72]	X	X	x	x	x	$-x$	$[x-m]_{+}$
Lifted structure [75]	X	X	$[x]_{+}$	$\log(x)$	$\log(x)$	e^{-x}	e^{x-m}
Binomial dev. [100]	X	X	x	$\log(1+x)$	$\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
Multi-similarity [69]	X	X	x	$\frac{1}{\beta}\log(1+x)$	$\frac{1}{\gamma}\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
Proxy Anchor [80]	proxy	X	x	$\frac{1}{\beta}\log(1+x)$	$\frac{1}{\gamma}\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
NCA [101]	X	X	x	$-\log(x)$	$\log(x)$	e^{x}	e^{x}
ProxyNCA [78]	X	proxy	x	$-\log(x)$	$\log(x)$	e^{x}	e^{x}
SoftTriple [79]	X	proxy	x	$-\log(x)$	$\log(x)$	$e^{\beta(x-m)}$	$e^{\beta(x-m)}+\sum e^{\beta x}$
EPSHN [102]	X	X	x	$-\log(x)$	$\log(x)$	e^{x}	$e^{x+}+e^{x}$
ProxyNCA++ [81]	X	proxy	x	$-\log(x)$	$\log(x)$	$e^{x/T}$	$e^{x/T}$

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^+\left(\sum_{(x,y)\in U(a)} y\rho^+(s(a,x))\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)\right)$$

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_{\lambda}(x,x'), \mathrm{mix}_{\lambda}(y,y')) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_\lambda\boxed{(x,x')}, \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

Pair of examples

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_\lambda(x,x'), \mathrm{mix}_\lambda\boxed{(y,y')}) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

Pair of labels

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(\boxed{f_\lambda(x,x')}, \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

The mixed embedding of the pair of examples
with interpolation factor λ

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_\lambda(x,x'), \boxed{\mathrm{mix}_\lambda(y,y')}) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

The interpolated label, which is no longer binary
but in [0,1]

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^+\left(\sum_{(x,y)\in U(a)} y\rho^+(s(a,x))\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_\lambda(x,x'), \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in \boxed{M(a)}, \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

The set of pairs of examples to mix

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(x,y)\in U(a)} y\rho^{+}(s(a,x))\right) + \sigma^{-}\left(\sum_{(x,y)\in U(a)} (1-y)\rho^{-}(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_{\lambda}(x,x'), \operatorname{mix}_{\lambda}(y,y')) : ((x,y),(x',y')) \in \boxed{M(a)}, \lambda \sim \operatorname{Beta}(\alpha,\alpha)\}$$

The set of pairs of examples to mix.

We allow mixing between:
positive – positive
positive – negative
negative – negative

The Mixed Loss Function: Achieving Mixup in Metric Learning

There is still no mixup happening here:

$$\ell(a;\theta) := \tau\left(\sigma^+\left(\sum_{(x,y)\in U(a)} y\rho^+(s(a,x))\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)\right)$$

We define the set of labeled mixed embeddings:

$$V(a) := \{(f_\lambda(x,x'), \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in M(a), \lambda \sim \mathrm{Beta}(\alpha,\alpha)\}$$

So, now, the mixed loss function:

$$\widetilde{\ell}(a;\theta) := \tau\left(\sigma^+\left(\sum_{(v,y)\in V(a)} y\rho^+(s(a,v))\right) + \sigma^-\left(\sum_{(v,y)\in V(a)} (1-y)\rho^-(s(a,v))\right)\right)$$

The Mixed Loss Function: Achieving Mixup in Metric Learning

Key Insights:

✓ With this formulation, every mixed embedding contributes to both positive and negative terms, especially in positive-negative pairs

So, now, the mixed loss function:

$$\widetilde{\ell}(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{\boxed{(v,y)\in V(a)}} y\rho^{+}(s(a,v))\right) + \sigma^{-}\left(\sum_{\boxed{(v,y)\in V(a)}} (1-y)\rho^{-}(s(a,v))\right)\right)$$

The Mixed Loss Function: Achieving Mixup in Metric Learning

Key Insights:

✓ With this formulation, every mixed embedding contributes to both positive and negative terms, especially in positive-negative pairs

✓ The label is interpolated (y in [0,1]) → both terms are non-zero → mixup achieved ☺

So, now, the mixed loss function:

$$\widetilde{\ell}(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(v,y)\in V(a)} \boxed{y}\,\rho^{+}(s(a,v))\right) + \sigma^{-}\left(\sum_{(v,y)\in V(a)} \boxed{(1-y)}\,\rho^{-}(s(a,v))\right)\right)$$

The Mixed Loss Function: Achieving Mixup in Metric Learning

Key Insights:

✓ With this formulation, every mixed embedding contributes to both positive and negative terms, especially in positive-negative pairs

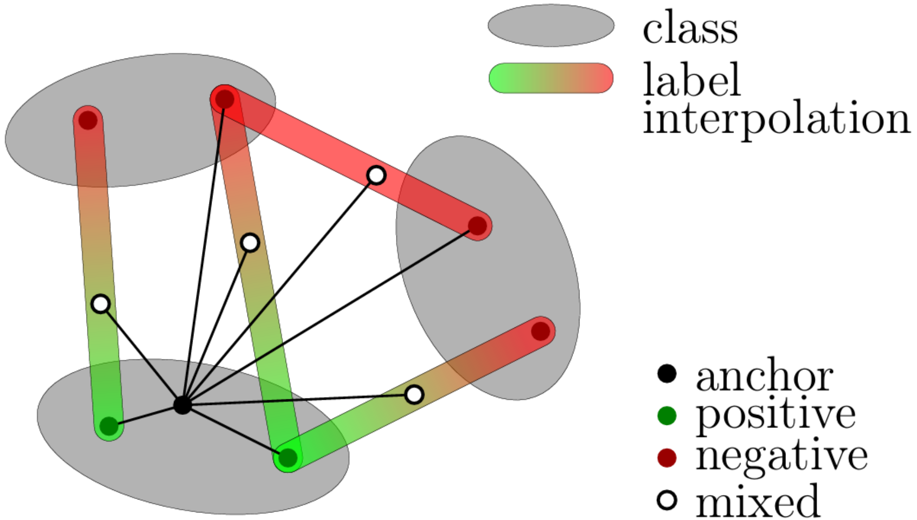
✓ The label is interpolated (y in [0,1]) → both terms are non-zero → mixup achieved ☺

✓ For positive-negative pairs, the mixed embedding behaves both positive and negative to varying degrees based on λ

So, now, the mixed loss function:

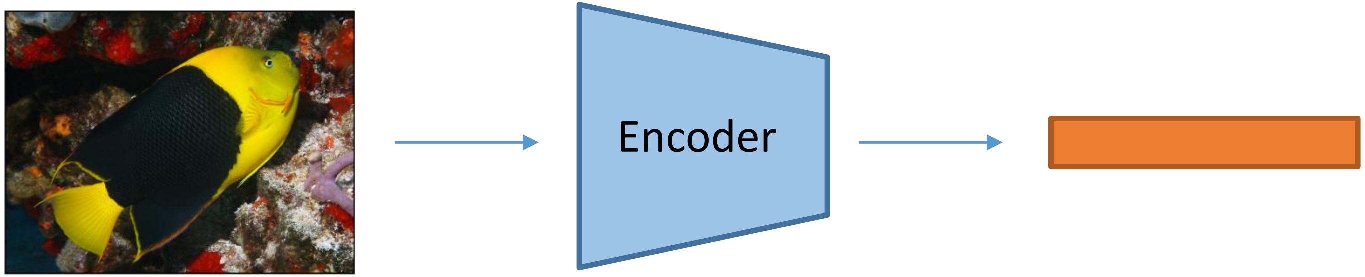
$$\widetilde{\ell}(a;\theta) := \tau\left(\sigma^{+}\left(\sum_{(v,y)\in V(a)} y\rho^{+}(s(a,\boxed{v}))\right) + \sigma^{-}\left(\sum_{(v,y)\in V(a)} (1-y)\rho^{-}(s(a,\boxed{v}))\right)\right)$$

Metrix: Mixup in Metric Learning



Metrix (= Metric Mix) allows an anchor to interact with positive, negative and interpolated examples, which also have interpolated labels

Metrix types



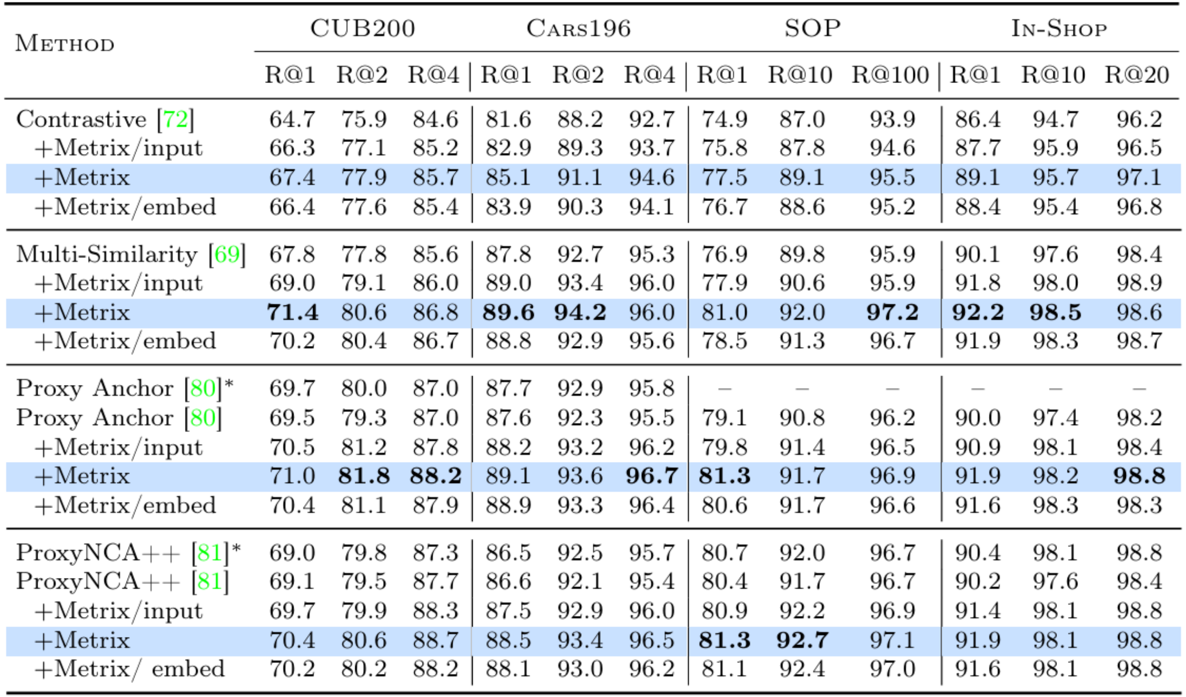
Encoder

Metrix/input

Metrix/feature
or Metrix

Metrix/embed

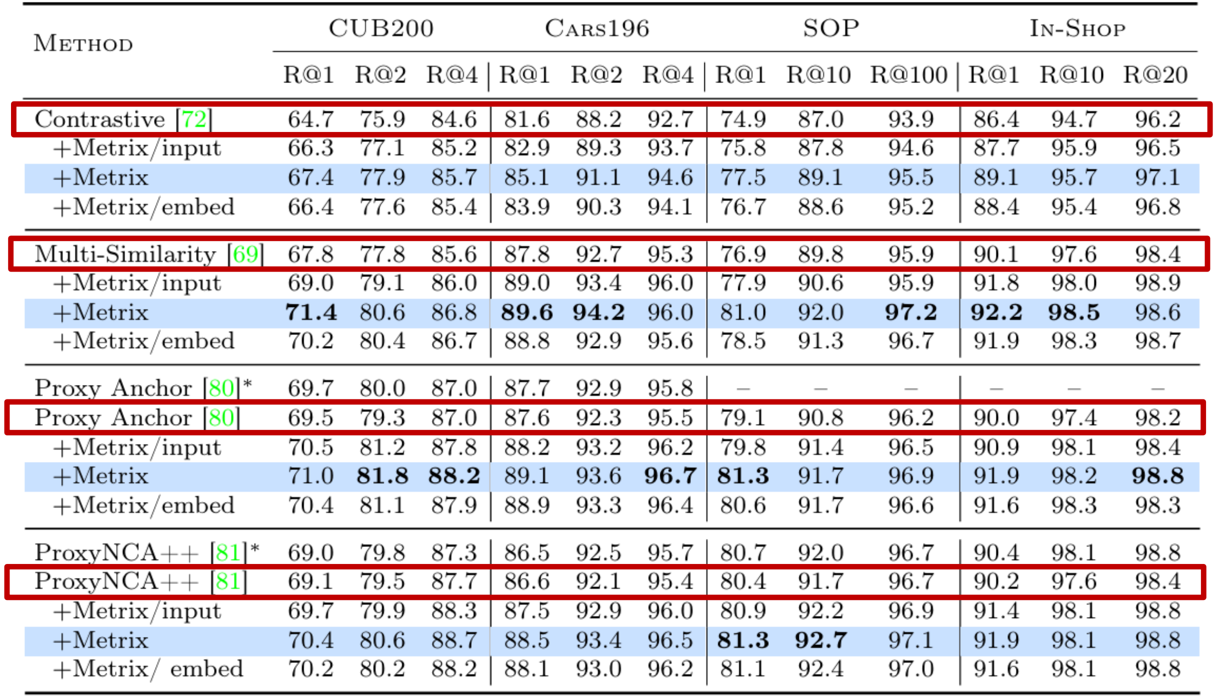
Quantitative Evaluation



Method	CUB200			Cars196			SOP			In-Shop		
	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
Contrastive [72]	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.2
+Metrix/input	66.3	77.1	85.2	82.9	89.3	93.7	75.8	87.8	94.6	87.7	95.9	96.5
+Metrix	67.4	77.9	85.7	85.1	91.1	94.6	77.5	89.1	95.5	89.1	95.7	97.1
+Metrix/embed	66.4	77.6	85.4	83.9	90.3	94.1	76.7	88.6	95.2	88.4	95.4	96.8
Multi-Similarity [69]	67.8	77.8	85.6	87.8	92.7	95.3	76.9	89.8	95.9	90.1	97.6	98.4
+Metrix/input	69.0	79.1	86.0	89.0	93.4	96.0	77.9	90.6	95.9	91.8	98.0	98.9
+Metrix	**71.4**	80.6	86.8	**89.6**	**94.2**	96.0	81.0	92.0	**97.2**	**92.2**	**98.5**	98.6
+Metrix/embed	70.2	80.4	86.7	88.8	92.9	95.6	78.5	91.3	96.7	91.9	98.3	98.7
Proxy Anchor [80]*	69.7	80.0	87.0	87.7	92.9	95.8	–	–	–	–	–	–
Proxy Anchor [80]	69.5	79.3	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2
+Metrix/input	70.5	81.2	87.8	88.2	93.2	96.2	79.8	91.4	96.5	90.9	98.1	98.4
+Metrix	71.0	**81.8**	**88.2**	89.1	93.6	**96.7**	**81.3**	91.7	96.9	91.9	98.2	**98.8**
+Metrix/embed	70.4	81.1	87.9	88.9	93.3	96.4	80.6	91.7	96.6	91.6	98.3	98.3
ProxyNCA++ [81]*	69.0	79.8	87.3	86.5	92.5	95.7	80.7	92.0	96.7	90.4	98.1	98.8
ProxyNCA++ [81]	69.1	79.5	87.7	86.6	92.1	95.4	80.4	91.7	96.7	90.2	97.6	98.4
+Metrix/input	69.7	79.9	88.3	87.5	92.9	96.0	80.9	92.2	96.9	91.4	98.1	98.8
+Metrix	70.4	80.6	88.7	88.5	93.4	96.5	**81.3**	**92.7**	97.1	91.9	98.1	98.8
+Metrix/ embed	70.2	80.2	88.2	88.1	93.0	96.2	81.1	92.4	97.0	91.6	98.1	98.8

Evaluating the impact of Metrix on four metric learning loss functions; ResNet-50 with embedding size d=512; Recall@K on four datasets.

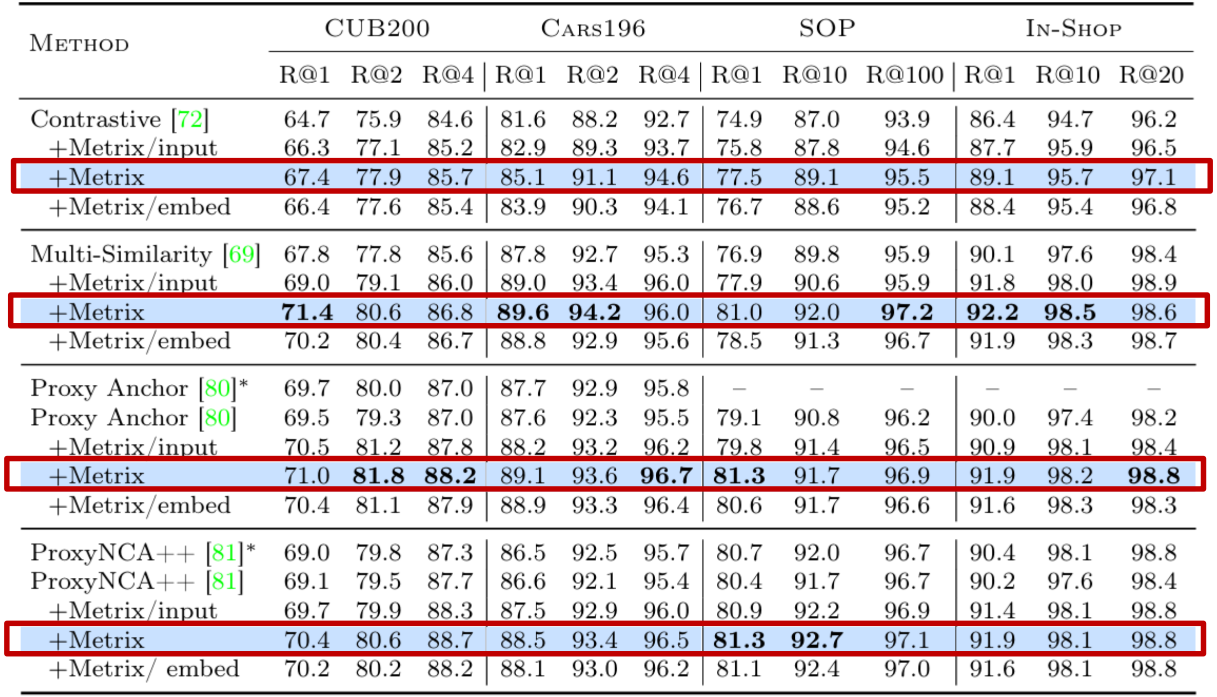
Quantitative Evaluation



Method	CUB200			Cars196			SOP			In-Shop		
	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
Contrastive [72]	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.2
+Metrix/input	66.3	77.1	85.2	82.9	89.3	93.7	75.8	87.8	94.6	87.7	95.9	96.5
+Metrix	67.4	77.9	85.7	85.1	91.1	94.6	77.5	89.1	95.5	89.1	95.7	97.1
+Metrix/embed	66.4	77.6	85.4	83.9	90.3	94.1	76.7	88.6	95.2	88.4	95.4	96.8
Multi-Similarity [69]	67.8	77.8	85.6	87.8	92.7	95.3	76.9	89.8	95.9	90.1	97.6	98.4
+Metrix/input	69.0	79.1	86.0	89.0	93.4	96.0	77.9	90.6	95.9	91.8	98.0	98.9
+Metrix	**71.4**	80.6	86.8	**89.6**	**94.2**	96.0	81.0	92.0	**97.2**	**92.2**	**98.5**	98.6
+Metrix/embed	70.2	80.4	86.7	88.8	92.9	95.6	78.5	91.3	96.7	91.9	98.3	98.7
Proxy Anchor [80]*	69.7	80.0	87.0	87.7	92.9	95.8	–	–	–	–	–	–
Proxy Anchor [80]	69.5	79.3	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2
+Metrix/input	70.5	81.2	87.8	88.2	93.2	96.2	79.8	91.4	96.5	90.9	98.1	98.4
+Metrix	71.0	**81.8**	**88.2**	89.1	93.6	**96.7**	**81.3**	91.7	96.9	91.9	98.2	**98.8**
+Metrix/embed	70.4	81.1	87.9	88.9	93.3	96.4	80.6	91.7	96.6	91.6	98.3	98.3
ProxyNCA++ [81]*	69.0	79.8	87.3	86.5	92.5	95.7	80.7	92.0	96.7	90.4	98.1	98.8
ProxyNCA++ [81]	69.1	79.5	87.7	86.6	92.1	95.4	80.4	91.7	96.7	90.2	97.6	98.4
+Metrix/input	69.7	79.9	88.3	87.5	92.9	96.0	80.9	92.2	96.9	91.4	98.1	98.8
+Metrix	70.4	80.6	88.7	88.5	93.4	96.5	**81.3**	**92.7**	97.1	91.9	98.1	98.8
+Metrix/ embed	70.2	80.2	88.2	88.1	93.0	96.2	81.1	92.4	97.0	91.6	98.1	98.8

Evaluating the impact of Metrix on four metric learning loss functions; ResNet-50 with embedding size d=512; Recall@K on four datasets.

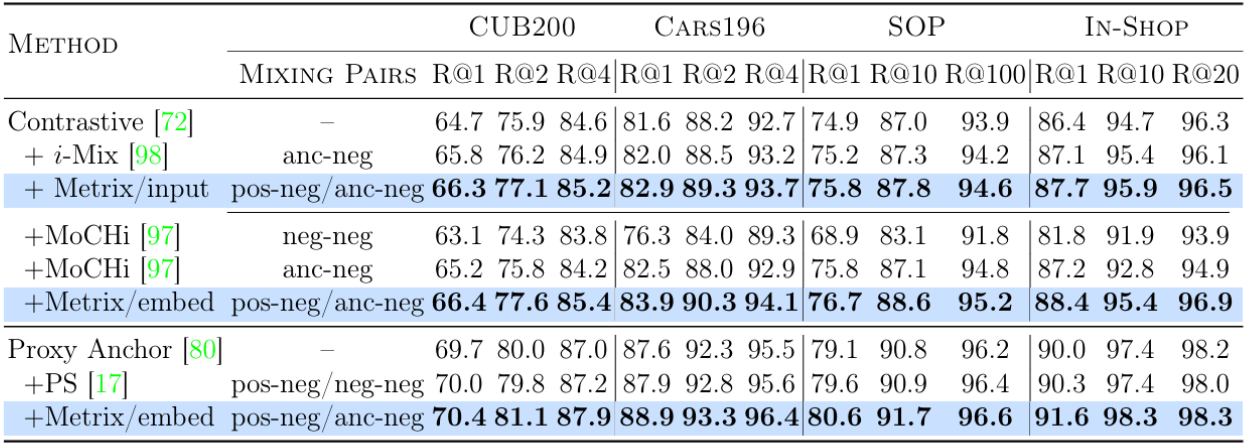
Quantitative Evaluation



Method	CUB200			Cars196			SOP			In-Shop			
	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20	
Contrastive [72]	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.2	
+Metrix/input	66.3	77.1	85.2	82.9	89.3	93.7	75.8	87.8	94.6	87.7	95.9	96.5	
+Metrix	67.4	77.9	85.7	85.1	91.1	94.6	77.5	89.1	95.5	89.1	95.7	97.1	+2.1%
+Metrix/embed	66.4	77.6	85.4	83.9	90.3	94.1	76.7	88.6	95.2	88.4	95.4	96.8	
Multi-Similarity [69]	67.8	77.8	85.6	87.8	92.7	95.3	76.9	89.8	95.9	90.1	97.6	98.4	
+Metrix/input	69.0	79.1	86.0	89.0	93.4	96.0	77.9	90.6	95.9	91.8	98.0	98.9	
+Metrix	**71.4**	80.6	86.8	**89.6**	**94.2**	96.0	81.0	92.0	**97.2**	**92.2**	**98.5**	98.6	+1.9%
+Metrix/embed	70.2	80.4	86.7	88.8	92.9	95.6	78.5	91.3	96.7	91.9	98.3	98.7	
Proxy Anchor [80]*	69.7	80.0	87.0	87.7	92.9	95.8	–	–	–	–	–	–	
Proxy Anchor [80]	69.5	79.3	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2	
+Metrix/input	70.5	81.2	87.8	88.2	93.2	96.2	79.8	91.4	96.5	90.9	98.1	98.4	
+Metrix	71.0	**81.8**	**88.2**	89.1	93.6	**96.7**	**81.3**	91.7	96.9	91.9	98.2	**98.8**	+1.4%
+Metrix/embed	70.4	81.1	87.9	88.9	93.3	96.4	80.6	91.7	96.6	91.6	98.3	98.3	
ProxyNCA++ [81]*	69.0	79.8	87.3	86.5	92.5	95.7	80.7	92.0	96.7	90.4	98.1	98.8	
ProxyNCA++ [81]	69.1	79.5	87.7	86.6	92.1	95.4	80.4	91.7	96.7	90.2	97.6	98.4	
+Metrix/input	69.7	79.9	88.3	87.5	92.9	96.0	80.9	92.2	96.9	91.4	98.1	98.8	
+Metrix	70.4	80.6	88.7	88.5	93.4	96.5	**81.3**	**92.7**	97.1	91.9	98.1	98.8	+1.1%
+Metrix/ embed	70.2	80.2	88.2	88.1	93.0	96.2	81.1	92.4	97.0	91.6	98.1	98.8	

Evaluating the impact of Metrix on four metric learning loss functions; ResNet-50 with embedding size d=512; Recall@K on four datasets.

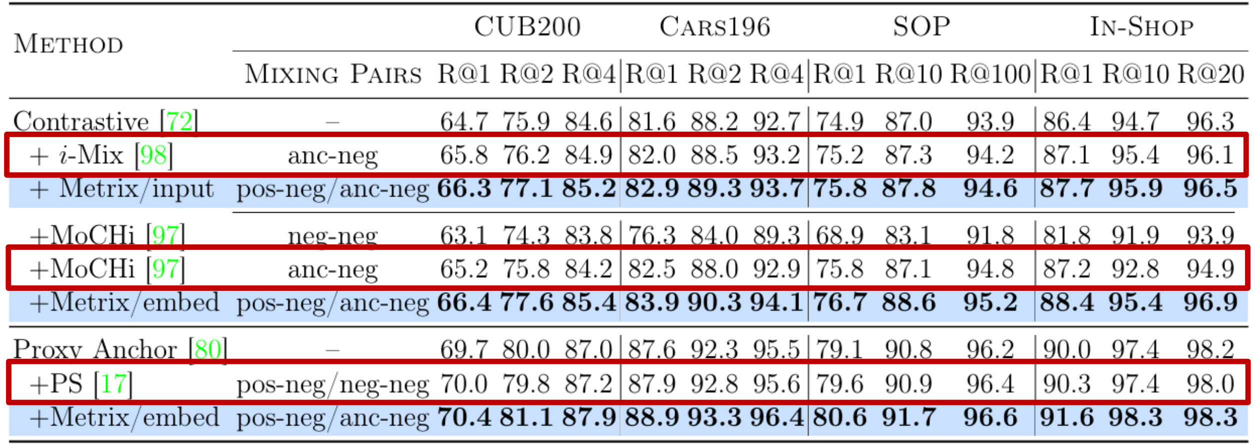
Quantitative Evaluation



METHOD		CUB200			CARS196			SOP			IN-SHOP		
	MIXING PAIRS	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
Contrastive [72]	–	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.3
+ *i*-Mix [98]	anc-neg	65.8	76.2	84.9	82.0	88.5	93.2	75.2	87.3	94.2	87.1	95.4	96.1
+ Metrix/input	pos-neg/anc-neg	**66.3**	**77.1**	**85.2**	**82.9**	**89.3**	**93.7**	**75.8**	**87.8**	**94.6**	**87.7**	**95.9**	**96.5**
+MoCHi [97]	neg-neg	63.1	74.3	83.8	76.3	84.0	89.3	68.9	83.1	91.8	81.8	91.9	93.9
+MoCHi [97]	anc-neg	65.2	75.8	84.2	82.5	88.0	92.9	75.8	87.1	94.8	87.2	92.8	94.9
+Metrix/embed	pos-neg/anc-neg	**66.4**	**77.6**	**85.4**	**83.9**	**90.3**	**94.1**	**76.7**	**88.6**	**95.2**	**88.4**	**95.4**	**96.9**
Proxy Anchor [80]	–	69.7	80.0	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2
+PS [17]	pos-neg/neg-neg	70.0	79.8	87.2	87.9	92.8	95.6	79.6	90.9	96.4	90.3	97.4	98.0
+Metrix/embed	pos-neg/anc-neg	**70.4**	**81.1**	**87.9**	**88.9**	**93.3**	**96.4**	**80.6**	**91.7**	**96.6**	**91.6**	**98.3**	**98.3**

Comparison of Metrix/input and Metrix/embed with other mixing methods; ResNet-50 with embedding size d=512; Recall@K on four datasets.

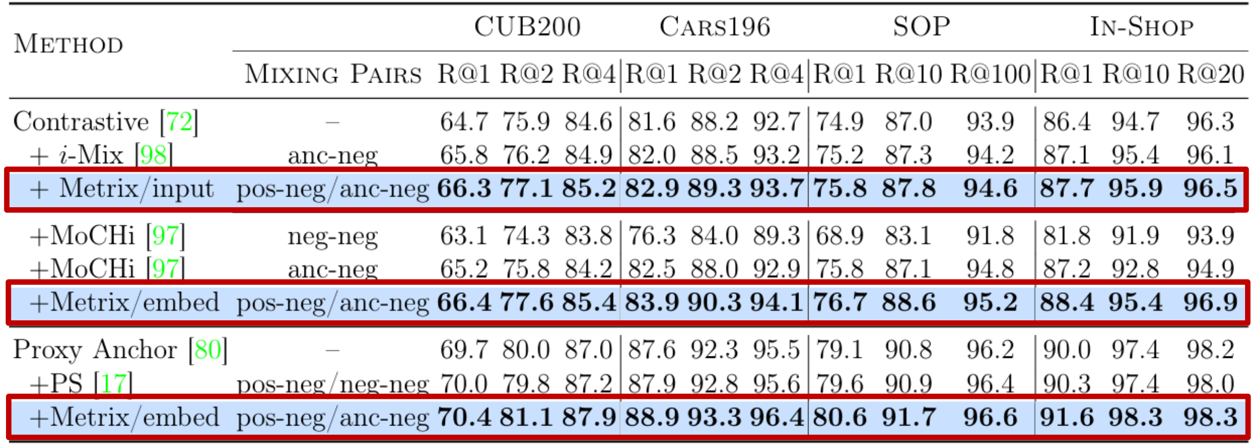
Quantitative Evaluation



METHOD		CUB200			CARS196			SOP			IN-SHOP		
	MIXING PAIRS	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
Contrastive [72]	–	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.3
+ *i*-Mix [98]	anc-neg	65.8	76.2	84.9	82.0	88.5	93.2	75.2	87.3	94.2	87.1	95.4	96.1
+ Metrix/input	pos-neg/anc-neg	**66.3**	**77.1**	**85.2**	**82.9**	**89.3**	**93.7**	**75.8**	**87.8**	**94.6**	**87.7**	**95.9**	**96.5**
+MoCHi [97]	neg-neg	63.1	74.3	83.8	76.3	84.0	89.3	68.9	83.1	91.8	81.8	91.9	93.9
+MoCHi [97]	anc-neg	65.2	75.8	84.2	82.5	88.0	92.9	75.8	87.1	94.8	87.2	92.8	94.9
+Metrix/embed	pos-neg/anc-neg	**66.4**	**77.6**	**85.4**	**83.9**	**90.3**	**94.1**	**76.7**	**88.6**	**95.2**	**88.4**	**95.4**	**96.9**
Proxy Anchor [80]	–	69.7	80.0	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2
+PS [17]	pos-neg/neg-neg	70.0	79.8	87.2	87.9	92.8	95.6	79.6	90.9	96.4	90.3	97.4	98.0
+Metrix/embed	pos-neg/anc-neg	**70.4**	**81.1**	**87.9**	**88.9**	**93.3**	**96.4**	**80.6**	**91.7**	**96.6**	**91.6**	**98.3**	**98.3**

Comparison of Metrix/input and Metrix/embed with other mixing methods; ResNet-50 with embedding size d=512; Recall@K on four datasets.

Quantitative Evaluation



METHOD		CUB200			CARS196			SOP			IN-SHOP		
	MIXING PAIRS	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
Contrastive [72]	–	64.7	75.9	84.6	81.6	88.2	92.7	74.9	87.0	93.9	86.4	94.7	96.3
+ *i*-Mix [98]	anc-neg	65.8	76.2	84.9	82.0	88.5	93.2	75.2	87.3	94.2	87.1	95.4	96.1
+ Metrix/input	pos-neg/anc-neg	**66.3**	**77.1**	**85.2**	**82.9**	**89.3**	**93.7**	**75.8**	**87.8**	**94.6**	**87.7**	**95.9**	**96.5**
+MoCHi [97]	neg-neg	63.1	74.3	83.8	76.3	84.0	89.3	68.9	83.1	91.8	81.8	91.9	93.9
+MoCHi [97]	anc-neg	65.2	75.8	84.2	82.5	88.0	92.9	75.8	87.1	94.8	87.2	92.8	94.9
+Metrix/embed	pos-neg/anc-neg	**66.4**	**77.6**	**85.4**	**83.9**	**90.3**	**94.1**	**76.7**	**88.6**	**95.2**	**88.4**	**95.4**	**96.9**
Proxy Anchor [80]	–	69.7	80.0	87.0	87.6	92.3	95.5	79.1	90.8	96.2	90.0	97.4	98.2
+PS [17]	pos-neg/neg-neg	70.0	79.8	87.2	87.9	92.8	95.6	79.6	90.9	96.4	90.3	97.4	98.0
+Metrix/embed	pos-neg/anc-neg	**70.4**	**81.1**	**87.9**	**88.9**	**93.3**	**96.4**	**80.6**	**91.7**	**96.6**	**91.6**	**98.3**	**98.3**

+0.6%

+1.5%

+0.8%

Comparison of Metrix/input and Metrix/embed with other mixing methods; ResNet-50 with embedding size d=512; Recall@K on four datasets.

Summarizing Insights

- ✓ Introduce a direct extension of mixup from classification to metric learning

- ✓ Develop a generic way of representing and interpolating labels, allowing the straightforward extension of any kind on mixup to metric learning for a large class of loss functions

- ✓ Propose and systematically evaluate a novel mixup method under different settings

ICCV23

PARIS

Learning Visual Representations via Model Architecture Component [SimPool]

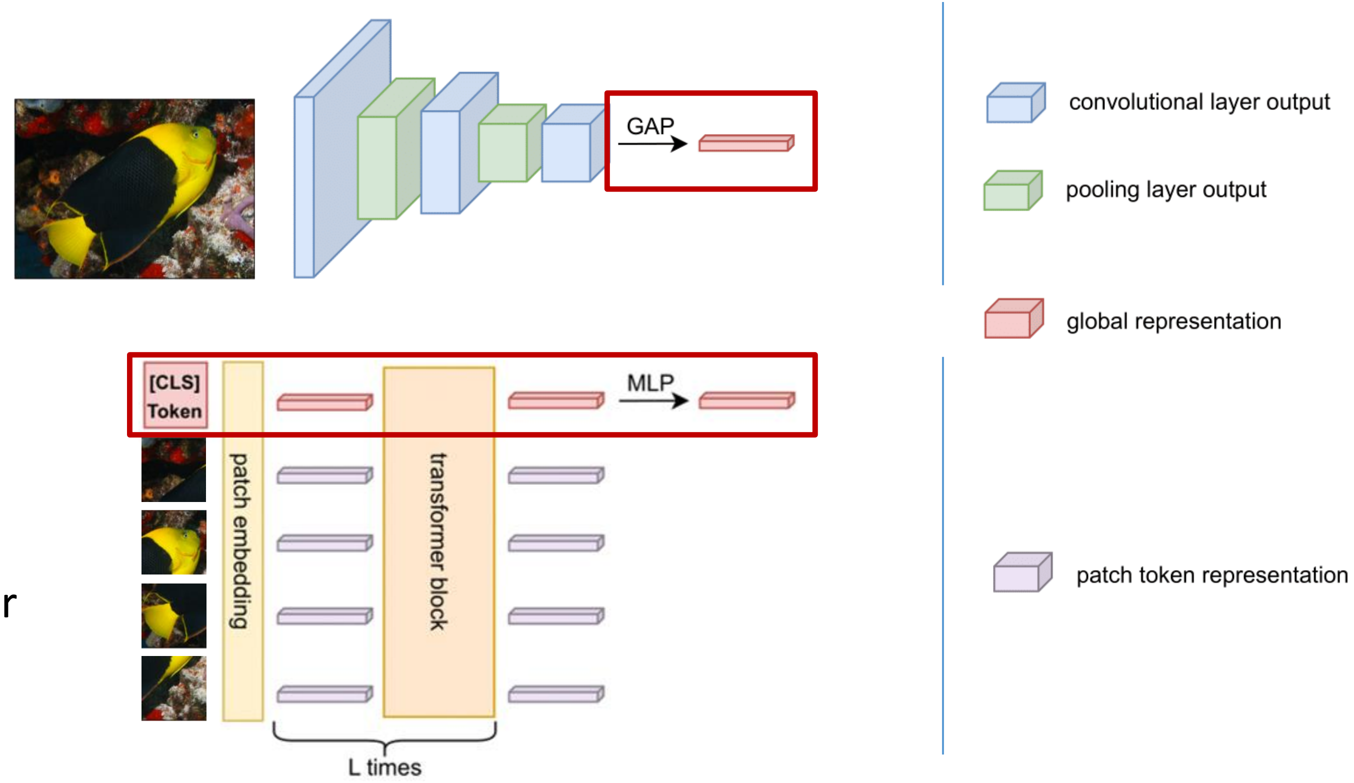
Psomas et al. *Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?*, **ICCV 2023**

Source code: https://github.com/billpsomas/simpool

Recap: Global Representation in CNNs vs. ViTs

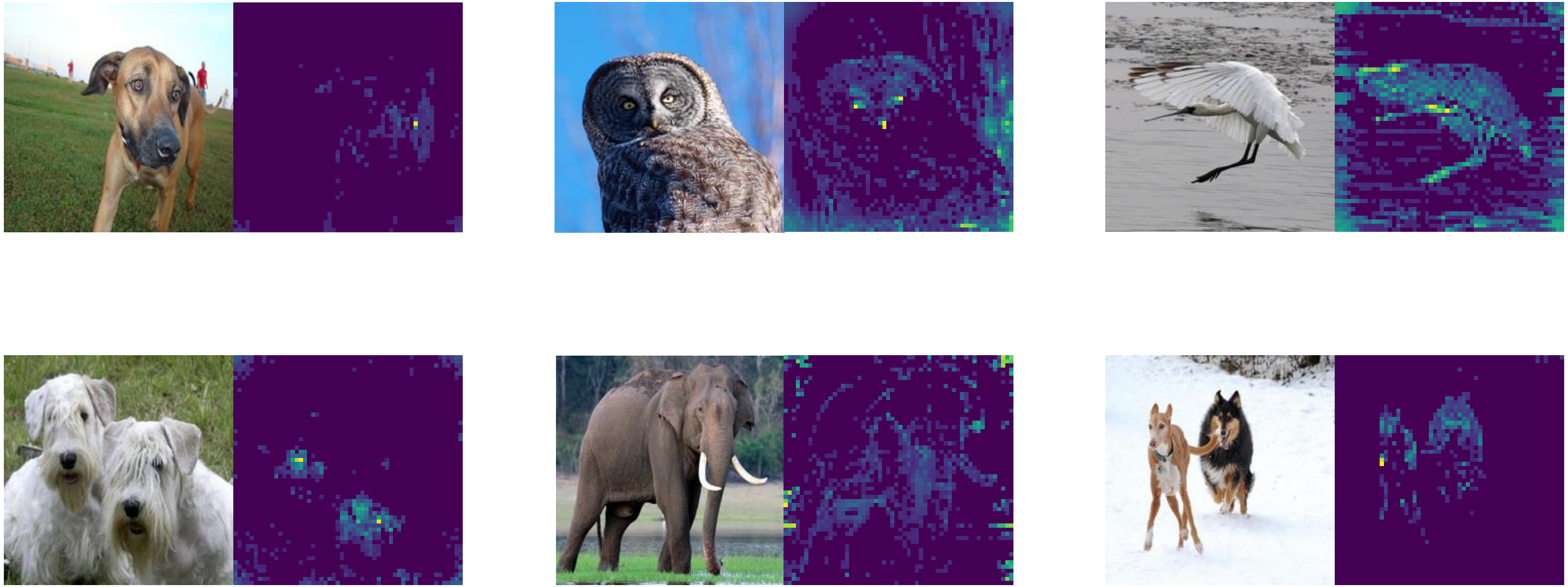
Convolutional Neural Network

Vision Transformer



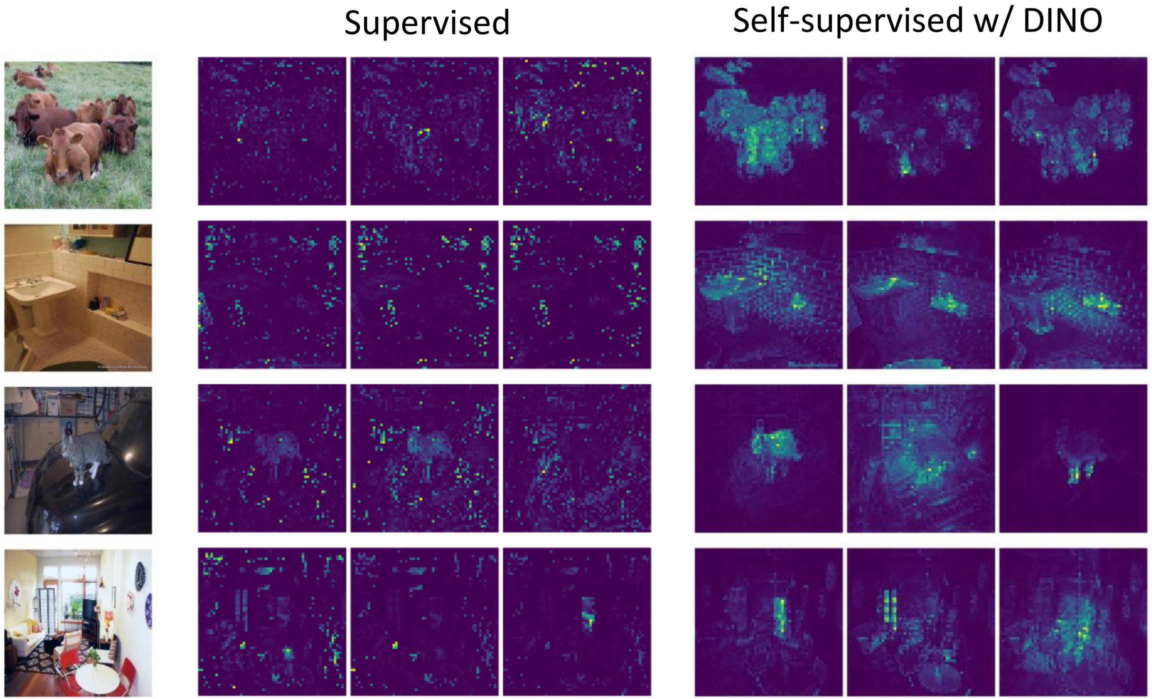
Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Related Work: Supervised ViTs have “low-quality” attention maps



ViT-S on Imagenet-1k; mean attention map of the [CLS]; final block

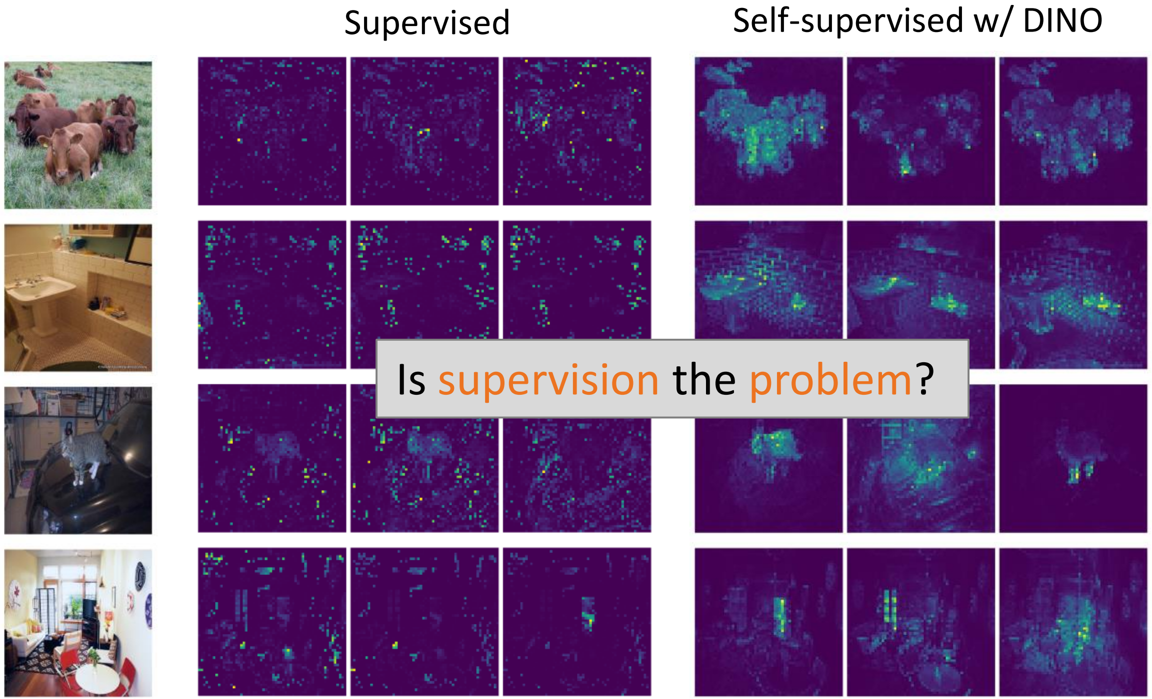
Related Work: DINO has “higher-quality” attention maps



ViT-S on Imagenet-1k; images from COCO val set;
attention maps of the [CLS] for 3 different heads; final block

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

Related Work: DINO has “higher-quality” attention maps

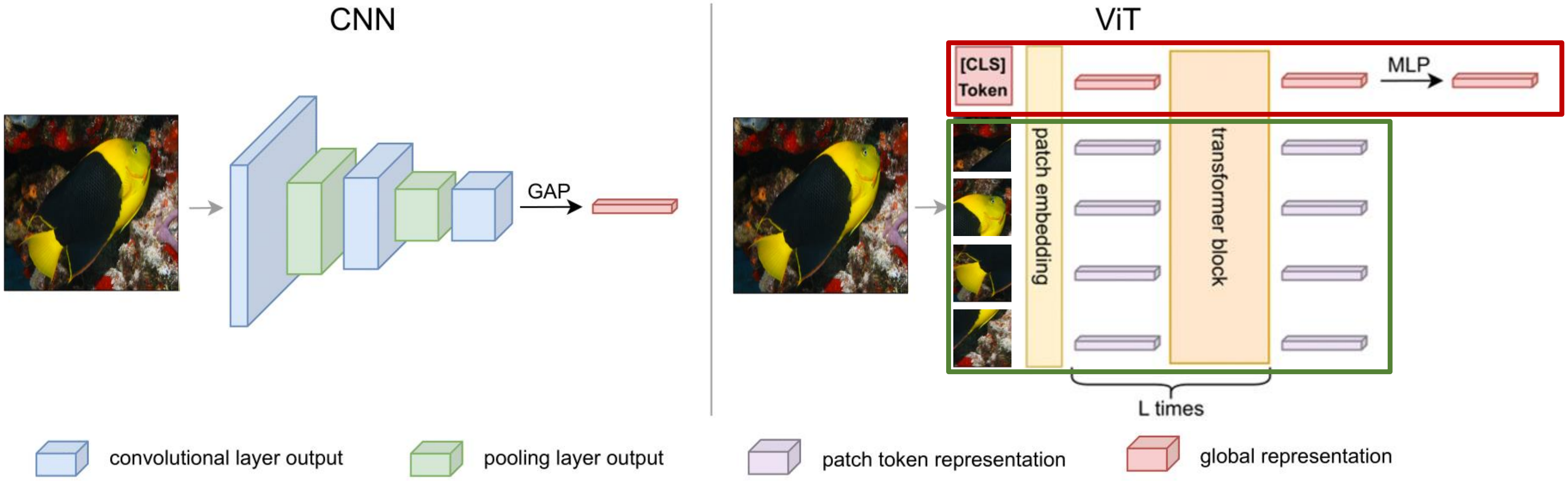


Is supervision the problem?

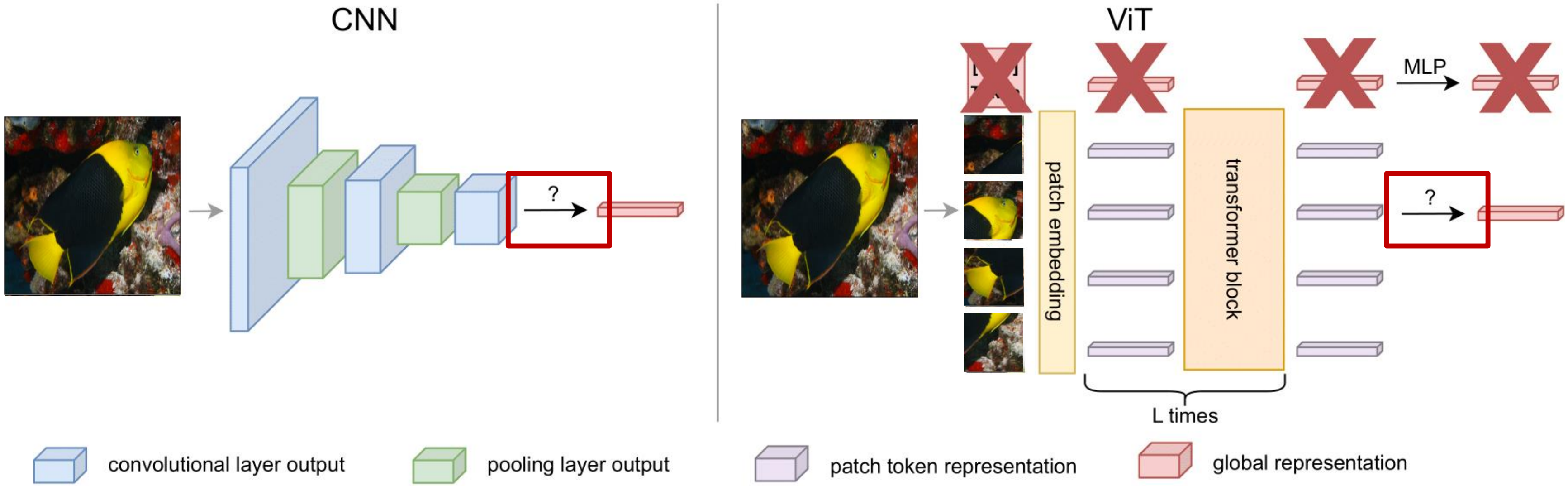
ViT-S on Imagenet-1k; images from COCO val set;
attention maps of the [CLS] for 3 different heads; final block

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

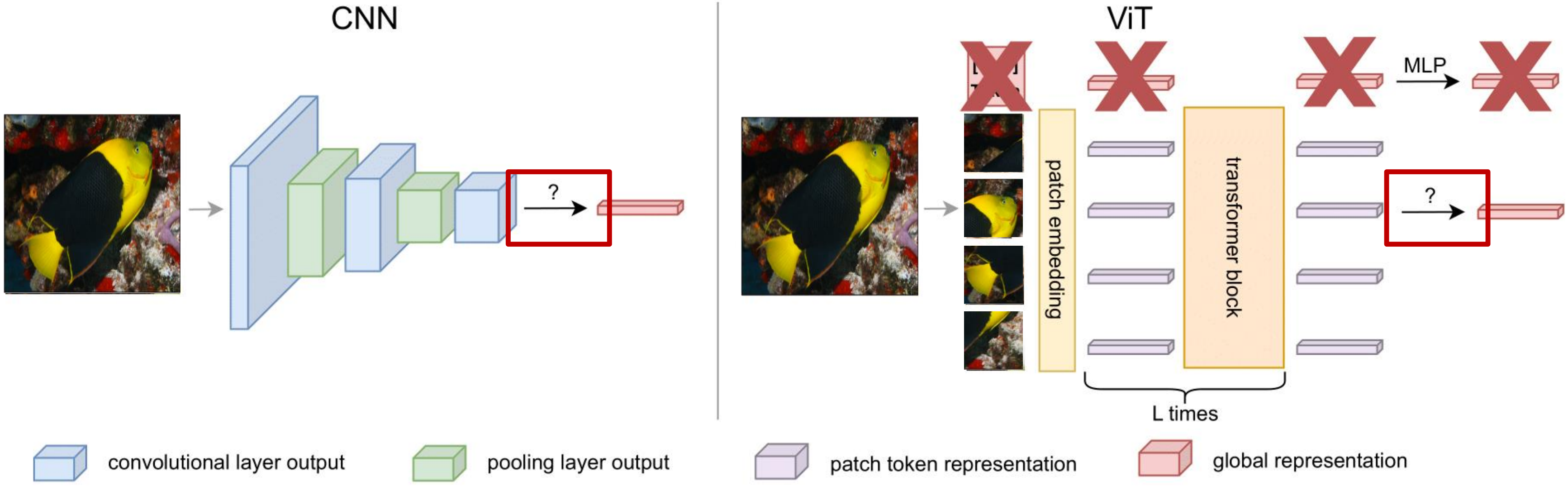
CNNs vs. ViTs


CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

“Universal” Pooling

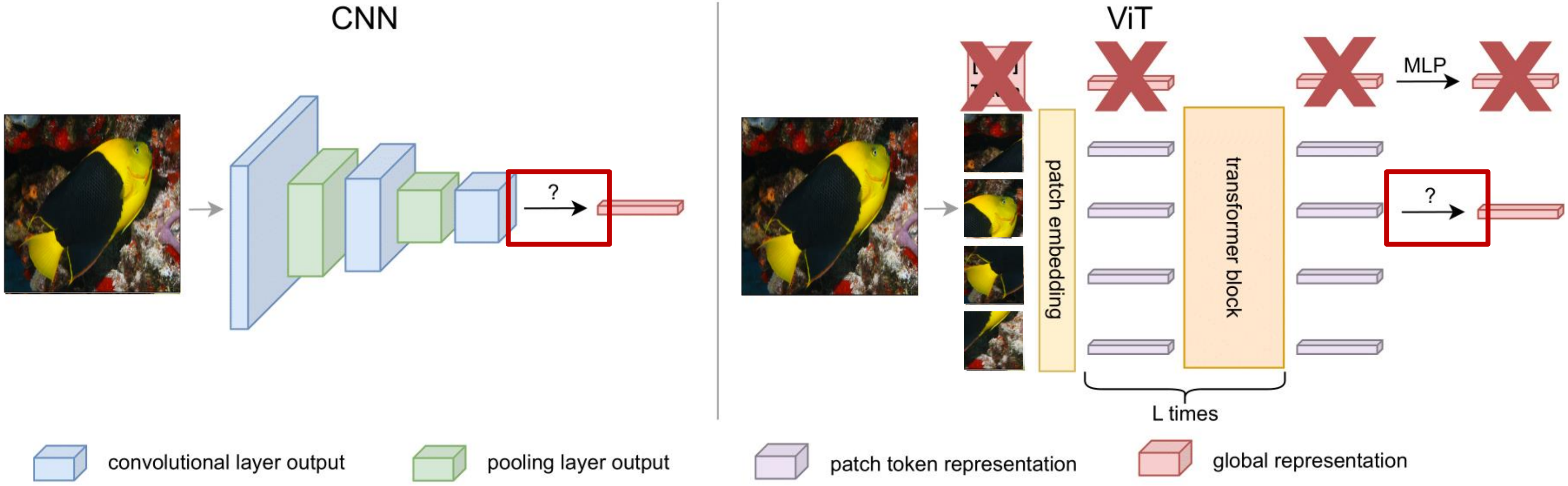


Focus



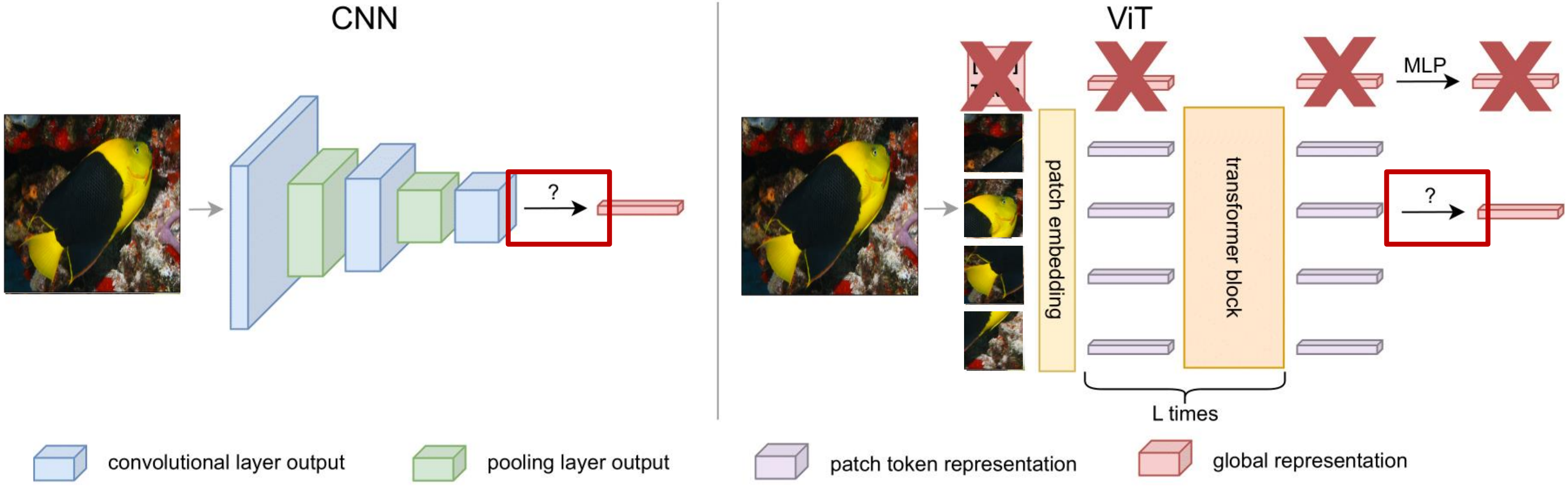
- Pooling at the very last step of both network types improving over default?

Focus



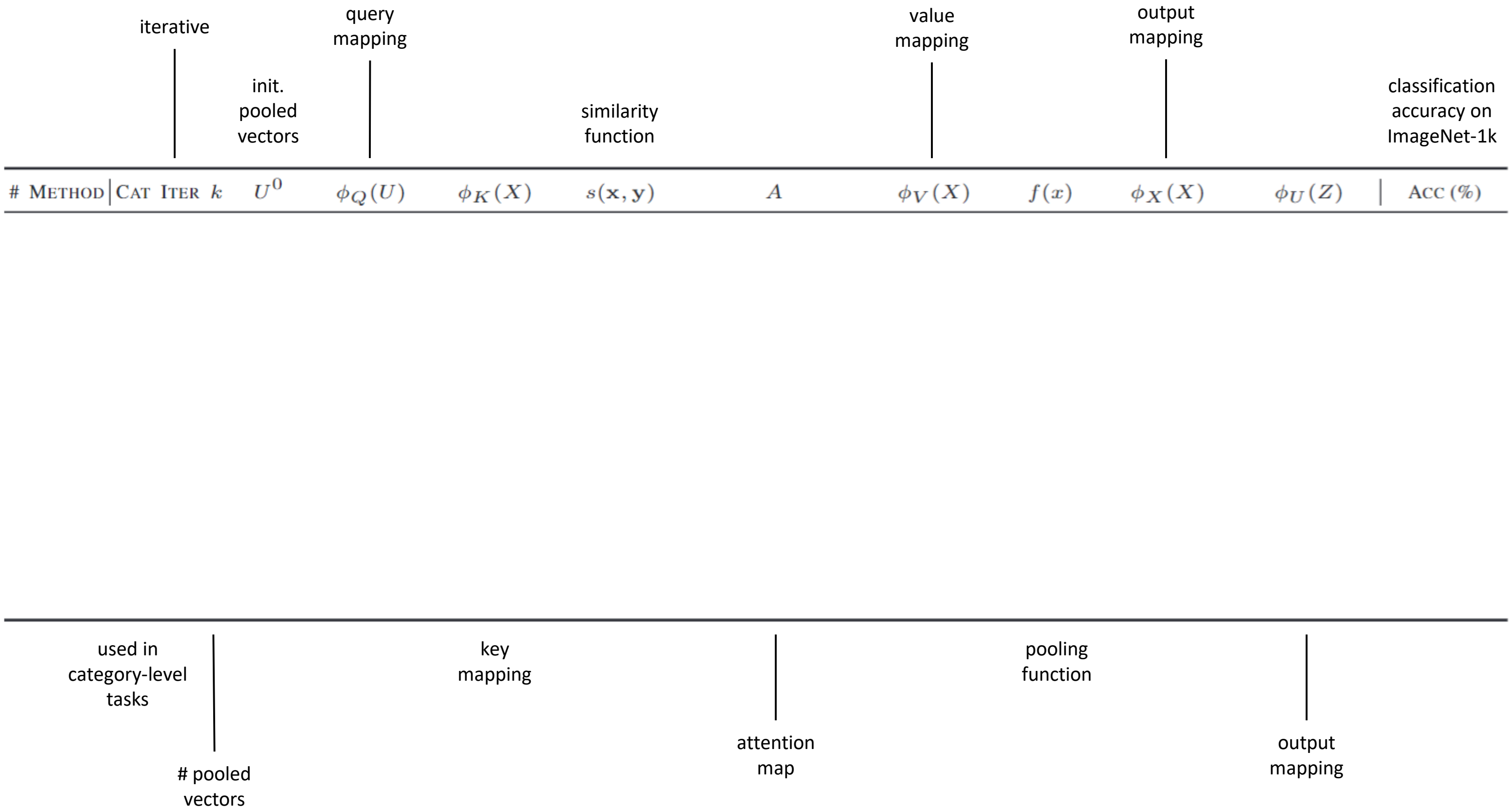
- Pooling at the very last step of both network types improving over default?
- Pooling for high-quality spatial attention?

Focus



- Pooling at the very last step of both network types improving over default?
- Pooling for high-quality spatial attention?
- Validity in both supervised and self-supervised settings?

Generic Pooling Framework



iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

used in category-level tasks

key mapping

pooling function

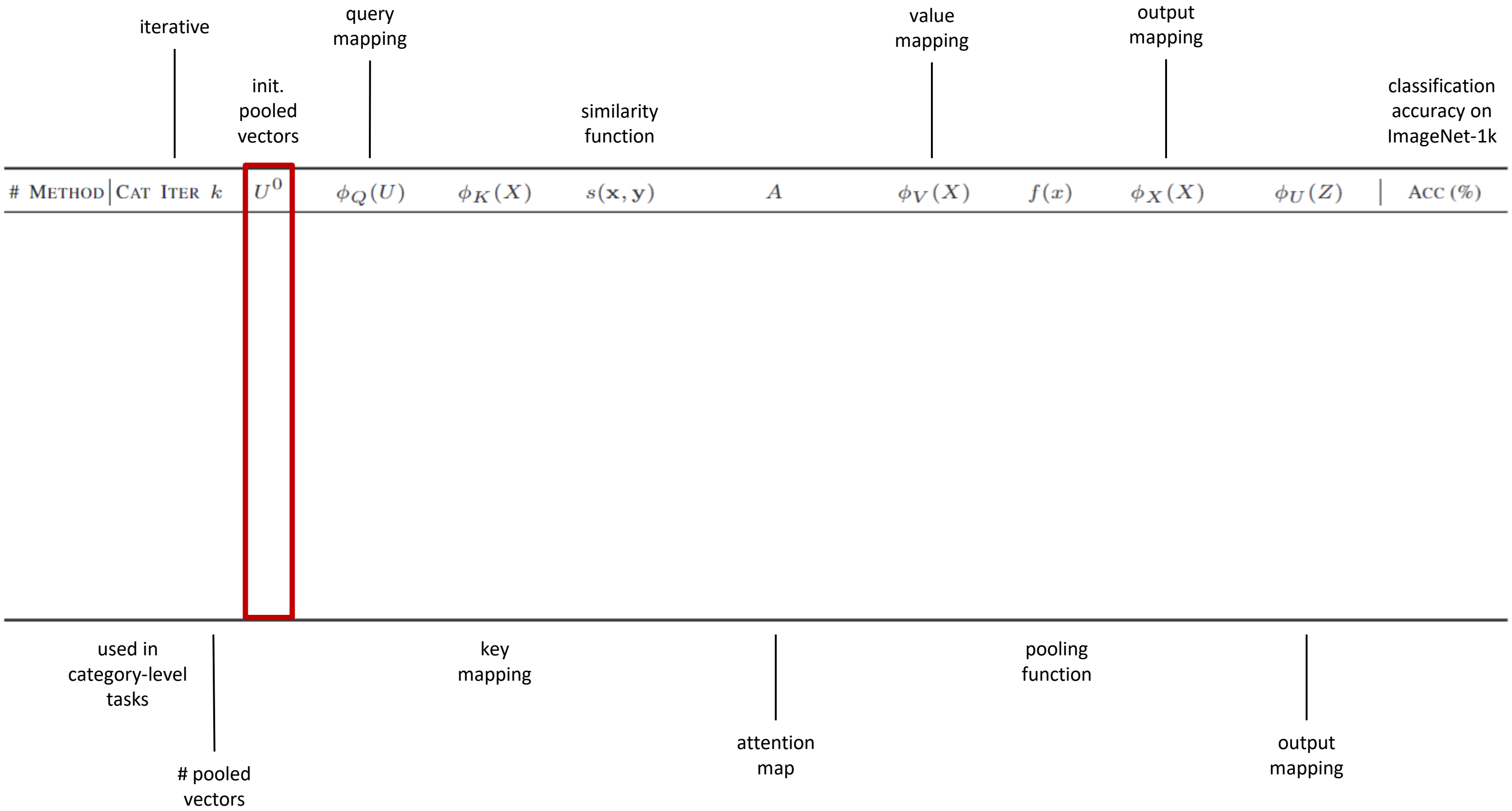
# pooled vectors

attention map

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

used in category-level tasks

# pooled vectors

key mapping

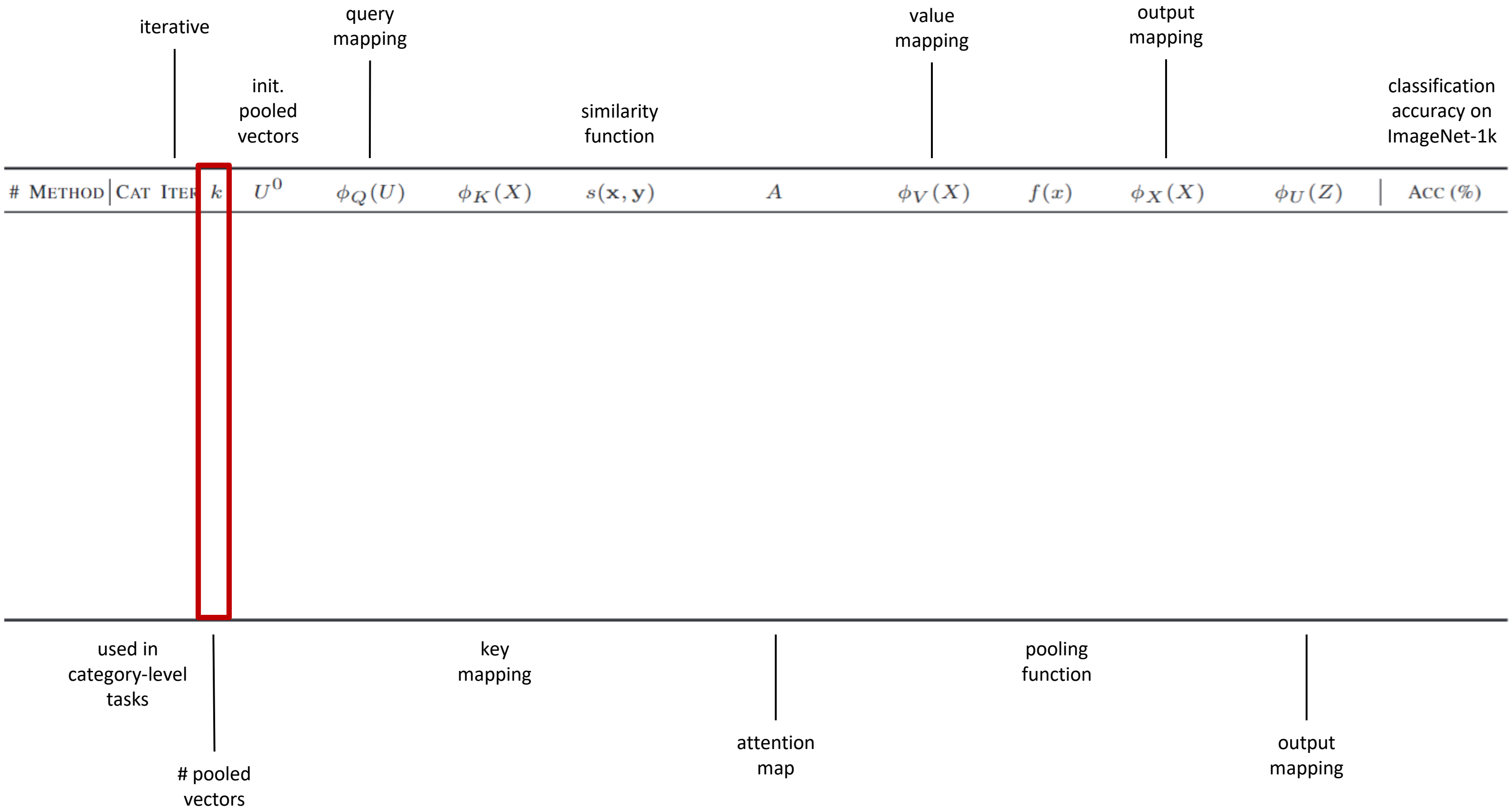
attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

used in category-level tasks

# pooled vectors

key mapping

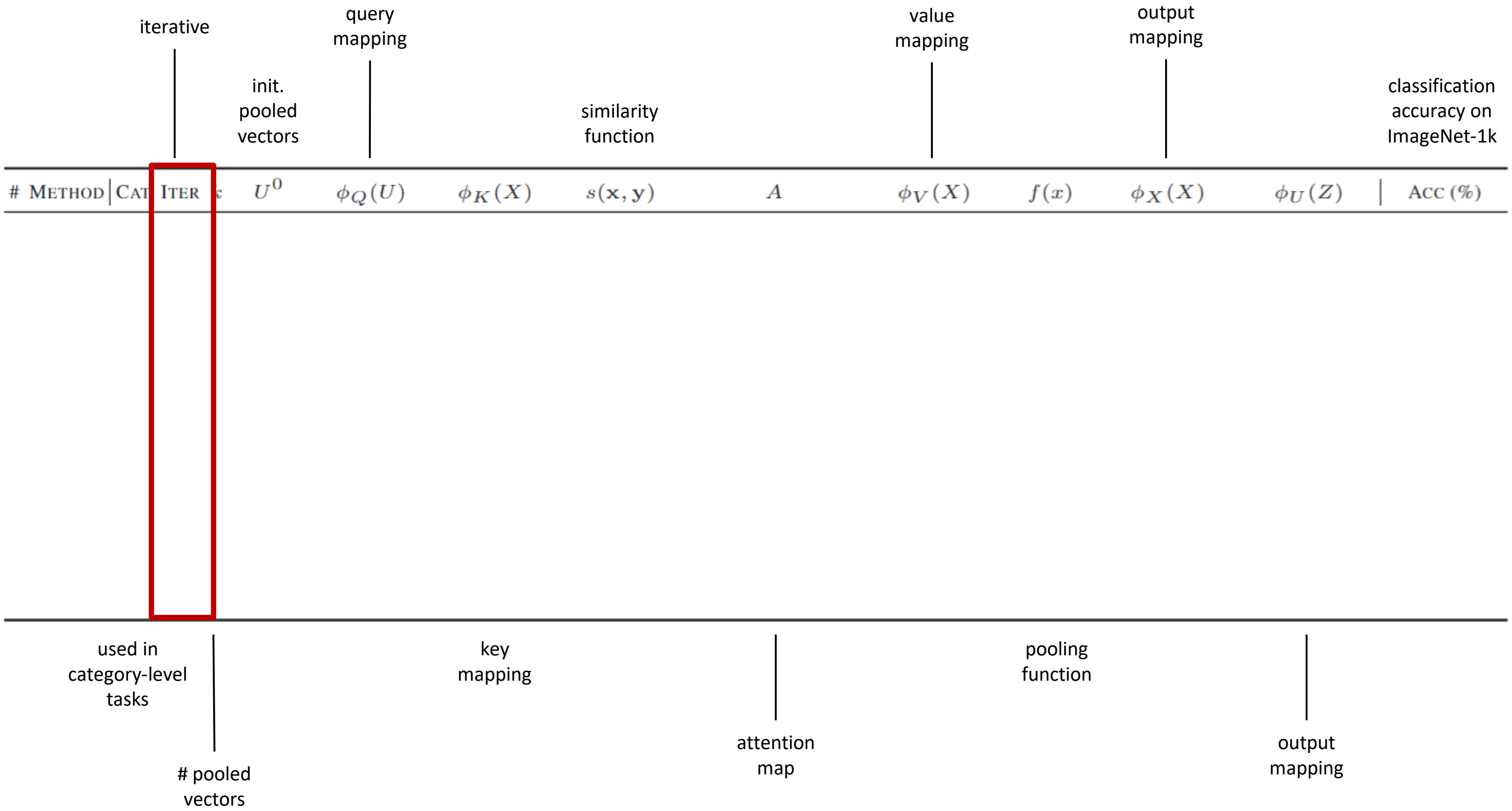
attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

used in category-level tasks

# pooled vectors

key mapping

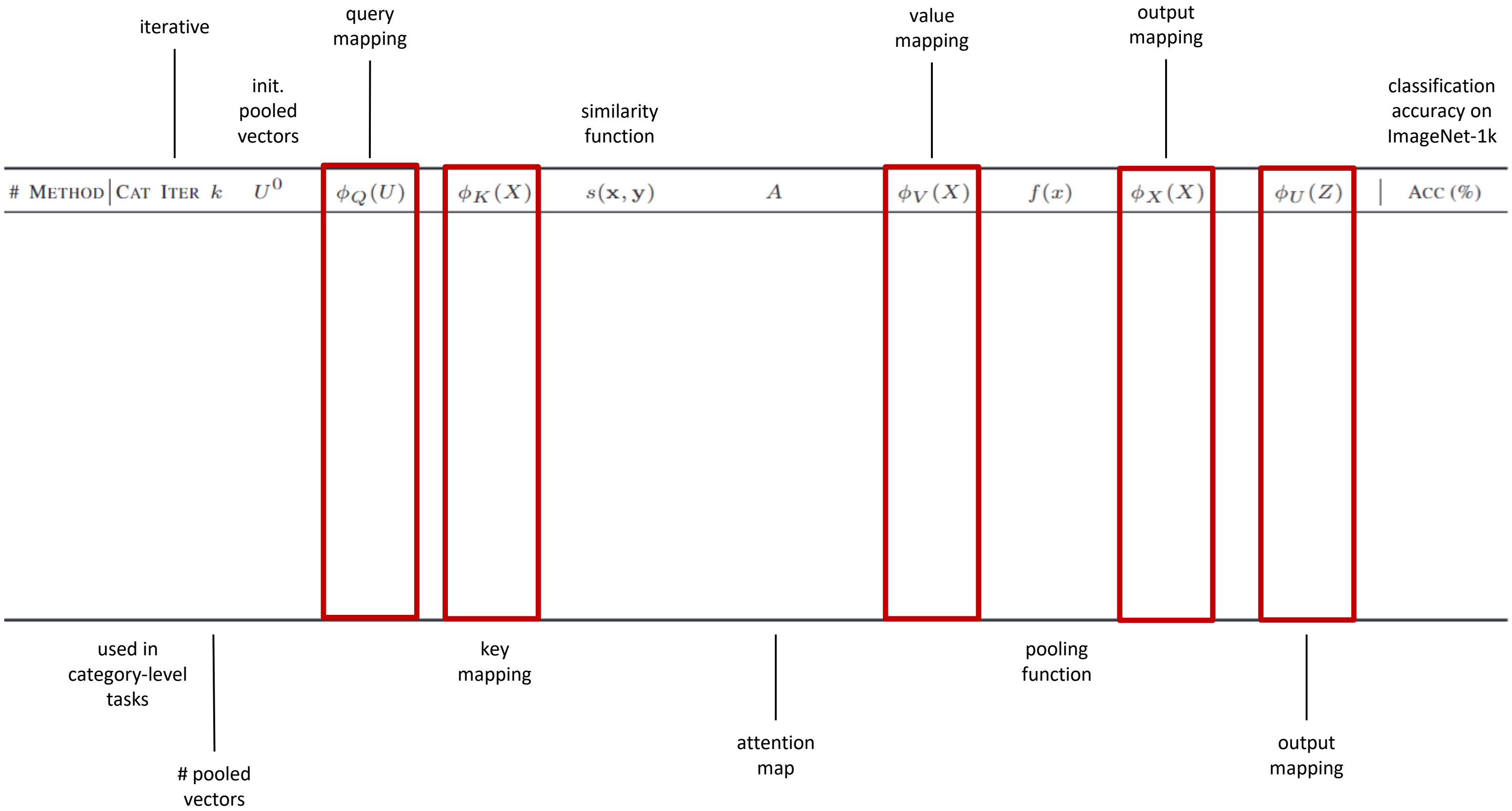
attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework

			iterative		init. pooled vectors	query mapping		similarity function		value mapping		output mapping		classification accuracy on ImageNet-1k
#	METHOD	CAT	ITER	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	ACC (%)
		used in category-level tasks		# pooled vectors			key mapping		attention map		pooling function		output mapping	

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Generic Pooling Framework

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x}, \mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Formulate methods as instantiations

iterative

query mapping

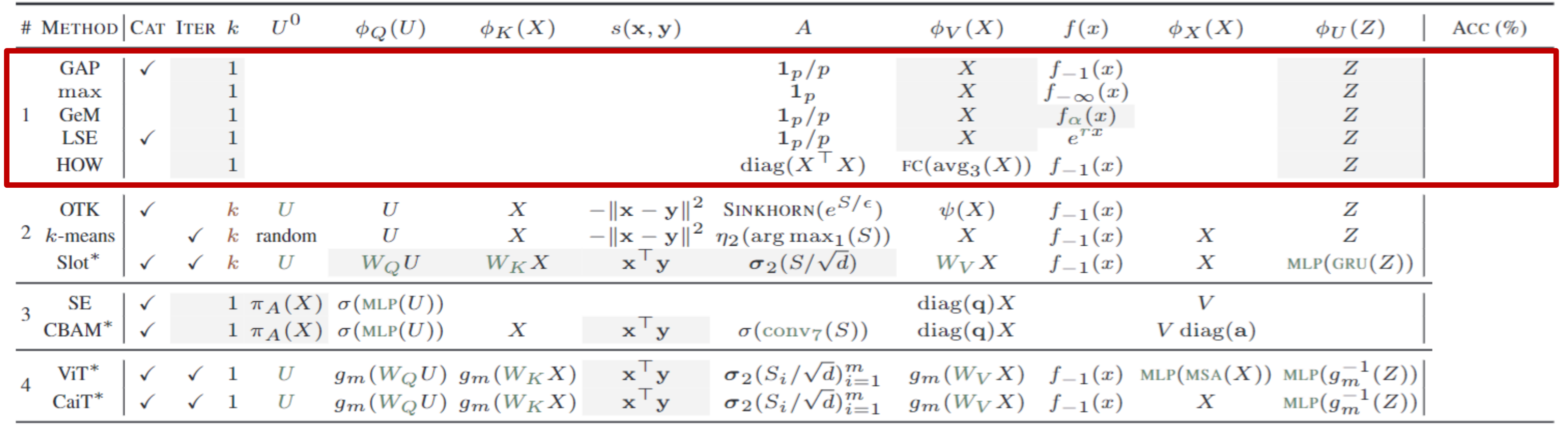
init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	
	GeM			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	
	HOW			1					$\operatorname{diag}(X^\top X)$	$\text{FC}(\operatorname{avg}_3(X))$	$f_{-1}(x)$		Z	
2	OTK	✓		k	U	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\text{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	
	k-means		✓	k	random	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\text{MLP}(\text{GRU}(Z))$	
3	SE	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$				$\operatorname{diag}(\mathbf{q})X$		V		
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\text{conv}_7(S))$	$\operatorname{diag}(\mathbf{q})X$		$V\operatorname{diag}(\mathbf{a})$		
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\text{MLP}(\text{MSA}(X))$	$\text{MLP}(g_m^{-1}(Z))$	
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\text{MLP}(g_m^{-1}(Z))$	

simple, k=1, non-attention

used in category-level tasks

key mapping

pooling function

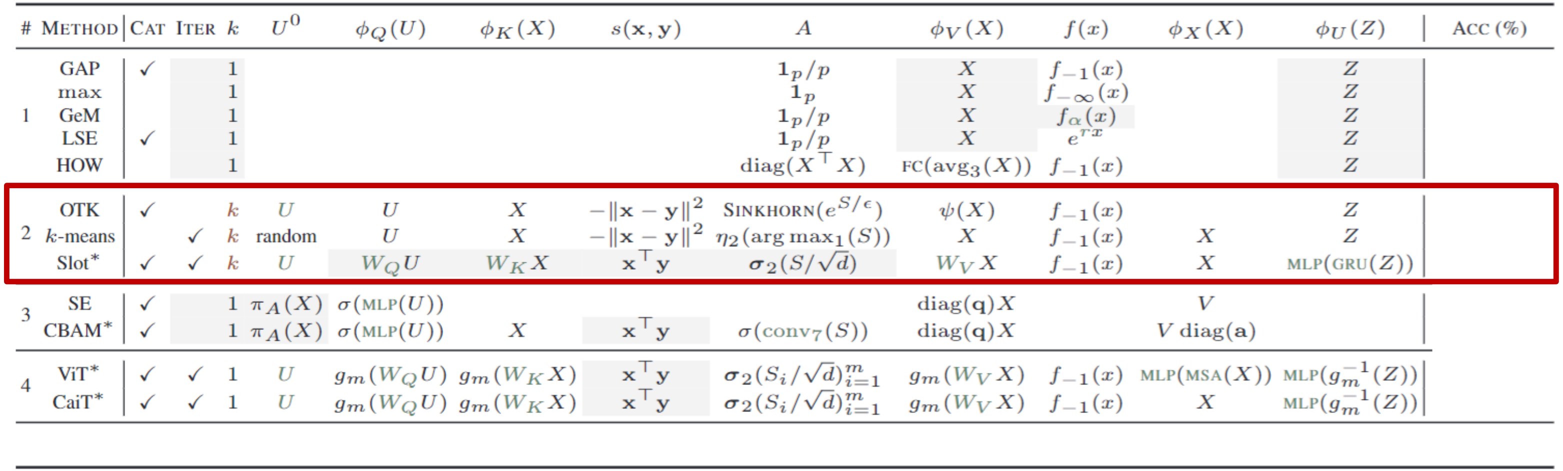
# pooled vectors

attention map

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Formulate methods as instantiations



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	
	GeM			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	
	HOW			1					$\operatorname{diag}(X^\top X)$	$\text{FC}(\text{avg}_3(X))$	$f_{-1}(x)$		Z	
2	OTK	✓		k	U	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\text{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	
	k-means		✓	k	random	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\text{MLP}(\text{GRU}(Z))$	
3	SE	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$				$\operatorname{diag}(\mathbf{q})X$		V		
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\text{conv}_7(S))$	$\operatorname{diag}(\mathbf{q})X$		$V\operatorname{diag}(\mathbf{a})$		
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\text{MLP}(\text{MSA}(X))$	$\text{MLP}(g_m^{-1}(Z))$	
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\text{MLP}(g_m^{-1}(Z))$	

iterative

init. pooled vectors

query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

simple, k=1, non-attention

k>1

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Formulate methods as instantiations

iterative

init. pooled vectors

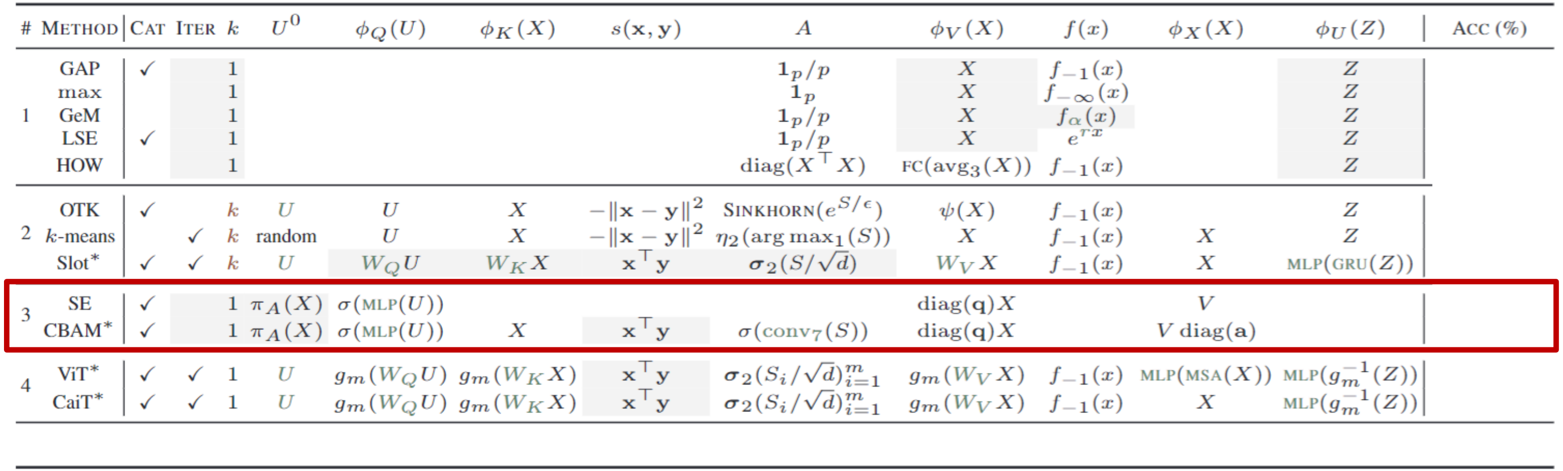
query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	
	GeM			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	
	HOW			1					$\operatorname{diag}(X^\top X)$	$\text{FC}(\text{avg}_3(X))$	$f_{-1}(x)$		Z	
2	OTK	✓		k	U	U	X	$-\Vert\mathbf{x}-\mathbf{y}\Vert^2$	$\text{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	
	k-means		✓	k	random	U	X	$-\Vert\mathbf{x}-\mathbf{y}\Vert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\text{MLP}(\text{GRU}(Z))$	
3	SE	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$				$\operatorname{diag}(\mathbf{q})X$		V		
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\text{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\text{conv}_7(S))$	$\operatorname{diag}(\mathbf{q})X$		$V\operatorname{diag}(\mathbf{a})$		
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\text{MLP}(\text{MSA}(X))$	$\text{MLP}(g_m^{-1}(Z))$	
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\text{MLP}(g_m^{-1}(Z))$	

simple, k=1, non-attention

k>1

modules within arch.

used in category-level tasks

key mapping

pooling function

# pooled vectors

attention map

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Formulate methods as instantiations

iterative

init. pooled vectors

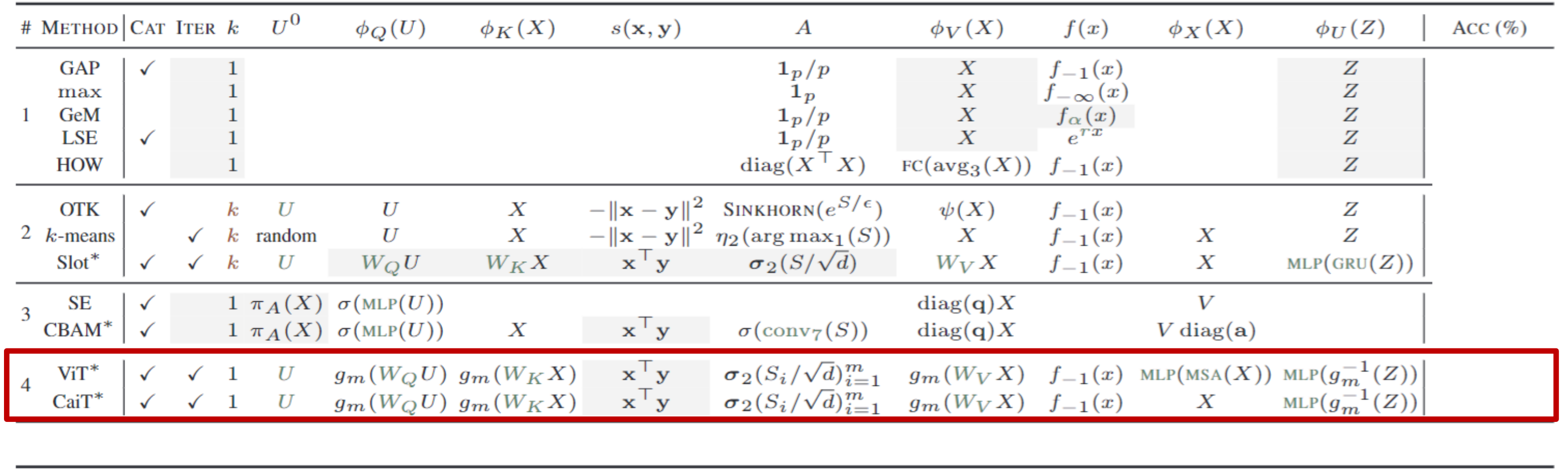
query mapping

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	
	GeM			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	
	HOW			1					$\mathrm{diag}(X^\top X)$	$\mathrm{FC}(\mathrm{avg}_3(X))$	$f_{-1}(x)$		Z	
2	OTK	✓		k	U	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\mathrm{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	
	k-means		✓	k	random	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\mathrm{MLP}(\mathrm{GRU}(Z))$	
3	SE	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$				$\mathrm{diag}(\mathbf{q})X$		V		
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\mathrm{conv}_7(S))$	$\mathrm{diag}(\mathbf{q})X$		$V\,\mathrm{diag}(\mathbf{a})$		
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\mathrm{MLP}(\mathrm{MSA}(X))$	$\mathrm{MLP}(g_m^{-1}(Z))$	
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\mathrm{MLP}(g_m^{-1}(Z))$	

simple, k=1, non-attention

k>1

modules within arch.

vision transformers

used in category-level tasks

key mapping

pooling function

output mapping

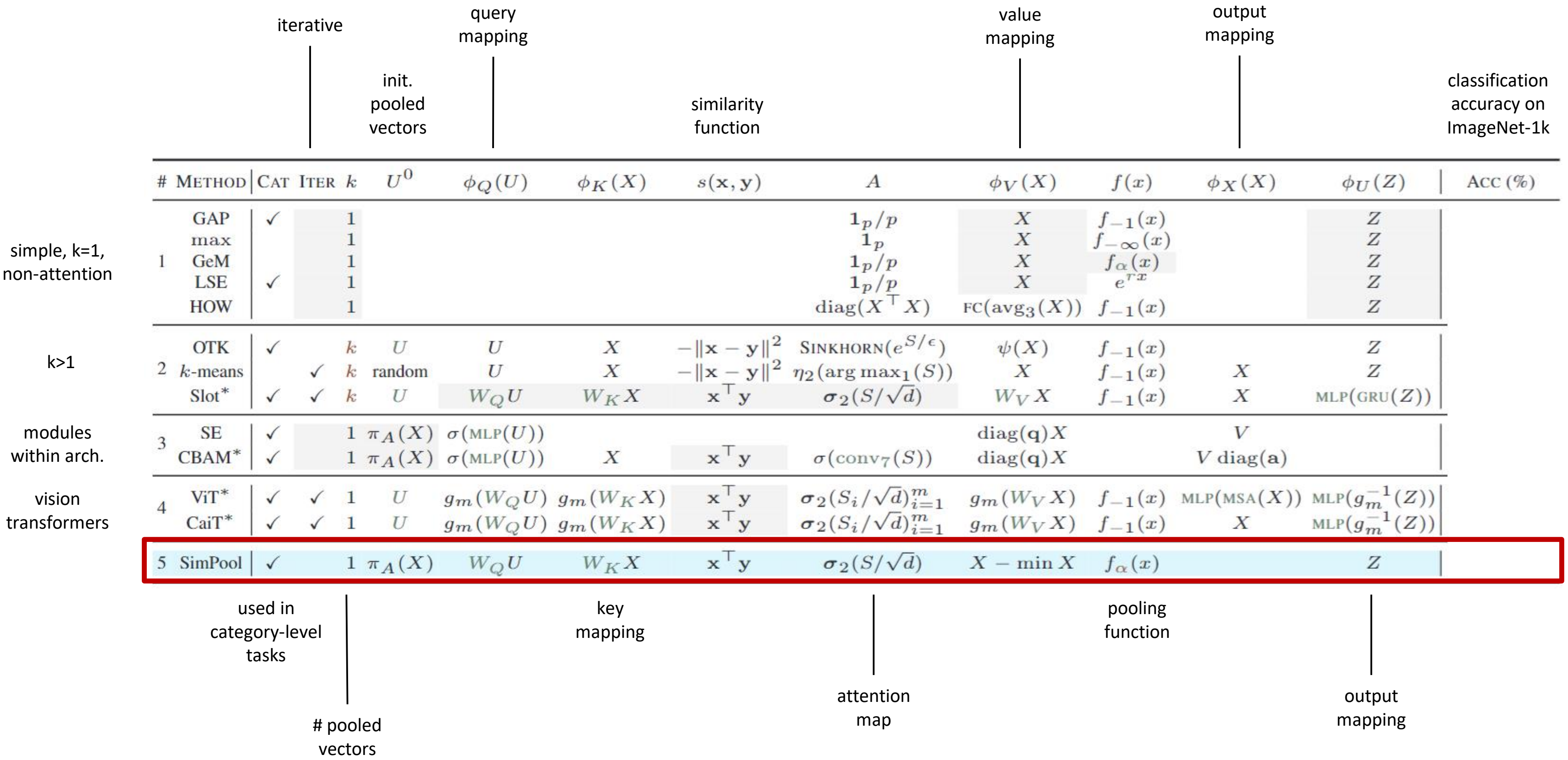
attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

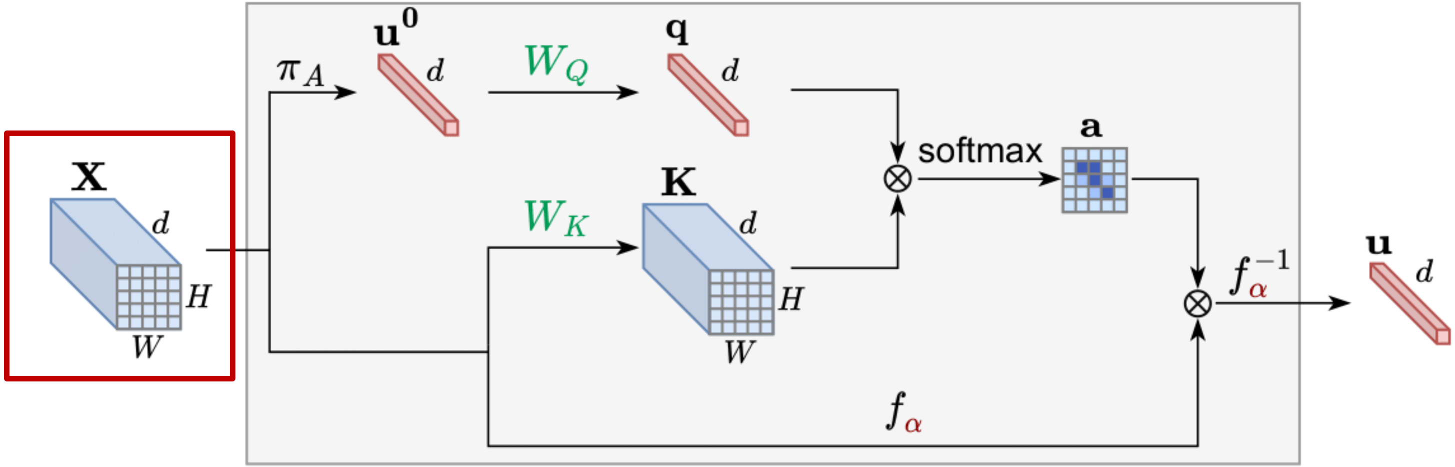
Discuss and derive



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	
	GeM			1					$\mathbf{1}_p/p$	X	$f_{\alpha}(x)$		Z	
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	
	HOW			1					$\mathrm{diag}(X^\top X)$	$\mathrm{FC}(\mathrm{avg}_3(X))$	$f_{-1}(x)$		Z	
2	OTK	✓		k	U	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\mathrm{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	
	k-means		✓	k	random	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\mathrm{MLP}(\mathrm{GRU}(Z))$	
3	SE	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$				$\mathrm{diag}(\mathbf{q})X$		V		
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\mathrm{conv}_7(S))$	$\mathrm{diag}(\mathbf{q})X$		$V\,\mathrm{diag}(\mathbf{a})$		
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\mathrm{MLP}(\mathrm{MSA}(X))$	$\mathrm{MLP}(g_m^{-1}(Z))$	
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\mathrm{MLP}(g_m^{-1}(Z))$	
5	SimPool	✓		1	$\pi_A(X)$	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\sigma_2(S/\sqrt{d})$	$X - \min X$	$f_{\alpha}(x)$		Z	

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

SimPool



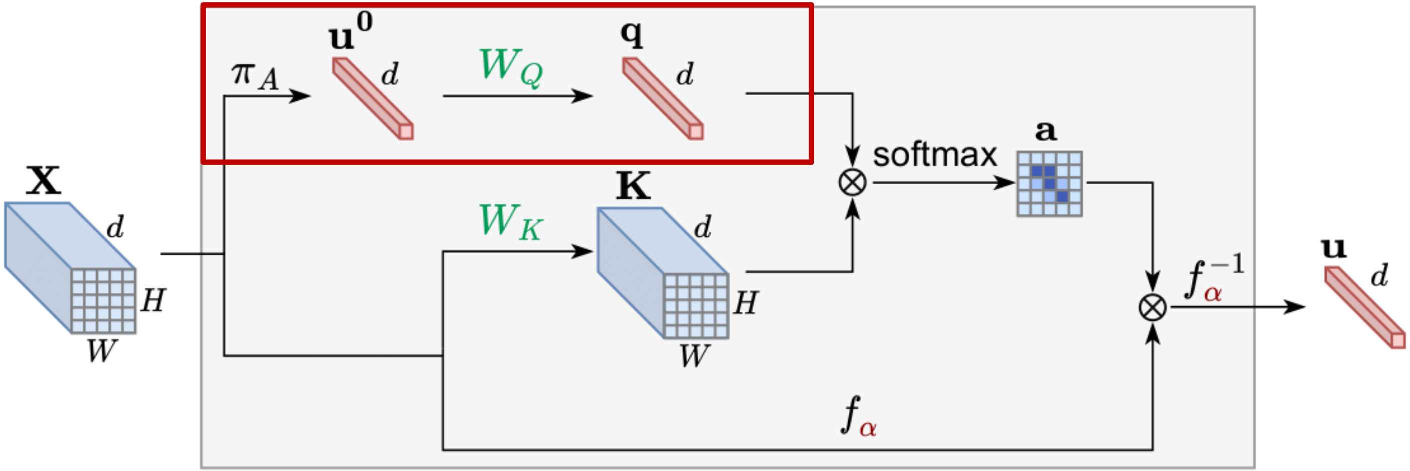
- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by W_Q (W_K) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left(K^\top \mathbf{q} / \sqrt{d} \right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

SimPool



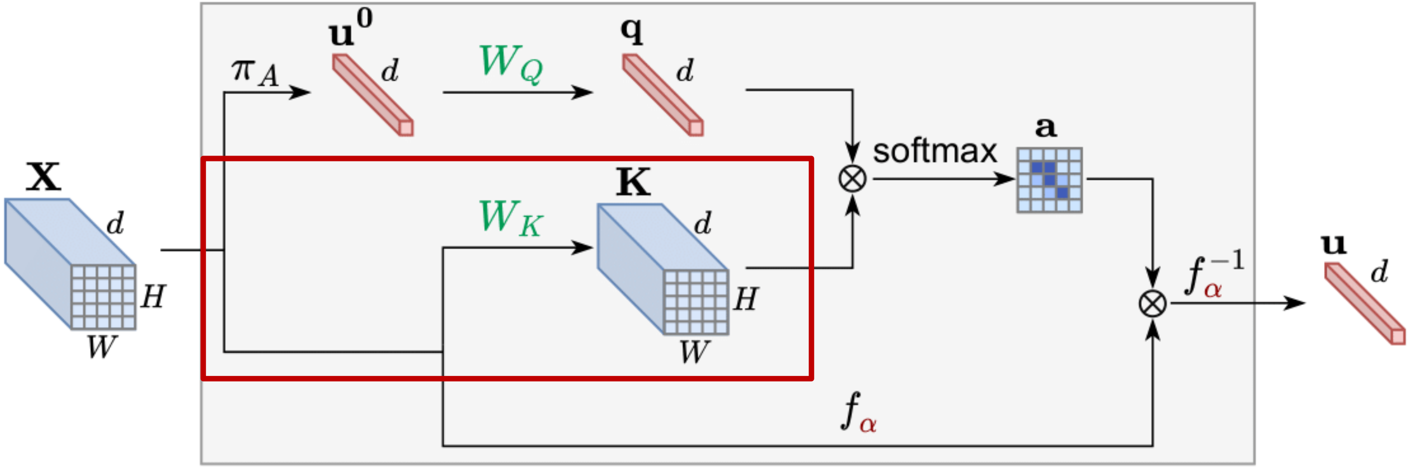
- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ $(\mathbf{X})$ mapped by W_Q (W_K) to form $\mathbf{q}$ $(\mathbf{K})$.
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

SimPool



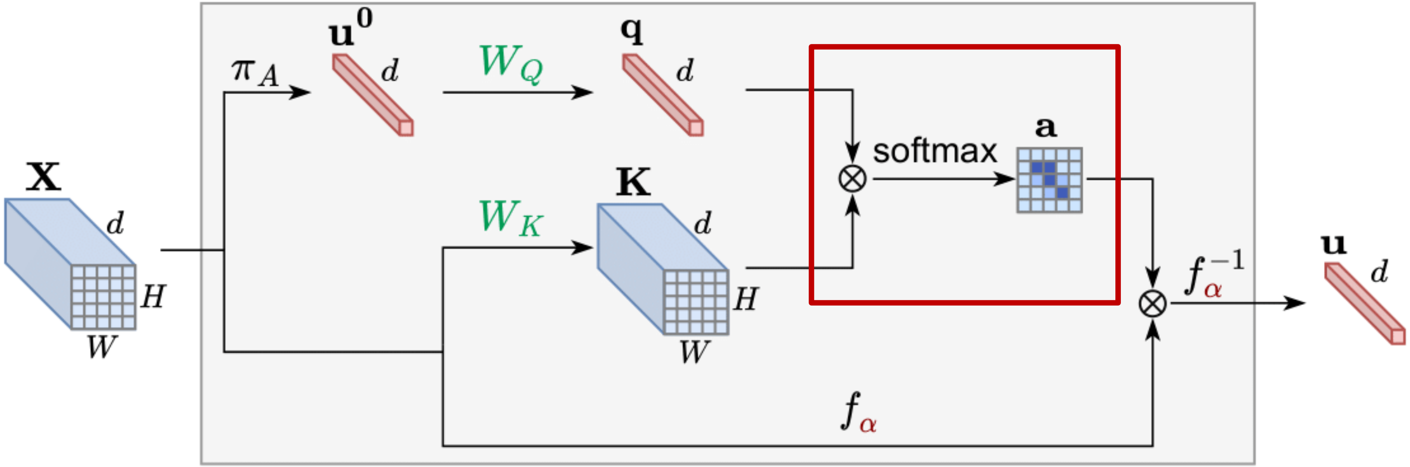
- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by W_Q (W_K) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

SimPool



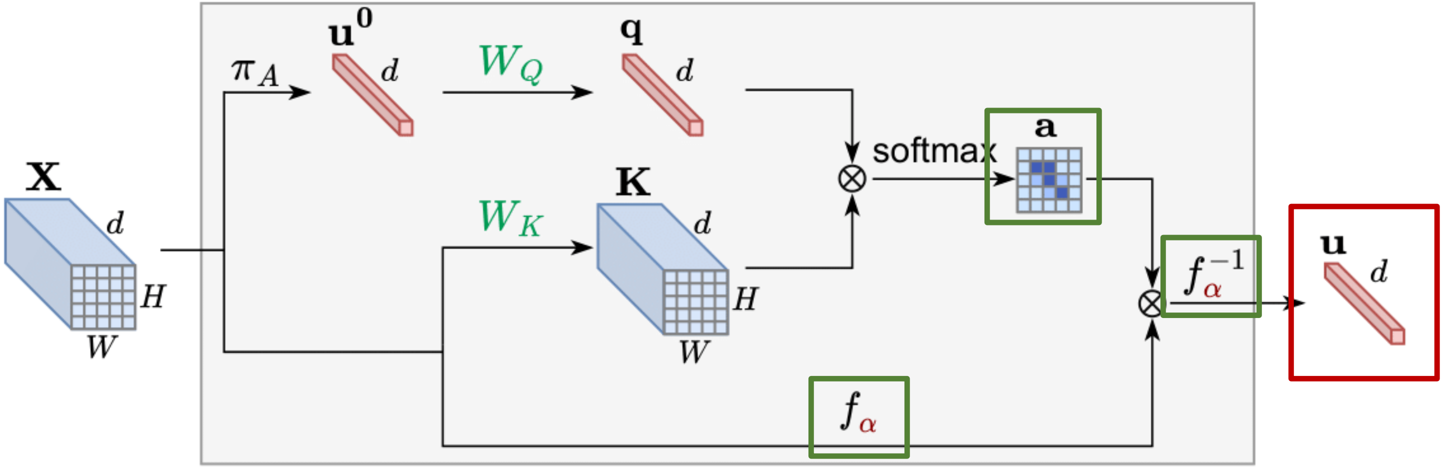
- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by W_Q (W_K) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left(K^\top \mathbf{q} / \sqrt{d} \right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

SimPool



- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by W_Q (W_K) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Benchmark

iterative

init. pooled vectors

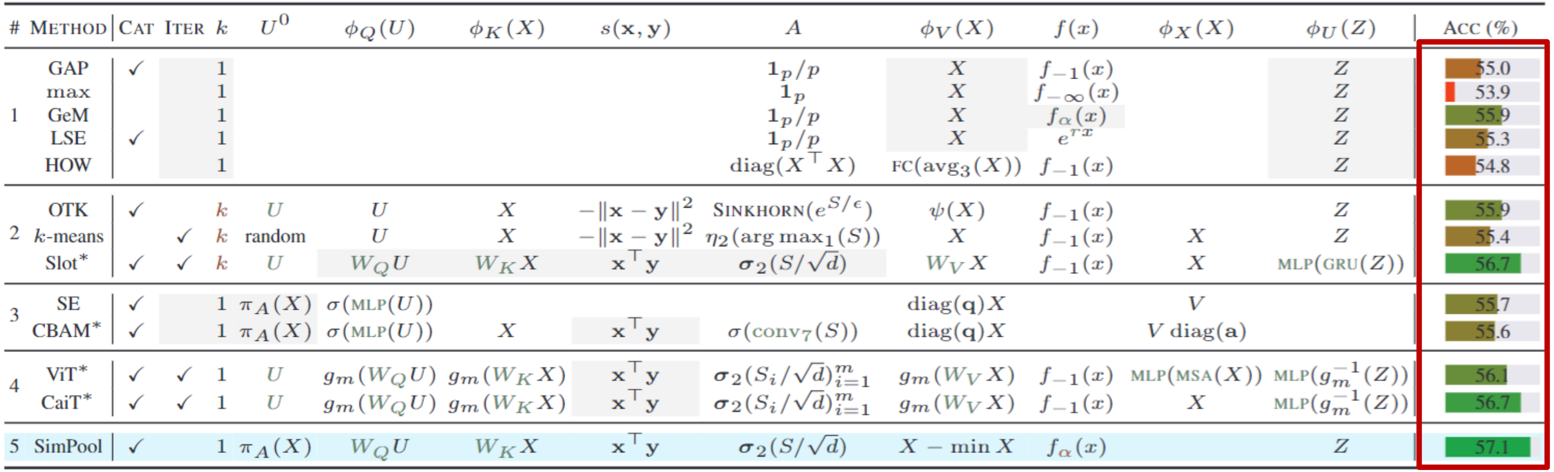
query mapping

similarity function

value mapping

output mapping

classification accuracy on 20% of ImageNet-1k



#	Method	Cat	Iter	k	U^0	$\phi_Q(U)$	$\phi_K(X)$	$s(\mathbf{x},\mathbf{y})$	A	$\phi_V(X)$	$f(x)$	$\phi_X(X)$	$\phi_U(Z)$	Acc (%)
1	GAP	✓		1					$\mathbf{1}_p/p$	X	$f_{-1}(x)$		Z	55.0
	max			1					$\mathbf{1}_p$	X	$f_{-\infty}(x)$		Z	53.9
	GeM			1					$\mathbf{1}_p/p$	X	$f_\alpha(x)$		Z	55.9
	LSE	✓		1					$\mathbf{1}_p/p$	X	e^{rx}		Z	55.3
	HOW			1					$\operatorname{diag}(X^\top X)$	$\mathrm{FC}(\operatorname{avg}_3(X))$	$f_{-1}(x)$		Z	54.8
2	OTK	✓		k	U	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\mathrm{Sinkhorn}(e^{S/\epsilon})$	$\psi(X)$	$f_{-1}(x)$		Z	55.9
	k-means		✓	k	random	U	X	$-\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\eta_2(\arg\max_1(S))$	X	$f_{-1}(x)$	X	Z	55.4
	Slot*	✓	✓	k	U	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$W_V X$	$f_{-1}(x)$	X	$\mathrm{MLP}(\mathrm{GRU}(Z))$	56.7
3	SE	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$				$\operatorname{diag}(\mathbf{q})X$		V		55.7
	CBAM*	✓		1	$\pi_A(X)$	$\sigma(\mathrm{MLP}(U))$	X	$\mathbf{x}^\top\mathbf{y}$	$\sigma(\operatorname{conv}_7(S))$	$\operatorname{diag}(\mathbf{q})X$		$V \operatorname{diag}(\mathbf{a})$		55.6
4	ViT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	$\mathrm{MLP}(\mathrm{MSA}(X))$	$\mathrm{MLP}(g_m^{-1}(Z))$	56.1
	CaiT*	✓	✓	1	U	$g_m(W_Q U)$	$g_m(W_K X)$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$	$g_m(W_V X)$	$f_{-1}(x)$	X	$\mathrm{MLP}(g_m^{-1}(Z))$	56.7
5	SimPool	✓		1	$\pi_A(X)$	$W_Q U$	$W_K X$	$\mathbf{x}^\top\mathbf{y}$	$\boldsymbol{\sigma}_2(S/\sqrt{d})$	$X - \min X$	$f_\alpha(x)$		Z	57.1

simple, k=1, non-attention

k>1

modules within arch.

vision transformers

used in category-level tasks

# pooled vectors

key mapping

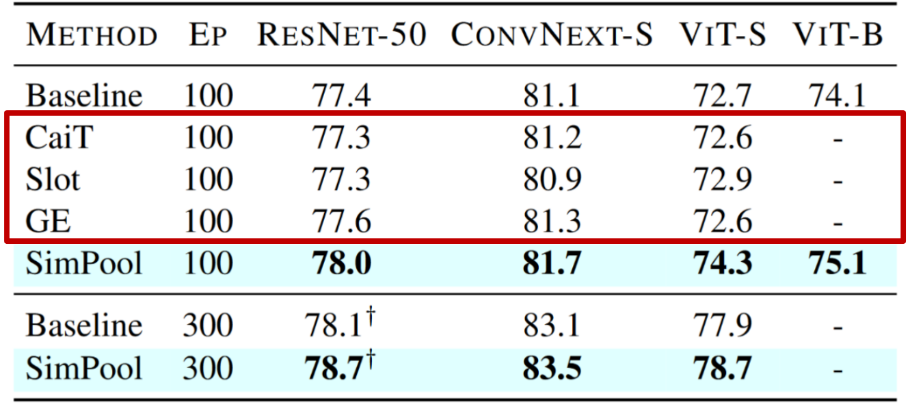
attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: “Universal” (Network & Settings)

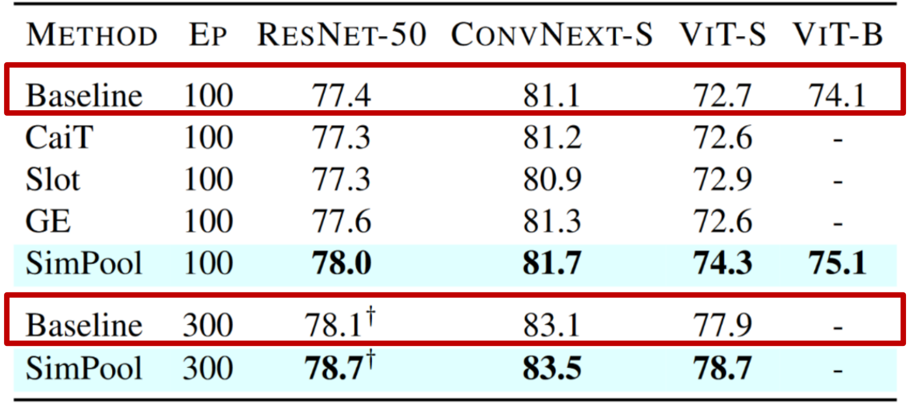


METHOD	EP	RESNET-50	CONVNEXT-S	VIT-S	VIT-B
Baseline	100	77.4	81.1	72.7	74.1
CaiT	100	77.3	81.2	72.6	-
Slot	100	77.3	80.9	72.9	-
GE	100	77.6	81.3	72.6	-
SimPool	100	**78.0**	**81.7**	**74.3**	**75.1**
Baseline	300	78.1†	83.1	77.9	-
SimPool	300	**78.7**†	**83.5**	**78.7**	-

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: “Universal” (Network & Settings)

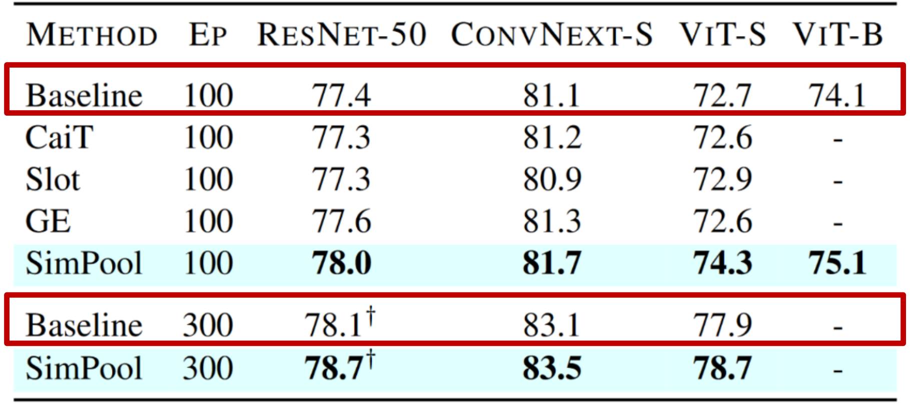


Method	Ep	ResNet-50	ConvNext-S	ViT-S	ViT-B
Baseline	100	77.4	81.1	72.7	74.1
CaiT	100	77.3	81.2	72.6	-
Slot	100	77.3	80.9	72.9	-
GE	100	77.6	81.3	72.6	-
SimPool	100	**78.0**	**81.7**	**74.3**	**75.1**
Baseline	300	78.1†	83.1	77.9	-
SimPool	300	**78.7**†	**83.5**	**78.7**	-

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

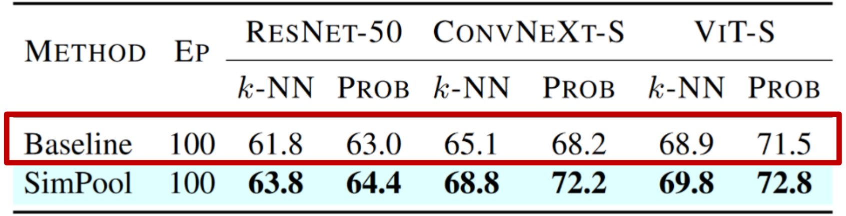
Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: “Universal” (Network & Settings)



METHOD	EP	RESNET-50	CONVNEXT-S	VIT-S	VIT-B
Baseline	100	77.4	81.1	72.7	74.1
CaiT	100	77.3	81.2	72.6	-
Slot	100	77.3	80.9	72.9	-
GE	100	77.6	81.3	72.6	-
SimPool	100	**78.0**	**81.7**	**74.3**	**75.1**
Baseline	300	78.1†	83.1	77.9	-
SimPool	300	**78.7**†	**83.5**	**78.7**	-

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

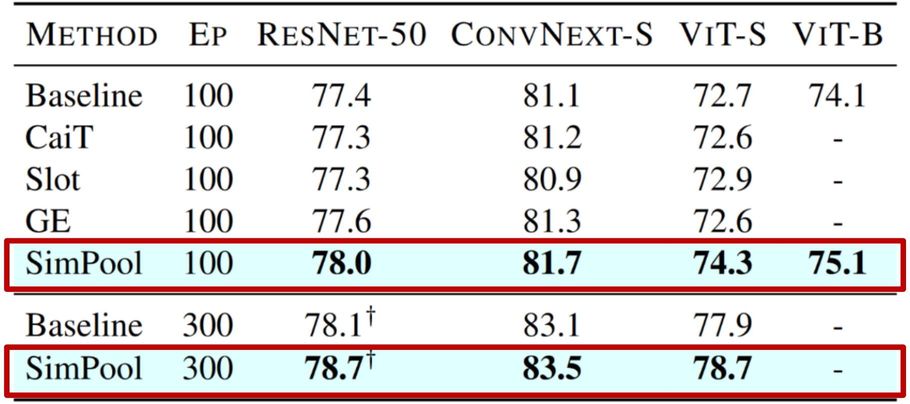


METHOD	EP	RESNET-50		CONVNEXT-S		VIT-S	
		k-NN	PROB	k-NN	PROB	k-NN	PROB
Baseline	100	61.8	63.0	65.1	68.2	68.9	71.5
SimPool	100	**63.8**	**64.4**	**68.8**	**72.2**	**69.8**	**72.8**

Classification accuracy on ImageNet-1k;
Self-supervised pre-training w/ DINO;
Baseline: GAP for convolutional, [CLS] for transformers.

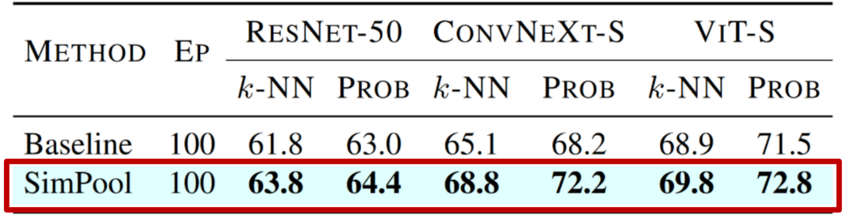
Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: “Universal” (Network & Settings)



Method	Ep	ResNet-50	ConvNeXt-S	ViT-S	ViT-B
Baseline	100	77.4	81.1	72.7	74.1
CaiT	100	77.3	81.2	72.6	-
Slot	100	77.3	80.9	72.9	-
GE	100	77.6	81.3	72.6	-
SimPool	100	**78.0**	**81.7**	**74.3**	**75.1**
Baseline	300	$78.1^\dagger$	83.1	77.9	-
SimPool	300	**$78.7^\dagger$**	**83.5**	**78.7**	-

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

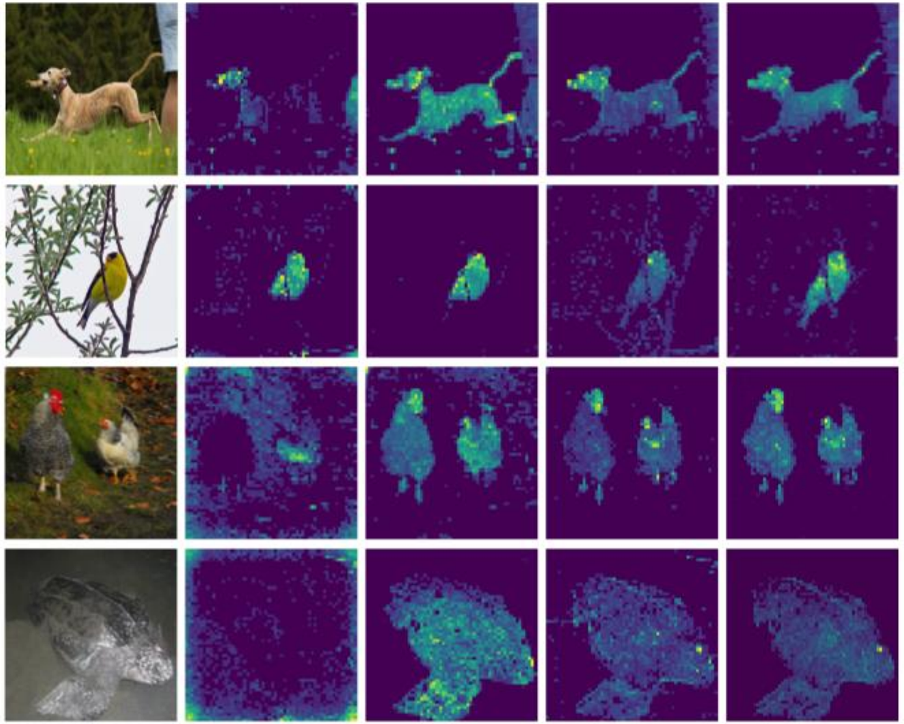


Method	Ep	ResNet-50		ConvNeXt-S		ViT-S	
		k-NN	Prob	k-NN	Prob	k-NN	Prob
Baseline	100	61.8	63.0	65.1	68.2	68.9	71.5
SimPool	100	**63.8**	**64.4**	**68.8**	**72.2**	**69.8**	**72.8**

Classification accuracy on ImageNet-1k;
Self-supervised pre-training w/ DINO;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

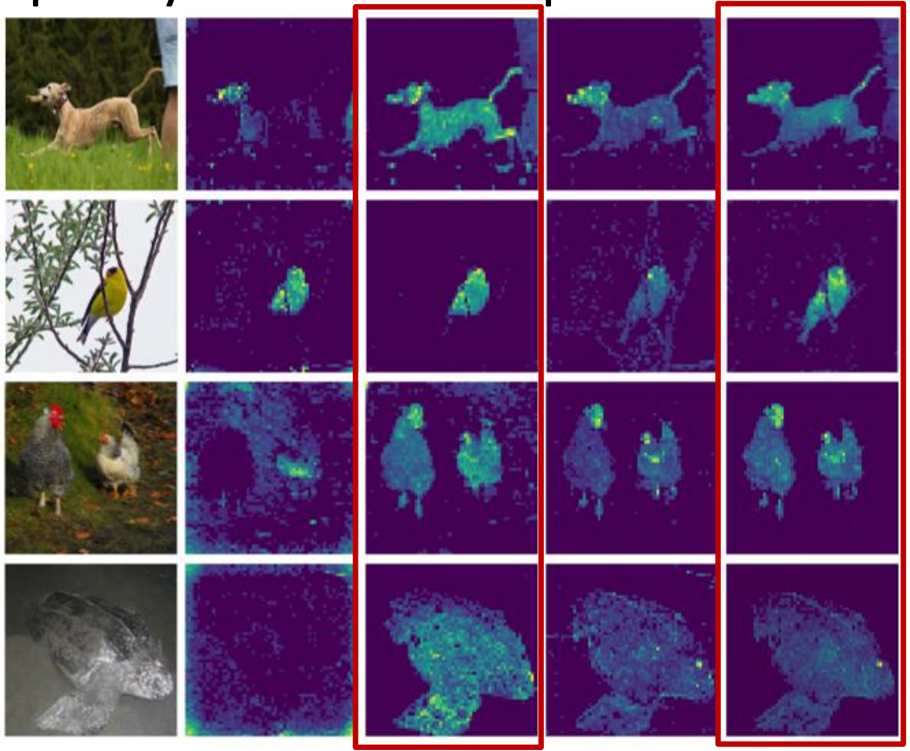
Property: High-quality attention maps from Transformers



ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

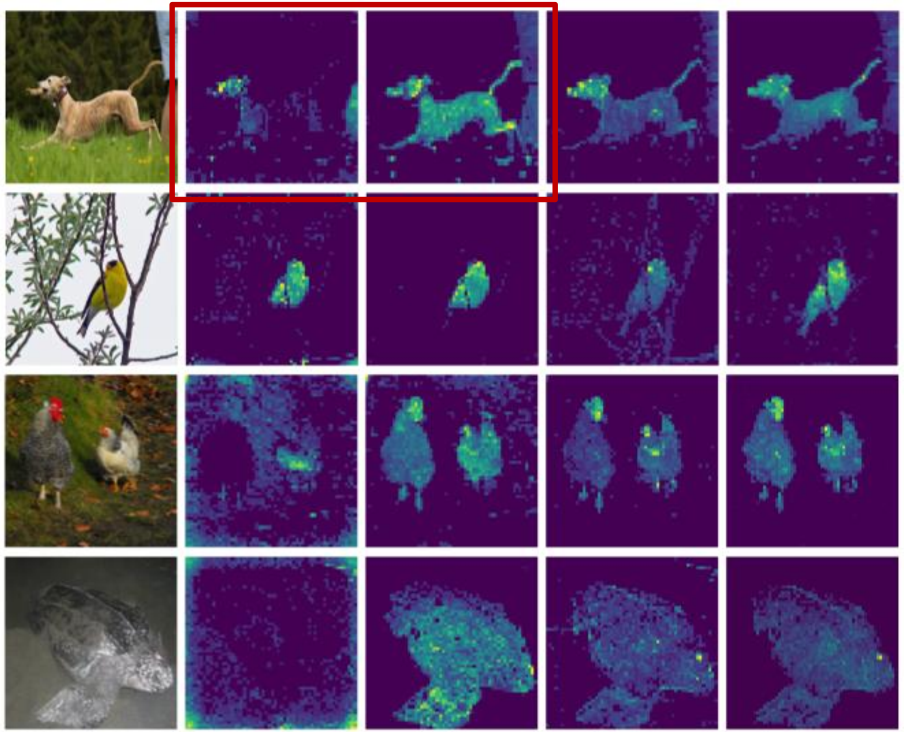
Property: High-quality attention maps from Transformers



ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

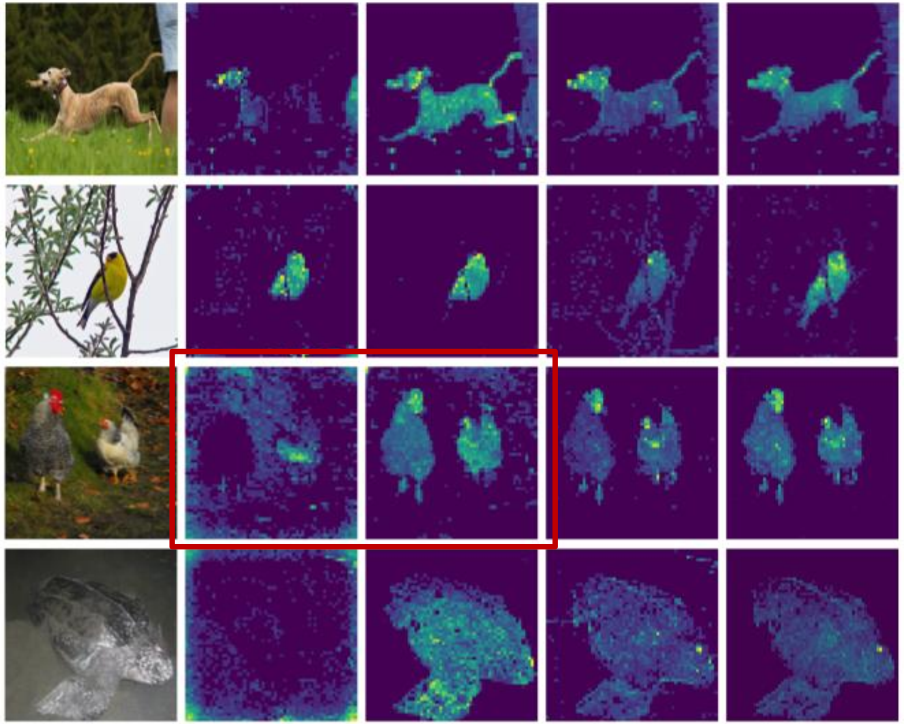
Property: High-quality attention maps from Transformers



ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

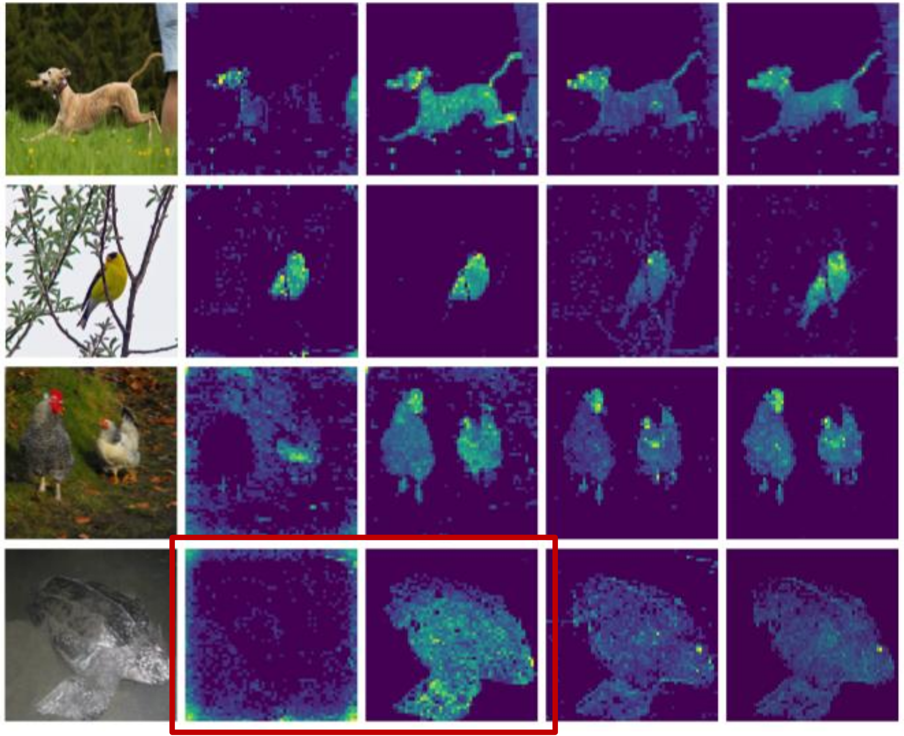
Property: High-quality attention maps from Transformers



ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

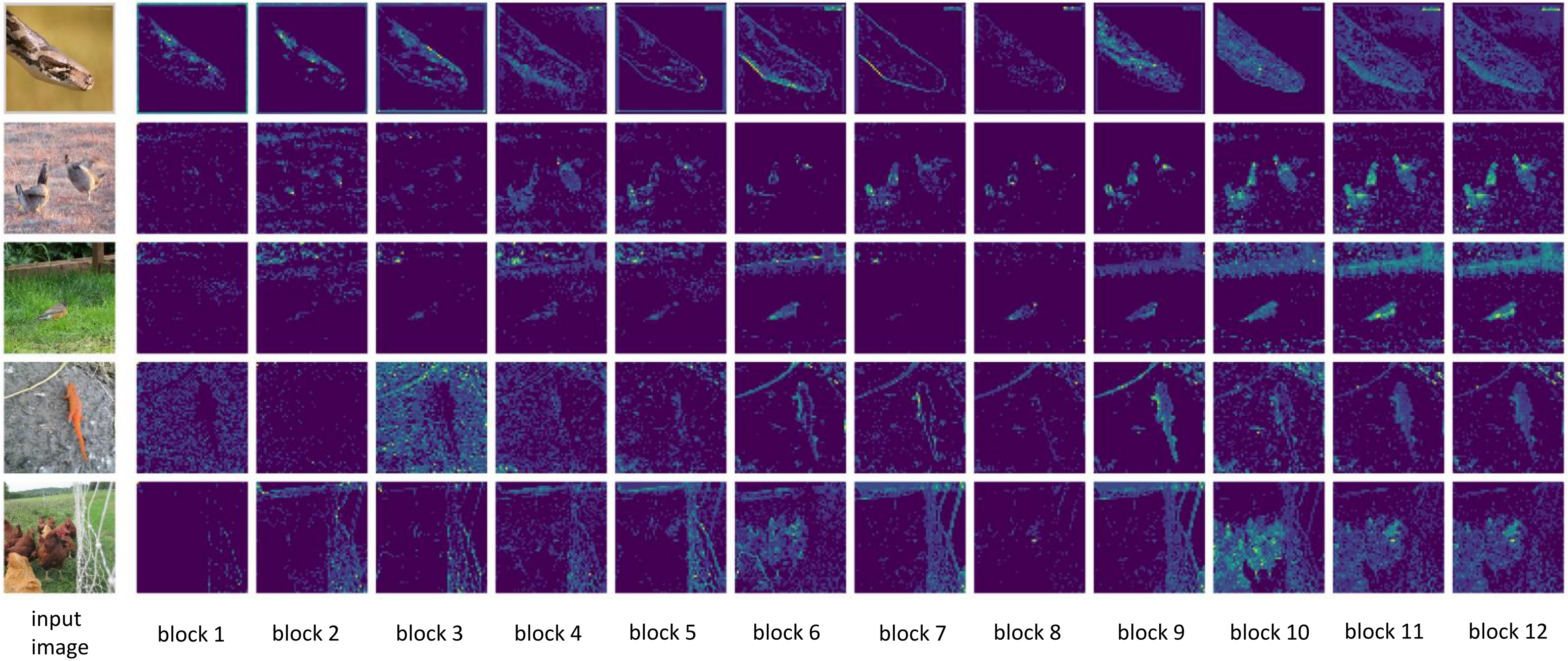
Property: High-quality attention maps from Transformers



ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

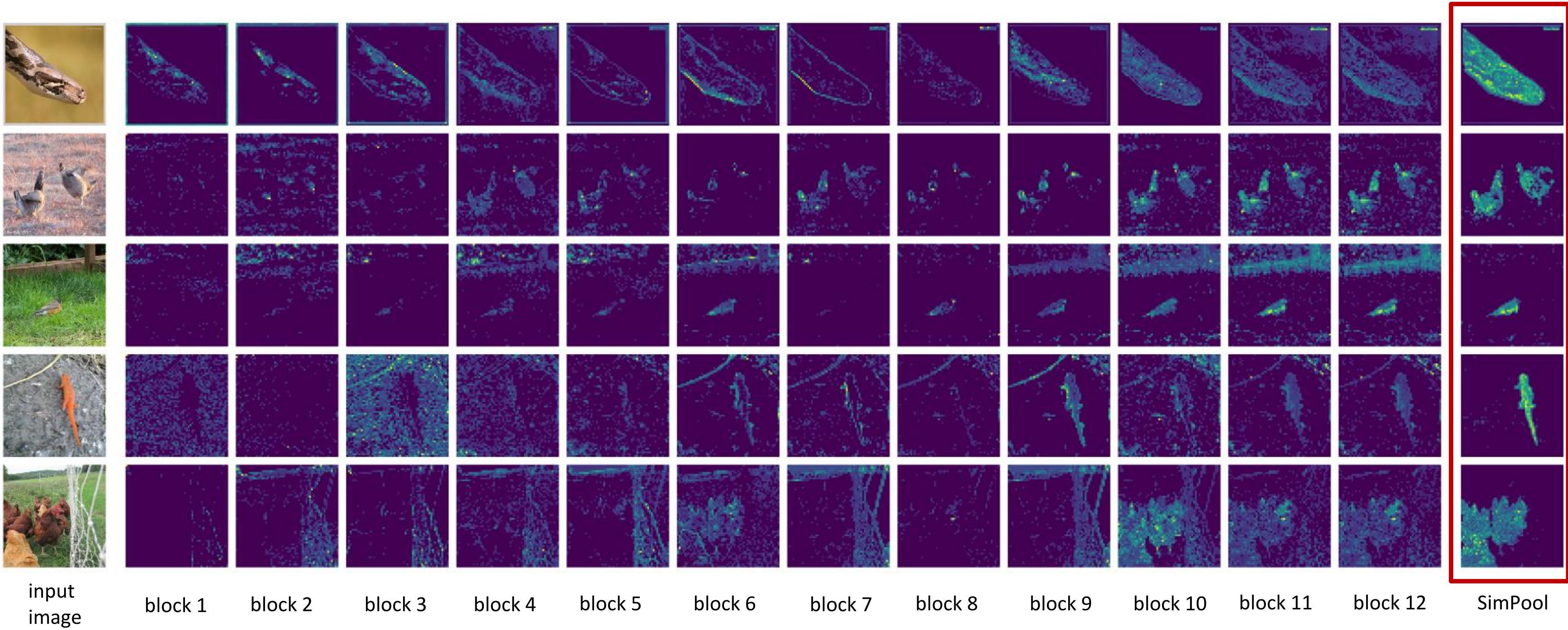
Property: Resolving the attention “deficit”



ViT-S on Imagenet-1k; supervised training;
mean attention map of the [CLS]

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: Resolving the attention "deficit"



ViT-S on Imagenet-1k; supervised training;
mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Property: Localization

What does "high-quality" attention maps mean?

How can we quantitatively evaluate the quality of the attention maps?

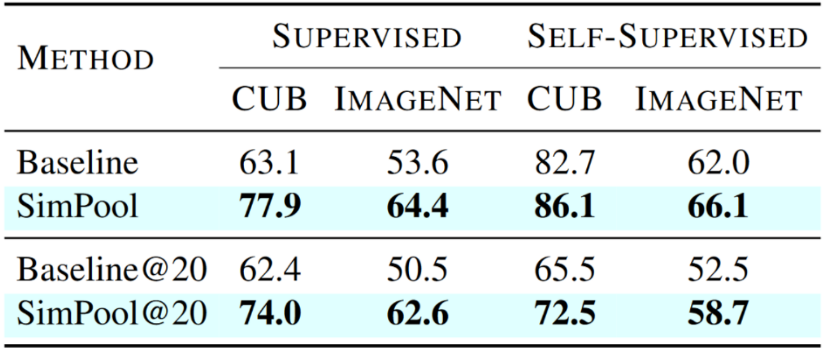
Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

Property: Localization

What does "high-quality" attention maps mean?
How can we quantitatively evaluate the quality of the attention maps?



Method	Supervised		Self-Supervised	
	CUB	ImageNet	CUB	ImageNet
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**

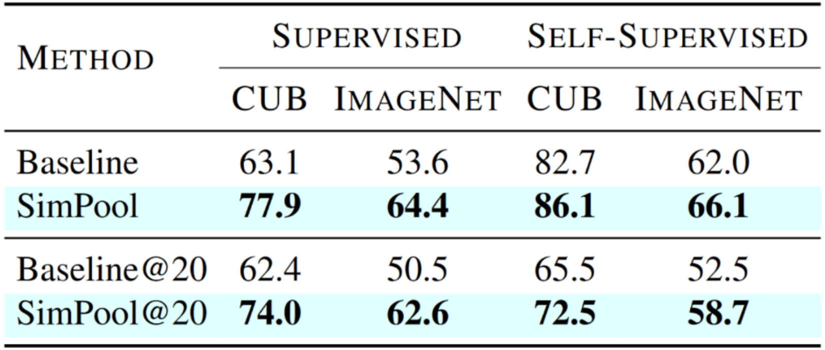
Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

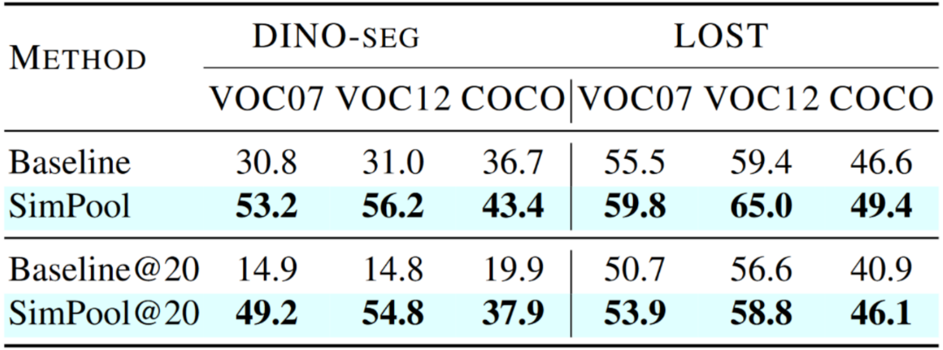
Property: Localization

What does “high-quality” attention maps mean?
How can we quantitatively evaluate the quality of the attention maps?



Method	Supervised		Self-Supervised	
	CUB	ImageNet	CUB	ImageNet
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20



Method	DINO-seg			LOST		
	VOC07	VOC12	COCO	VOC07	VOC12	COCO
Baseline	30.8	31.0	36.7	55.5	59.4	46.6
SimPool	**53.2**	**56.2**	**43.4**	**59.8**	**65.0**	**49.4**
Baseline@20	14.9	14.8	19.9	50.7	56.6	40.9
SimPool@20	**49.2**	**54.8**	**37.9**	**53.9**	**58.8**	**46.1**

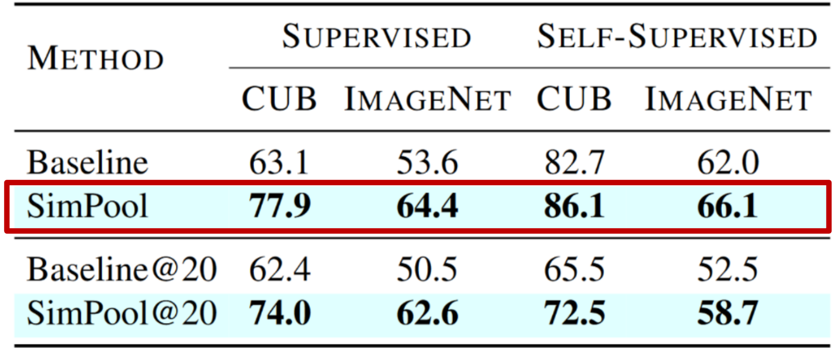
Unsupervised object discovery CorLoc with ViT-S;
DINO-seg uses attention maps;
LOST uses raw features;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

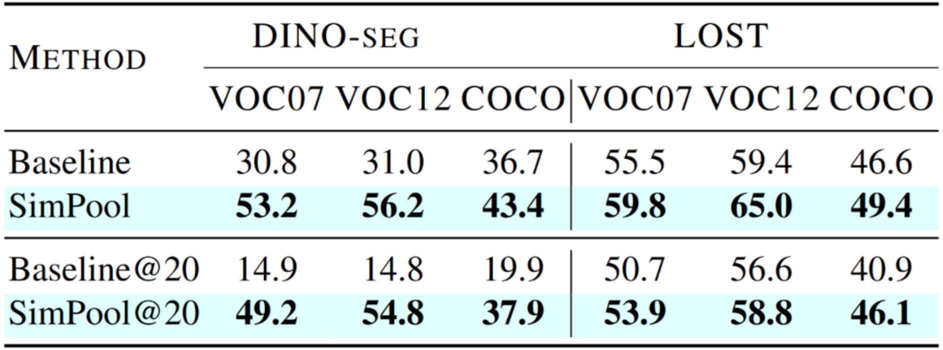
Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

Property: Localization

What does “high-quality” attention maps mean?
How can we quantitatively evaluate the quality of the attention maps?



METHOD	SUPERVISED		SELF-SUPERVISED	
	CUB	IMAGENET	CUB	IMAGENET
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**



METHOD	DINO-SEG			LOST		
	VOC07	VOC12	COCO	VOC07	VOC12	COCO
Baseline	30.8	31.0	36.7	55.5	59.4	46.6
SimPool	**53.2**	**56.2**	**43.4**	**59.8**	**65.0**	**49.4**
Baseline@20	14.9	14.8	19.9	50.7	56.6	40.9
SimPool@20	**49.2**	**54.8**	**37.9**	**53.9**	**58.8**	**46.1**

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

Unsupervised object discovery CorLoc with ViT-S;
DINO-SEG uses attention maps;
LOST uses raw features;
@20: at epoch 20

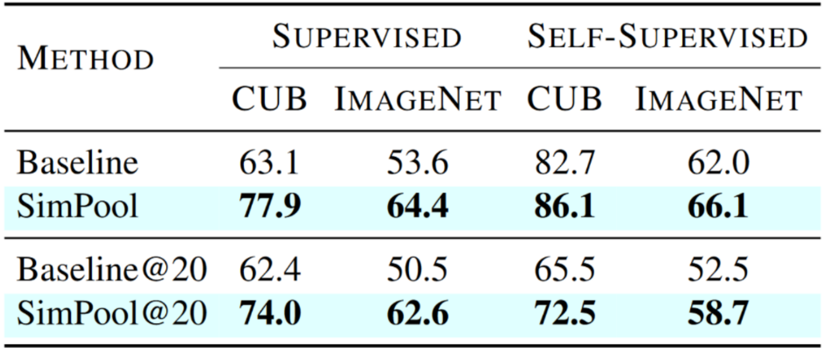
✓ Up to +14% when supervised and up to +7% when self-supervised

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

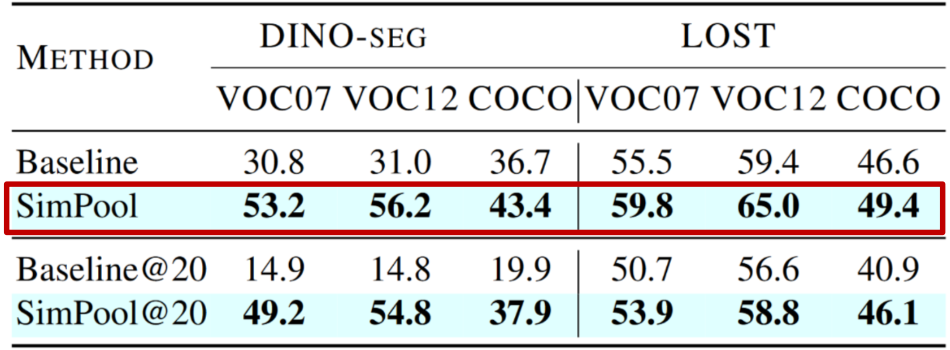
Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

Property: Localization

What does “high-quality” attention maps mean?
How can we quantitatively evaluate the quality of the attention maps?



METHOD	SUPERVISED		SELF-SUPERVISED	
	CUB	IMAGENET	CUB	IMAGENET
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**



METHOD	DINO-SEG			LOST		
	VOC07	VOC12	COCO	VOC07	VOC12	COCO
Baseline	30.8	31.0	36.7	55.5	59.4	46.6
SimPool	**53.2**	**56.2**	**43.4**	**59.8**	**65.0**	**49.4**
Baseline@20	14.9	14.8	19.9	50.7	56.6	40.9
SimPool@20	**49.2**	**54.8**	**37.9**	**53.9**	**58.8**	**46.1**

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

Unsupervised object discovery CorLoc with ViT-S;
DINO-SEG uses attention maps;
LOST uses raw features;
@20: at epoch 20

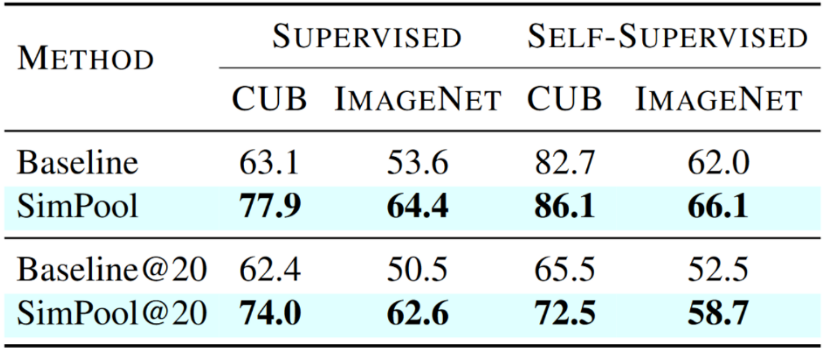
✓ Up to +14% when supervised and up to +7% when self-supervised

✓ Up to +25% for DINO-seg and up to +6% for LOST

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

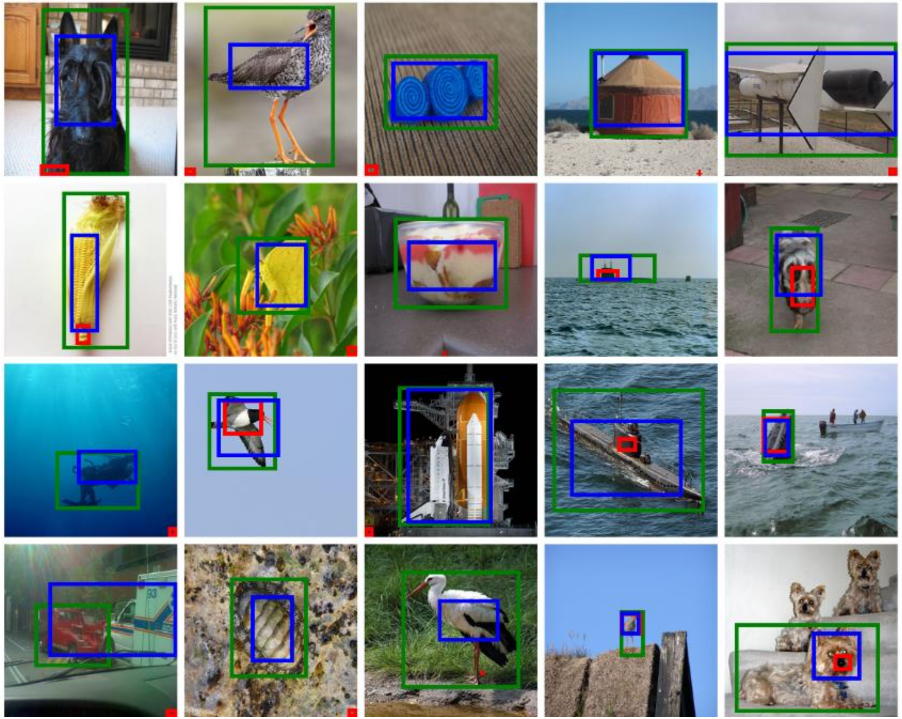
Property: Localization



METHOD	SUPERVISED		SELF-SUPERVISED	
	CUB	IMAGENET	CUB	IMAGENET
Baseline	63.1	53.6	82.7	62.0
SimPool	**77.9**	**64.4**	**86.1**	**66.1**
Baseline@20	62.4	50.5	65.5	52.5
SimPool@20	**74.0**	**62.6**	**72.5**	**58.7**

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020



Object localization on ImageNet-1k;
green: ground-truth; red: baseline; blue: SimPool

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

Summarizing Insights

SimPool:

- ✓ Improves performance of convolutional networks and transformers under supervised or self-supervised setting
- ✓ Outperforms the other pooling methods
- ✓ Incurs low additional cost
- ✓ Produces high-quality attention maps that delineate object boundaries
- ✓ Presents strong localization properties

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

3. Multimodal Representations

IEEE
IGARSS
2024

July 7-12
ATHENS
GREECE

Extracting Multimodal Representations for Remote Sensing Composed Image Retrieval [WeiCom]

Psomas et al. *Composed Image Retrieval for Remote Sensing,* **IGARSS 2024**

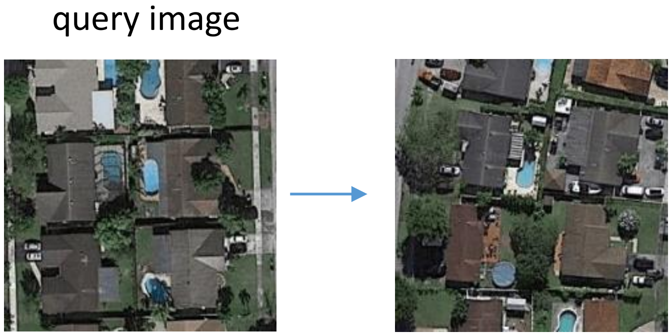
Source code: https://github.com/billpsomas/rscir

Related Work: Remote Sensing Image Retrieval

query image

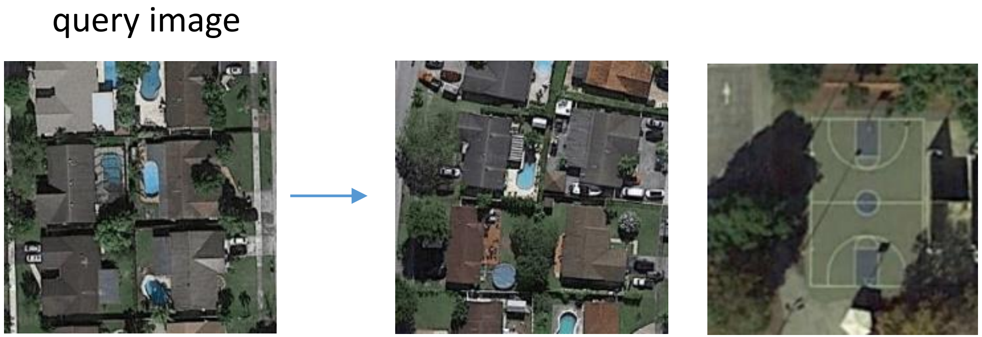
Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval



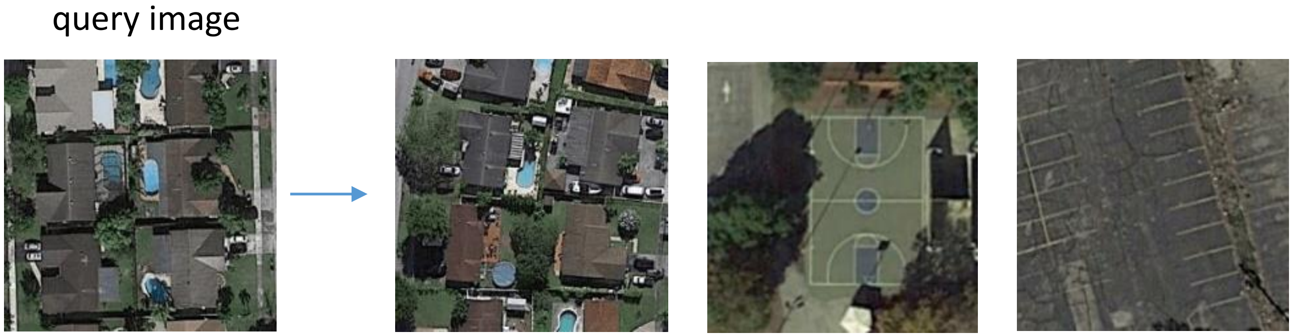
Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval



Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval

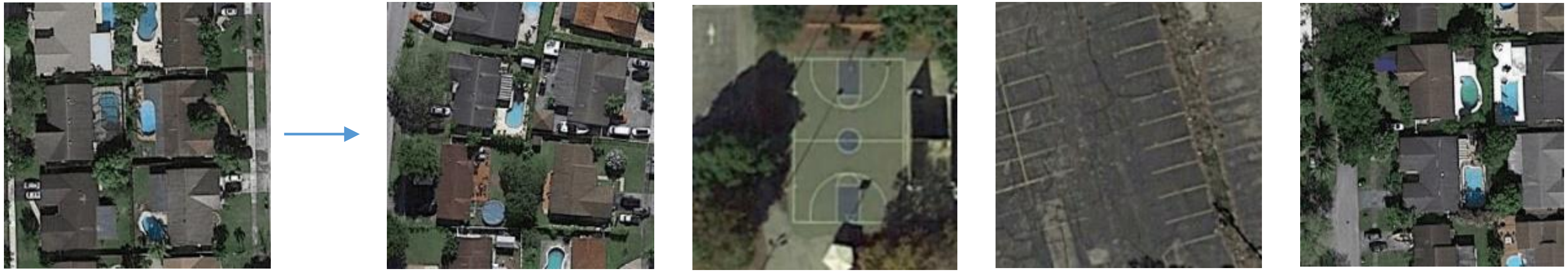


Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval

query image

top-k retrieved images



descending order of
similarity to the query image

Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval

query image

residential

top-k retrieved images

residential

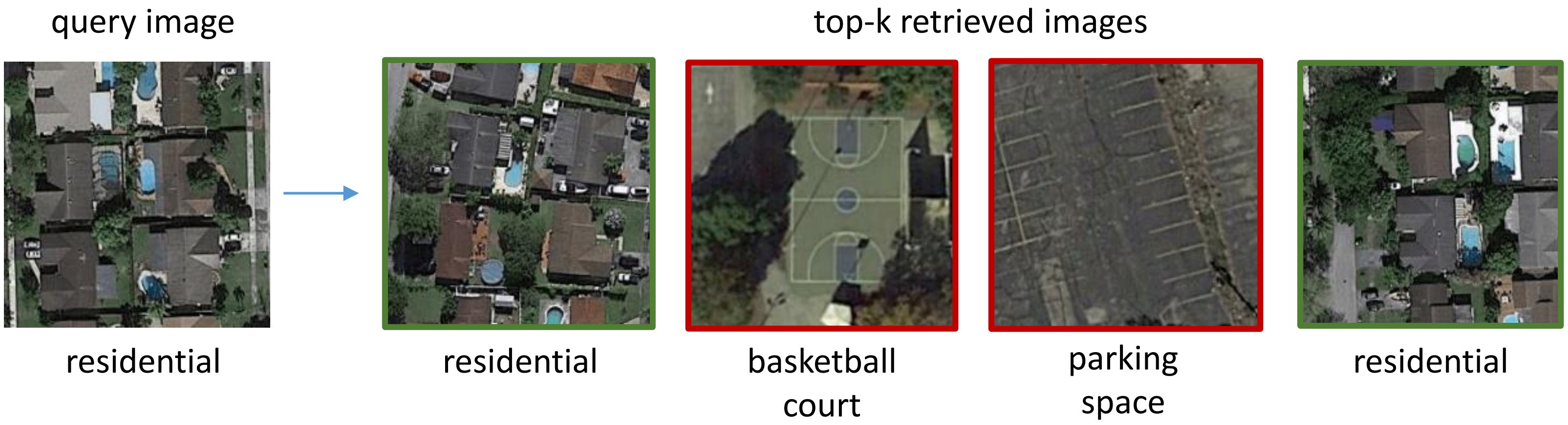
basketball court

parking space

residential

Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

Related Work: Remote Sensing Image Retrieval



query image

top-k retrieved images

residential

residential

basketball court

parking space

residential

Dongyang et al., Exploiting low dimensional features from the mobilenets for remote sensing image retrieval, Earth Science Informatics, 2020

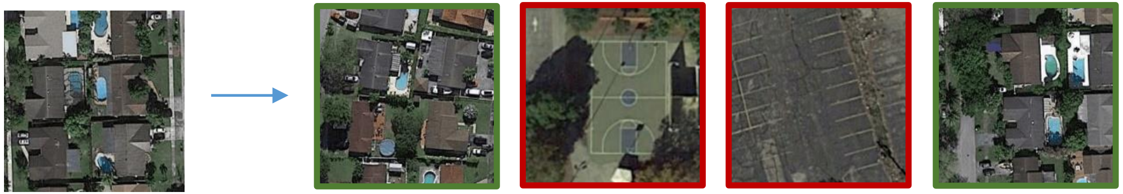
Related Work: Remote Sensing Image Retrieval

unisource

single-label

query image

top-k retrieved images



residential

residential

basketball court

parking space

residential

cross-source

Zhou et al., Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions, IEEE J-STARS, 2023

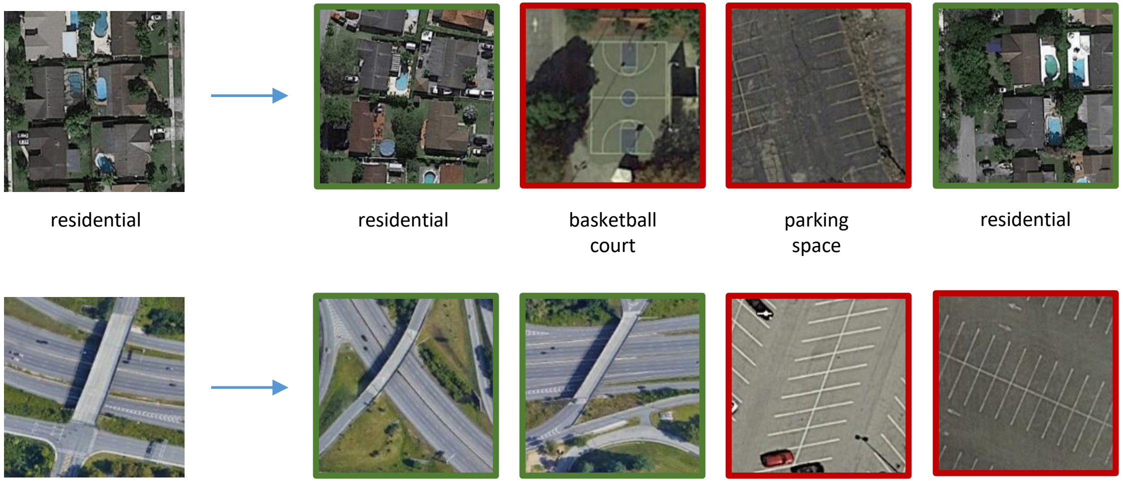
Related Work: Remote Sensing Image Retrieval

unisource

single-label

query image

top-k retrieved images



residential

residential

basketball court

parking space

residential

multi-label

cars, grass, pavement

cars, grass, pavement

cars, grass, pavement

cars, pavement

pavement

cross-source

Zhou et al., Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions, IEEE J-STARS, 2023

Related Work: Remote Sensing Image Retrieval

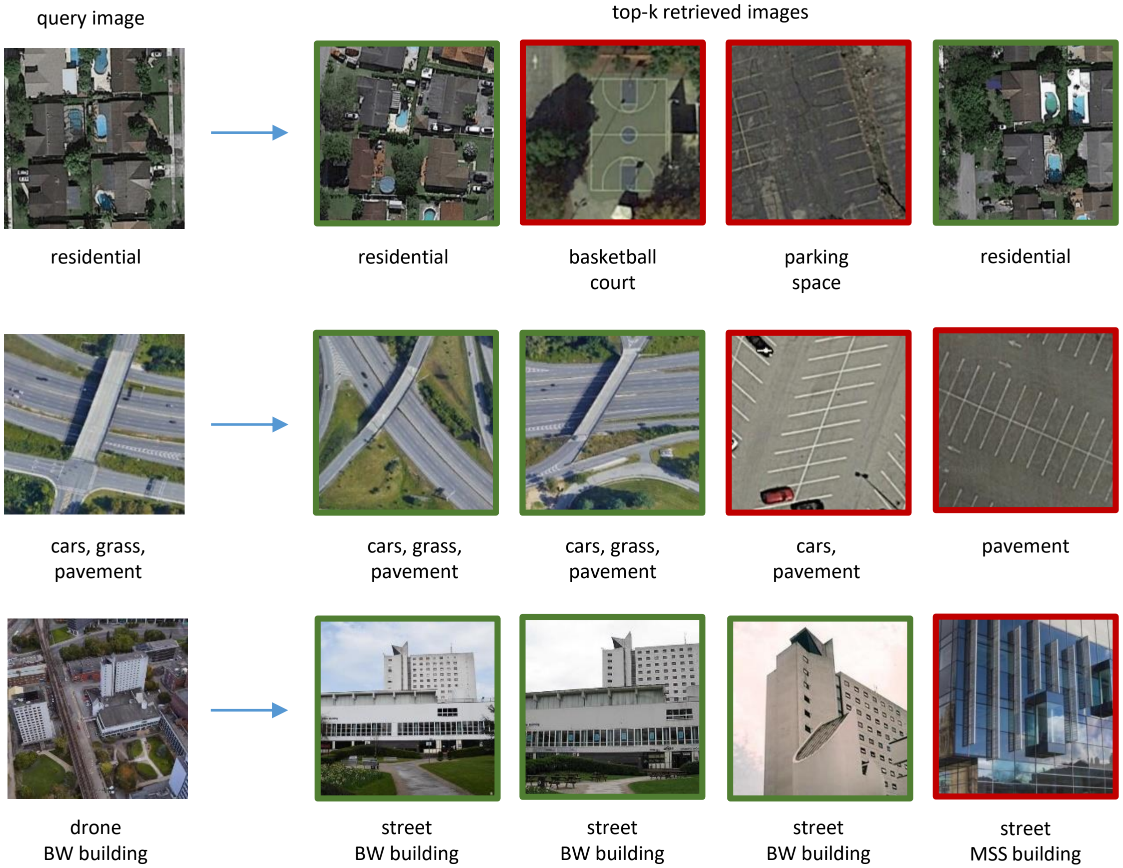
unisource

single-label

multi-label

cross-source

view



query image

top-k retrieved images

residential

residential

basketball court

parking space

residential

cars, grass, pavement

cars, grass, pavement

cars, grass, pavement

cars, pavement

pavement

drone BW building

street BW building

street BW building

street BW building

street MSS building

Limitation of Remote Sensing Image Retrieval

unisource

single-label

query image

multi-label

query image

cross-source

view

query image

query of
single modality!

Limitation of Remote Sensing Image Retrieval

unisource

single-label

query image

multi-label

query image

cross-source

view

query image

query of
single modality!

restricts users
from expressing
specific
requirements…

Remote Sensing Composed Image Retrieval

Remote Sensing Composed Image Retrieval

query image

Remote Sensing Composed Image Retrieval

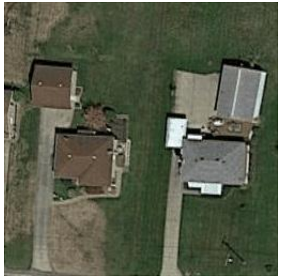
query image

query text

“dense”

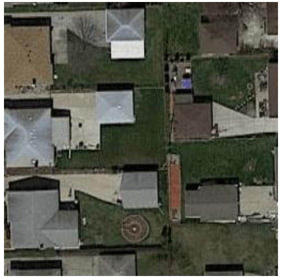
Remote Sensing Composed Image Retrieval

query image



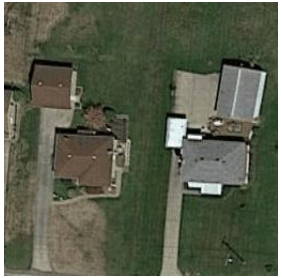
query text

“dense”



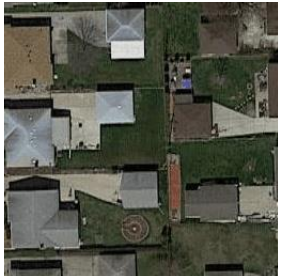
Remote Sensing Composed Image Retrieval

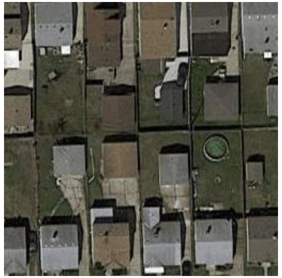
query image



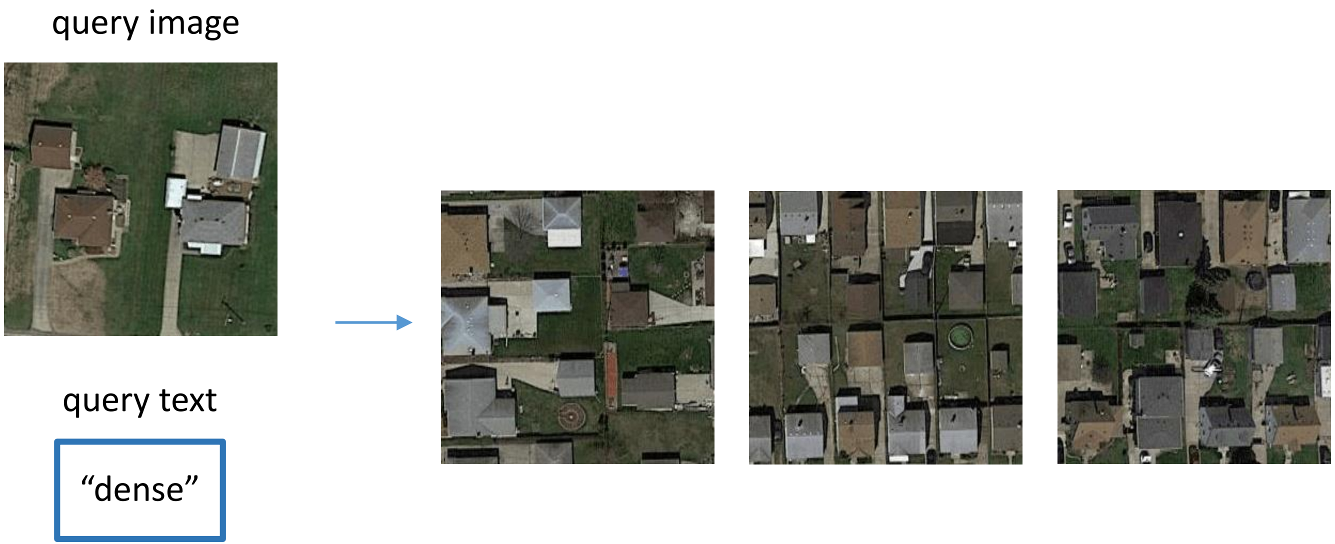
query text

“dense”





Remote Sensing Composed Image Retrieval

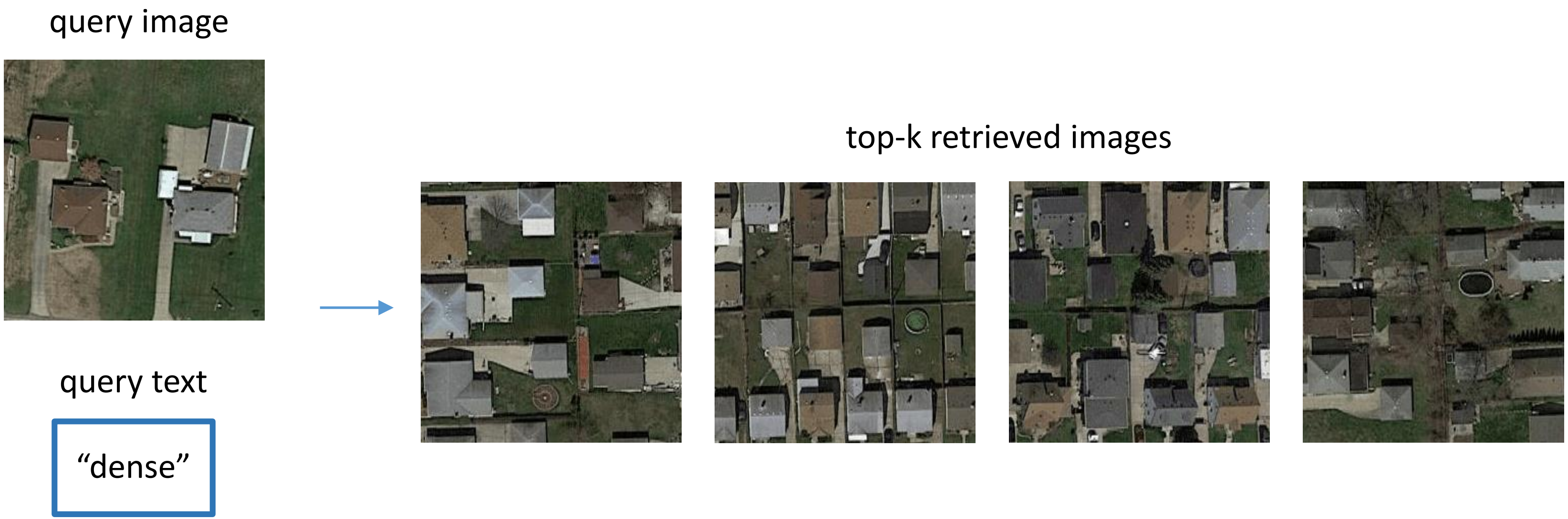


query image

query text

“dense”

Remote Sensing Composed Image Retrieval



query image

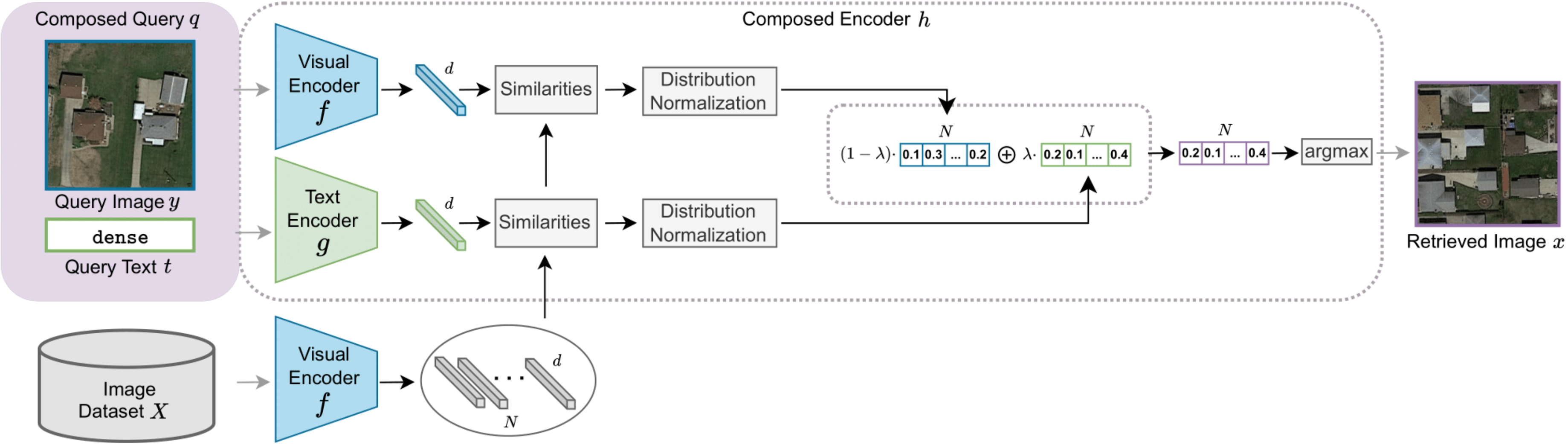
query text

“dense”

top-k retrieved images

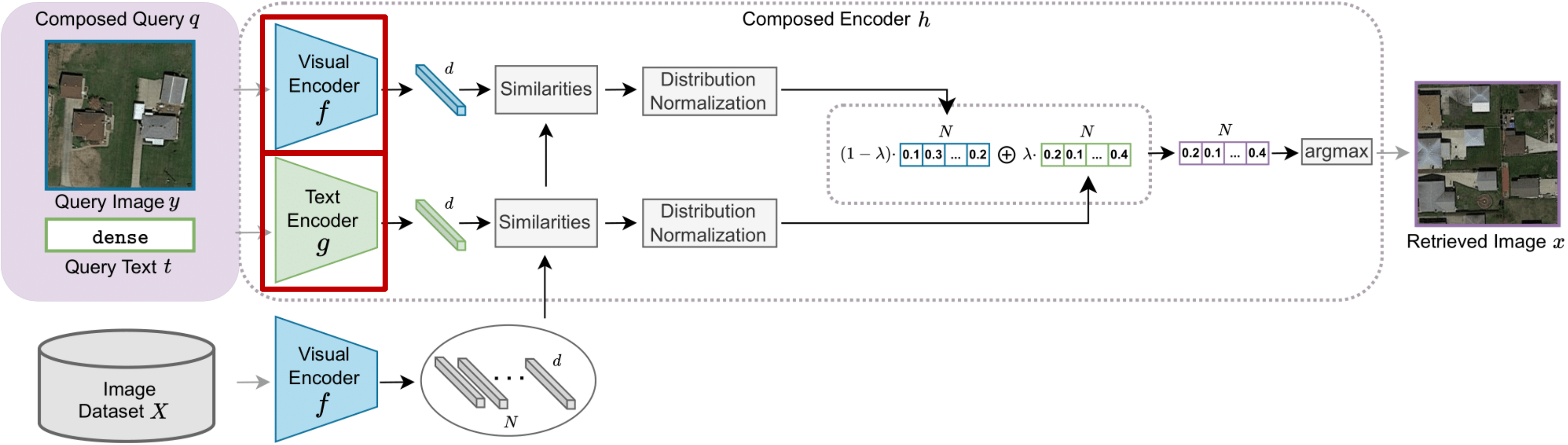
expressive and flexible search!

Our method, WeiCom

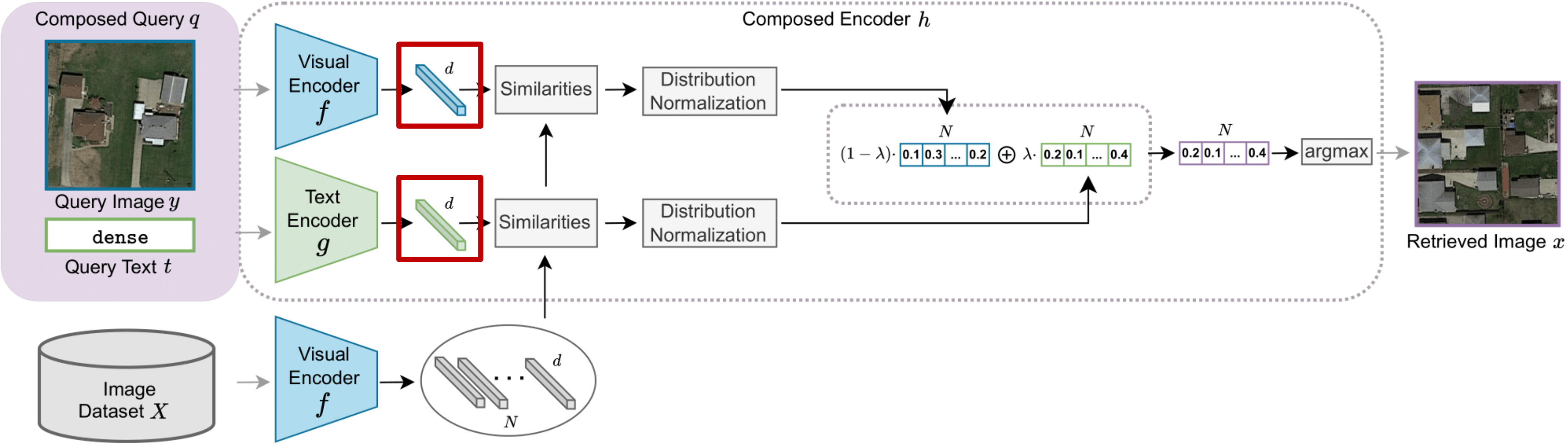


Radford et al., Learning transferable visual models from natural language supervision, ICML, 2021
Liu et al., Remoteclip: A vision language foundation model for remote sensing, IEEE TGRS, 2024

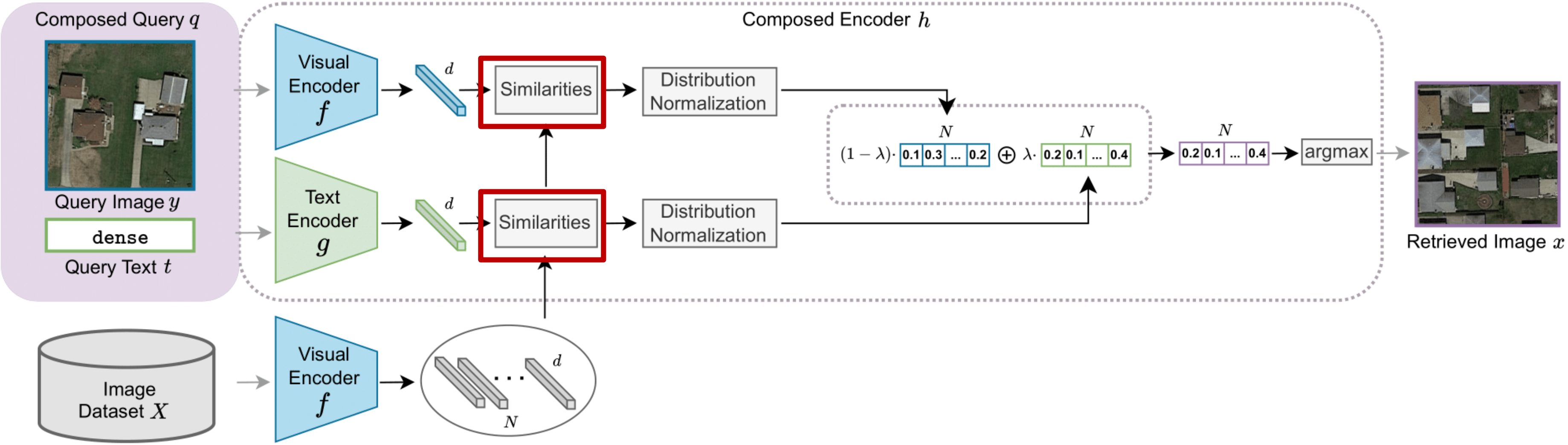
Our method, WeiCom



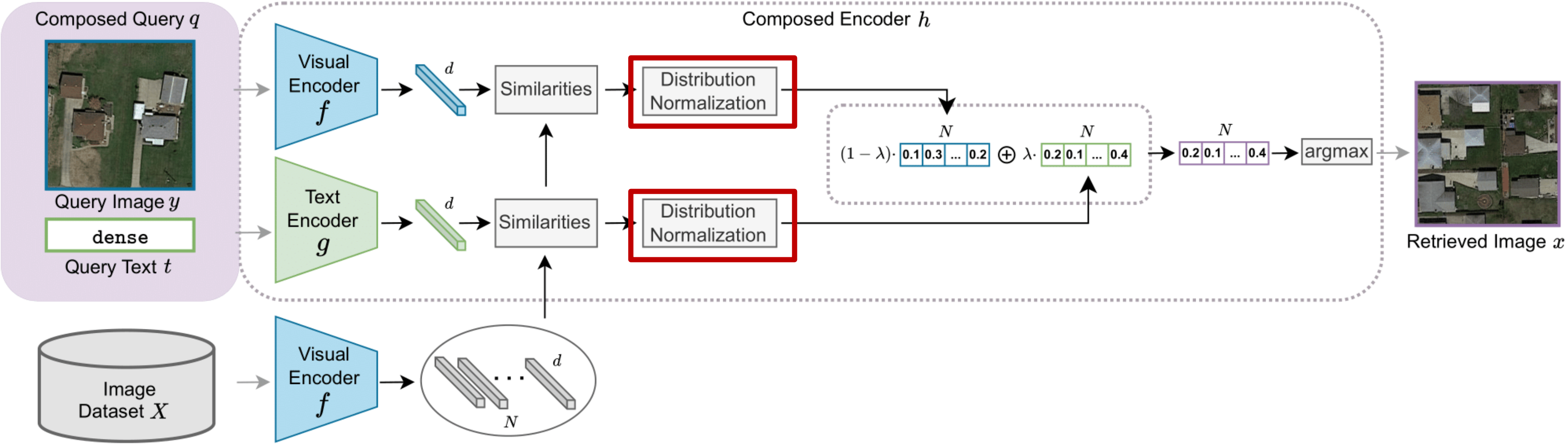
Our method, WeiCom



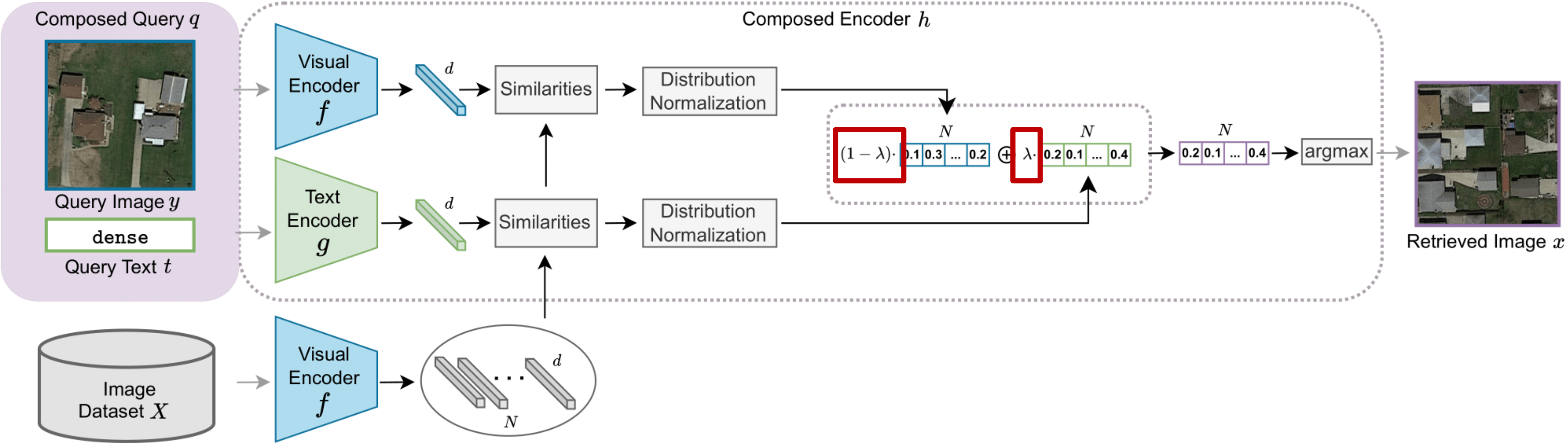
Our method, WeiCom



Our method, WeiCom

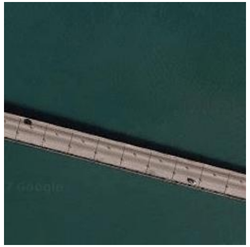


Our method, WeiCom



WeiCom’s control parameter λ

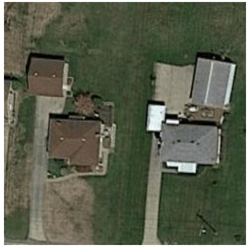
query image



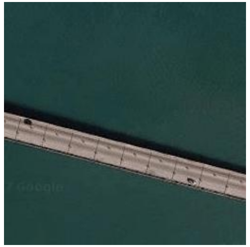
WeiCom’s control parameter λ

query image

query text



“dense”



“concrete”

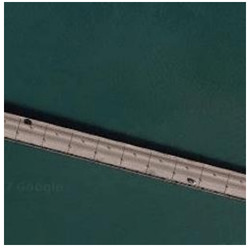
WeiCom’s control parameter λ

query image

retrieved images

query text

“dense”



“concrete”

image only
λ=0

λ=0.5

λ=0.75

λ=0.95

text only
λ=1

WeiCom’s control parameter λ

query image

retrieved images

query text

“dense”

“concrete”

image only
λ=0

λ=0.5

λ=0.75

λ=0.95

text only
λ=1

WeiCom’s control parameter λ

query image

retrieved images

query text

“dense”

“concrete”

image only
λ=0

λ=0.5

λ=0.75

λ=0.95

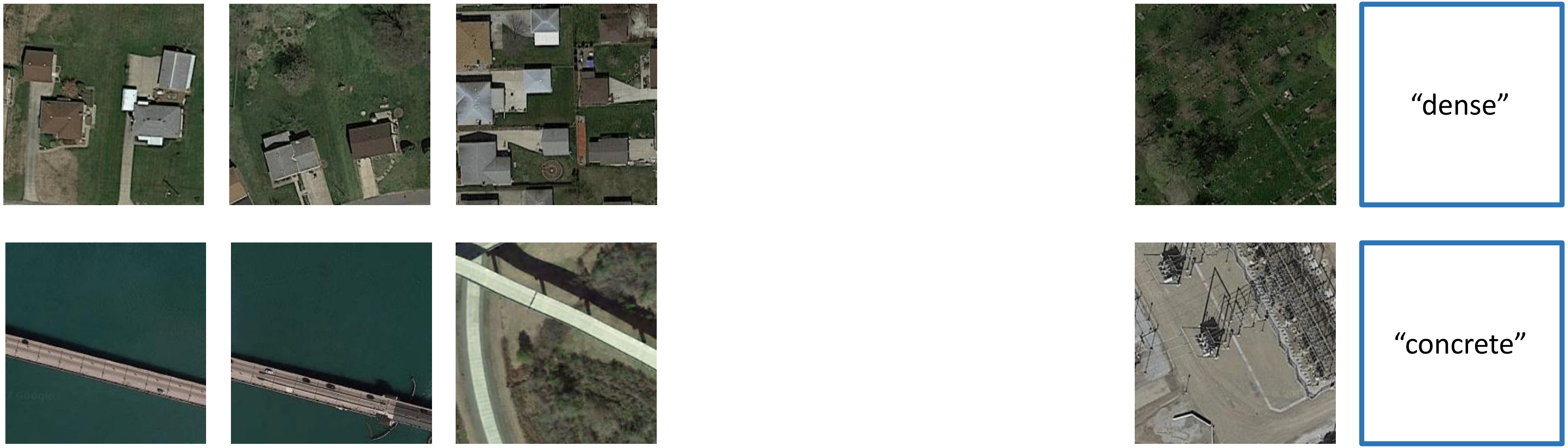
text only
λ=1

WeiCom’s control parameter λ

query image

retrieved images

query text



“dense”

“concrete”

image only
λ=0

λ=0.5

λ=0.75

λ=0.95

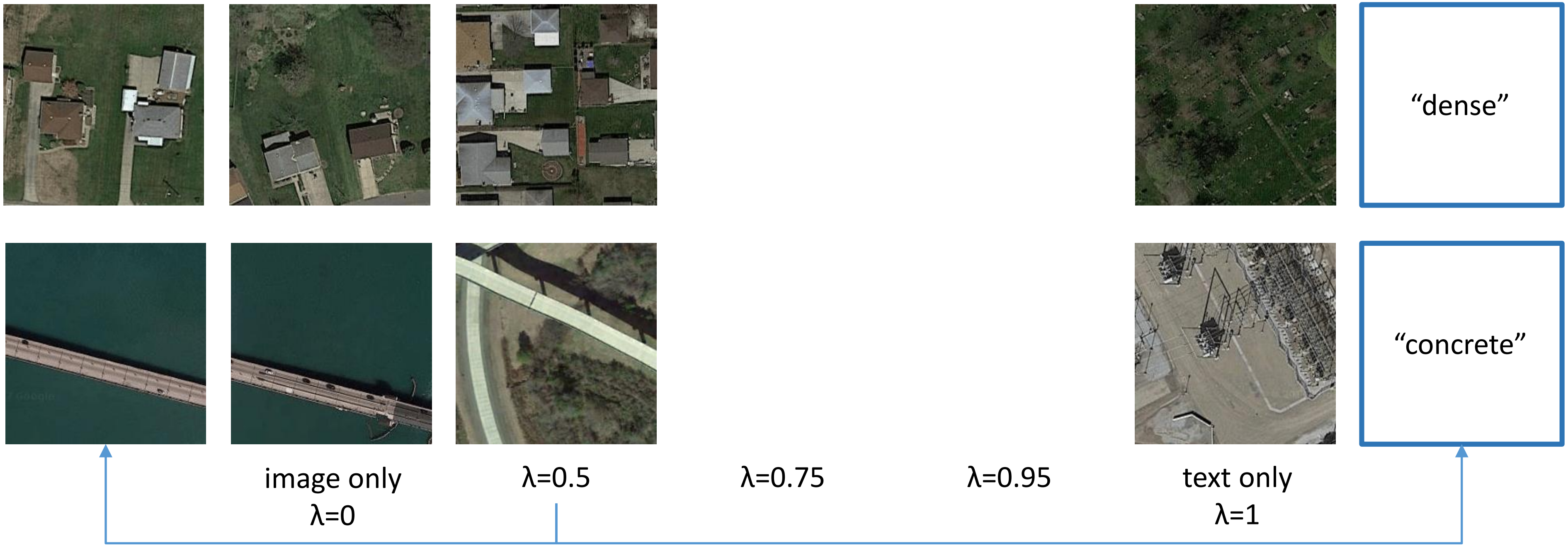
text only
λ=1

WeiCom's control parameter λ

query image

retrieved images

query text



“dense”

“concrete”

image only
λ=0

λ=0.5

λ=0.75

λ=0.95

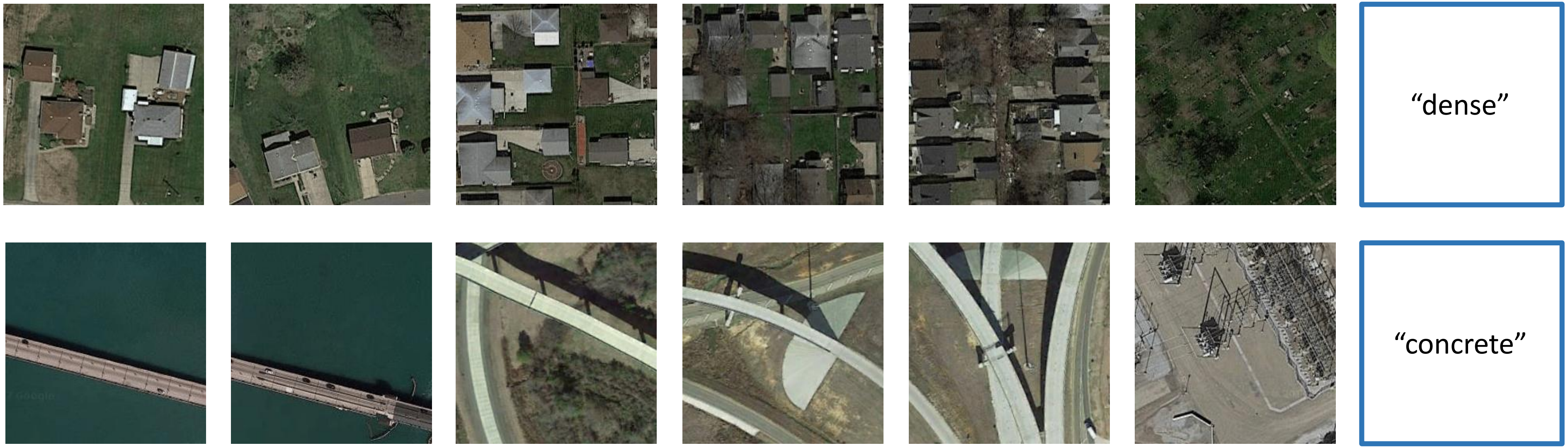
text only
λ=1

WeiCom’s control parameter λ

query image

retrieved images

query text



“dense”

“concrete”

image only
λ=0

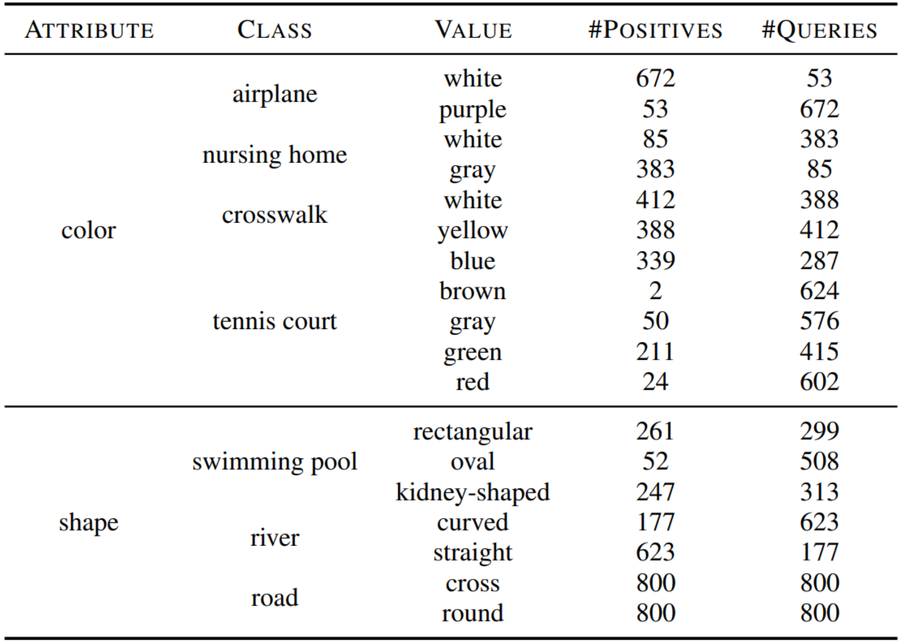
λ=0.5

λ=0.75

λ=0.95

text only
λ=1

PatternCom, our benchmark dataset

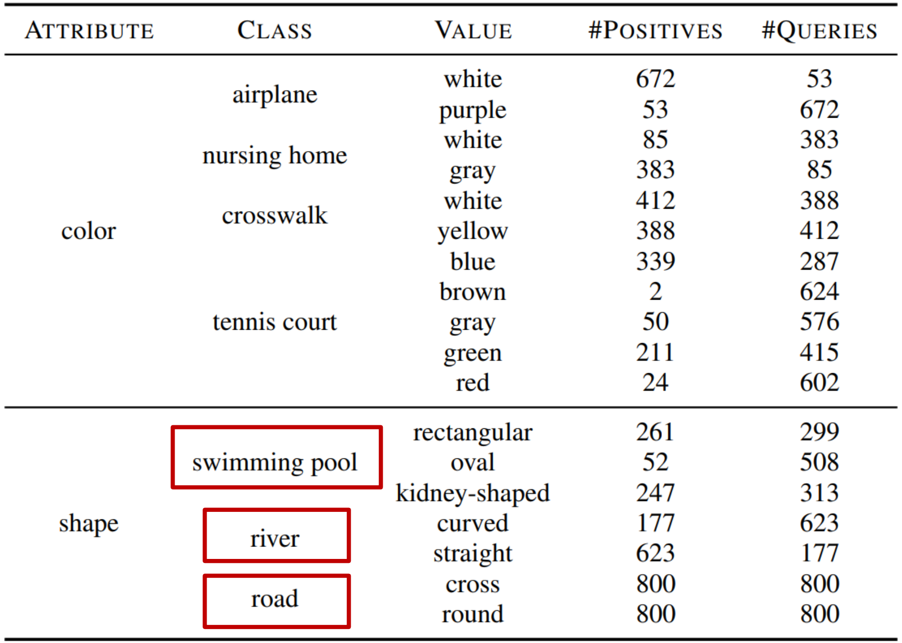


ATTRIBUTE	CLASS	VALUE	#POSITIVES	#QUERIES
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Statistics for color and shape attributes of PatternCom

Zhou et al., Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval, ISPRS Journal, 2018

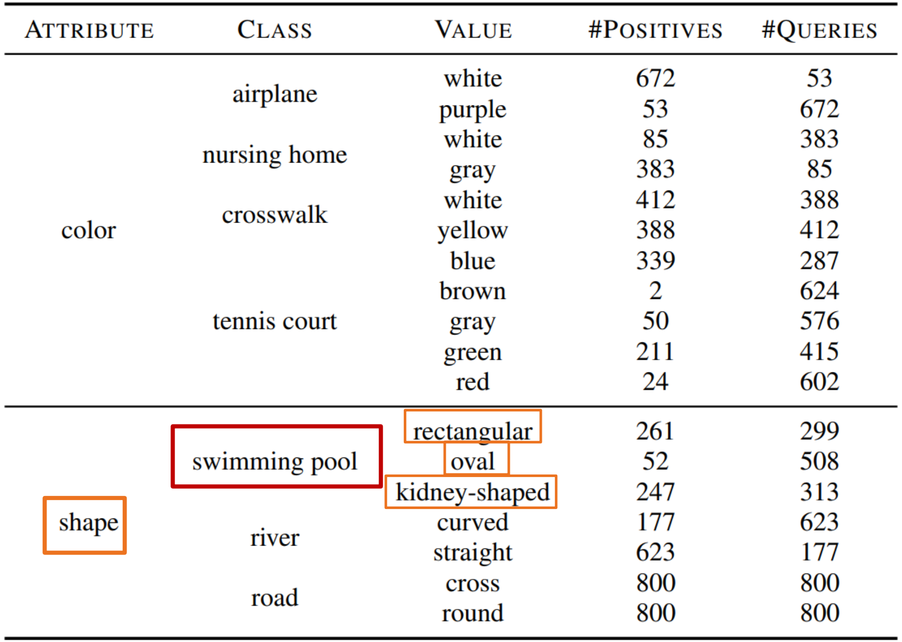
PatternCom, our benchmark dataset



ATTRIBUTE	CLASS	VALUE	#POSITIVES	#QUERIES
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Statistics for color and shape attributes of PatternCom

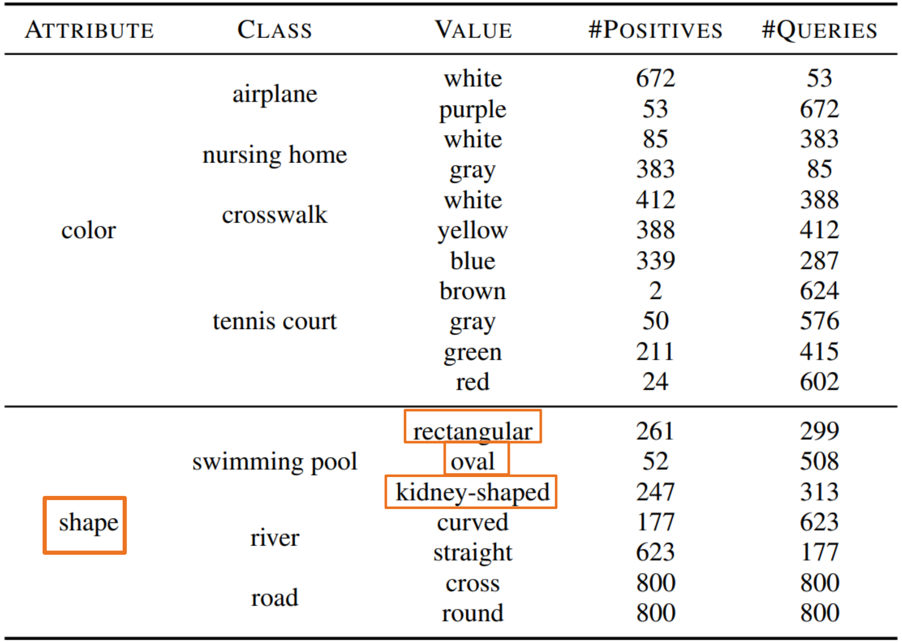
PatternCom, our benchmark dataset



Attribute	Class	Value	#Positives	#Queries
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

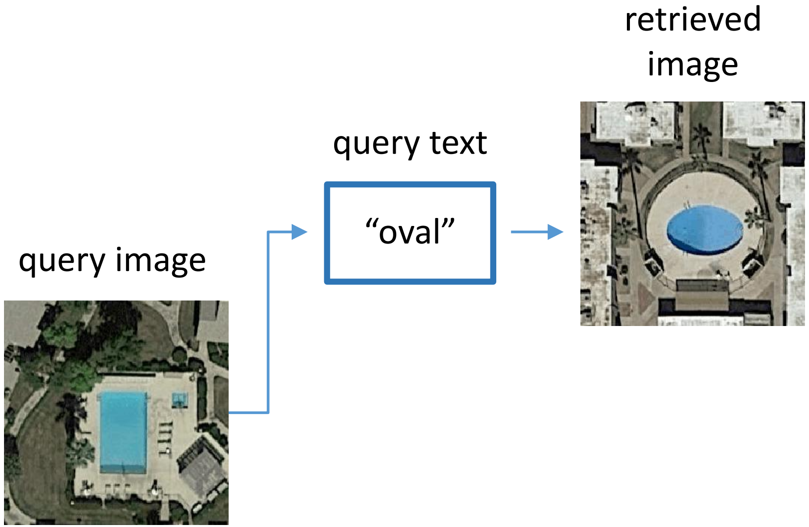
Statistics for color and shape attributes of PatternCom

PatternCom, our benchmark dataset

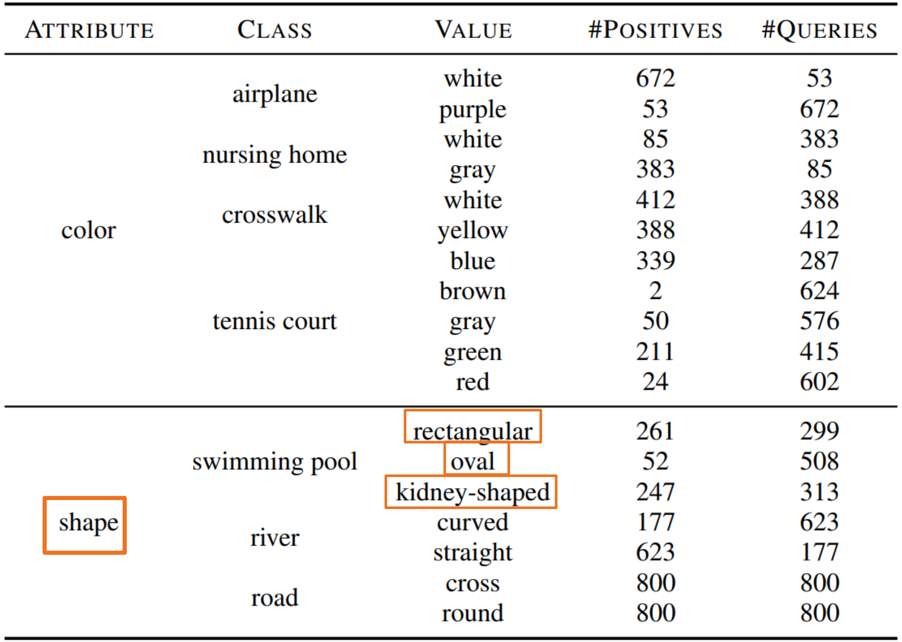


ATTRIBUTE	CLASS	VALUE	#POSITIVES	#QUERIES
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Statistics for color and shape attributes of PatternCom

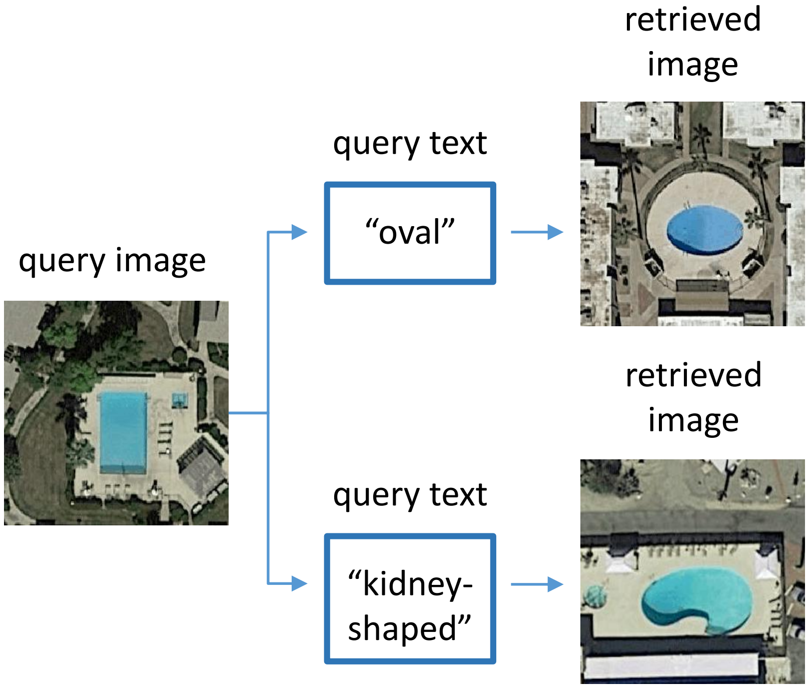


PatternCom, our benchmark dataset



Attribute	Class	Value	#Positives	#Queries
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Statistics for color and shape attributes of PatternCom



query image

query text

"oval"

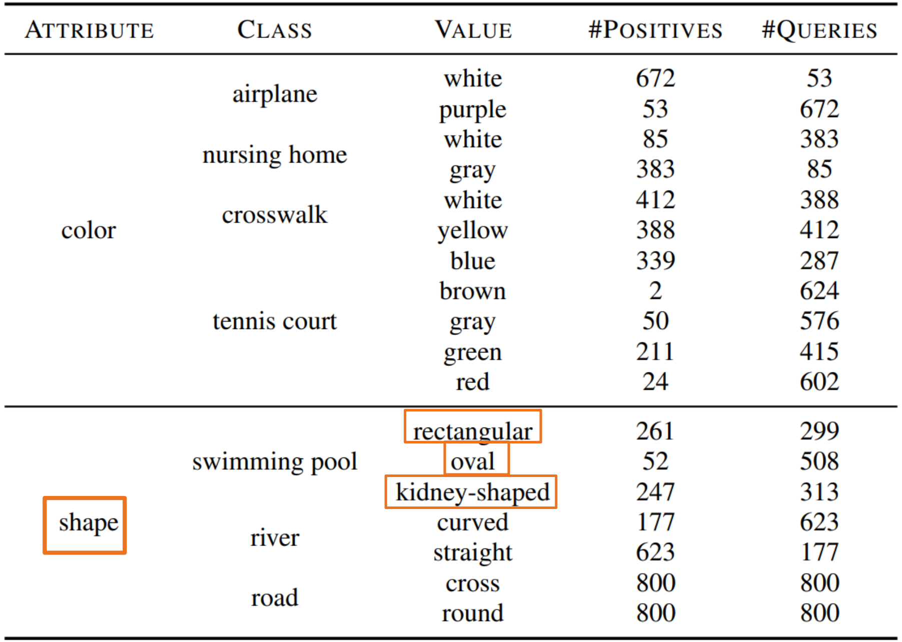
retrieved image

query text

"kidney-shaped"

retrieved image

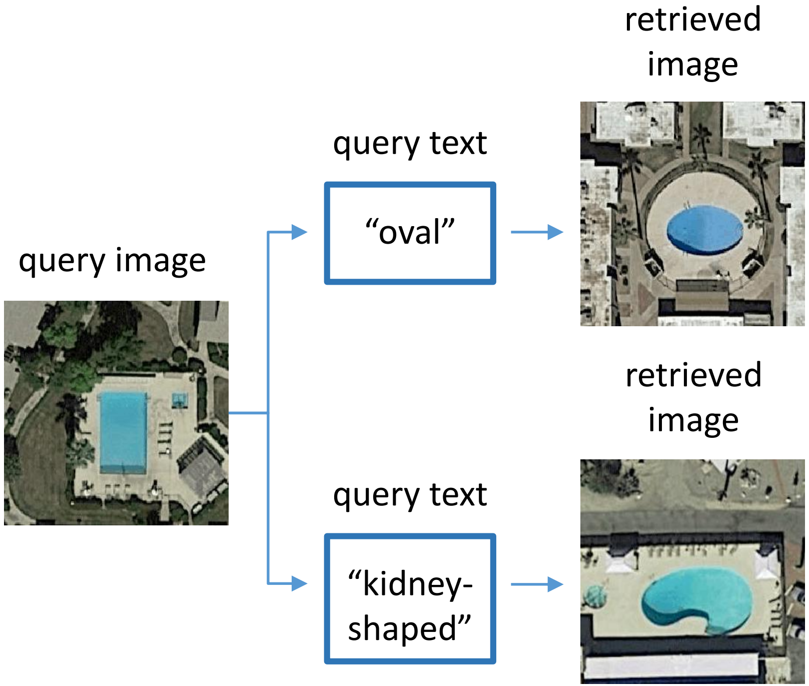
PatternCom, our benchmark dataset



Attribute	Class	Value	#Positives	#Queries
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Statistics for color and shape attributes of PatternCom

>21k queries in total!



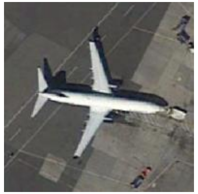
PatternCom: attributes

query image

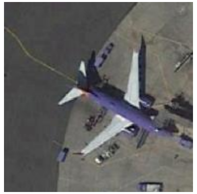
query text

retrieved image

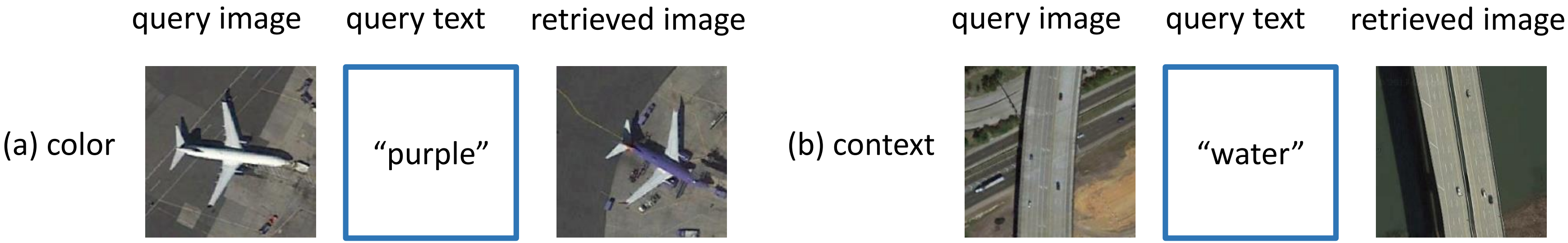
(a) color



“purple”



PatternCom: attributes



query image query text retrieved image

(a) color

“purple”

query image query text retrieved image

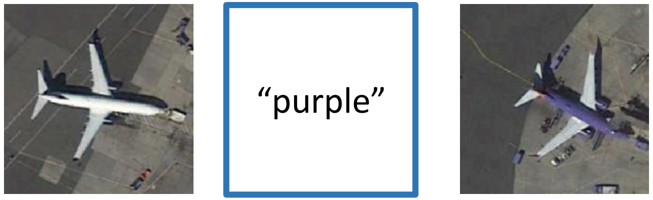
(b) context

“water”

PatternCom: attributes

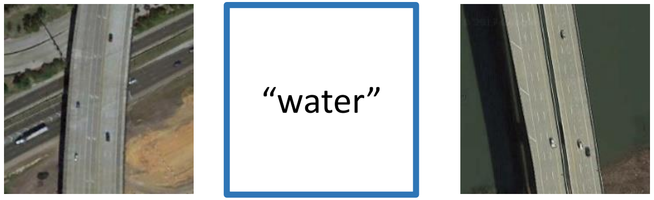
query image query text retrieved image

(a) color

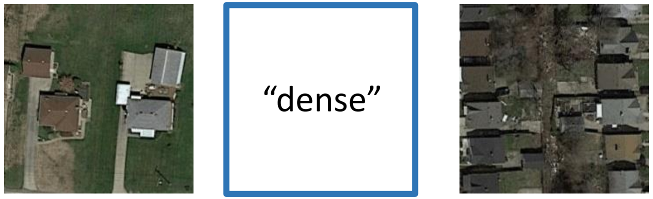

“purple”

query image query text retrieved image

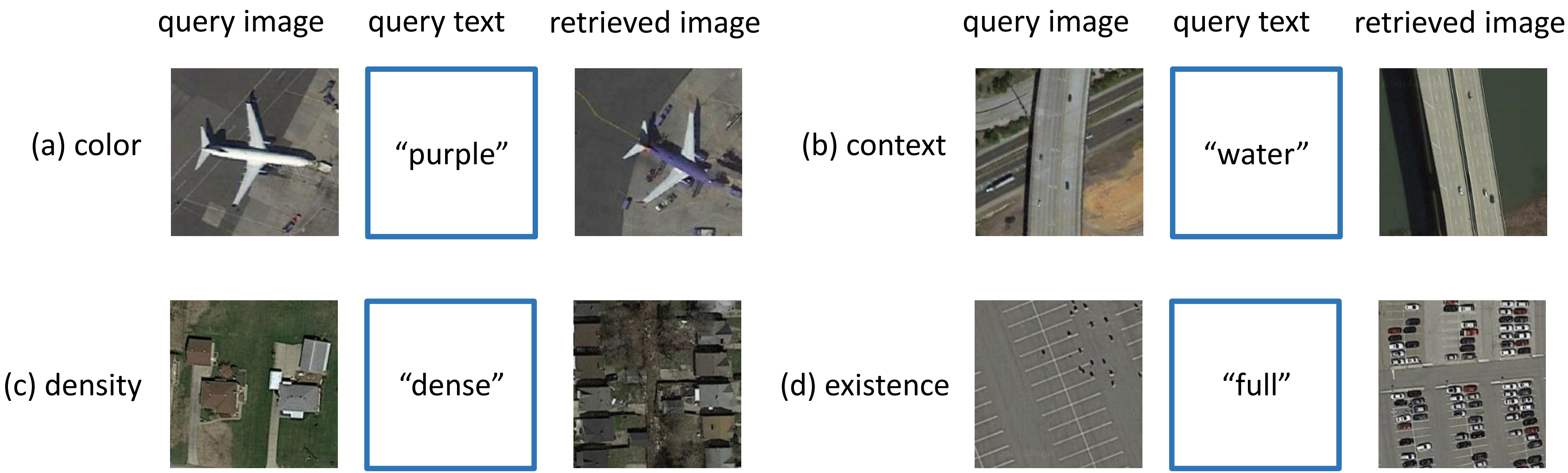
(b) context


“water”

(c) density


“dense”

PatternCom: attributes



query image query text retrieved image

(a) color “purple”

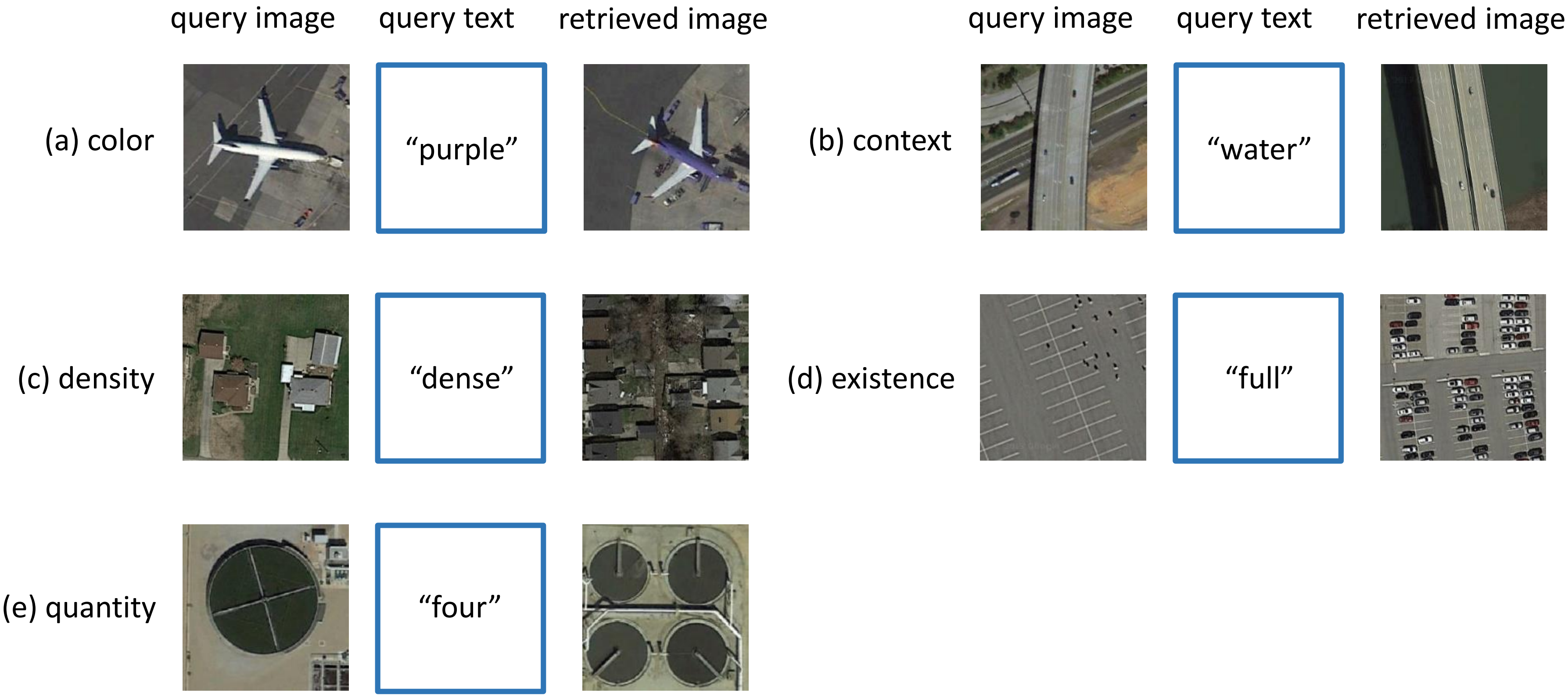
(c) density “dense”

query image query text retrieved image

(b) context “water”

(d) existence “full”

PatternCom: attributes



query image query text retrieved image

(a) color "purple"

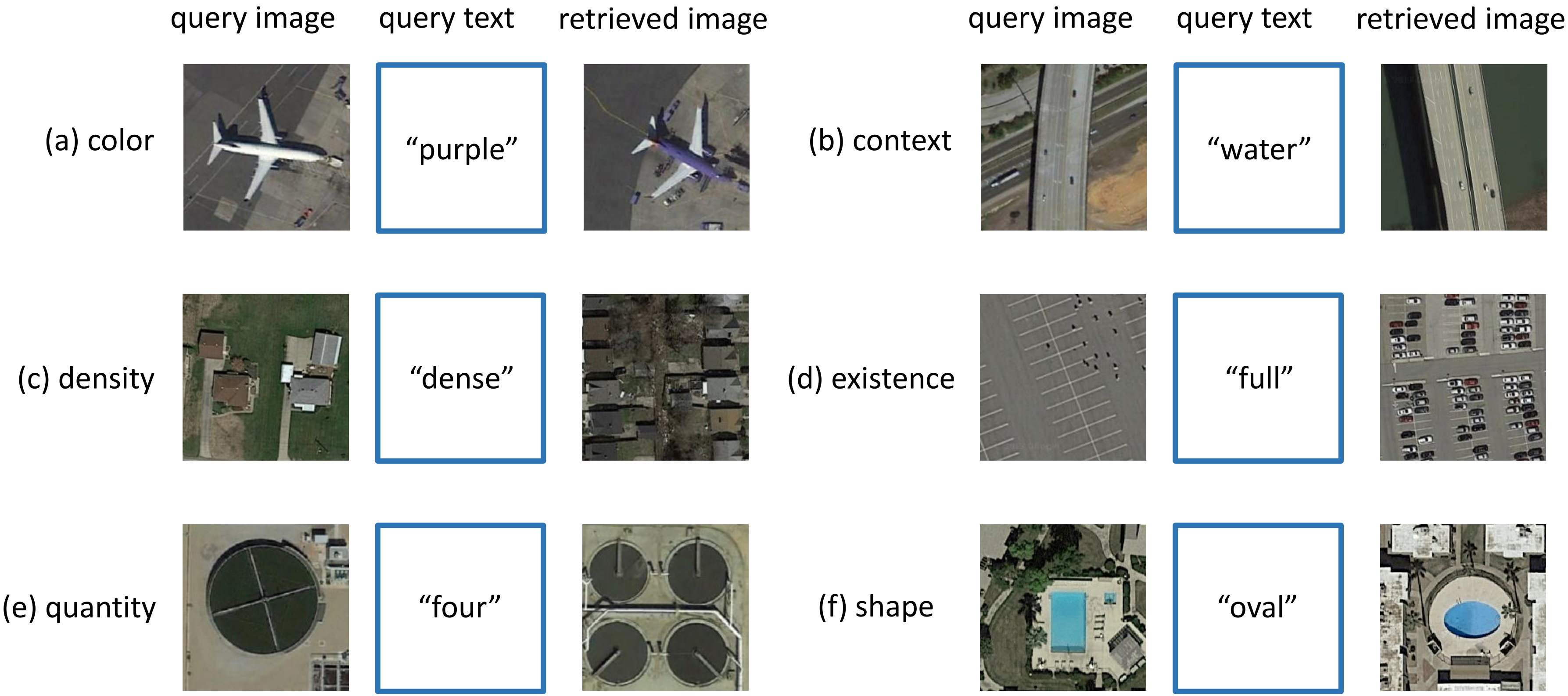
(b) context "water"

(c) density "dense"

(d) existence "full"

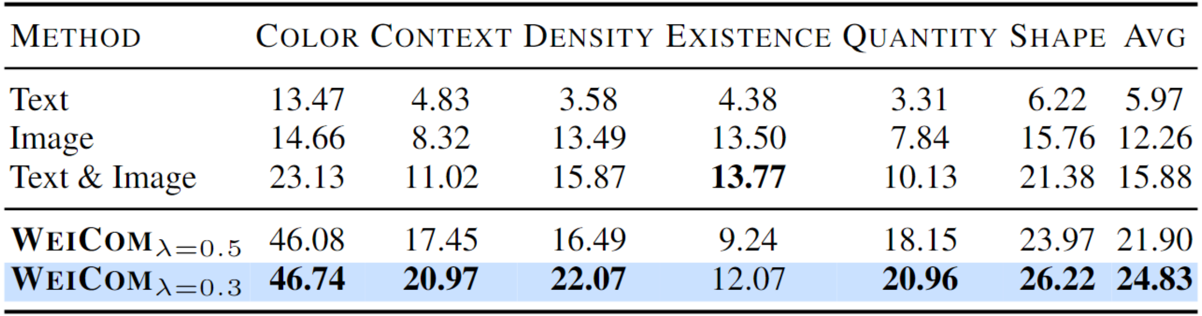
(e) quantity "four"

PatternCom: attributes

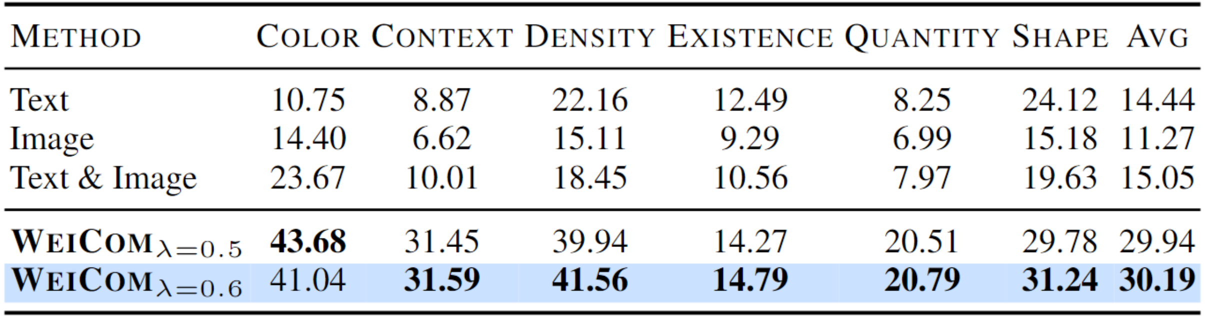


Quantitative Evaluation

CLIP



Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WeiCom$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WeiCom$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

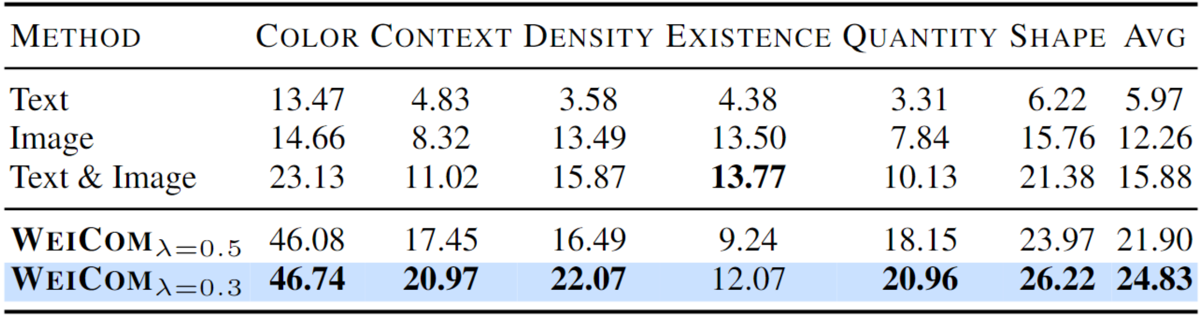


Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WeiCom$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WeiCom$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

Radford et al., Learning transferable visual models from natural language supervision, ICML, 2021

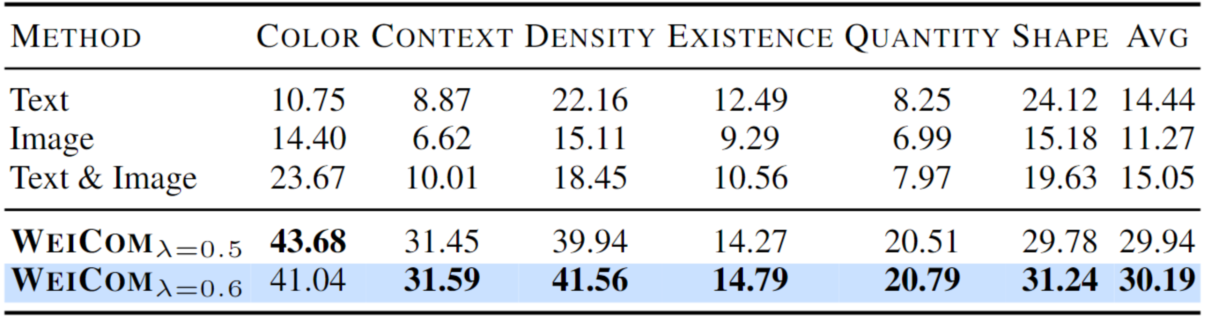
Quantitative Evaluation

CLIP



Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WeiCom$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WeiCom$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

RemoteCLIP



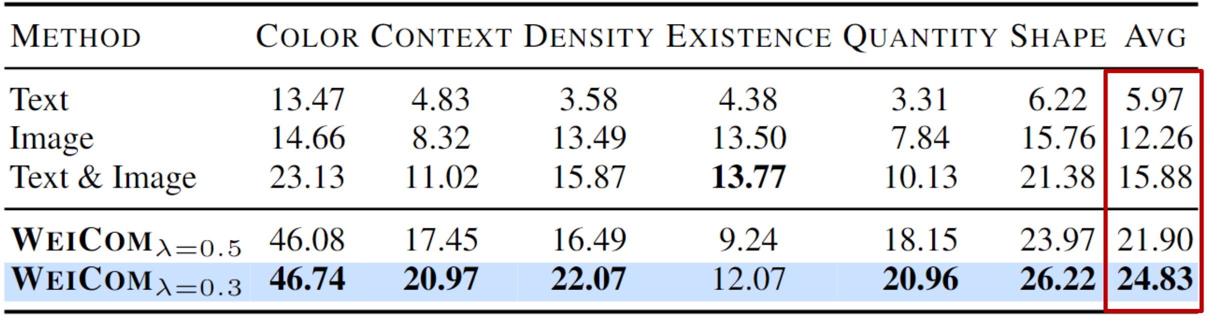
Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WeiCom$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WeiCom$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

Attribute modification mAP (%); comparison of WeiCom with baselines.
For each attribute value, average mAP over all the rest attribute values.

Liu et al., Remoteclip: A vision language foundation model for remote sensing, IEEE TGRS, 2024

Quantitative Evaluation

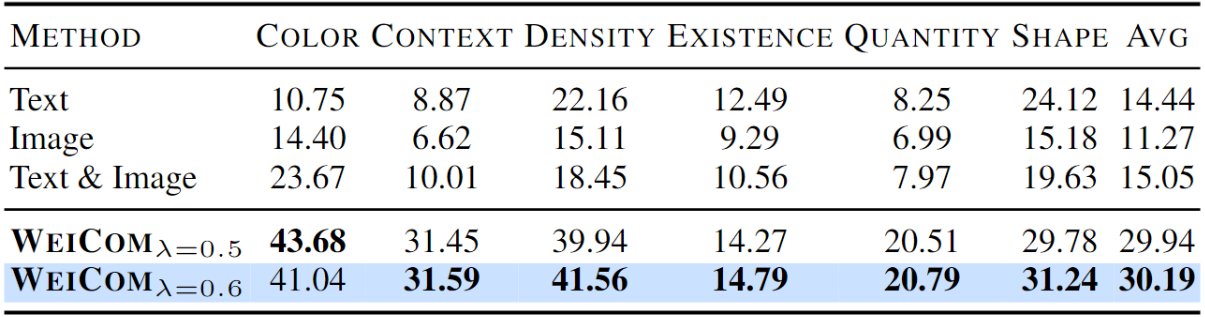
CLIP



Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WeiCom$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WeiCom$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

+9%

RemoteCLIP

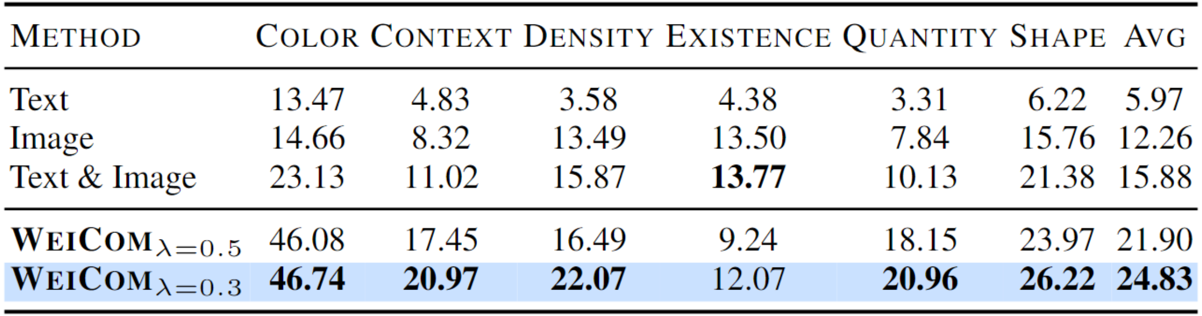


Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WeiCom$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WeiCom$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

Attribute modification mAP (%); comparison of WeiCom with baselines.
For each attribute value, average mAP over all the rest attribute values.

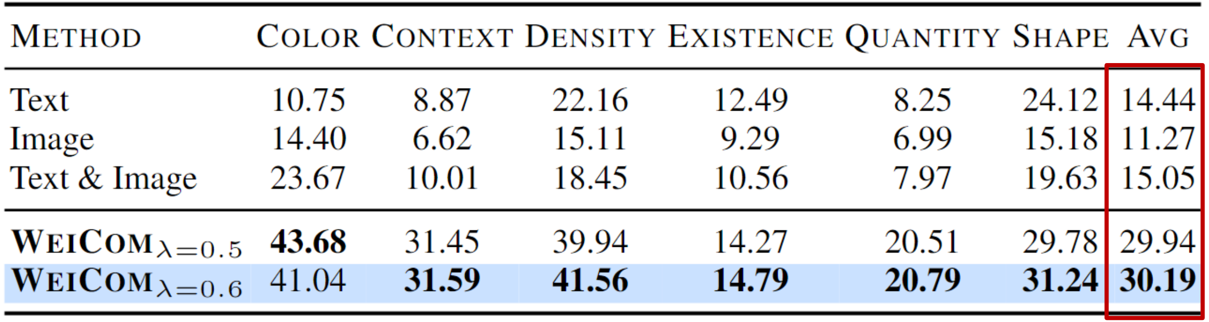
Quantitative Evaluation

CLIP



METHOD	COLOR	CONTEXT	DENSITY	EXISTENCE	QUANTITY	SHAPE	AVG
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WEICOM$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WEICOM$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

RemoteCLIP



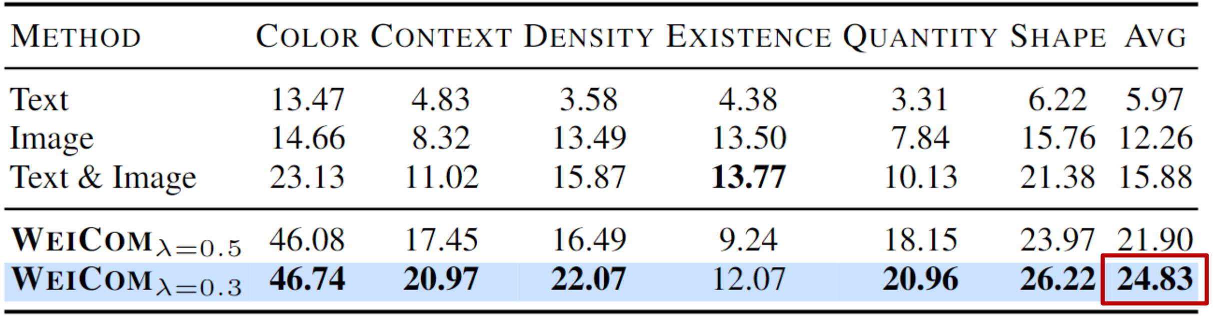
METHOD	COLOR	CONTEXT	DENSITY	EXISTENCE	QUANTITY	SHAPE	AVG
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WEICOM$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WEICOM$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

+15.1%

Attribute modification mAP (%); comparison of WeiCom with baselines.
For each attribute value, average mAP over all the rest attribute values.

Quantitative Evaluation

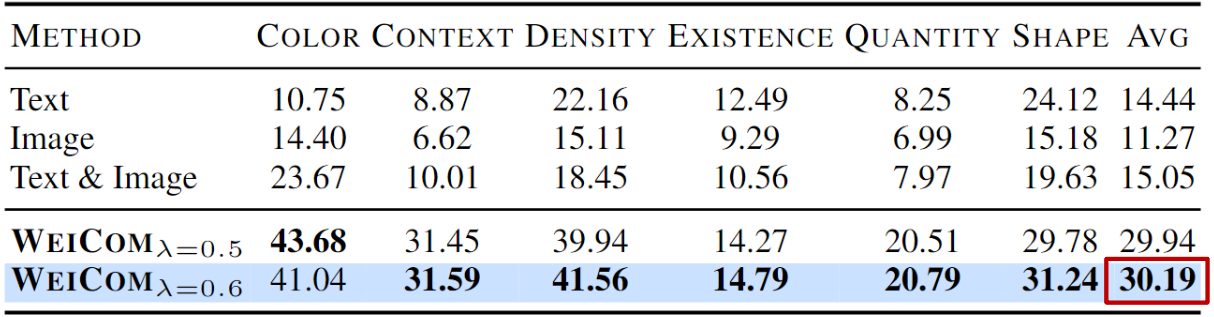
CLIP



Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WeiCom$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WeiCom$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

+5.4%

RemoteCLIP



Method	Color	Context	Density	Existence	Quantity	Shape	Avg
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WeiCom$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WeiCom$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

Attribute modification mAP (%); comparison of WeiCom with baselines.
For each attribute value, average mAP over all the rest attribute values.

Summarizing Insights

✓ Introduce Remote Sensing Composed Image Retrieval, accompanied with PatternCom, a benchmark dataset

✓ Demonstrate its versatility through use cases modifying attributes like color and shape

✓ Introduce WeiCom, a training-free method utilizing a modality control parameter λ

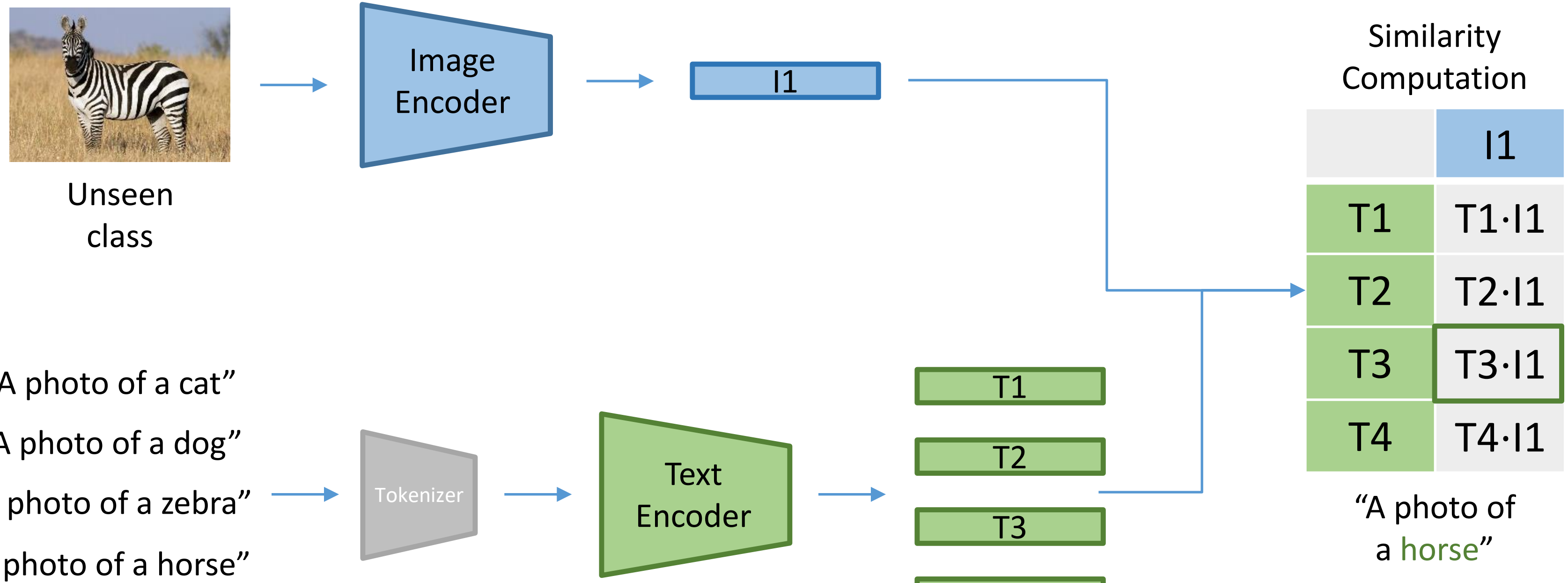
WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Extracting Multimodal Representations via Discrete-Space Textual Inversion [FreeDom]

Efthymiadis, **Psomas** et al. *Composed Image Retrieval for Training-Free Domain Conversion*, **WACV 2025** [under review]

Source code: TBA

Recap: Zero-Shot Recognition with CLIP



Unseen class

"A photo of a cat"

"A photo of a dog"

"A photo of a zebra"

"A photo of a horse"

Text descriptions

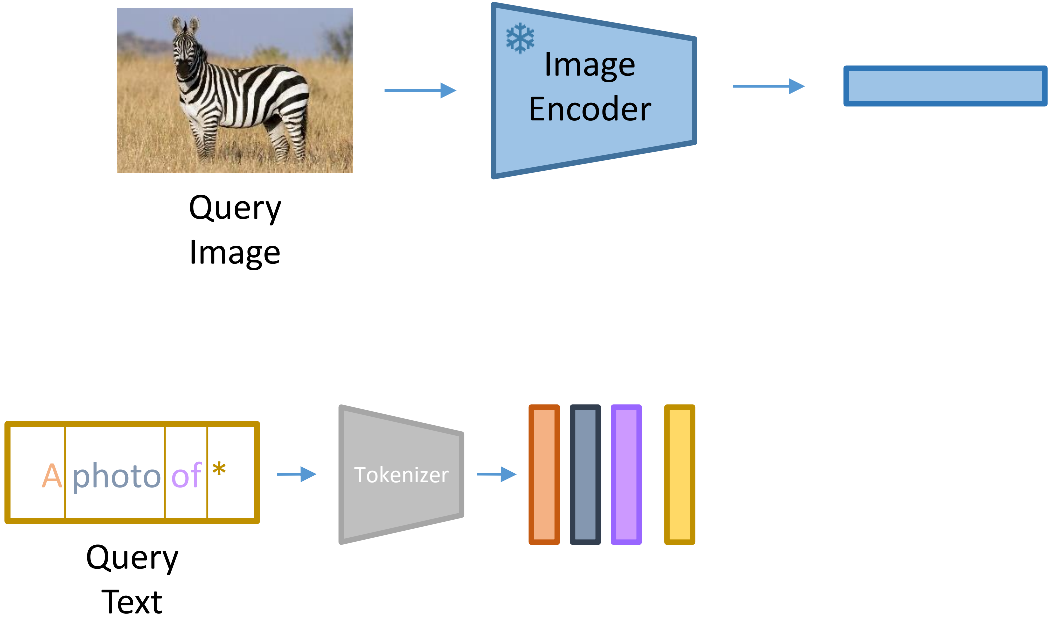
Similarity Computation

	I1
T1	T1·I1
T2	T2·I1
T3	**T3·I1**
T4	T4·I1

"A photo of a horse"

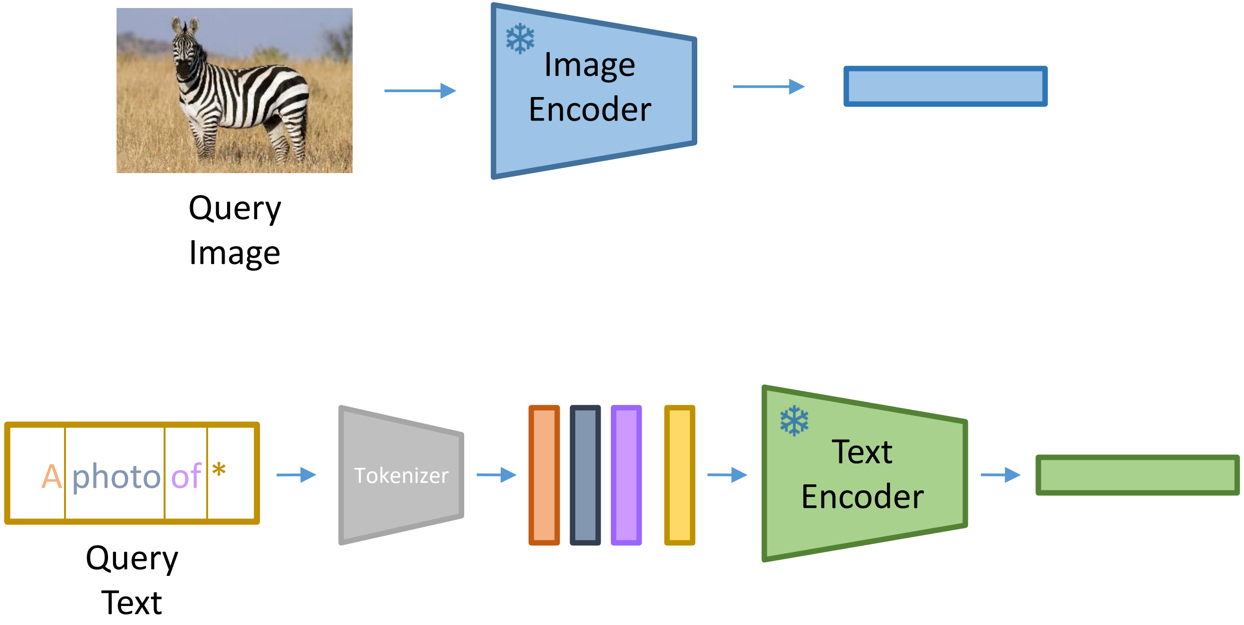
Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR 2021

Related Work: Continuous-Space Textual Inversion



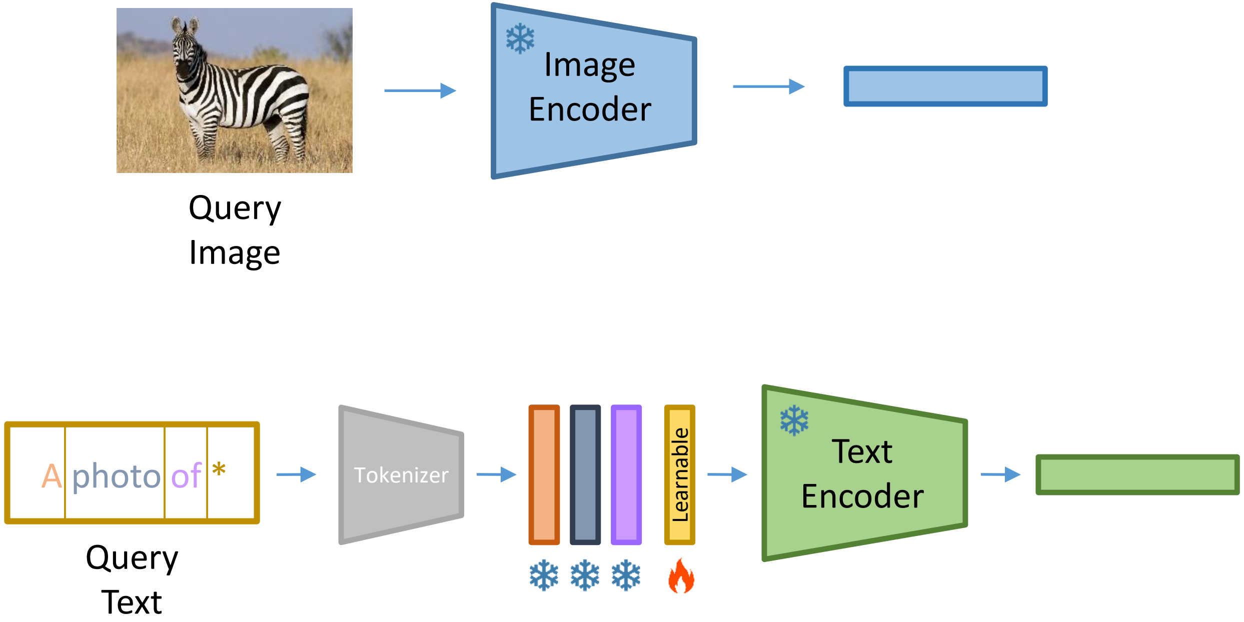
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



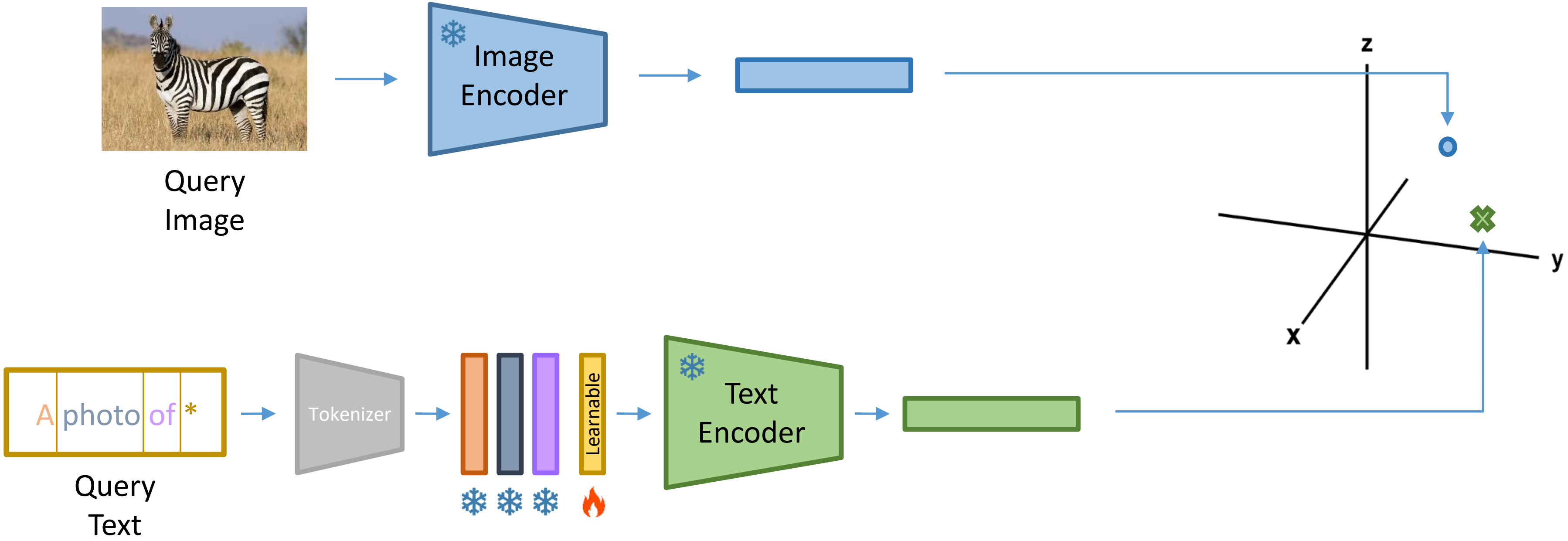
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



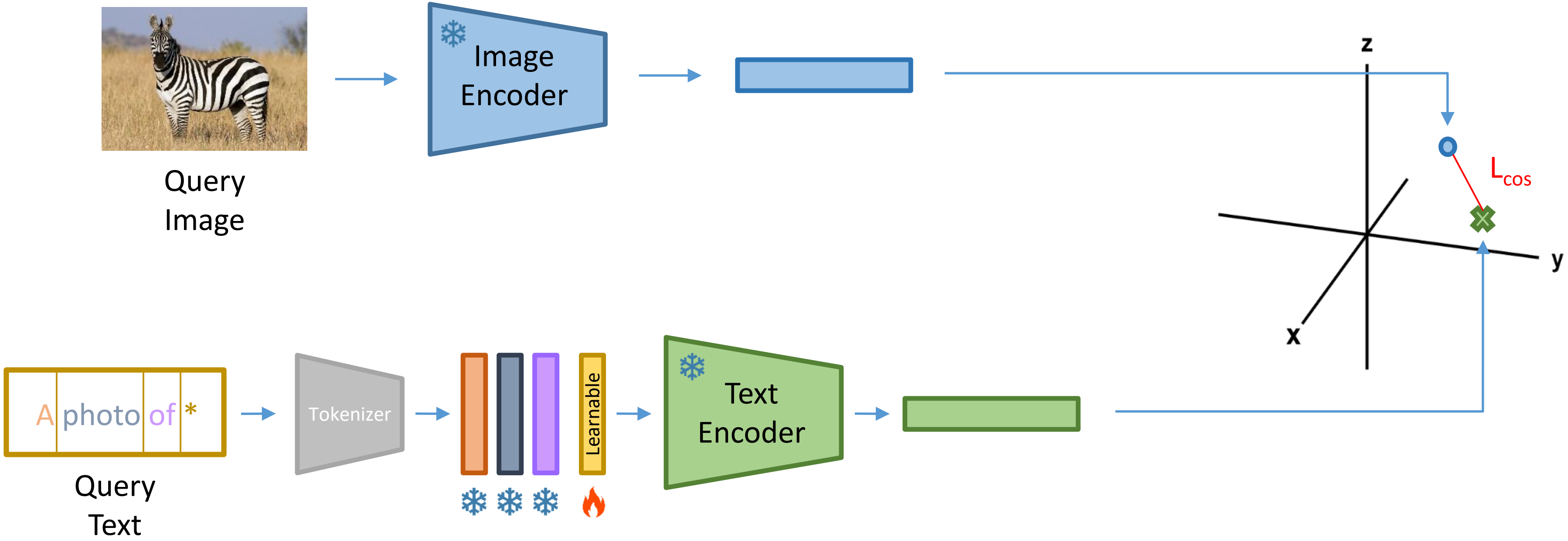
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



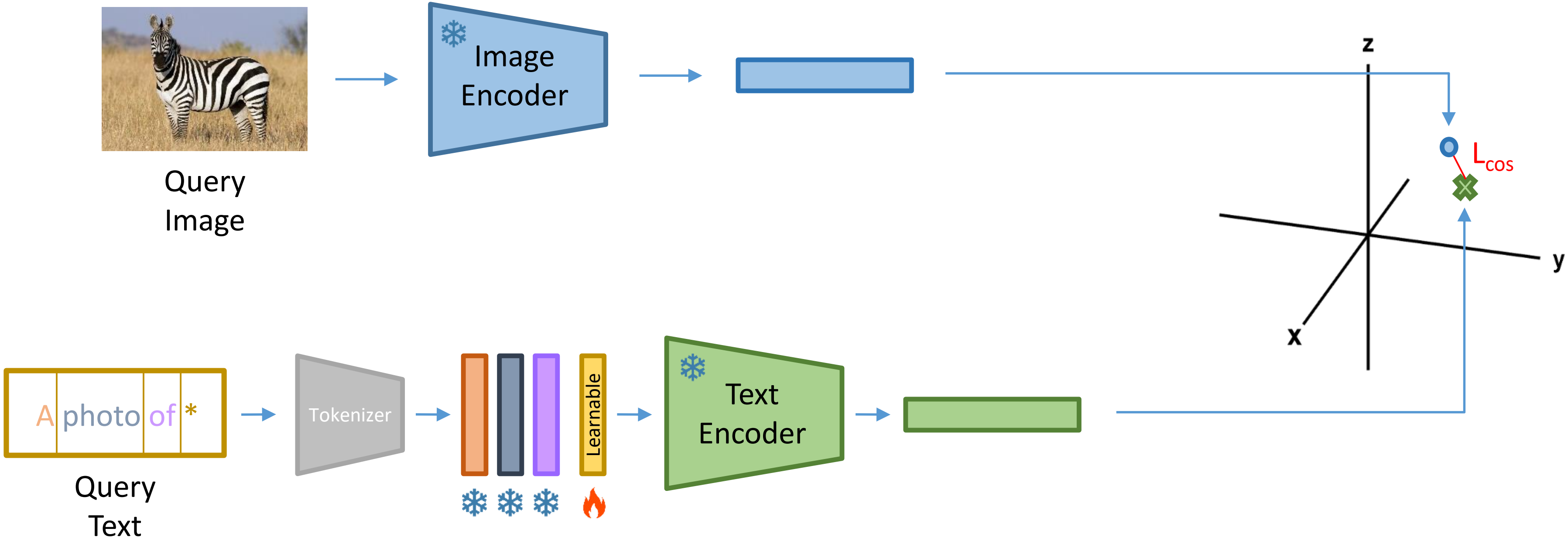
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



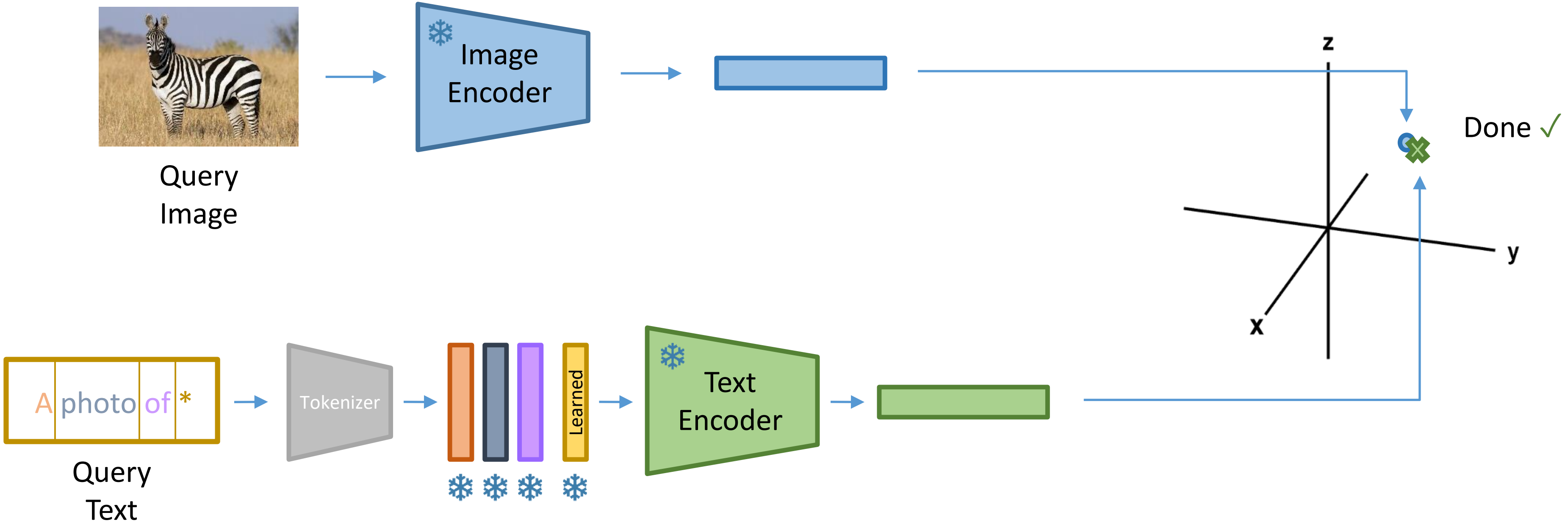
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



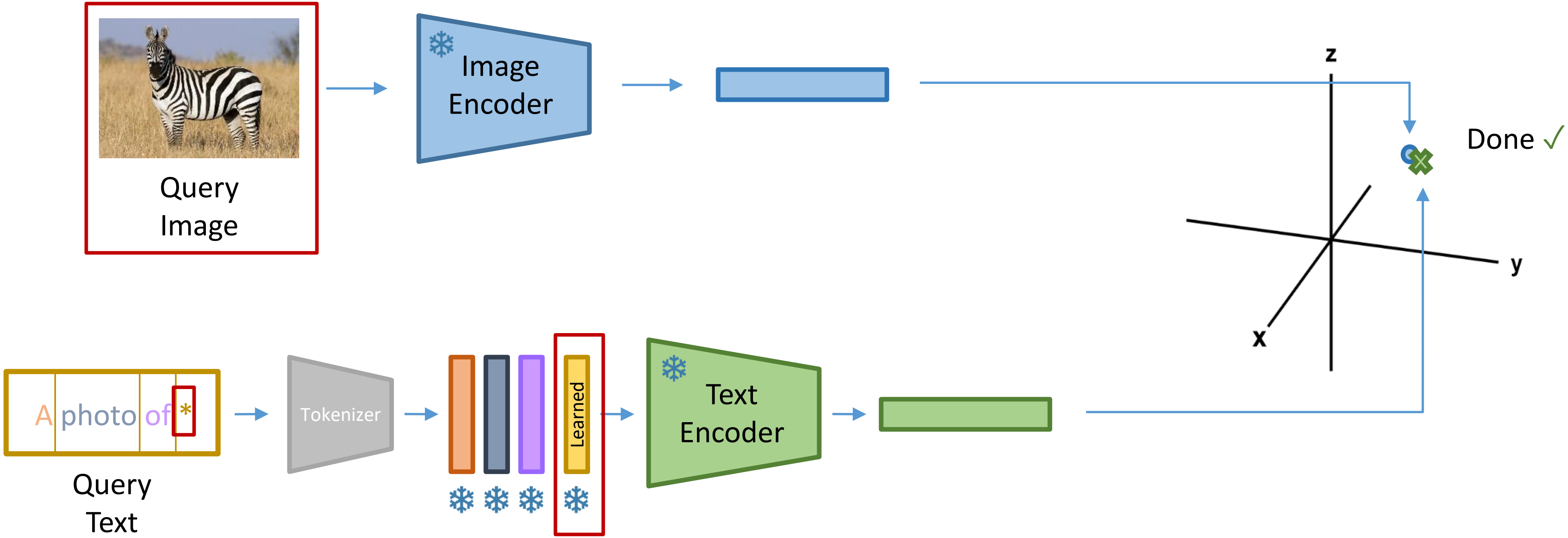
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



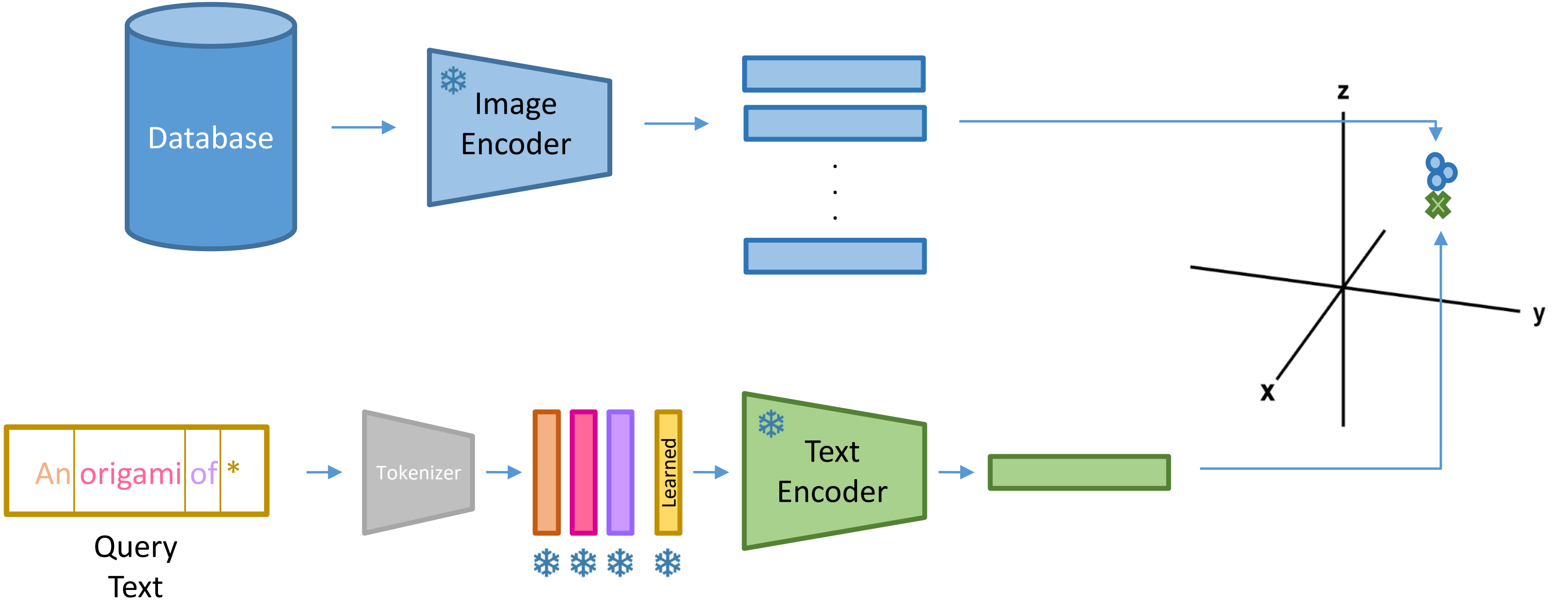
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



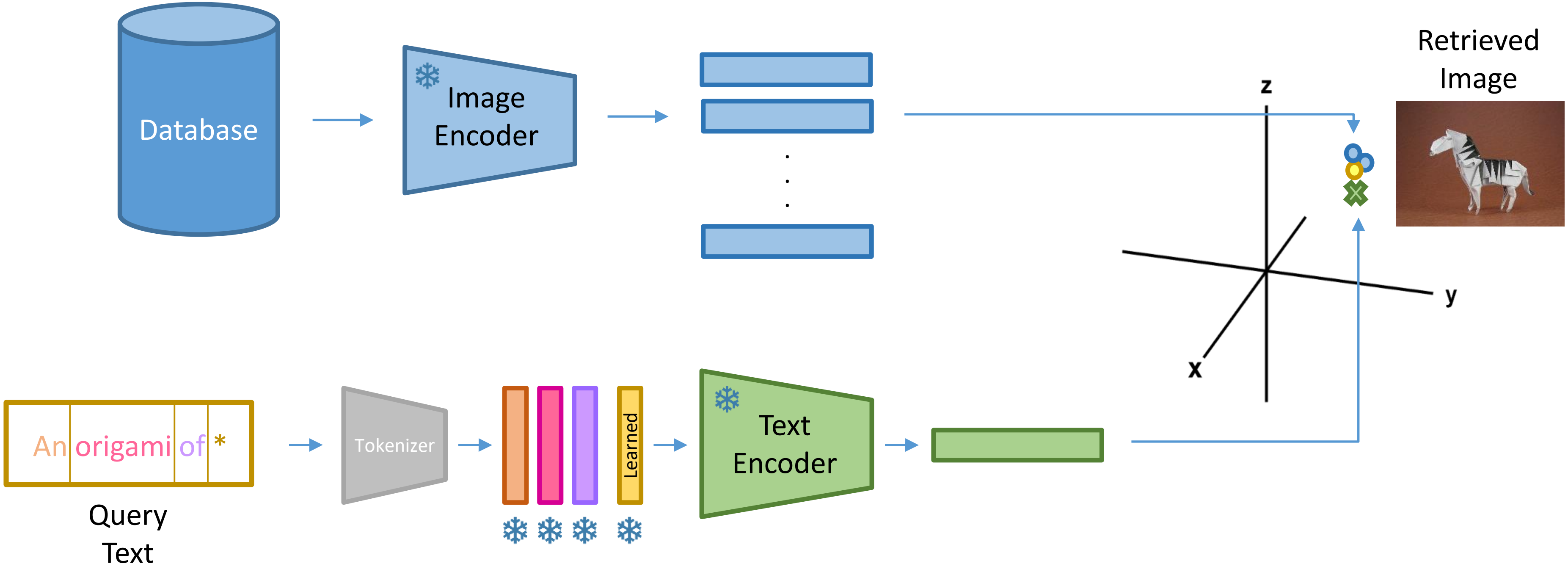
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



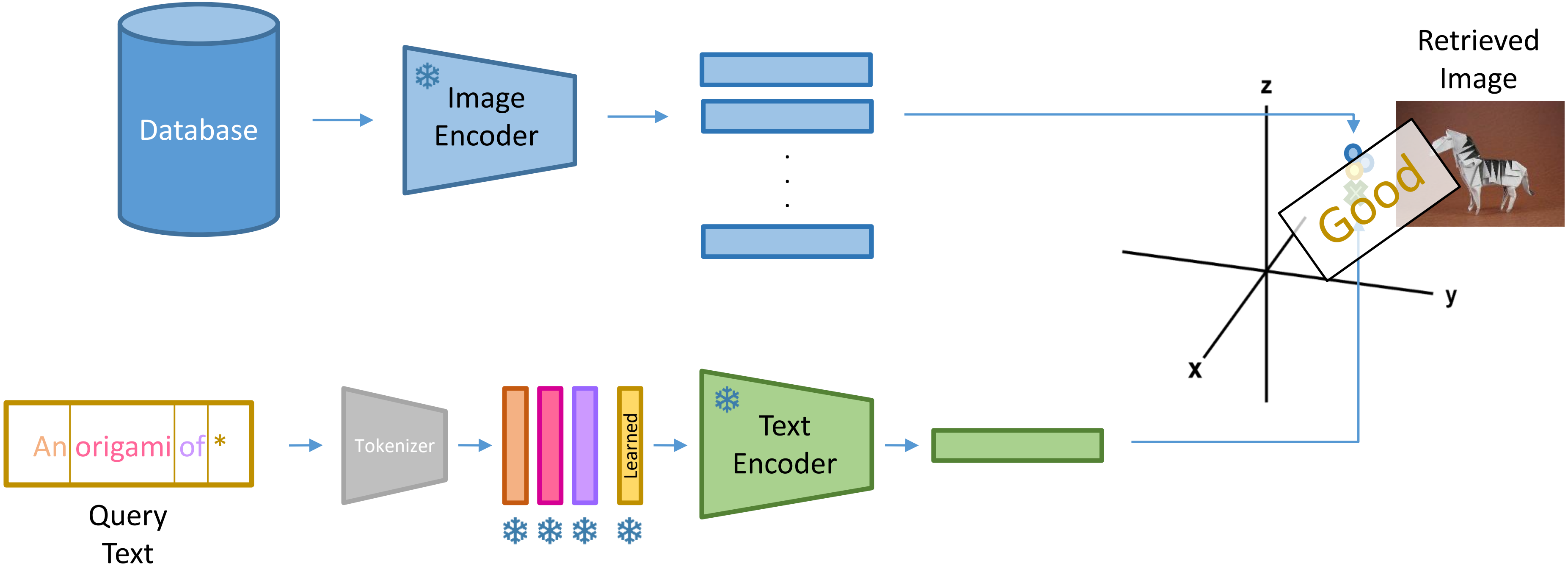
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



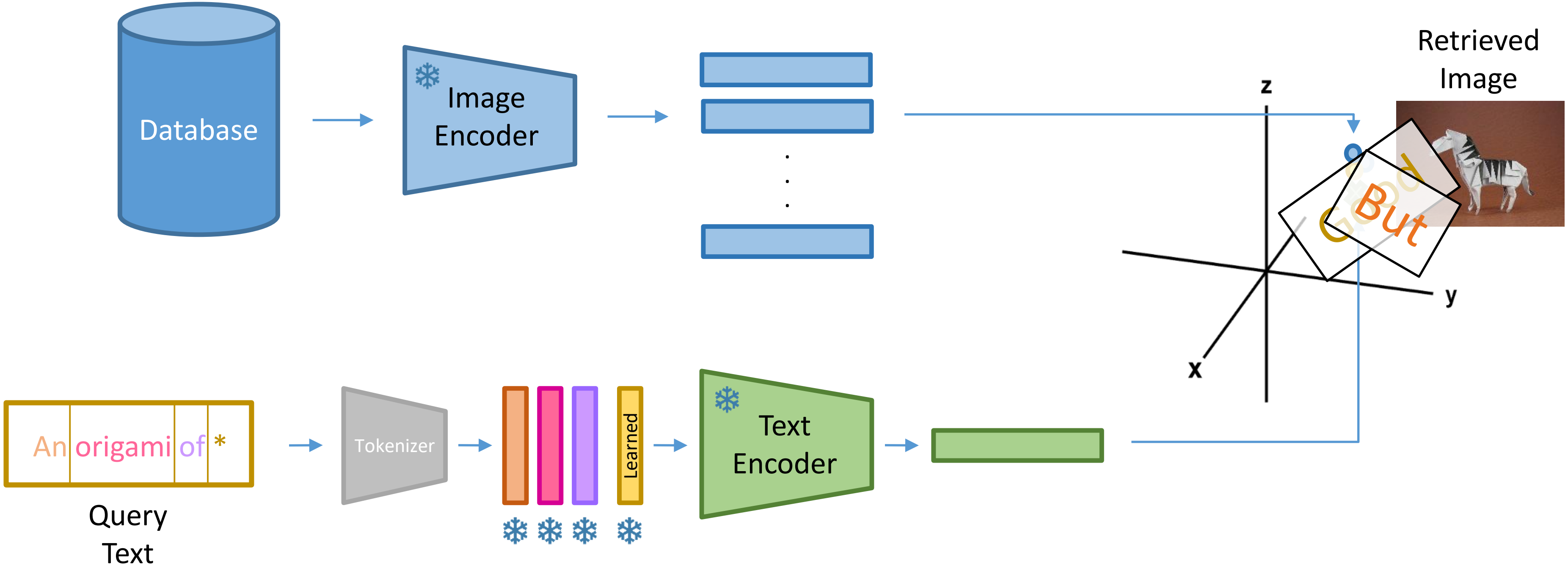
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



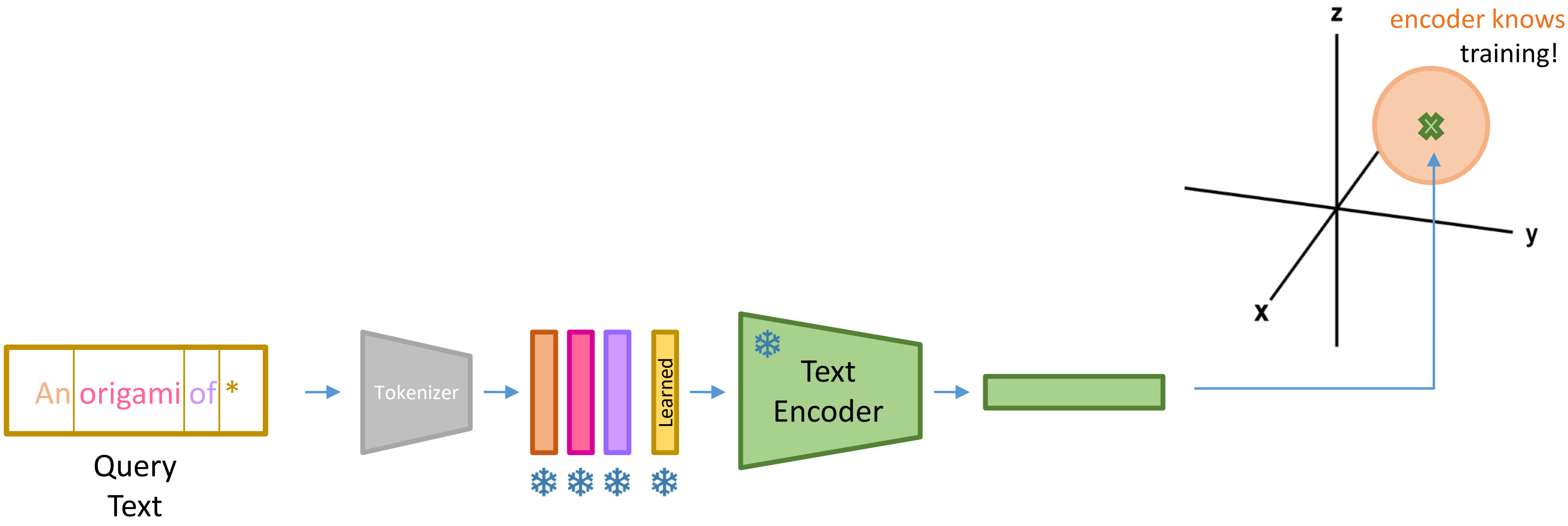
Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Related Work: Continuous-Space Textual Inversion



Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

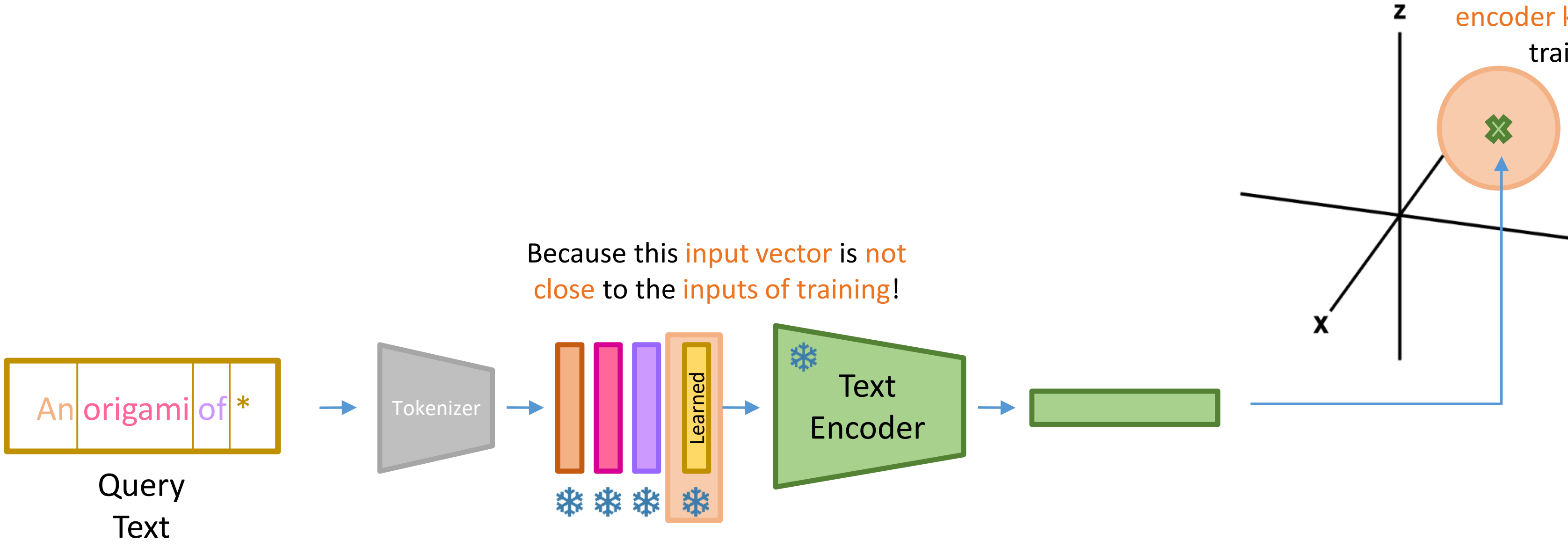
Continuous-Space Textual Inversion: The Issue



This neighborhood is not something the text encoder knows from training!

Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Continuous-Space Textual Inversion: The Issue

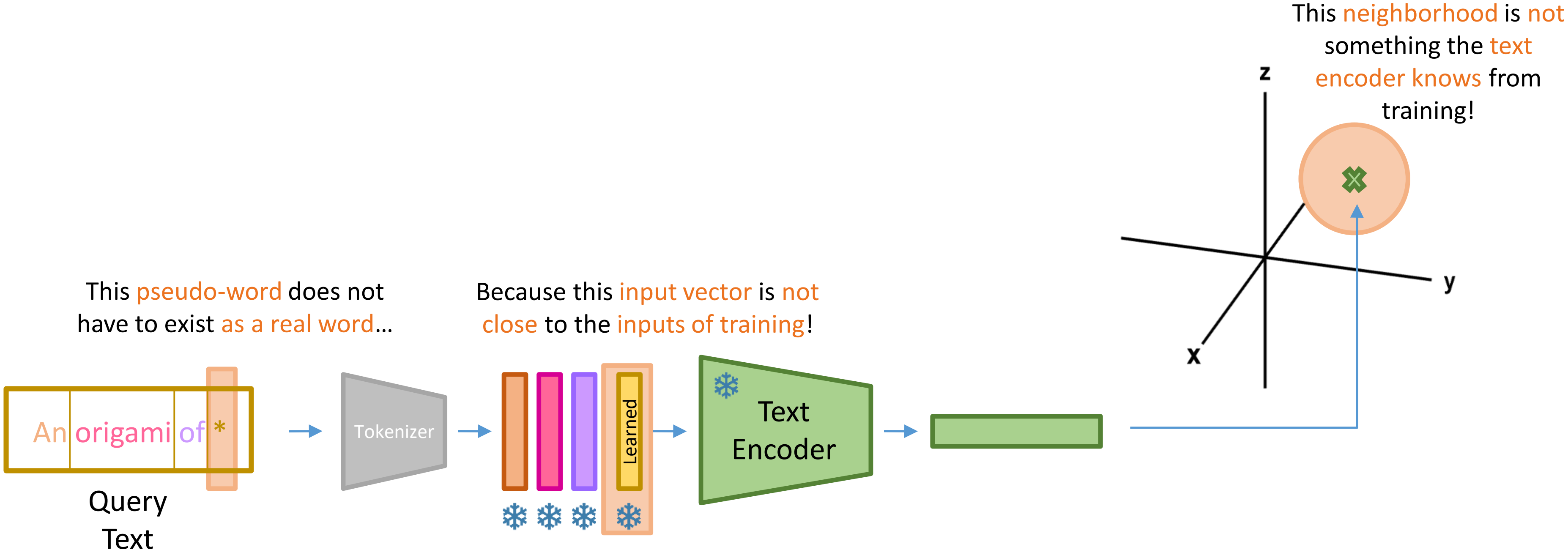


Because this input vector is not close to the inputs of training!

This neighborhood is not something the text encoder knows from training!

Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

Continuous-Space Textual Inversion: The Issue



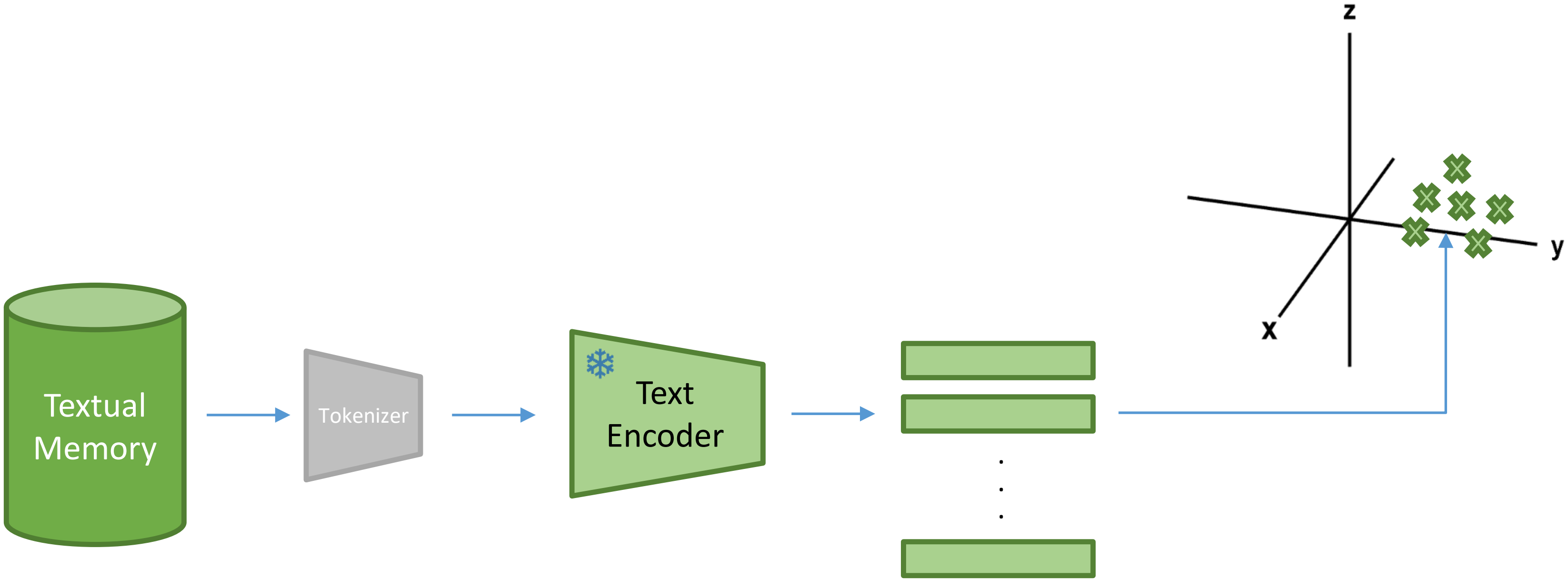
This pseudo-word does not have to exist as a real word…

Because this input vector is not close to the inputs of training!

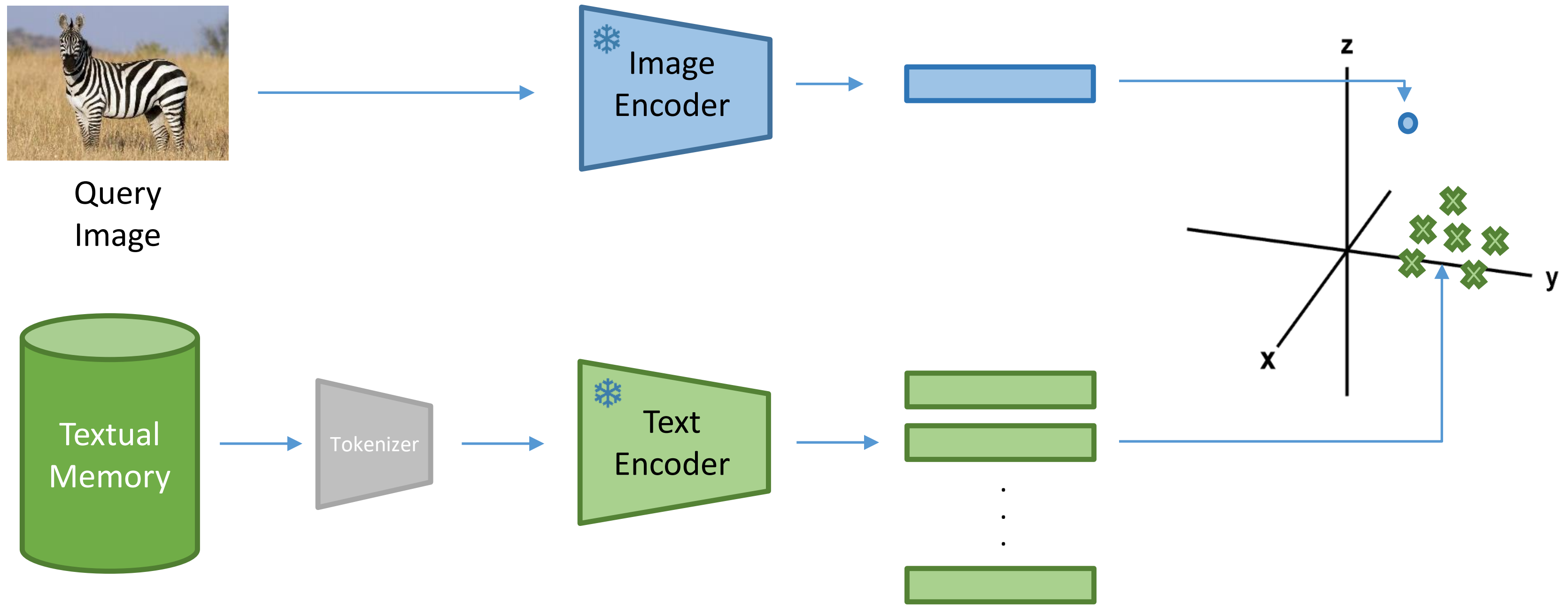
This neighborhood is not something the text encoder knows from training!

Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023

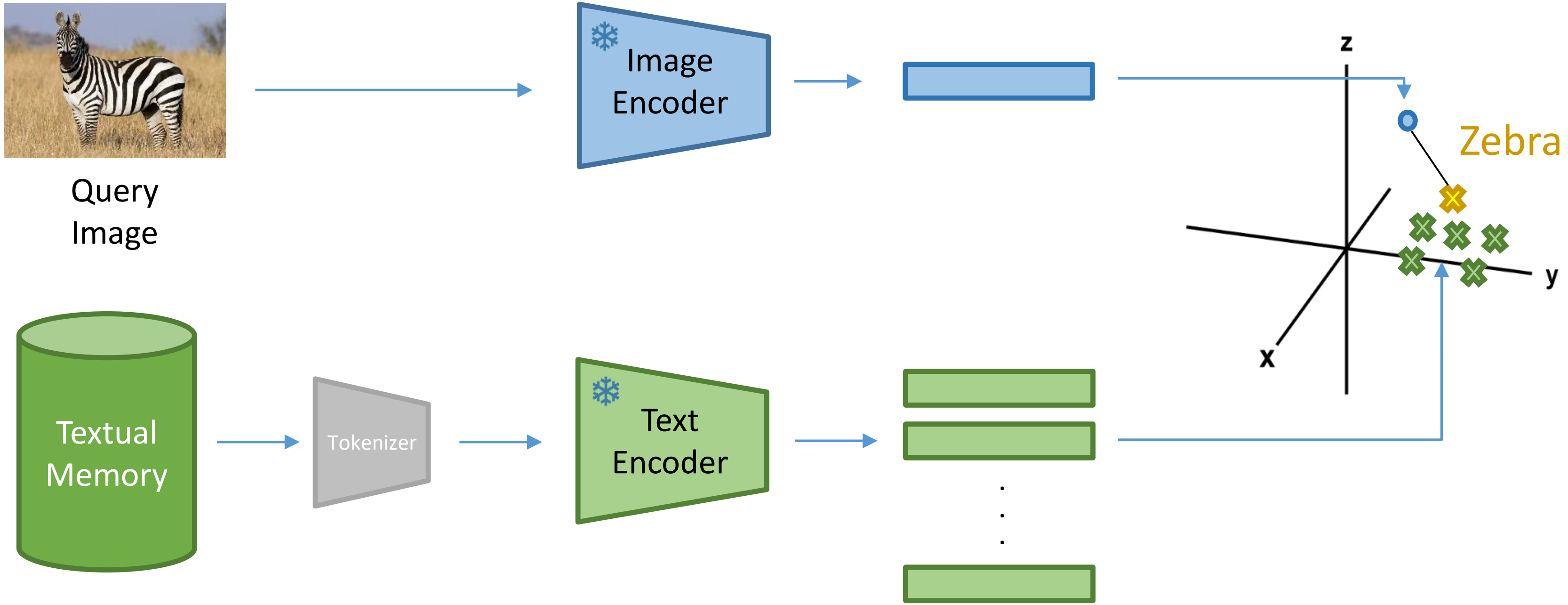
Solution: Discrete-Space Memory-Based Textual Inversion



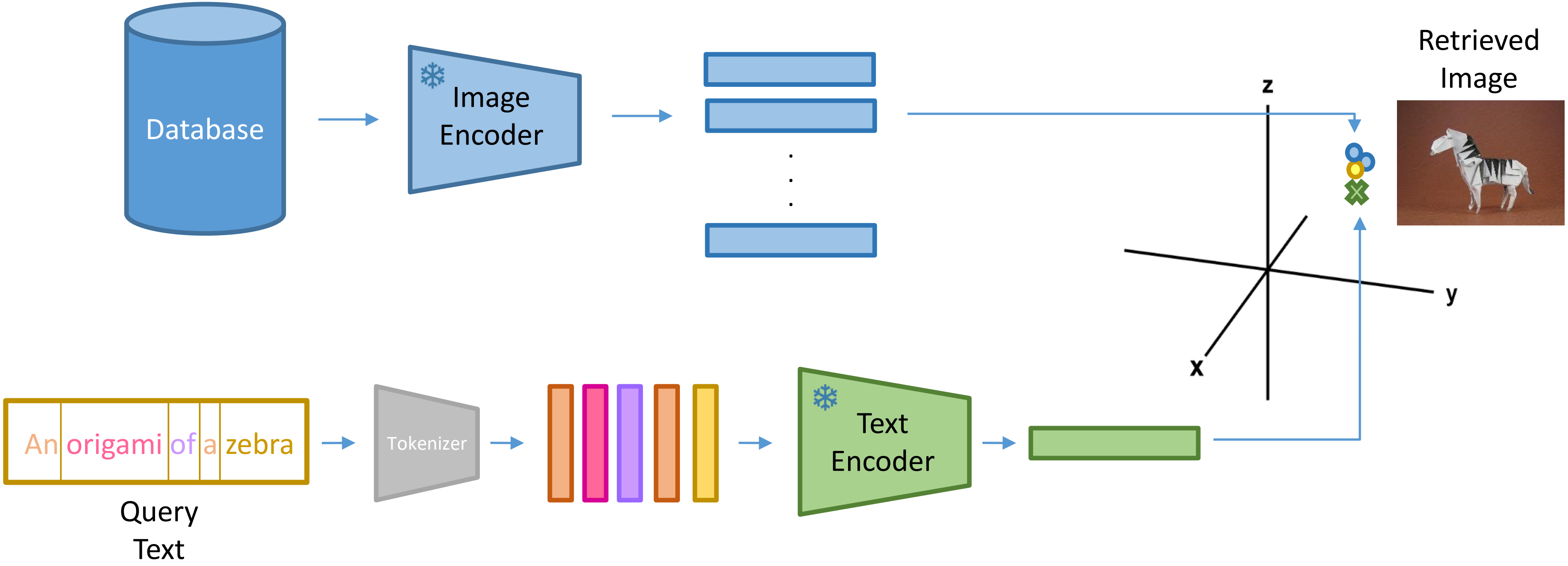
Solution: Discrete-Space Memory-Based Textual Inversion



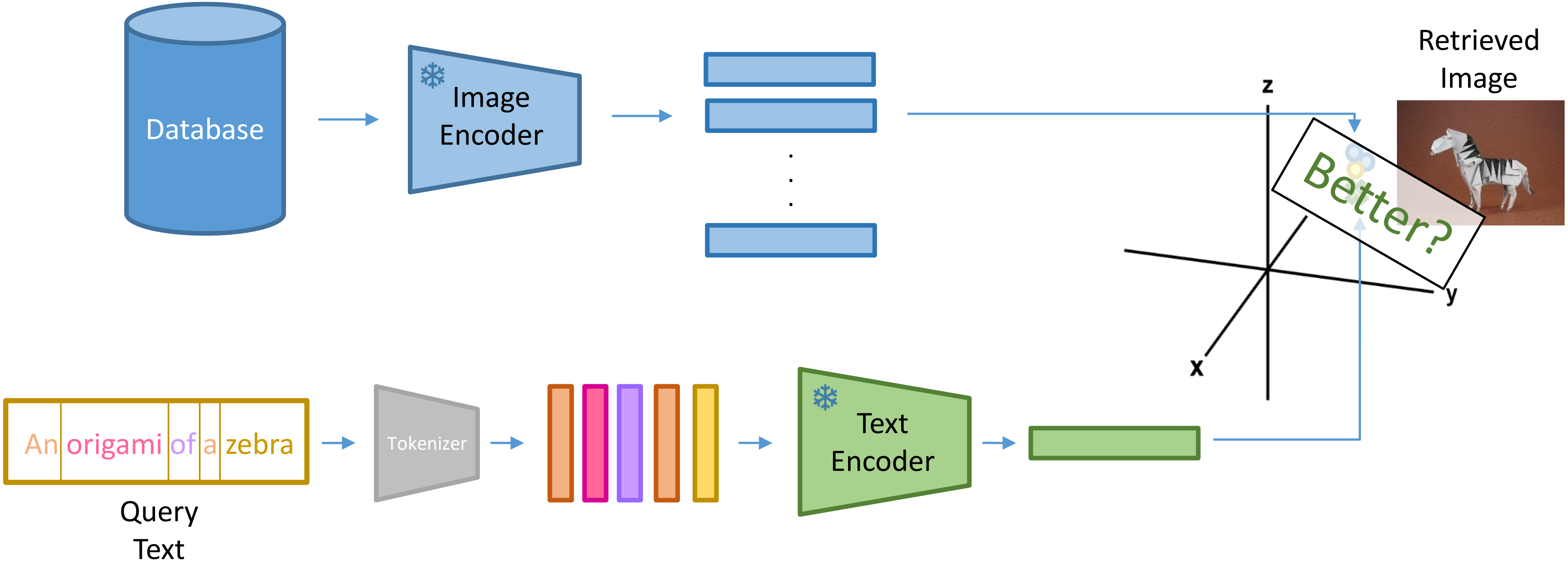
Solution: Discrete-Space Memory-Based Textual Inversion



Solution: Discrete-Space Memory-Based Textual Inversion



Solution: Discrete-Space Memory-Based Textual Inversion



Composed Image Retrieval for Domain Conversion

image query

text query: “cartoon”

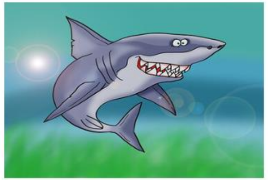
text query: “origami”

text query: “toy”

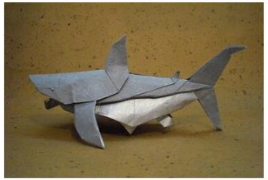
Saito et al., Pic2word: Mapping pictures to words for zero-shot composed image retrieval, CVPR 2023

Composed Image Retrieval for Domain Conversion

image query



text query: “cartoon”



text query: “origami”

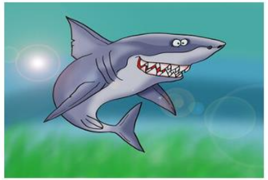
text query: “toy”

Saito et al., Pic2word: Mapping pictures to words for zero-shot composed image retrieval, CVPR 2023

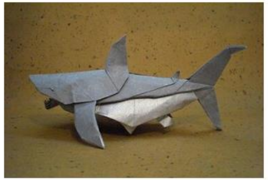
Composed Image Retrieval for Domain Conversion

image: category
domain: style

image query



text query: “cartoon”



text query: “origami”

text query: “toy”

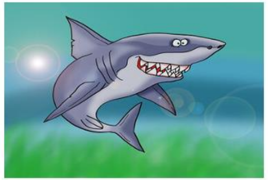
ImageNet-R

Saito et al., Pic2word: Mapping pictures to words for zero-shot composed image retrieval, CVPR 2023

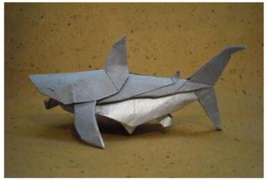
Domain Conversion: Our Benchmarks

image: category
domain: style

image query



text query: “cartoon”



text query: “origami”

text query: “toy”

ImageNet-R

image: category
domain: context

image query

text query: “grass”

text query: “autumn”

text query: “rock”

NICO++

Domain Conversion: Our Benchmarks

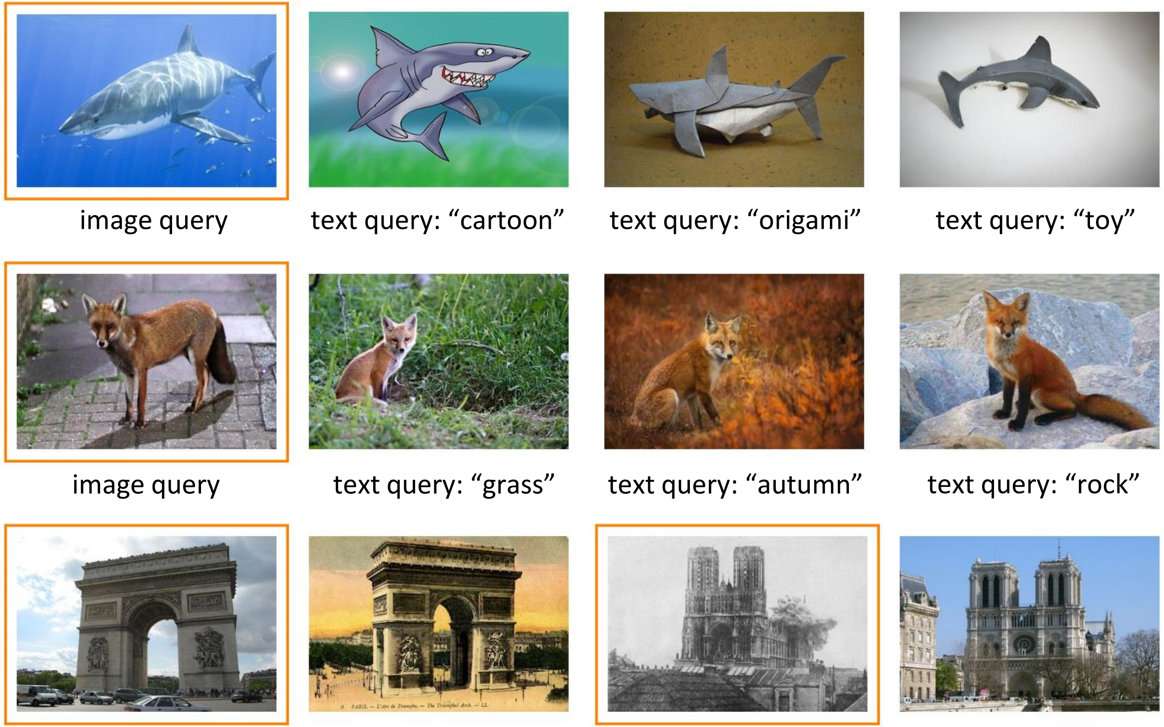


image: category
domain: style

image query

text query: “cartoon”

text query: “origami”

text query: “toy”

ImageNet-R

image: category
domain: context

image query

text query: “grass”

text query: “autumn”

text query: “rock”

NICO++

image: instance
domain: style

image query

text query: “archive”

image query

text query: “today”

LTLL

Domain Conversion: Our Benchmarks

image: category
domain: style

image query

text query: “cartoon”

text query: “origami”

text query: “toy”

+1 more benchmark

ImageNet-R

image: category
domain: context

image query

text query: “grass”

text query: “autumn”

text query: “rock”

NICO++

image: instance
domain: style

image query

text query: “archive”

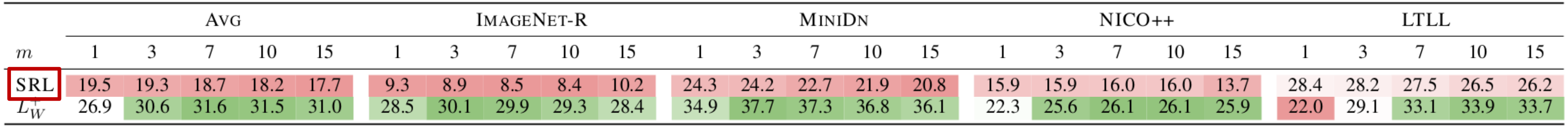
image query

text query: “today”

LTLL

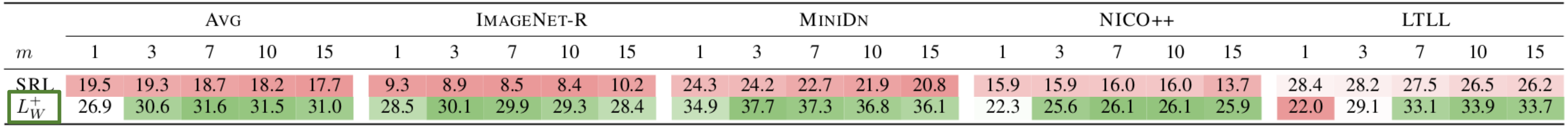
Continuous-Space vs. Discrete-Space Memory-Based

Continuous-Space Textual Inversion with SEARLE



	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

Continuous-Space vs. Discrete-Space Memory-Based

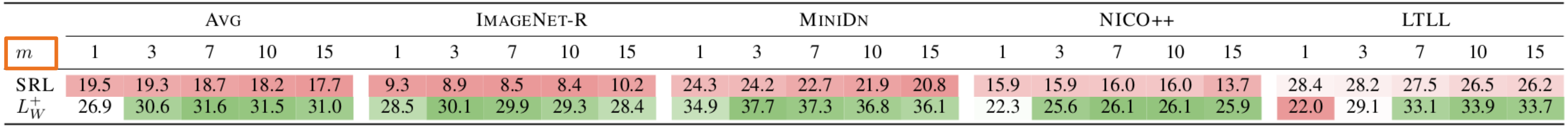


	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

Discrete-Space Memory-Based Textual Inversion with our method, FreeDom

Continuous-Space vs. Discrete-Space Memory-Based

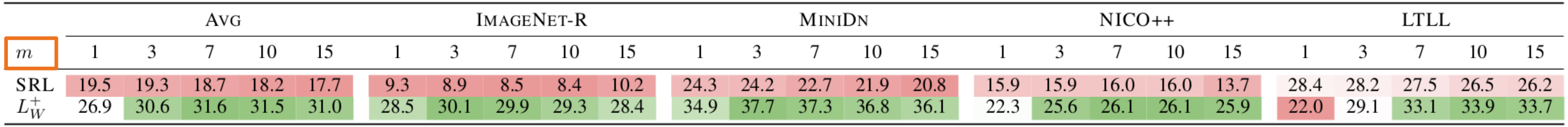
Number of words retrieved
from textual memory



	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

Continuous-Space vs. Discrete-Space Memory-Based

Number of words retrieved
from textual memory



	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

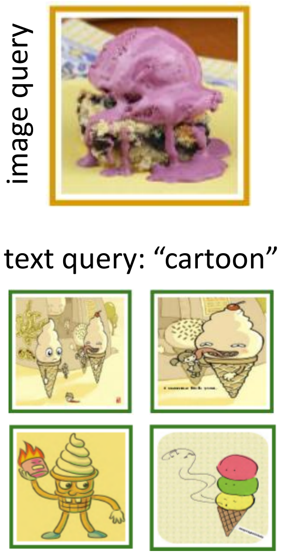


image query

text query: "cartoon"

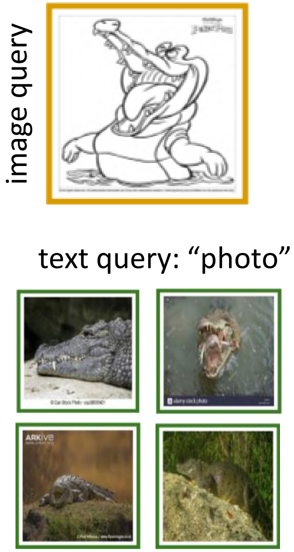


image query

text query: "photo"

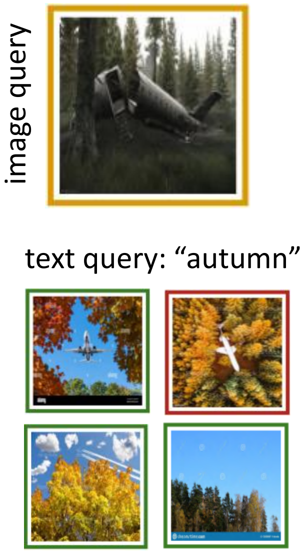


image query

text query: "autumn"

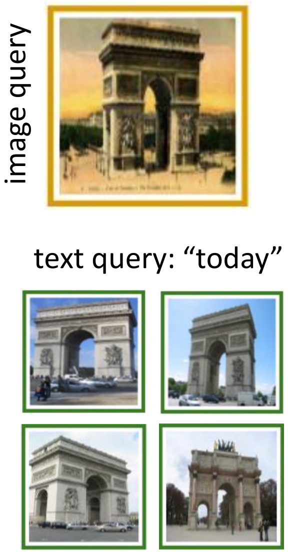


image query

text query: "today"

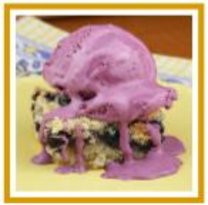
Continuous-Space vs. Discrete-Space Memory-Based

Number of words retrieved from textual memory

Problem with our instance-level LTLL dataset

	Avg					ImageNet-R					MiniDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

image query



text query: “cartoon”

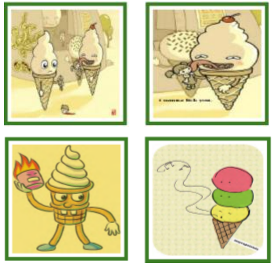
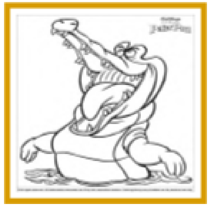


image query



text query: “photo”

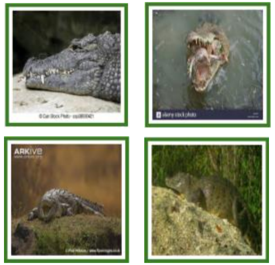
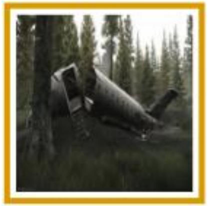


image query



text query: “autumn”

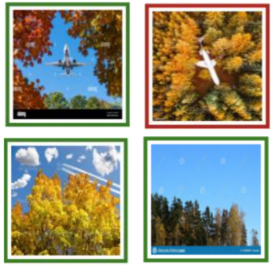
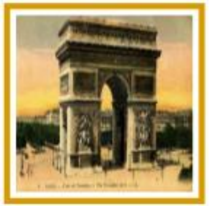
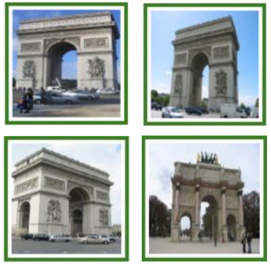


image query



text query: “today”

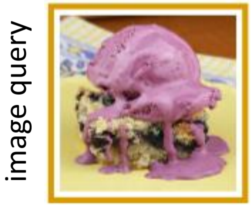


Continuous-Space vs. Discrete-Space Memory-Based

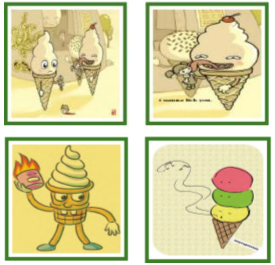
Number of words retrieved
from textual memory

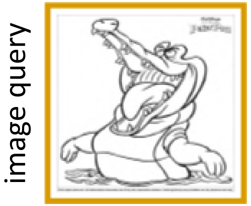
Makes sense: A single word
cannot describe an instance well…

	Avg					ImageNet-R					MiniDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

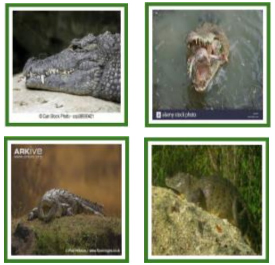


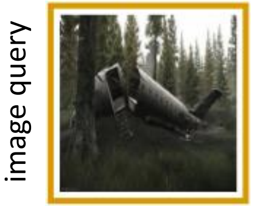
text query: “cartoon”



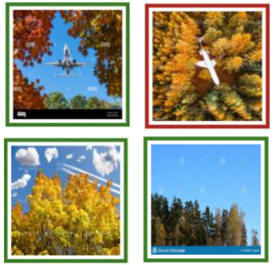


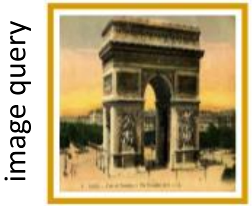
text query: “photo”



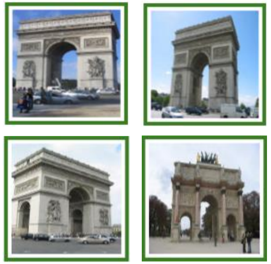


text query: “autumn”





text query: “today”

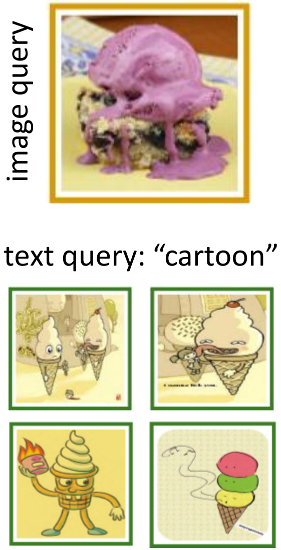


Continuous-Space vs. Discrete-Space Memory-Based

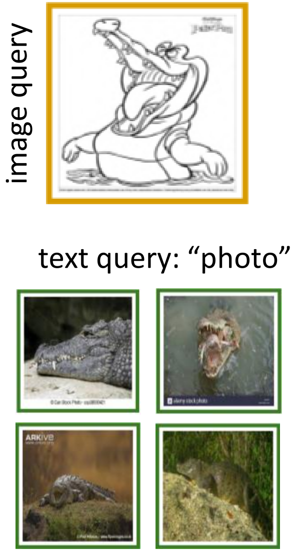
Number of words retrieved from textual memory

This can be fixed using more words ☺

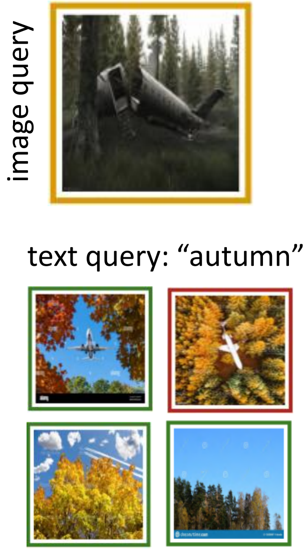
	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7



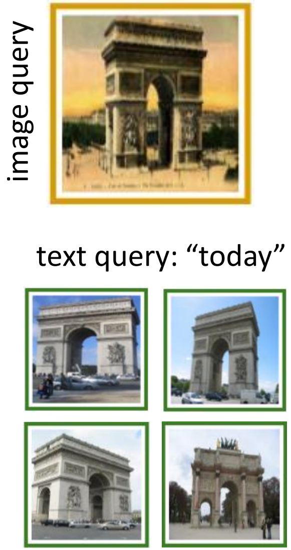
text query: "cartoon"



text query: "photo"



text query: "autumn"



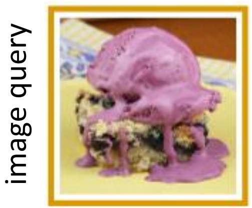
text query: "today"

Continuous-Space vs. Discrete-Space Memory-Based

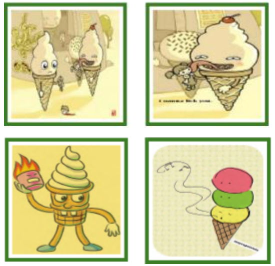
Number of words retrieved from textual memory

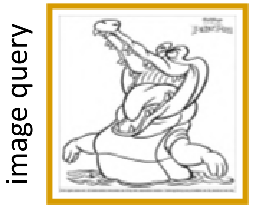
This can be fixed using more words ☺

	AVG					IMAGENET-R					MINIDN					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
L_W^+	26.9	30.6	31.6	31.5	31.0	28.5	30.1	29.9	29.3	28.4	34.9	37.7	37.3	36.8	36.1	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.1	33.9	33.7

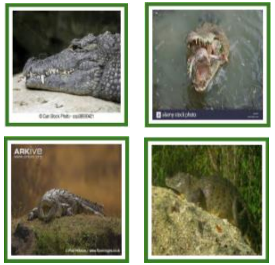


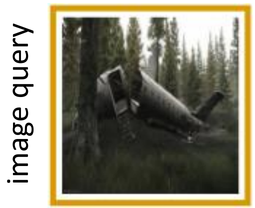
text query: “cartoon”



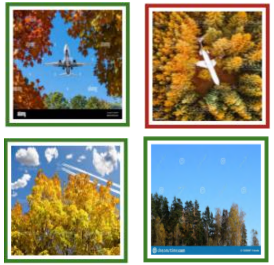


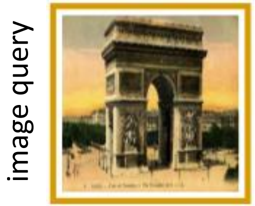
text query: “photo”



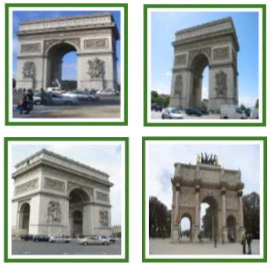


text query: “autumn”



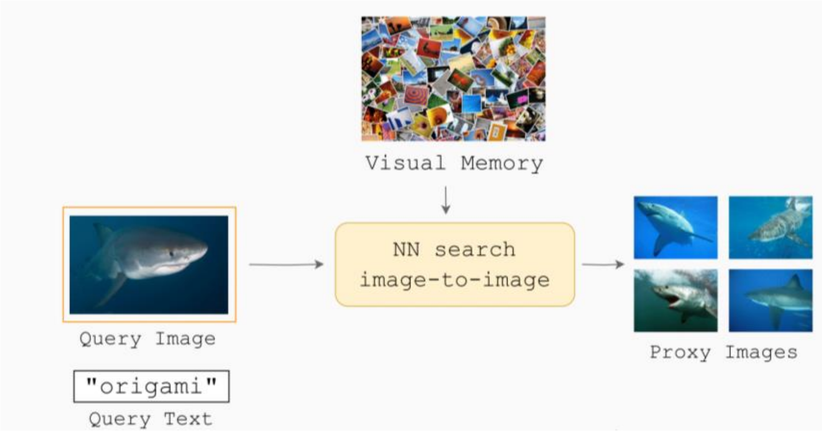


text query: “today”

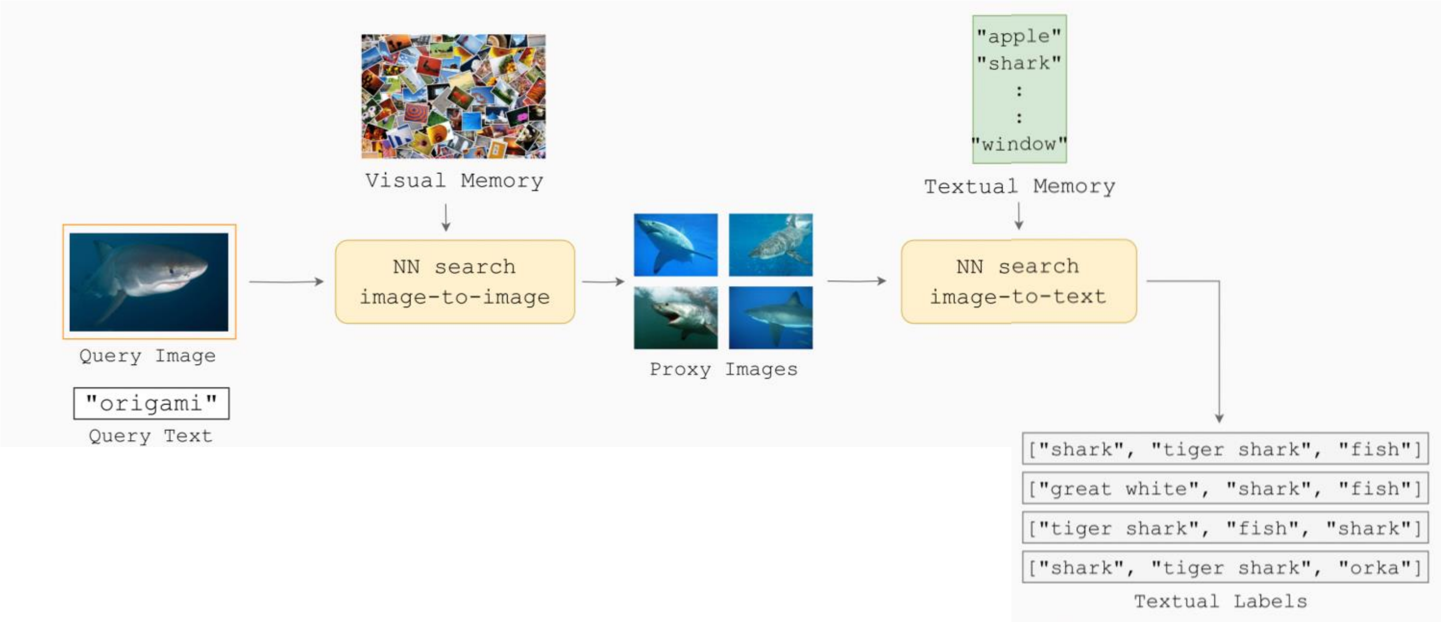


Our Full Method: FreeDom

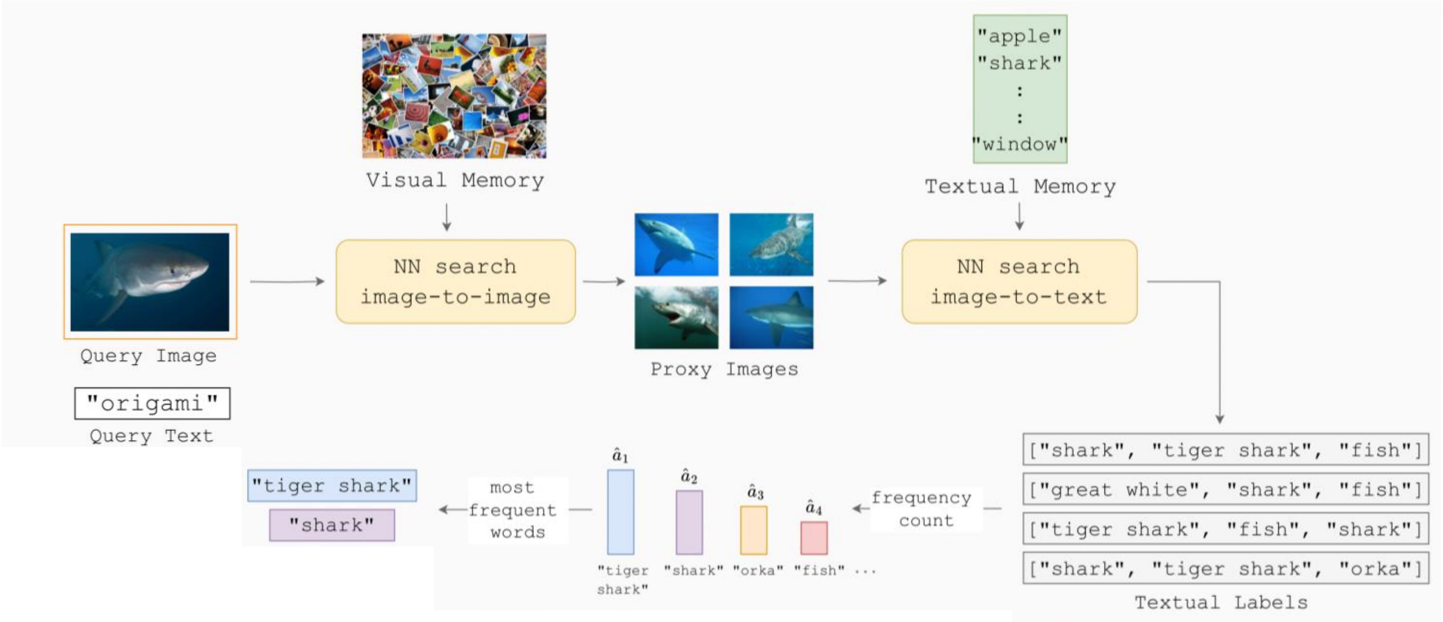
Our Full Method: FreeDom



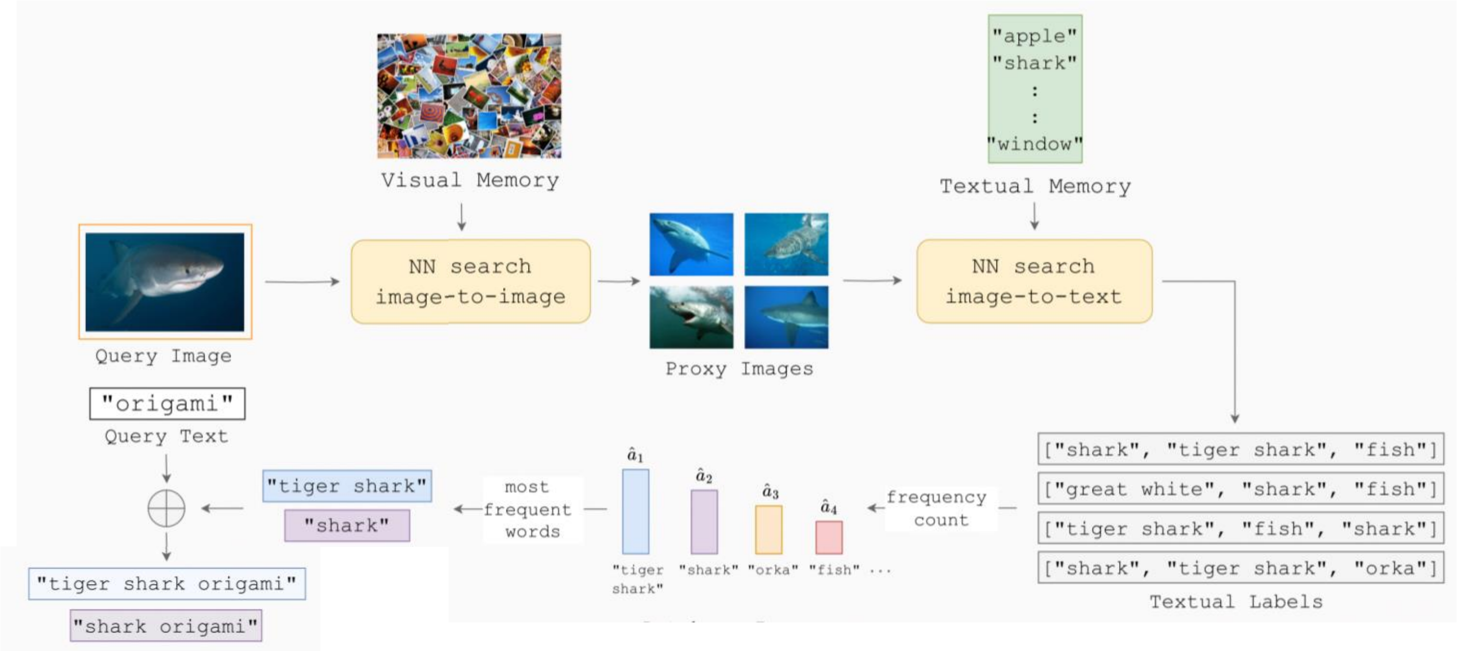
Our Full Method: FreeDom



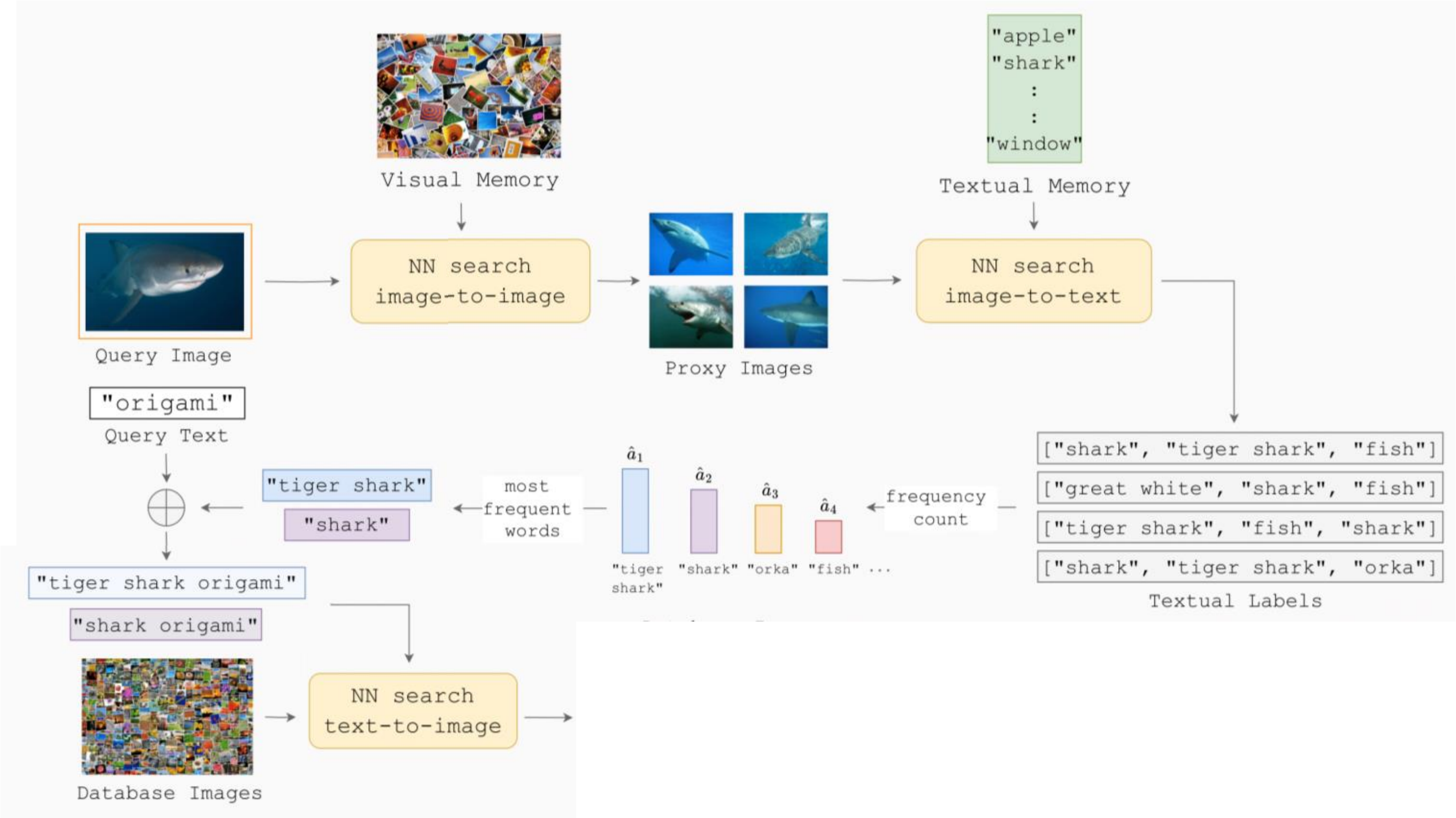
Our Full Method: FreeDom



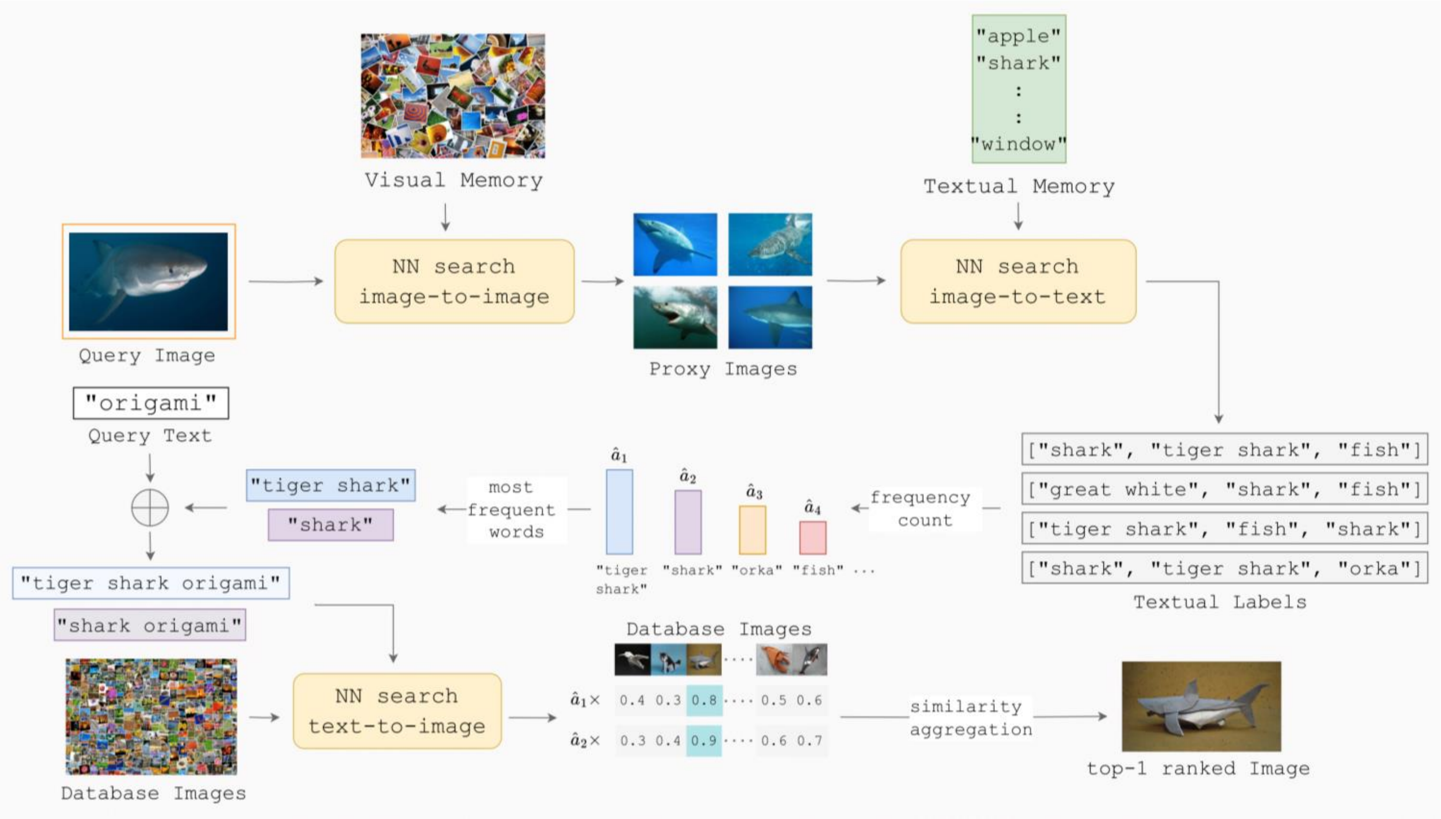
Our Full Method: FreeDom



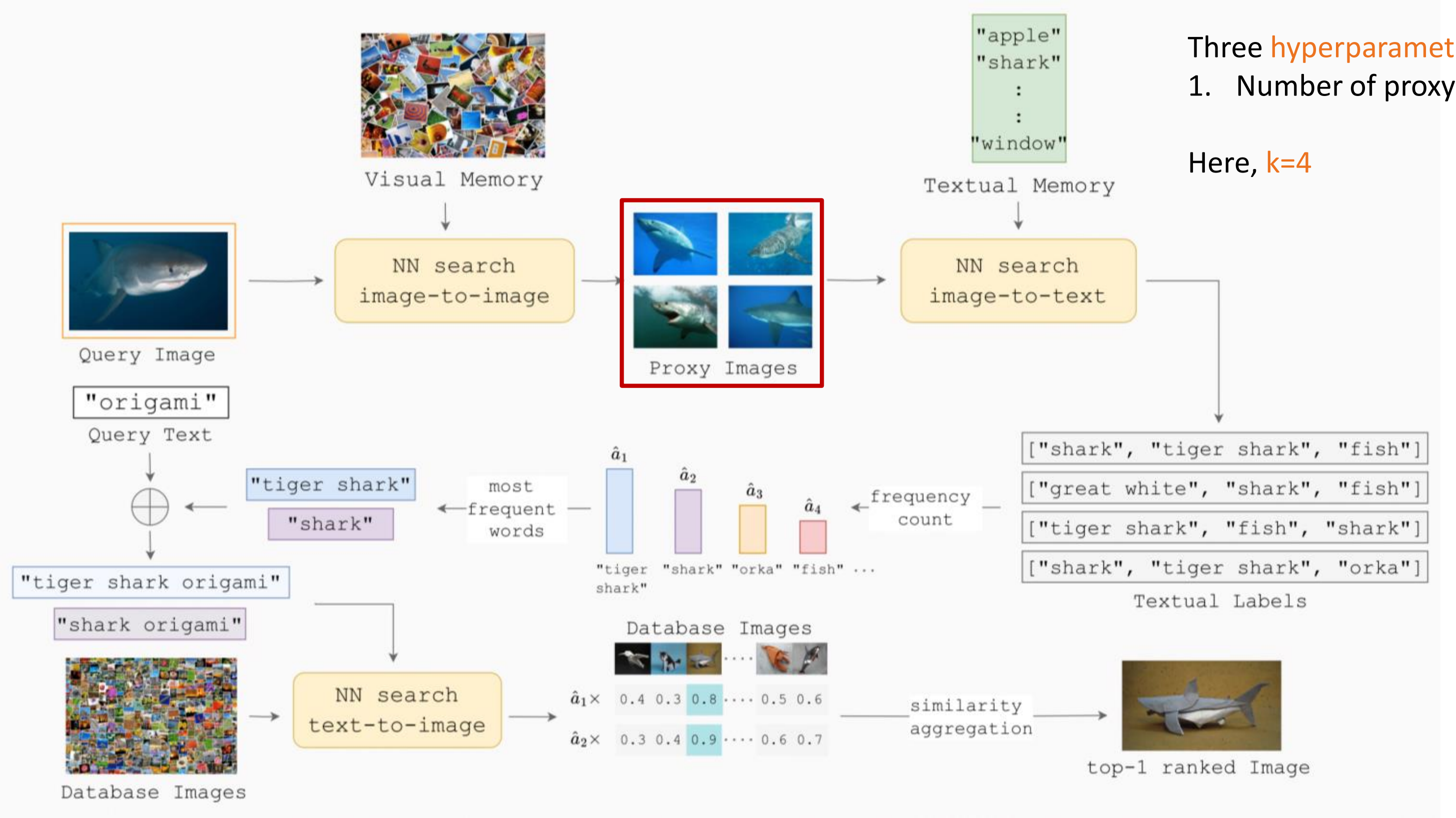
Our Full Method: FreeDom



Our Full Method: FreeDom



Our Full Method: FreeDom

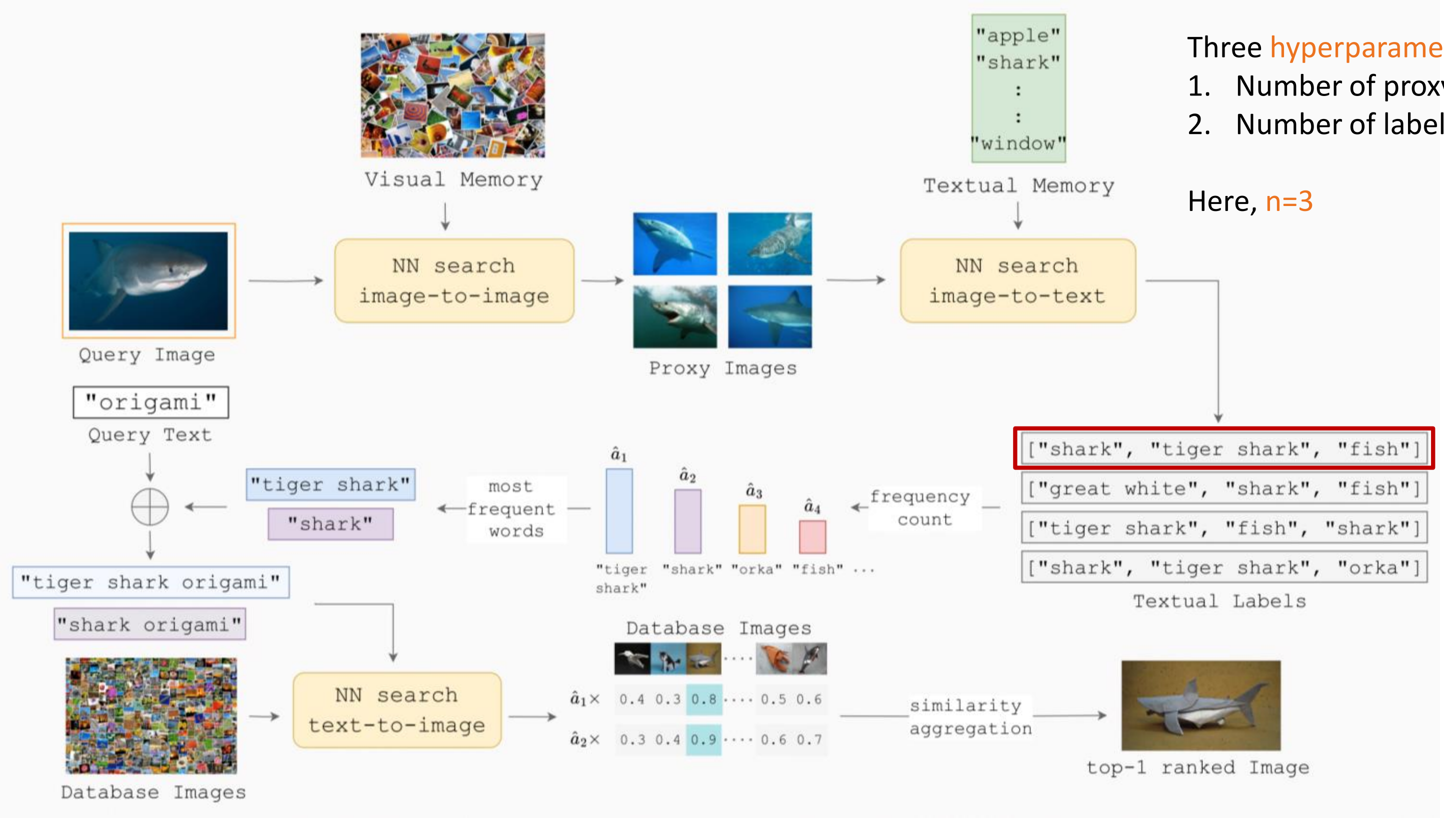


Three hyperparameters:

1. Number of proxy images (k)

Here, k=4

Our Full Method: FreeDom

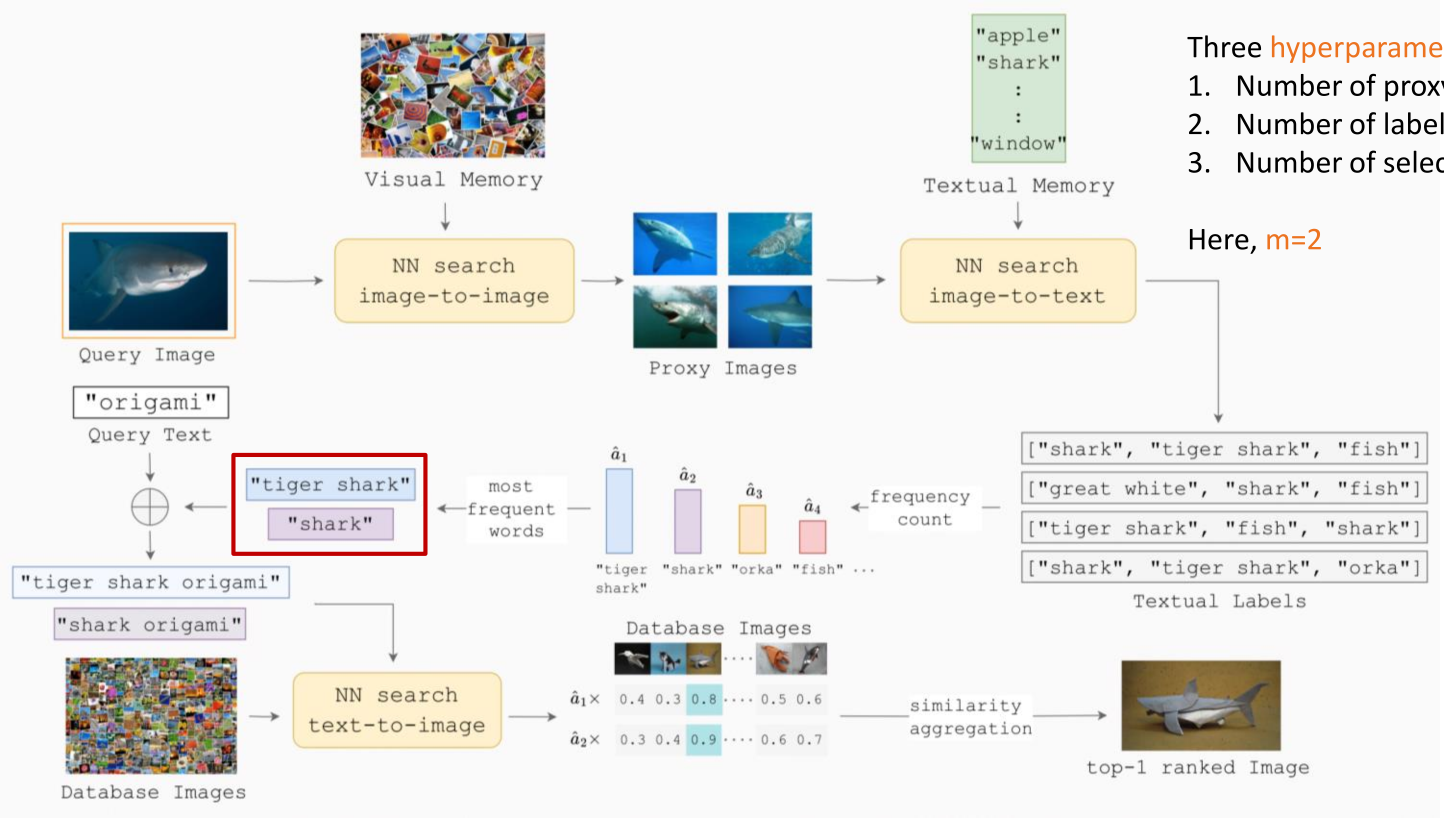


Three hyperparameters:

1. Number of proxy images (k)
2. Number of labels per proxy (n)

Here, n=3

Our Full Method: FreeDom



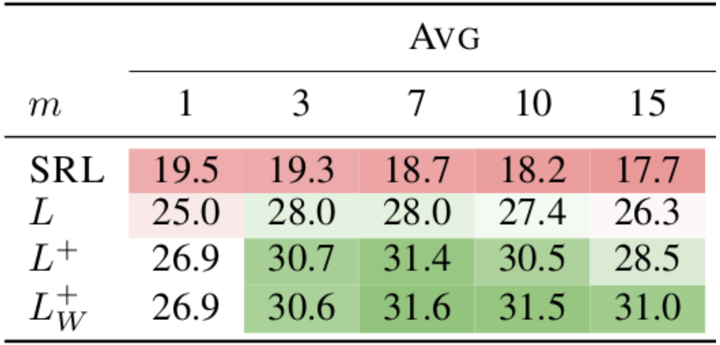
Three hyperparameters:

1. Number of proxy images (k)
2. Number of labels per proxy (n)
3. Number of selected words (m)

Here, m=2

FreeDom: Ablations

Three hyperparameters → Three method components



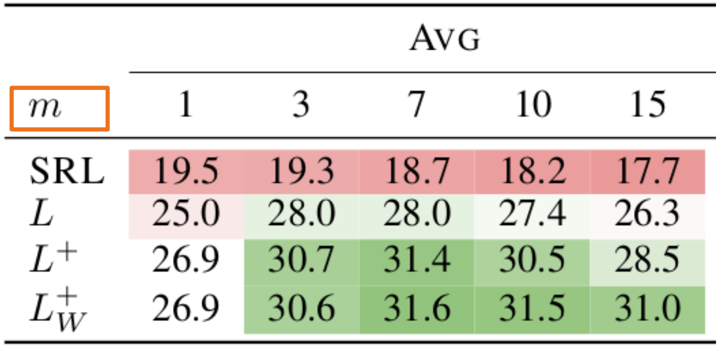
	Avg				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L^+_W	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

Three hyperparameters → Three method components

Continuous-Space Textual Inversion with SEARLE



	AVG				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L^+_W	26.9	30.6	31.6	31.5	31.0

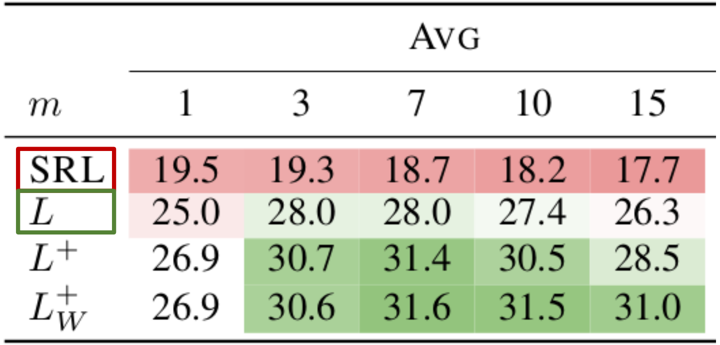
The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

Three hyperparameters → Three method components

Continuous-Space Textual Inversion with SEARLE

Discrete-Space Textual Inversion with FreeDom (Textual Memory)



	AVG				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L^+_W	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

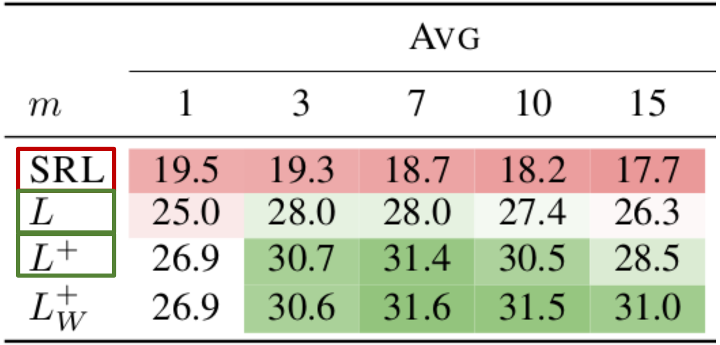
FreeDom: Ablations

Three hyperparameters → Three method components

Continuous-Space Textual Inversion with SEARLE

Discrete-Space Textual Inversion with FreeDom (Textual Memory)

+ Visual Memory



	AVG				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L_W^+	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

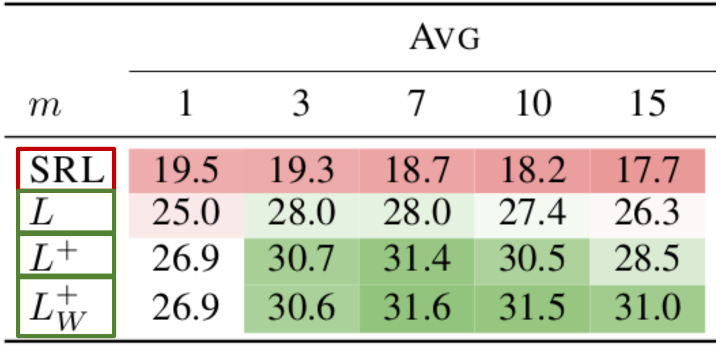
Three hyperparameters → Three method components

Continuous-Space Textual Inversion with SEARLE

Discrete-Space Textual Inversion with FreeDom (Textual Memory)

+ Visual Memory

+ Frequencies as Weights



	AVG				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L_W^+	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

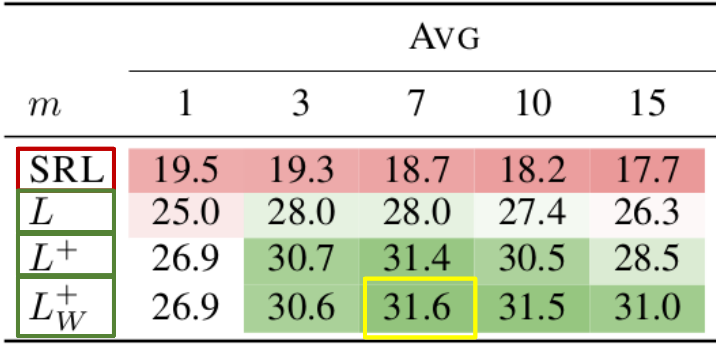
Three hyperparameters → Three method components

Continuous-Space Textual Inversion with SEARLE

Discrete-Space Textual Inversion with FreeDom (Textual Memory)

+ Visual Memory

+ Frequencies as Weights



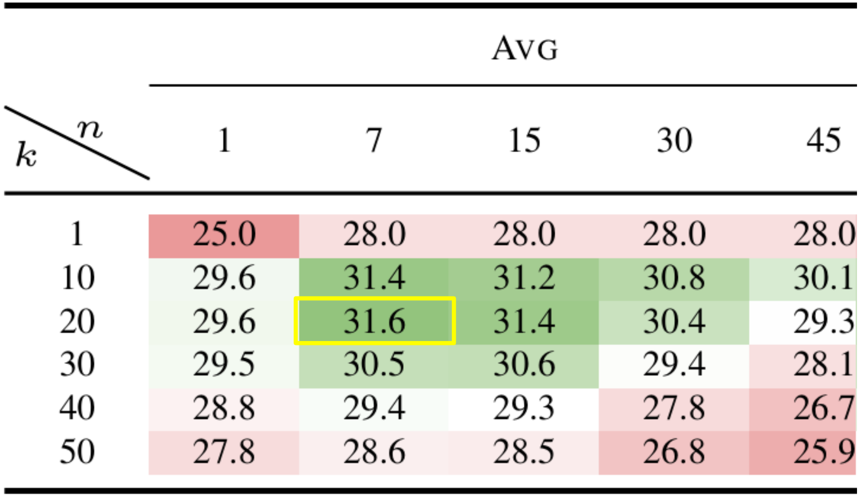
	AVG				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L_W^+	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

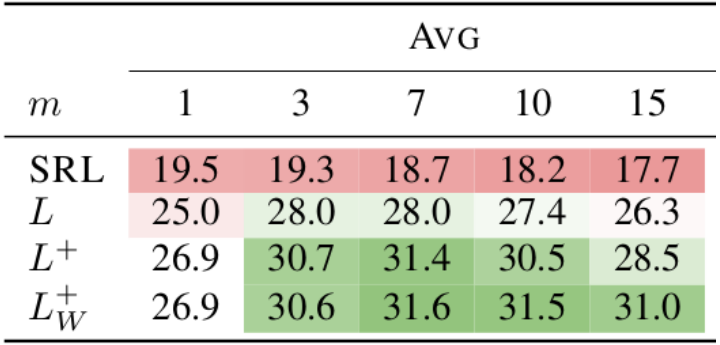
Three hyperparameters:

1. Number of proxy images (k)
2. Number of labels per proxy (n)
3. Number of selected words (m)



	Avg				
$k \backslash n$	1	7	15	30	45
1	25.0	28.0	28.0	28.0	28.0
10	29.6	31.4	31.2	30.8	30.1
20	29.6	31.6	31.4	30.4	29.3
30	29.5	30.5	30.6	29.4	28.1
40	28.8	29.4	29.3	27.8	26.7
50	27.8	28.6	28.5	26.8	25.9

The effect of number of proxy images (k) and number of labels per proxy (n) in FreeDom



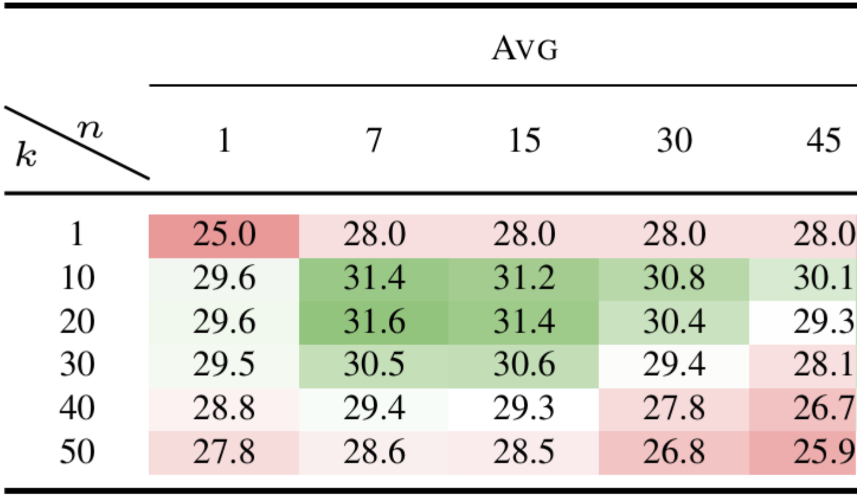
	Avg				
m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L_W^+	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

FreeDom: Ablations

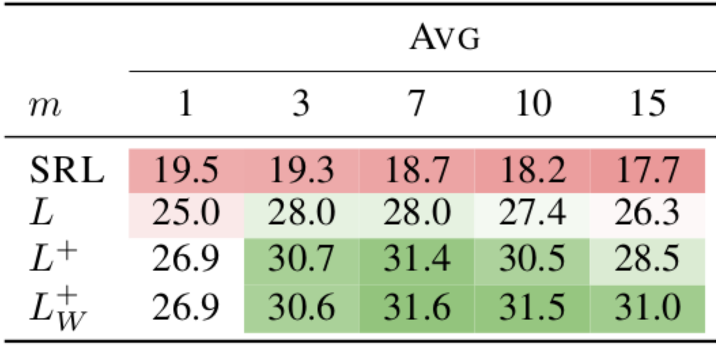
Best hyperparameters:

1. Number of proxy images k = 20
2. Number of labels per proxy n = 7
3. Number of selected words m = 7



k \ n	AVG 1	7	15	30	45
1	25.0	28.0	28.0	28.0	28.0
10	29.6	31.4	31.2	30.8	30.1
20	29.6	31.6	31.4	30.4	29.3
30	29.5	30.5	30.6	29.4	28.1
40	28.8	29.4	29.3	27.8	26.7
50	27.8	28.6	28.5	26.8	25.9

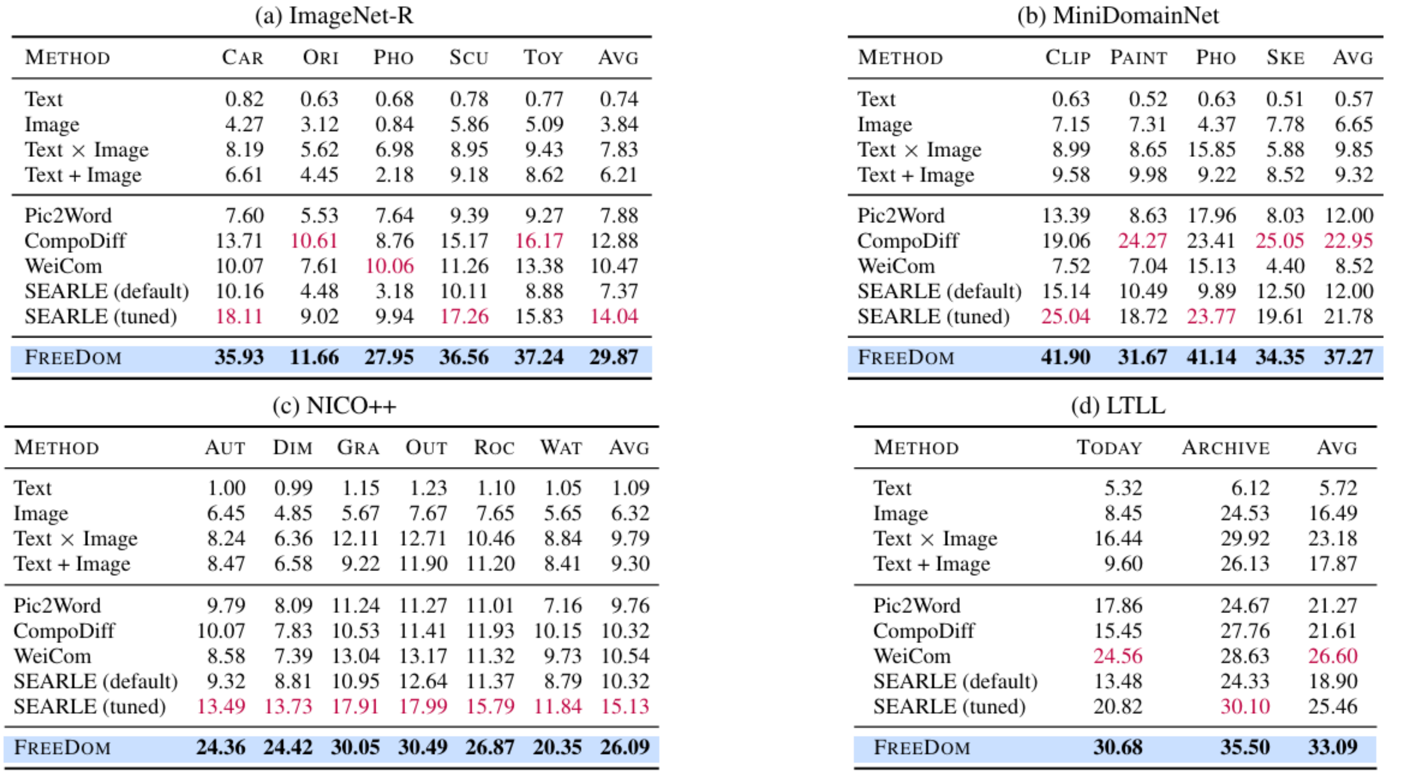
The effect of number of proxy images (k) and number of labels per proxy (n) in FreeDom



m	AVG 1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.0	28.0	28.0	27.4	26.3
L^+	26.9	30.7	31.4	30.5	28.5
L_W^+	26.9	30.6	31.6	31.5	31.0

The effect of number of selected words (m) on each FreeDom component

Quantitative Evaluation



(a) ImageNet-R

Method	Car	Ori	Pho	Scu	Toy	Avg
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FreeDom	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

Method	Clip	Paint	Pho	Ske	Avg
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FreeDom	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

Method	Aut	Dim	Gra	Out	Roc	Wat	Avg
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FreeDom	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

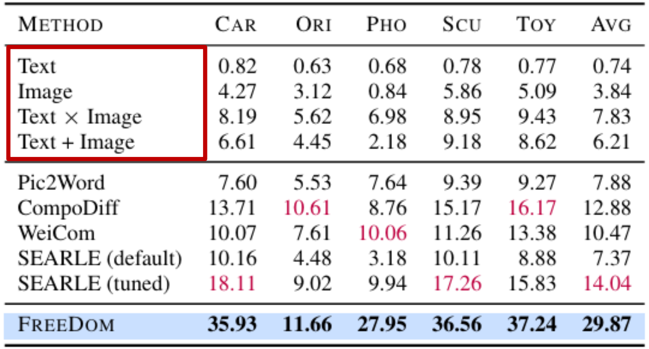
(d) LTLL

Method	Today	Archive	Avg
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FreeDom	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

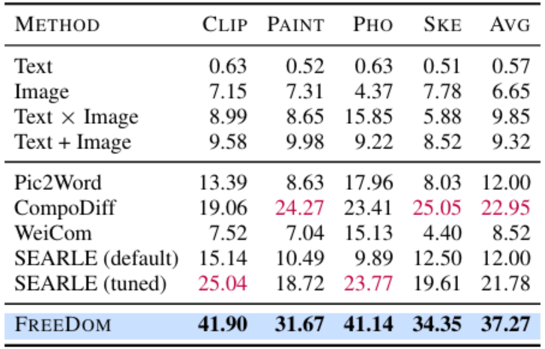
Quantitative Evaluation

(a) ImageNet-R



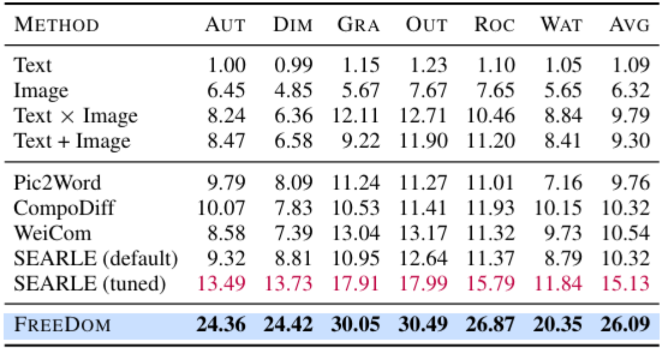
METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FREEDOM	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet



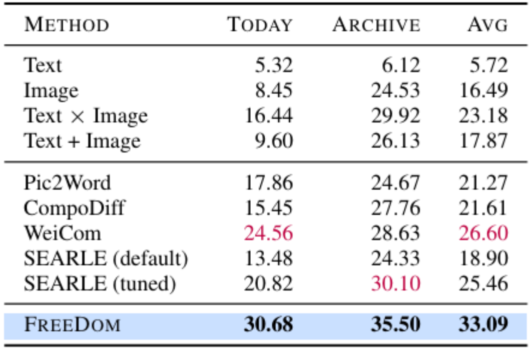
METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FREEDOM	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++



METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FREEDOM	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

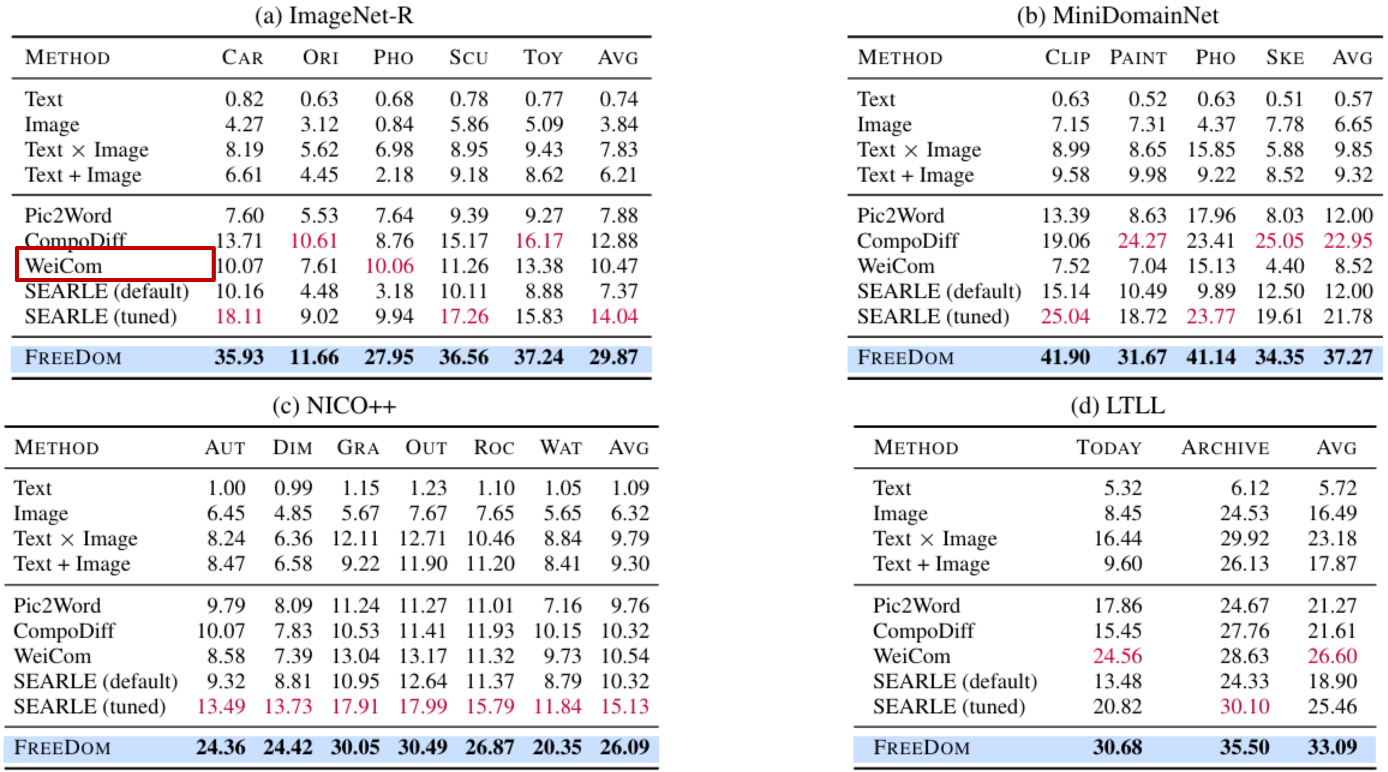
(d) LTLL



METHOD	TODAY	ARCHIVE	AVG
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FREEDOM	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Quantitative Evaluation



(a) ImageNet-R

Method	Car	Ori	Pho	Scu	Toy	Avg
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FreeDom	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

Method	Clip	Paint	Pho	Ske	Avg
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FreeDom	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

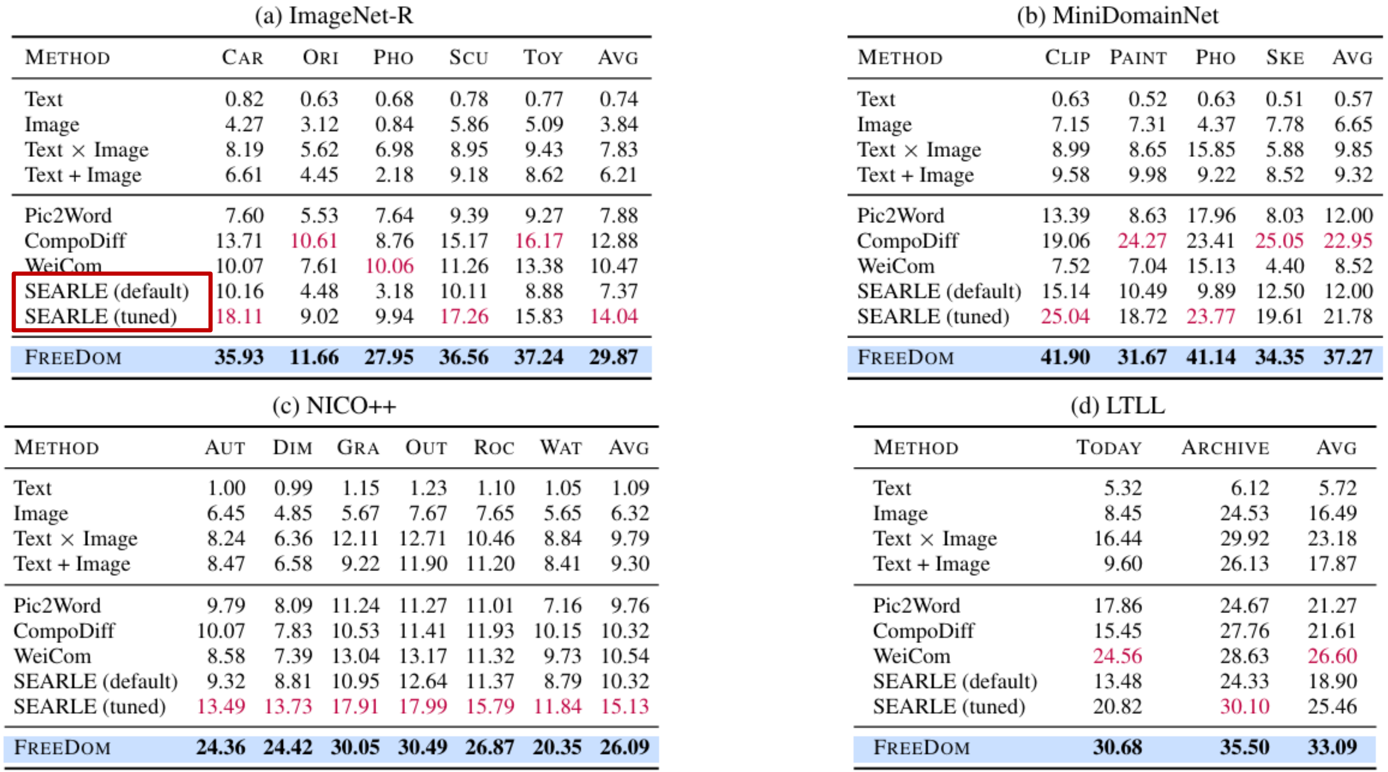
Method	Aut	Dim	Gra	Out	Roc	Wat	Avg
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FreeDom	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

(d) LTLL

Method	Today	Archive	Avg
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FreeDom	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Quantitative Evaluation



(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FREEDOM	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FREEDOM	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

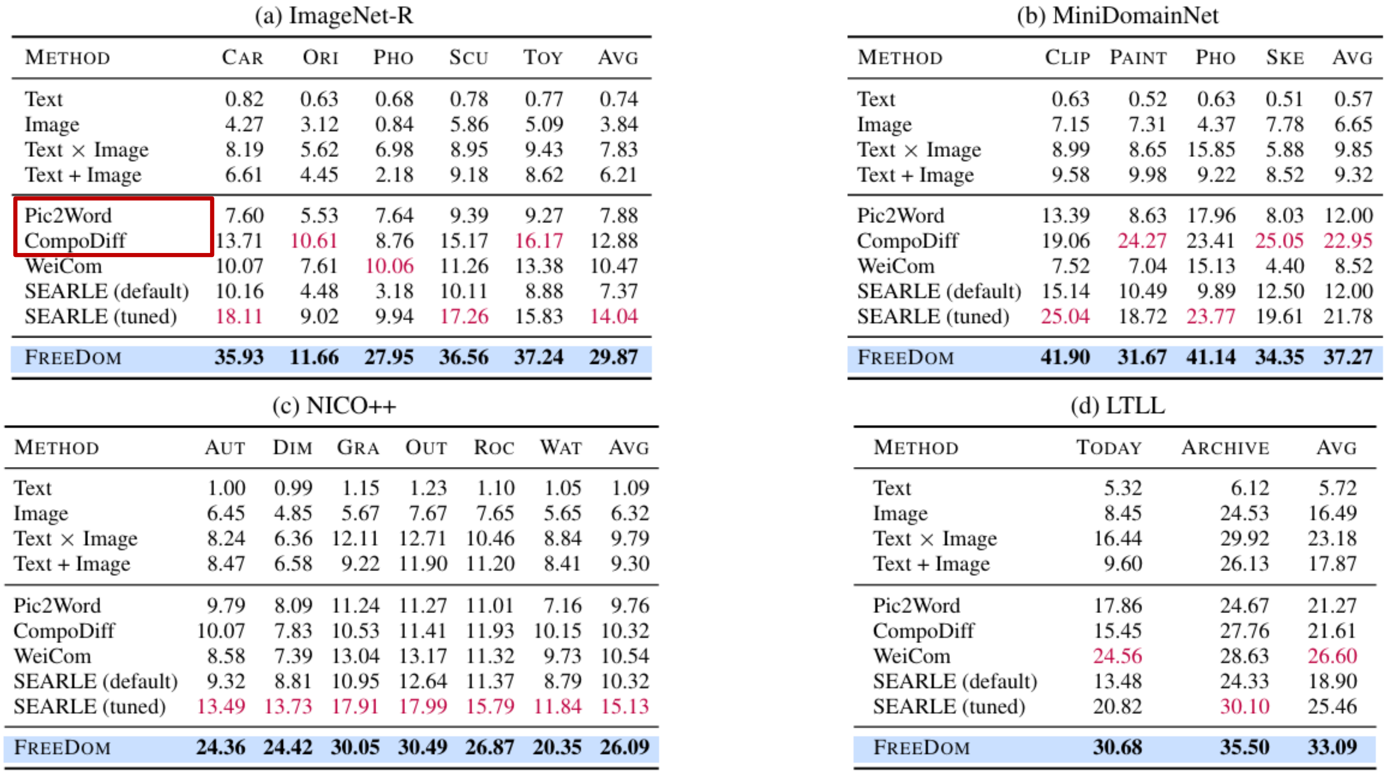
METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FREEDOM	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FREEDOM	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Quantitative Evaluation



(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FREEDOM	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FREEDOM	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

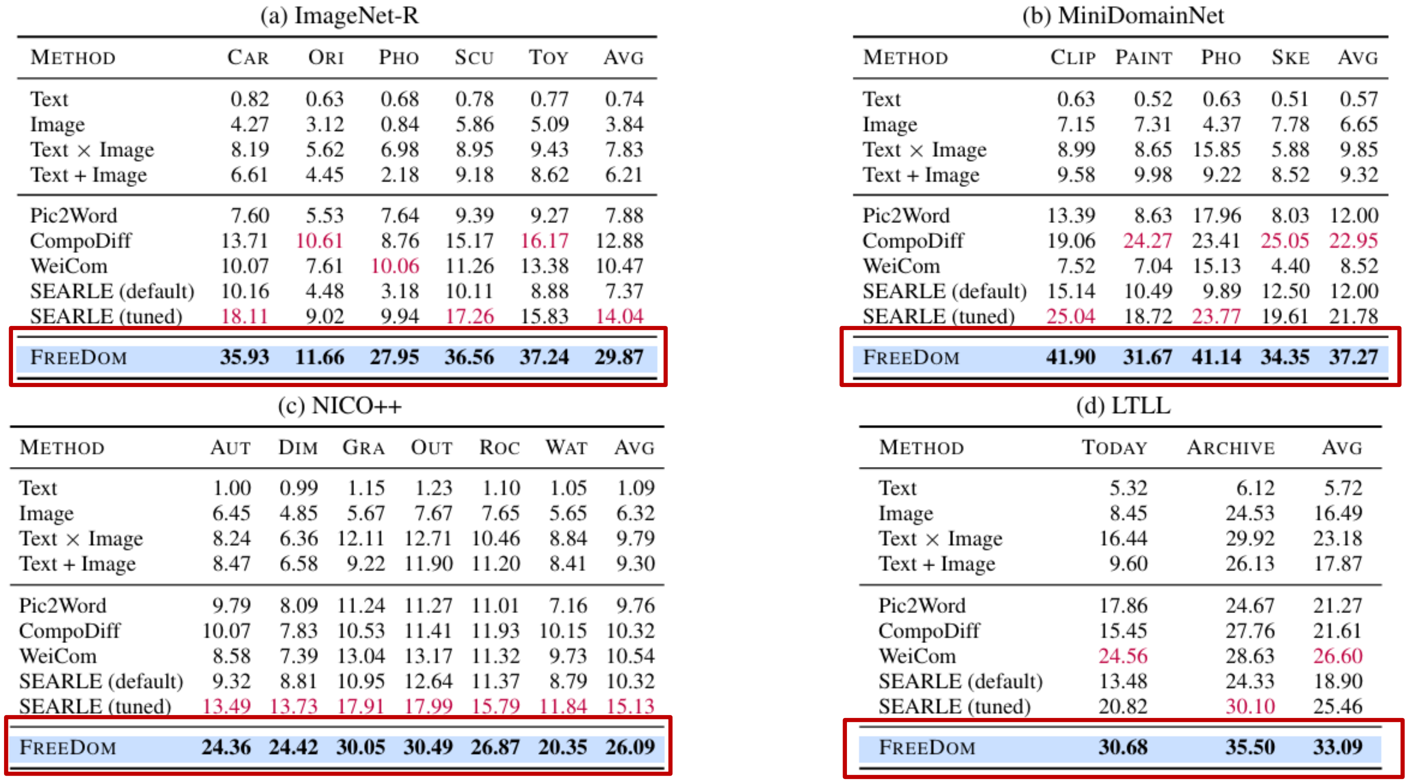
METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FREEDOM	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FREEDOM	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Quantitative Evaluation



(a) ImageNet-R

Method	Car	Ori	Pho	Scu	Toy	Avg
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FreeDom	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

Method	Clip	Paint	Pho	Ske	Avg
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FreeDom	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

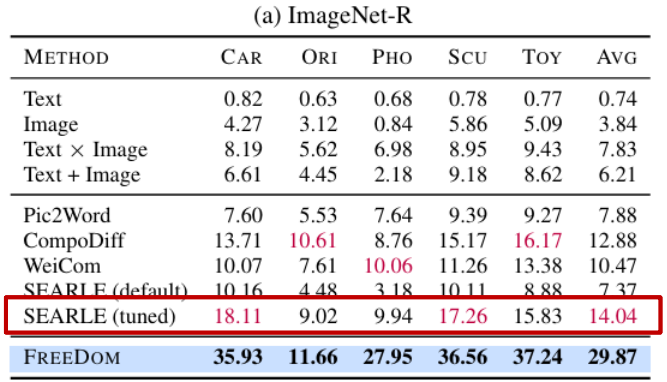
Method	Aut	Dim	Gra	Out	Roc	Wat	Avg
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FreeDom	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

(d) LTLL

Method	Today	Archive	Avg
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FreeDom	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

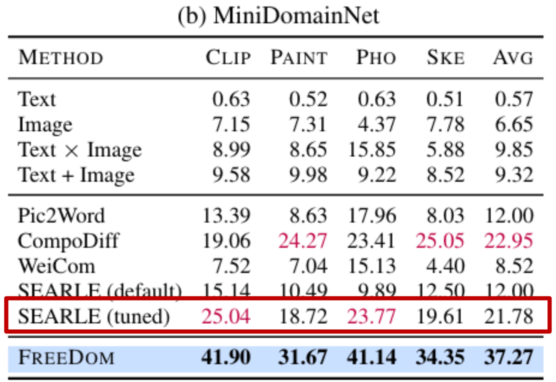
Quantitative Evaluation



(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FREEDOM	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

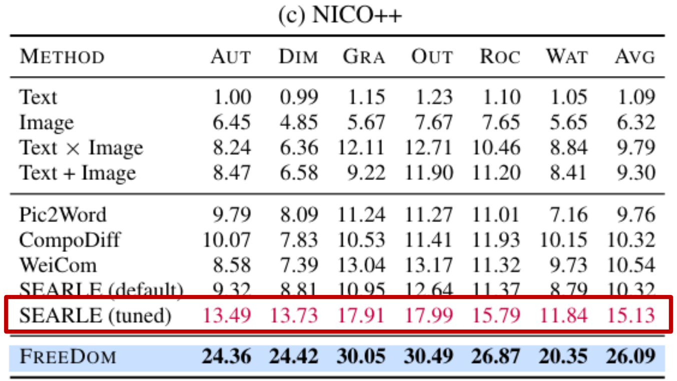
+15.9%



(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FREEDOM	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

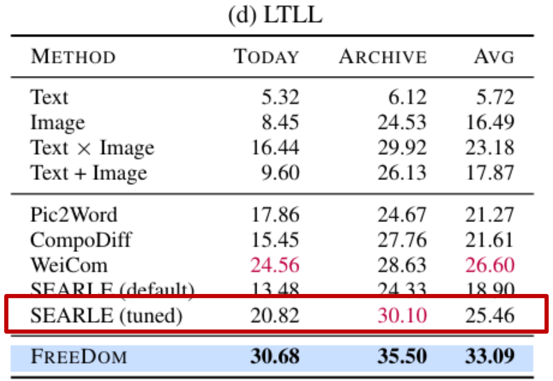
+15.5%



(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FREEDOM	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

+11.0%



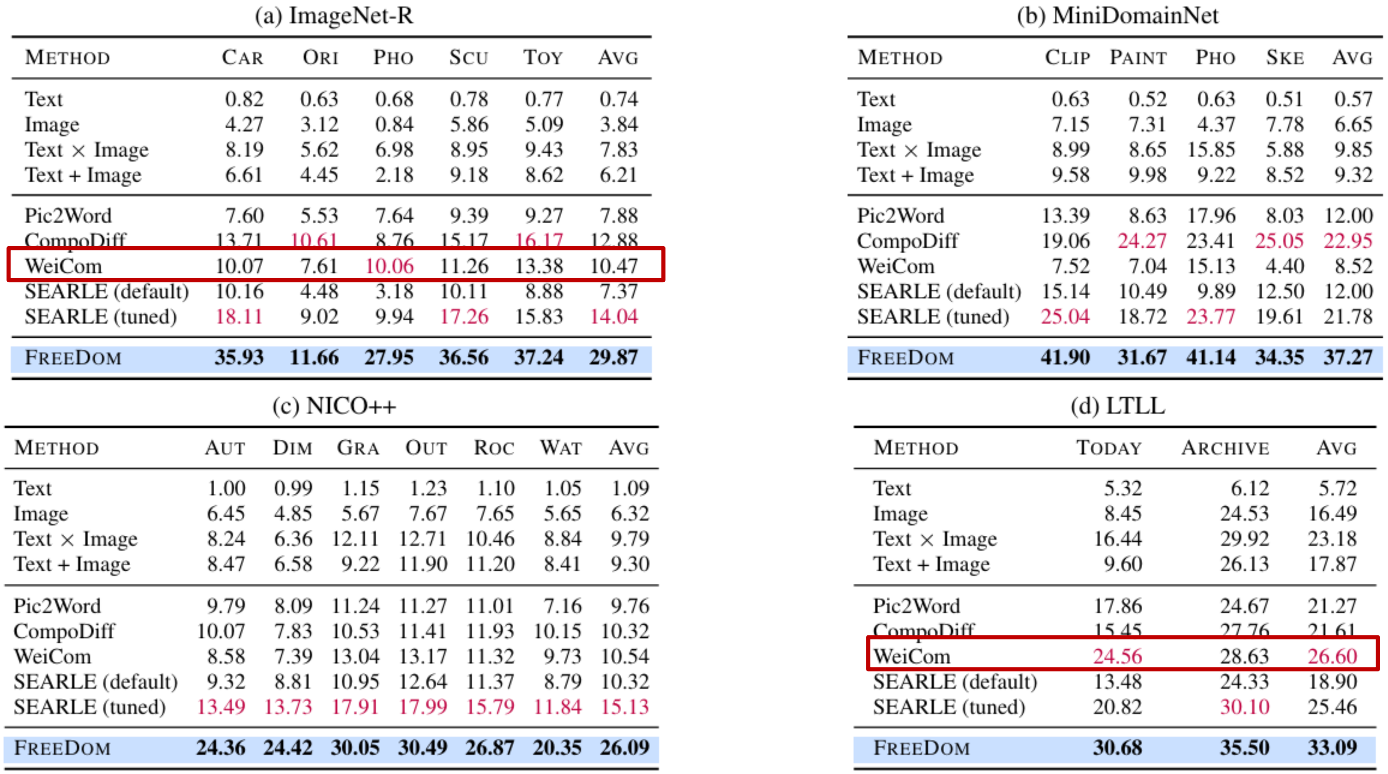
(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FREEDOM	**30.68**	**35.50**	**33.09**

+7.6%

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Quantitative Evaluation



(a) ImageNet-R

Method	Car	Ori	Pho	Scu	Toy	Avg
Text	0.82	0.63	0.68	0.78	0.77	0.74
Image	4.27	3.12	0.84	5.86	5.09	3.84
Text × Image	8.19	5.62	6.98	8.95	9.43	7.83
Text + Image	6.61	4.45	2.18	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE (default)	10.16	4.48	3.18	10.11	8.88	7.37
SEARLE (tuned)	18.11	9.02	9.94	17.26	15.83	14.04
FreeDom	**35.93**	**11.66**	**27.95**	**36.56**	**37.24**	**29.87**

(b) MiniDomainNet

Method	Clip	Paint	Pho	Ske	Avg
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.37	7.78	6.65
Text × Image	8.99	8.65	15.85	5.88	9.85
Text + Image	9.58	9.98	9.22	8.52	9.32
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE (default)	15.14	10.49	9.89	12.50	12.00
SEARLE (tuned)	25.04	18.72	23.77	19.61	21.78
FreeDom	**41.90**	**31.67**	**41.14**	**34.35**	**37.27**

(c) NICO++

Method	Aut	Dim	Gra	Out	Roc	Wat	Avg
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.47	6.58	9.22	11.90	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE (default)	9.32	8.81	10.95	12.64	11.37	8.79	10.32
SEARLE (tuned)	13.49	13.73	17.91	17.99	15.79	11.84	15.13
FreeDom	**24.36**	**24.42**	**30.05**	**30.49**	**26.87**	**20.35**	**26.09**

(d) LTLL

Method	Today	Archive	Avg
Text	5.32	6.12	5.72
Image	8.45	24.53	16.49
Text × Image	16.44	29.92	23.18
Text + Image	9.60	26.13	17.87
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE (default)	13.48	24.33	18.90
SEARLE (tuned)	20.82	30.10	25.46
FreeDom	**30.68**	**35.50**	**33.09**

Domain Conversion mAP (%) on four datasets; comparison of FreeDom with baselines and competitors.

Summarizing Insights

FreeDom:

- ✓ Training-free composed image retrieval method for domain conversion, based on a pre-trained and frozen CLIP
- ✓ Key component: discrete-space memory-based textual inversion
- ✓ Outperforms SOTA methods by a large margin on the task
- ✓ Robust to the choice of hyper-parameters

- ✓ Introduced new benchmarks

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

4. Conclusion

Conclusion

- ✓ Metrix achieved state-of-the-art results across multiple benchmarks in metric learning.
- ✓ SimPool improved performance across convolutional and transformer encoders on various benchmarks under different settings.
- ✓ SimPool provided high-quality attention maps, presenting strong localization properties.
- ✓ WeiCom proved to be an effective, efficient and flexible method for RSCIR.
- ✓ FreeDom outperformed state-of-the-art methods in domain conversion.

5. Future Work

Future Work

Visual Representations:

- ✓ Develop a mixup method for metric learning that leverages attention mechanisms to identify and align semantic correspondences between images before interpolation.
- ✓ Evaluate the effectiveness of Metrix on remote sensing data.
- ✓ Implement SimPool in pre-trained and frozen encoders to avoid costly training, positioning it as an intermediate step between linear probing and full fine-tuning in self-supervised learning scenarios and beyond.
- ✓ Apply SimPool iteratively, across different layers of the network, or modify SimPool to generate multiple local representations instead of a single global one.

Future Work

Multimodal Representations:

- ✓ Investigate the use of Remote Sensing Composed Image Retrieval for change detection. This would allow users to query a scene and a specific type of change.
- ✓ Create a new method based on VLMs and train it directly on PatternCom or an expanded version, encompassing a broader range of remote sensing scenes and attributes.
- ✓ Explore transitioning from text-to-image to image-to-image search in composed image retrieval, potentially leveraging synthetic image generation via Stable Diffusion or by combining features from both the image and text queries.
- ✓ Integrate gradient-based discrete optimization techniques into FreeDom to generate more contextually relevant text prompts, improving precision and robustness in complex or nuanced cases.

6. Publications

Publications

- S. Venkataramanan*, **B. Psomas***, E. Kijak, L. Amsaleg, K. Karantzalos, Y. Avrithis, «It Takes Two to Tango: Mixup for Deep Metric Learning», in International Conference on Learning Representations (**ICLR**), 2022

- **B. Psomas**, I. Kakogeorgiou, K. Karantzalos, Y. Avrithis, «Keep It Simpool: Who Said Supervised Transformers Suffer from Attention Deficit?», in International Conference on Computer Vision (**ICCV**), 2023

- **B. Psomas**, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y.Avrithis, K. Karantzalos, «Composed Image Retrieval for Remote Sensing», in IEEE International Geoscience and Remote Sensing Symposium (**IGARSS**), 2024

- N. Efthymiadis, **B. Psomas**, Z. Laskar, K. Karantzalos, Y. Avrithis, O. Chum, G. Tolias, «Composed Image Retrieval for Training-Free Domain Conversion», **under review** in Winter Conference on Applications of Computer Vision (**WACV**), 2024

More publications

- I. Kakogeorgiou, S. Gidaris, **B. Psomas**, Y. Avrithis, A. Bursuc, K. Karantzalos, N. Komodakis, «What to Hide from Your Students: Attention-Guided Masked Image Modeling», in European Conference on Computer Vision (**ECCV**), 2022

- P. Riccio, **B. Psomas**, F. Galati, F. Escolano, T. Hofmann, N. Oliver, «Openfilter: A Framework to Democratize Research Access to Social Media AR Filters», Advances in Neural Information Processing Systems (**NeurIPS**), 2022

- S. Vellas, **B. Psomas**, K. Karadima, D. Danopoulos, A. Paterakis, G. Lentaris, D. Soudris, K. Karantzalos, «Evaluation of Resource-Efficient Crater Detectors on Embedded Systems», in IEEE International Geoscience and Remote Sensing Symposium (**IGARSS**), 2024

- M. Sdraka, I. Papoutsis, **B. Psomas**, K. Vlachos, K. Ioannidis, K. Karantzalos, I. Gialampoukidis, S. Vrochidis, «Deep Learning for Downscaling Remote Sensing Images: Fusion and Super-Resolution», IEEE Geoscience and Remote Sensing Magazine (**GRSM**), 2022

Open source code

- **Metrix**

  https://github.com/billpsomas/metrix

- **SimPool**

  https://github.com/billpsomas/simpool

- **WeiCom**

  https://github.com/billpsomas/rscir

Acknowledgements

Konstantinos
Karantzalos

Yannis
Avrithis

Giorgos
Tolias

Demetre
Argialas

Ioannis
Kakogeorgiou

Shashank
Venkataramanan

Acknowledgements

Nikos
Efthymiadis

Spyros
Gidaris

Andrei
Bursuc

Ondrej
Chum

Zakaria
Laskar

Nikos
Komodakis

Ioannis
Papoutsis

Piera
Riccio

Nuria
Oliver

Francesco
Galati

Simon Vellas

George
Lentaris

Milly Vasileiou

Maria Sdraka

Dimitris Danopoulos

Kalliopi Karadima

George Ouzounidis

Christos Anagnostopoulos

Pol Kolokousis

Antonia Kournopoulou

Eleni Sofikiti

Jason Manesis

Ioannis Tsiotas

Sotiris Spanos

Vassilis Andronis

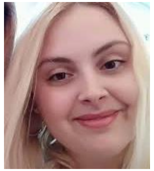
Kleanthis Karamvasis

Maria Adepli

Zacharias Kandylakis

Christina Karakizi

Bill Tsironis



Athena Psalta

Evi Mikeli

Konstantinos Tertikas

Olyna Gounari

Alekos Falagas

Katerina Kikaki

Katerina Adam

Makis Douskos

Eirini Baltzi

Dionysis Christopoulos

Thanks for your attention!

Attention map of SimPool