Paper Review

Active Token Mixer

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Active Token Mixer

Guoqiang Wei[1]*, Zhizheng Zhang[2]*, Cuiling Lan[2], Yan Lu[2], Zhibo Chen[1]

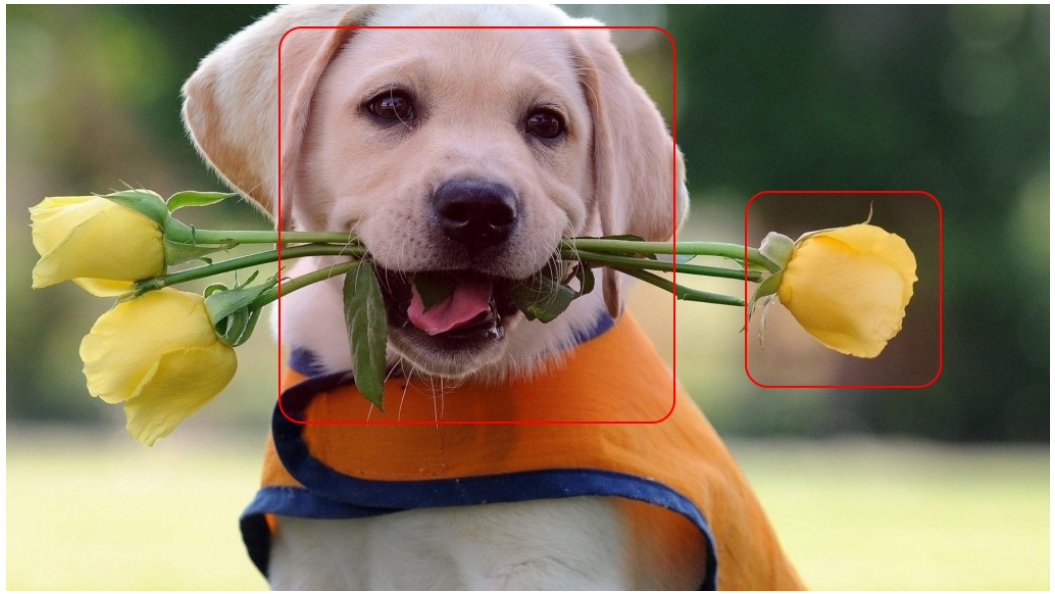
[1]University of Science and Technology of China
[2]Microsoft Research Asia
wgq7441@mail.ustc.edu.cn,{zhizzhang, culan, yanlu}@microsoft.com, chenzhibo@ustc.edu.cn

Warm Up

Receptive Field [5/5]

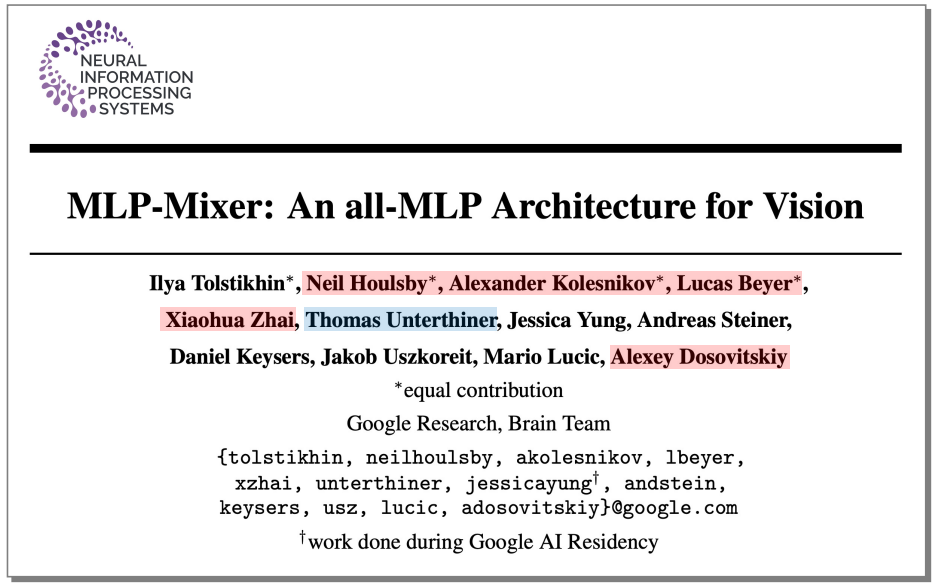
1025

577



https://yeonghyeon.medium.com/deep-learning-why-graph-neural-network-cd1b071c25ed

Related Work



NEURAL INFORMATION PROCESSING SYSTEMS

MLP-Mixer: An all-MLP Architecture for Vision

Ilya Tolstikhin*, Neil Houlsby*, Alexander Kolesnikov*, Lucas Beyer*,
Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner,
Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy

*equal contribution

Google Research, Brain Team

{tolstikhin, neilhoulsby, akolesnikov, lbeyer,
xzhai, unterthiner, jessicayung†, andstein,
keysers, usz, lucic, adosovitskiy}@google.com

†work done during Google AI Residency

- **Publication**
  - Vision Transformer (ViT)
    - arXiv: October 2020
    - ICLR: May 2021
    - 3383 citations
  - MLP-Mixer
    - arXiv: May 2021 (after 7 months of ViT)
    - NeurIPS: December 2021
    - 265 citations

- **Authors**
  - Half (6/12) of the authors are ViT authors
  - Five authors are the first author of ViT
  - One author is the second author of ViT

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *International Conference on Learning Representations. 2021*. 3383 citations
Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." *Advances in Neural Information Processing Systems.*. 2021. 265 citations

ATM

Active Token Mixer

Summaries

Motivation and solution

- **Motivation**: Capture useful information by feature fusion
  - Non-constraint feature fusion module in receptive field perspective
- **Solution**: Feature fusion by offset selection
  - Information will be aggregated for each spatial axis

Contributions

- Content adaptivity
  - Context selection/localization by offset operation (compared to deformation method)
- Flexibility
  - Dynamic selection of context tokens
- Efficiency
  - Linear complexity according to the input resolution
  - No constraint to the receptive fields

Paper Strengths

- Sufficient explanation of literature studies (CNNs, Transformers, and MLPs)
- Eliminates the effort to extract feature maps containing rich information
  - Some layers of neural network
  - Specific channels from each layer

Weaknesses

- Marginal performance gap compared to recent similar scale models
- Some details of ATM are skipped in the manuscript
  - Only can be seen in the supplementary material of arXiv version

Overview

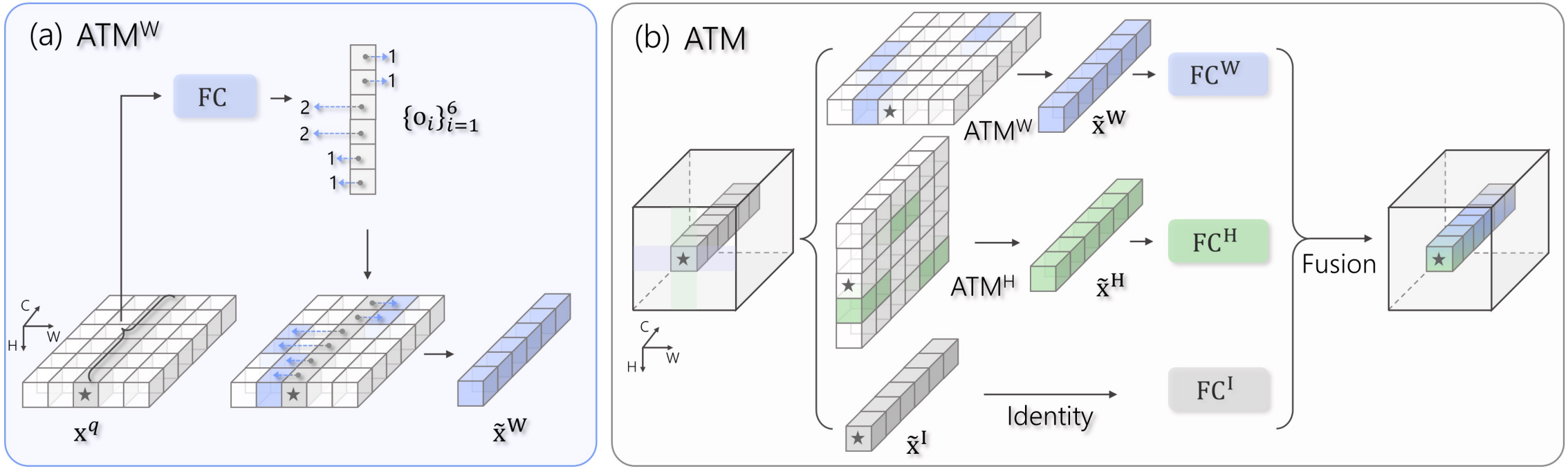


Figure 1: Illustration of our proposed Active Token Mixer (ATM). (a) ATM along the horizontal (width) dimension. For a query $\mathbf{x}^q$, ATM actively captures the useful contexts by recomposing the elements from selected tokens into $\tilde{\mathbf{x}}^W \in \mathbb{R}^C$ based on the learned channel-wise offsets. (b) ATM module consisting of ATM^W along horizontal dimension, ATM^H along vertical dimension, and the identity branch ATM^I. The two recomposed tokens ($\tilde{\mathbf{x}}^W$, $\tilde{\mathbf{x}}^H$) and the original $\tilde{\mathbf{x}}^I$ are then adaptively fused after being embedded by $FC^{\{W,H,I\}}$.

Overview

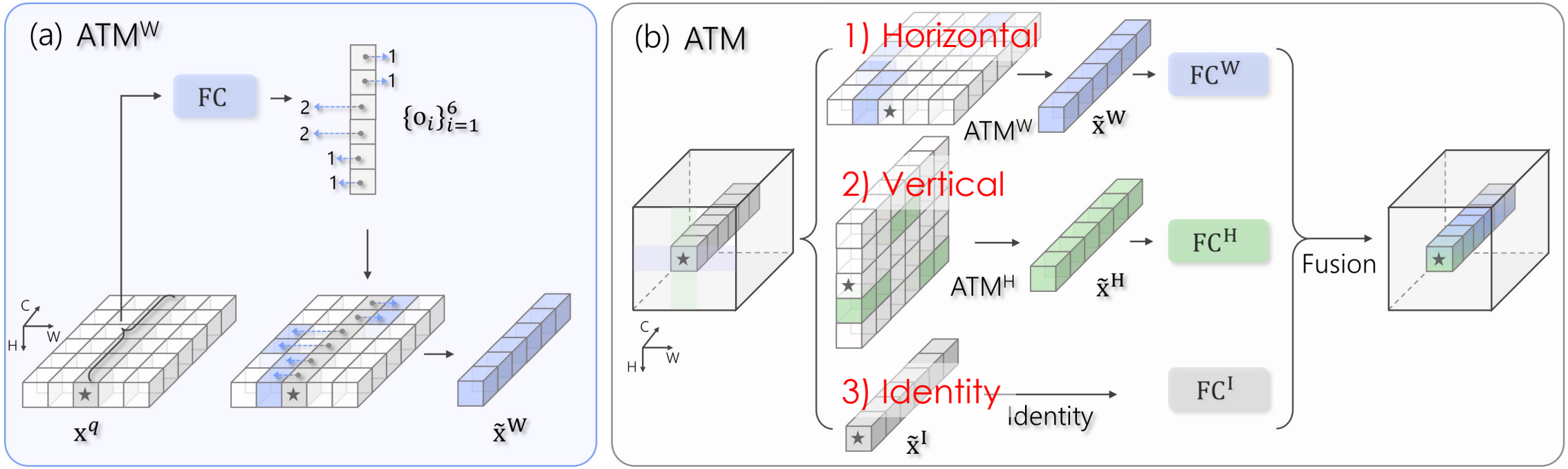
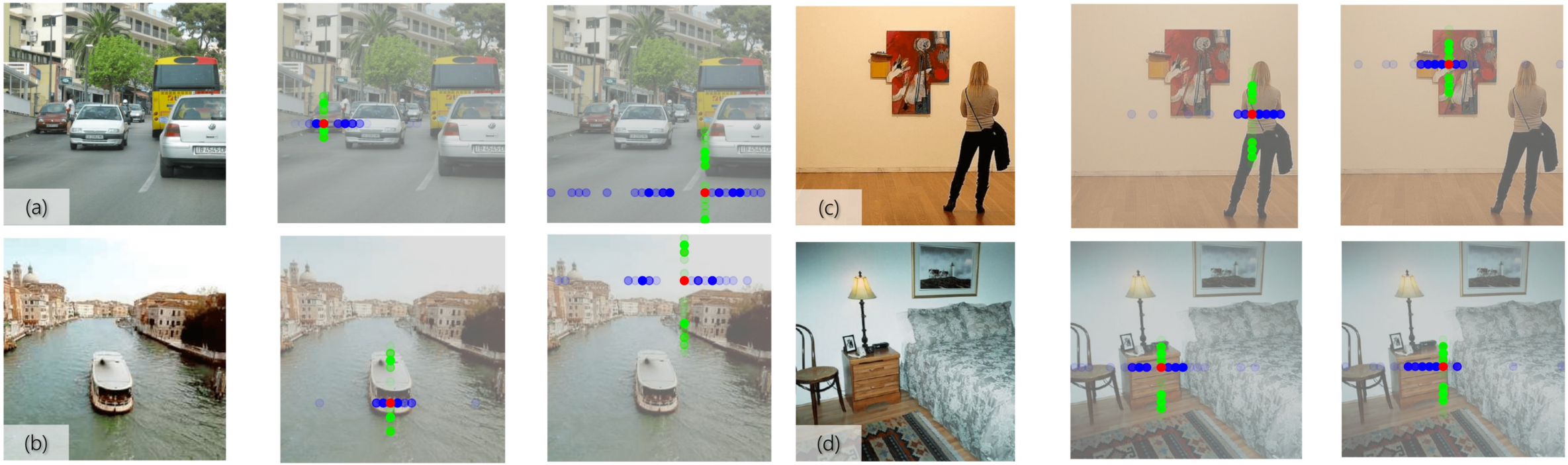


Figure 1: Illustration of our proposed Active Token Mixer (ATM). (a) ATM along the horizontal (width) dimension. For a query $\mathbf{x}^q$, ATM actively captures the useful contexts by recomposing the elements from selected tokens into $\tilde{\mathbf{x}}^W \in \mathbb{R}^C$ based on the learned channel-wise offsets. (b) ATM module consisting of ATM^W along horizontal dimension, ATM^H along vertical dimension, and the identity branch ATM^I. The two recomposed tokens ($\tilde{\mathbf{x}}^W$, $\tilde{\mathbf{x}}^H$) and the original $\tilde{\mathbf{x}}^I$ are then adaptively fused after being embedded by $FC^{\{W,H,I\}}$.

Overview



Semantic Segmentation

Figure 5: Illustration of the horizontal (●) and vertical (●) offsets for the given query token (●) on ADE20K (input size: 512×512). The transparency of each circle corresponds to how many times the token at this position is sampled, *i.e.*, the more transparent the circle is, the corresponding offset values appear less in $\mathcal{O} = \{o_i\}_{i=1}^{C}$. The visualized offsets are from a randomly sampled layer `layer_3_18`. Similar phenomenons can be observed at other layers.

Deformable Convolution

Figure 6: Each image triplet shows the sampling locations ($9^3 = 729$ red points in each image) in three levels of 3×3 deformable filters (see Figure 5 as a reference) for three activation units (green points) on the background (left), a small object (middle), and a large object (right), respectively.

Kernel offsets are shared over all channels
without considering spatial semantics!

Key Observations

Spatial
- Visual objects present diverse shapes and deformations

Channel
- Multiple semantic attributes are distributed in different channels

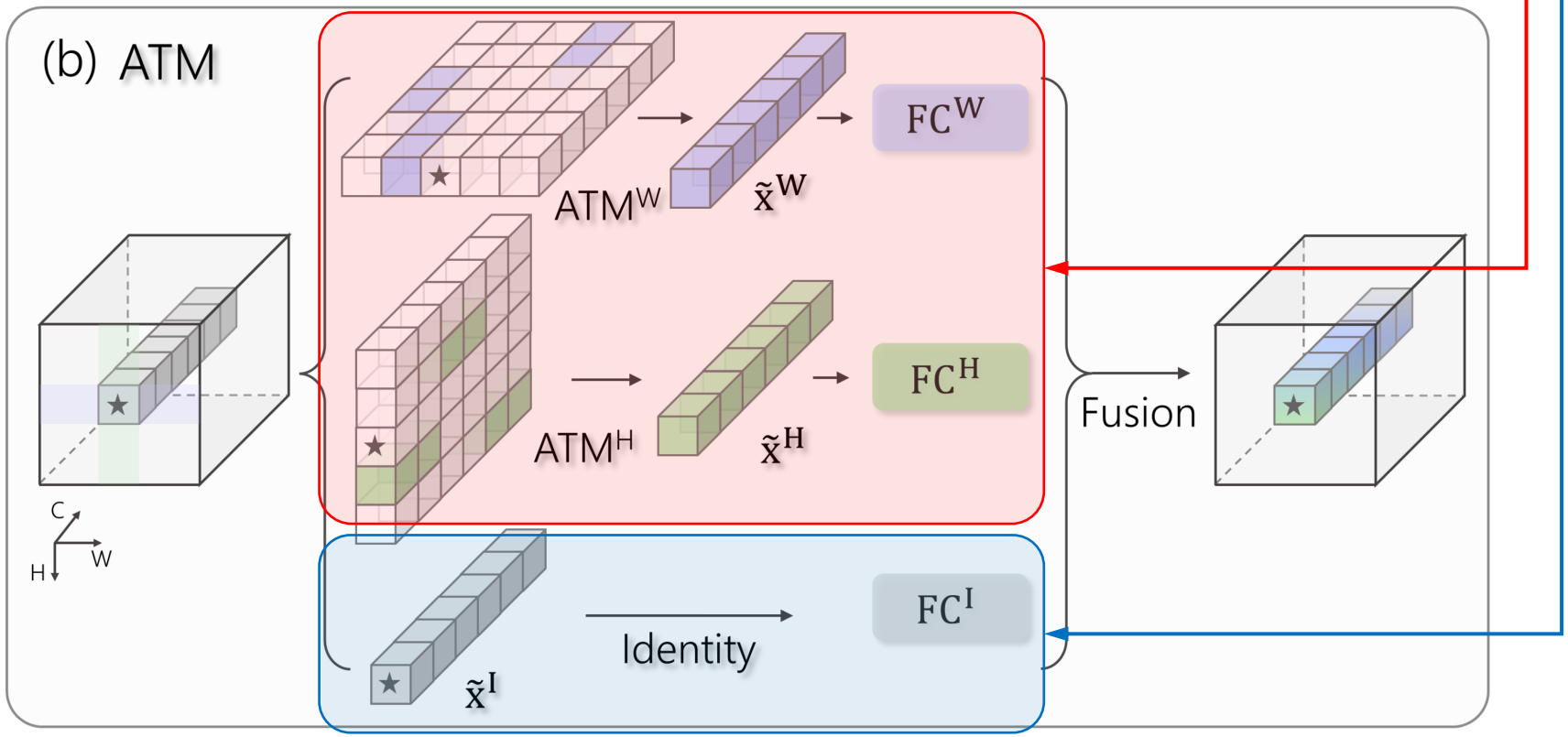
Key Observations

Spatial

- Visual objects present diverse shapes and deformations

Channel

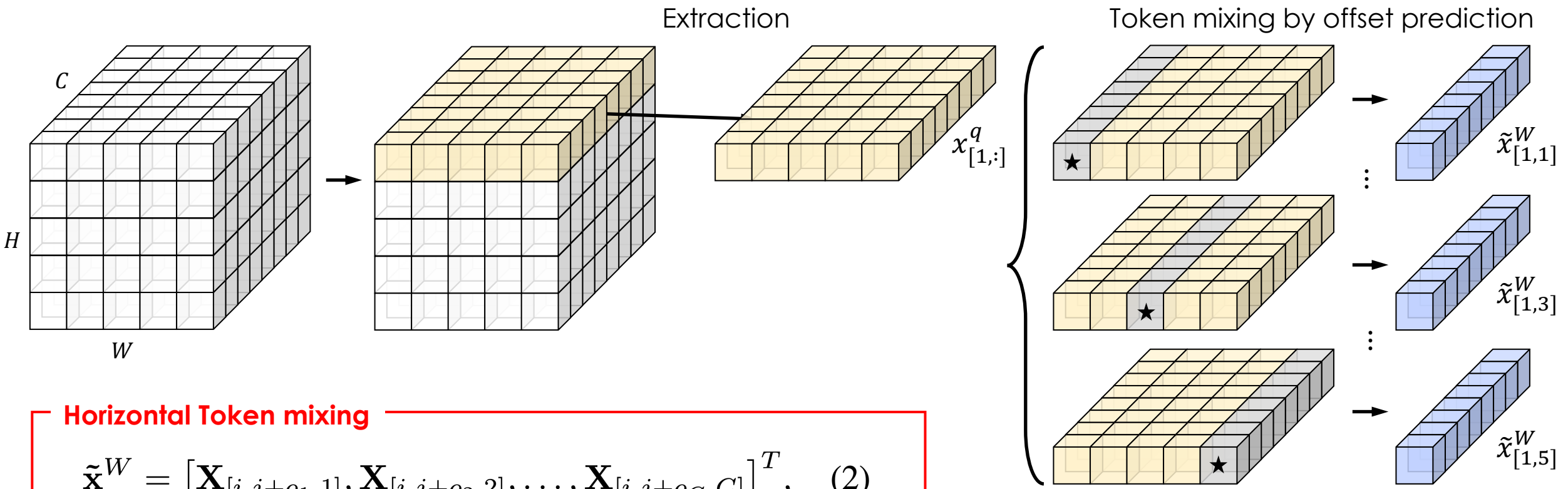
- Multiple semantic attributes are distributed in different channels



Active Token Mixer

Along the Horizontal Dimension

★ Query for token mixing

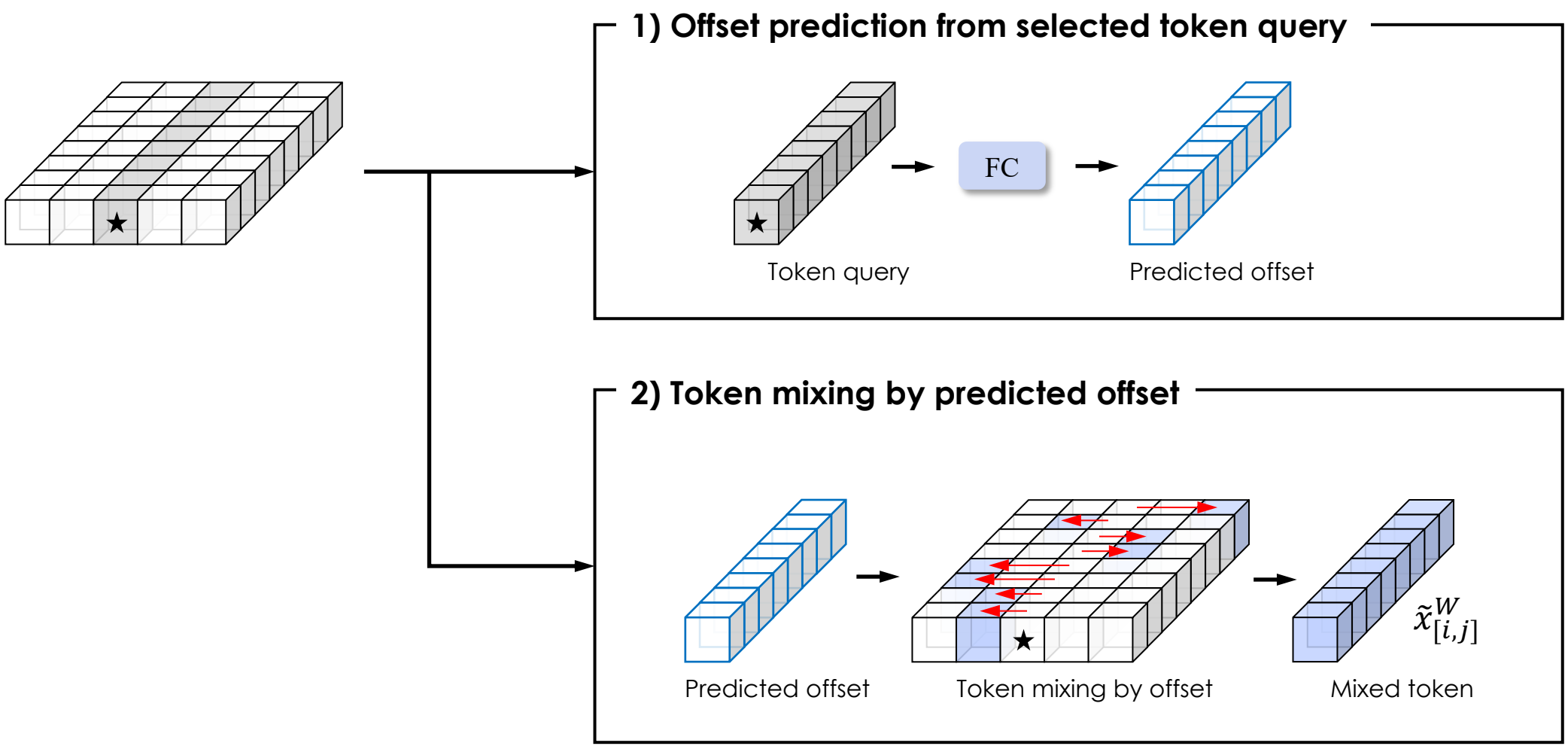


Horizontal Token mixing

$$\tilde{\mathbf{x}}^W = \left[\mathbf{X}_{[i,j+o_1,1]}, \mathbf{X}_{[i,j+o_2,2]}, \ldots, \mathbf{X}_{[i,j+o_C,C]}\right]^T, \quad (2)$$

- i: index of vertical-axis (fixed value for each horizontal mixing)
- j: index of horizontal-axis (linear searching for mixing)
- o_c: channel-wise offset for token mixing (predicted by FC layer)

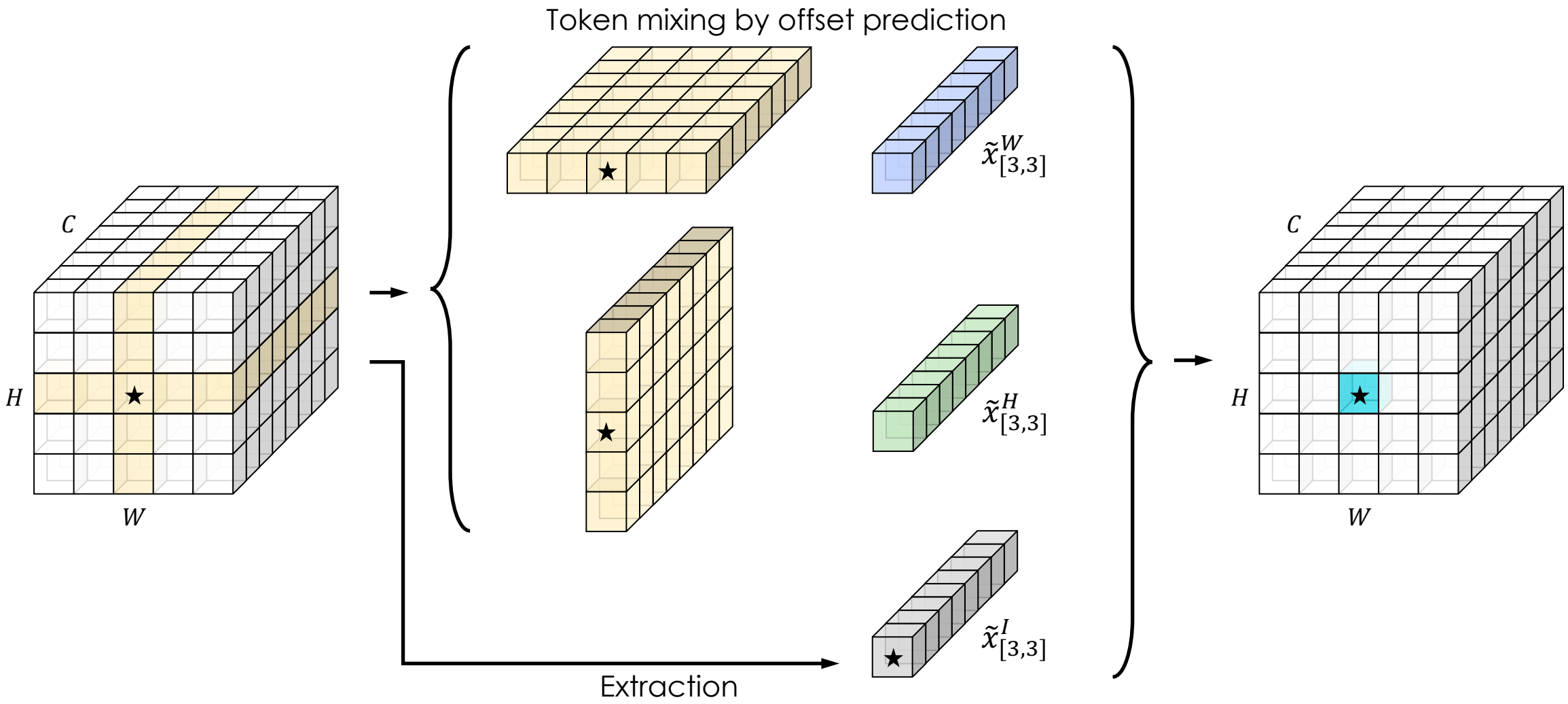
Token Mixing by Offset Prediction



Active Token Mixer

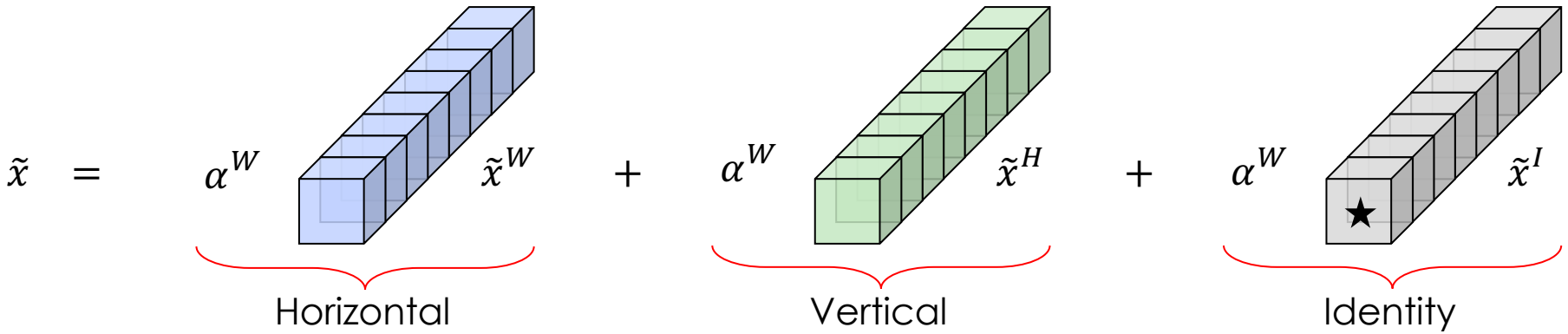
Feature Fusion

★ Query for token mixing



Feature Fusion

α: learnable parameter



$$\hat{\mathbf{x}} = \boldsymbol{\alpha}^W \odot \hat{\mathbf{x}}^W + \boldsymbol{\alpha}^H \odot \hat{\mathbf{x}}^H + \boldsymbol{\alpha}^I \odot \hat{\mathbf{x}}^I, \qquad (3)$$

$$[\boldsymbol{\alpha}^W, \boldsymbol{\alpha}^H, \boldsymbol{\alpha}^I] = \sigma([W^W \cdot \hat{\mathbf{x}}^\Sigma, W^H \cdot \hat{\mathbf{x}}^\Sigma, W^I \cdot \hat{\mathbf{x}}^\Sigma]), \quad (4)$$

summation $\hat{\mathbf{x}}^\Sigma$ of $\hat{\mathbf{x}}^{\{W,H,I\}}$

ATM Networks

$$\hat{\mathbf{X}}^l = ATM^l(LN(\mathbf{X}^{l-1})) + \mathbf{X}^{l-1}, \quad (5)$$

$$\mathbf{X}^l = MLP^l(LN(\hat{\mathbf{X}}^l)) + \hat{\mathbf{X}}^l, \quad (6)$$

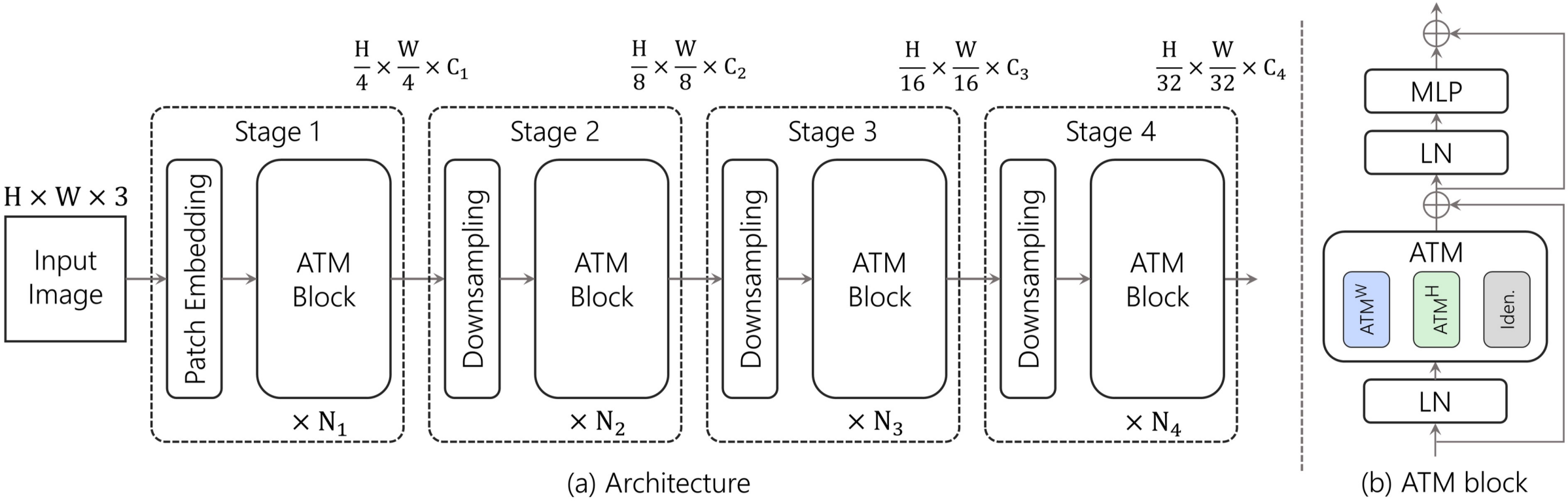


Figure 4: a) The overall architecture of ATMNet. b) The ATM block.

Experiments

ImageNet-1k: https://www.image-net.org/download.php
ADE20k: https://groups.csail.mit.edu/vision/datasets/ADE20K/
COCO: https://cocodataset.org/#home

List of Experiments

- **Classification**
  - ImageNet-1k dataset without extra data
  - 224×224 and 384×384 resolutions

- **Semantic Segmentation**
  - ADE20k dataset
  - Semantic FPN and UperNet frameworks

- **Object Detection**
  - COCO dataset
  - Mask R-CNN, RetinaNet, and cascade Mask R-CNN frameworks

Performance
Classification

Model	Size	#P.(M)	FLOPs(G)	Top-1(%)
ResNet18	224^2	12	1.8	69.8
ResMLP-S12	224^2	15	3.0	76.6
CycleMLP-B1	224^2	15	2.1	78.9
ATMNet-xT	224^2	15	2.2	**79.7**
ResNet50	224^2	26	4.1	78.5
Deit-S	224^2	22	4.6	79.8
PVT-S	224^2	25	3.8	79.8
Swin-T	224^2	29	4.6	81.2
TwinsP-S	224^2	24	3.8	81.2
Twins-S	224^2	24	2.9	81.7
ResMLP-S24	224^2	30	6.0	79.4
ASMLP-T	224^2	28	4.4	81.3
ViP-S	224^2	25	6.9	81.5
MorphMLP-T	224^2	23	3.9	81.6
CycleMLP-B2	224^2	27	3.9	81.6
Shift-T	224^2	29	4.5	81.7
ATMNet-T	224^2	27	4.0	**82.0**
PVT-M	224^2	44	6.7	81.2
TwinsP-B	224^2	44	6.7	82.7
MorphMLP-S	224^2	38	7.0	82.6
CycleMLP-B3	224^2	38	6.9	82.6
ATMNet-S	224^2	39	6.9	**83.1**

Model	Size	#P.(M)	FLOPs(G)	Top-1(%)
PVT-L	224^2	61	9.8	81.7
Swin-S	224^2	50	8.7	83.2
Twins-B	224^2	56	8.6	83.2
ViP-M	224^2	55	16.3	82.7
Shift-S	224^2	50	8.8	82.8
CycleMLP-B4	224^2	52	10.1	83.0
ASMLP-S	224^2	50	8.5	83.1
MorphMLP-B	224^2	58	10.2	83.2
ATMNet-B	224^2	52	10.1	**83.5**
Deit-B	224^2	86	17.5	81.8
Swin-B	224^2	88	15.4	83.5
S^2MLP-W	224^2	71	14.0	80.0
CycleMLP-B5	224^2	76	15.3	83.1
ViP-L	224^2	88	24.4	83.2
Shift-B	224^2	89	15.6	83.3
ASMLP-B	224^2	88	15.2	83.3
MorphMLP-L	224^2	76	12.5	83.4
ATMNet-L	224^2	76	12.3	**83.8**
ViT-B/16↑	384^2	86	55.4	77.9
Deit-B↑	384^2	86	55.4	83.1
Swin-B↑	384^2	88	47.1	84.5
ATMNet-L↑	384^2	76	36.4	**84.8**

Table 1: Comparisons with state-of-the-art models on ImageNet-1K without extra data. All models are trained with input size of 224×224, except ↑ with 384×384.

[1] Tete Xiao, et al. "Unified perceptual parsing for scene understanding." ECCV. 2018.
[2] Alexander Kirillo, et al. "Panoptic feature pyramid networks." CVPR. 2019.

Performance
Semantic Segmentation

UperNet [1]			
Model	#P	FLOPs	mIoU/mIoU(ms)
Swin-T	60	945	44.5 / 45.8
Twins-S	54	931	46.2 / 47.1
ConvNeXt-T	60	939	- / 46.7
ASMLP-T	60	937	- / 46.5
CycleMLP-T	60	937	- / 47.1
ATMNet-T	57	927	**46.5 / 47.6**
Swin-B	121	1188	48.1 / 49.7
Twins-L	133	1236	48.8 / 50.2
ConvNeXt-T	122	1170	- / 49.9
ASMLP-B	121	1166	- / 49.5
CycleMLP-B	121	1166	- / 49.7
ATMNet-S	69	988	48.4 / 49.5
ATMNet-L	108	1106	**50.1 / 51.1**

Light model

Semantic FPN [2]			
Model	#P	FLOPs	mIoU
Swin-T	31.9	48	41.5
Twins-S	28.3	37	43.2
TwinsP-S	28.4	40	44.3
MorphMLP-T	26.4	-	43.0
CycleMLP-B2	30.6	42	43.4
Wave-MLP-S	31.2	-	44.4
ATMNet-T	30.9	42.4	**45.8**
Swin-B	91.2	107	46.0
TwinsP-L	65.3	71	46.4
Twins-L	103.7	102	46.7
CycleMLP-B5	79.4	86	45.5
MorphMLP-B	59.3	-	45.9
ATMNet-L	79.8	86.6	**48.1**

Table 2: Semantic segmentation results on ADE20K `val` with UperNet and Semantic FPN. FLOPs are evaluated on 512×2048 for UperNet and 512×512 for Semantic FPN. All backbones are pretrained on ImageNet-1K. mIoU(ms): mIoU with multi-scale inference. The results of other variants are in the Supplementary.

[3] Kaiming He, et al. "Mask R-CNN." ICCV. 2017.
[4] Tsung-Yi Lin, et al. "Focal loss for dense object detection." ICCV. 2017.

Performance
Object Detection

Backbone	#Params. (M)	FLOPs (G)	Mask R-CNN 1× [3] AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	RetinaNet 1× [4] AP^b	AP^b_{50}	AP^b_{75}	~~Mask R-CNN 3× MS~~ AP^m	AP^m_{50}	AP^m_{75}
ResNet-50	44	260	38.0	58.6	41.4	34.4	55.1	36.7	41.0	61.7	44.9	37.1	58.4	40.1
Swin-T	48	264	42.2	64.6	46.2	39.1	61.6	42.0	46.0	68.2	50.2	41.6	65.1	44.8
ConvNext-T	48	262	-	-	-	-	-	-	46.2	67.9	50.8	41.7	65.0	44.9
ASMLP-T	48	260	-	-	-	-	-	-	46.0	67.5	50.7	41.5	64.6	44.5
CycleMLP-B2	47	250	42.1	64.0	45.7	38.9	61.2	41.8	-	-	-	-	-	-
WaveMLP-S	47	250	44.0	65.8	48.2	40.0	63.1	42.9	-	-	-	-	-	-
ATMNet-xT	35	215	42.8	64.9	46.9	39.5	62.1	42.5	45.0	67.4	49.5	41.1	64.4	44.2
ATMNet-T	47	251	**44.8**	**66.9**	**49.0**	**41.0**	**64.2**	**44.3**	**47.1**	**69.0**	**51.7**	**42.7**	**66.5**	**46.0**
X101-64	102	493	42.8	63.8	47.3	38.4	60.6	41.3	44.4	64.9	48.8	39.7	61.9	42.6
Twins-L	120	474	45.9	-	-	41.6	-	-	-	-	-	-	-	-
Swin-B	107	496	45.5	-	-	41.3	-	-	48.5	69.8	53.2	43.4	66.8	46.9
MViT-B	73	438	-	-	-	-	-	-	48.8	71.2	53.5	44.2	68.4	47.6
CycleMLP-B5	95	421	44.1	65.5	48.4	40.1	62.8	43.0	-	-	-	-	-	-
WaveMLP-B	75	353	45.7	67.5	50.1	27.8	49.2	59.7	-	-	-	-	-	-
ATMNet-B	72	377	46.5	68.6	51.0	42.5	66.1	45.8	49.0	70.7	54.0	43.9	67.7	47.5
ATMNet-L	96	424	**47.4**	**69.9**	**52.0**	**43.2**	**67.3**	**46.5**	**49.5**	**71.5**	**54.3**	**44.5**	**68.7**	**48.1**

Table 3: Object detection results on COCO `val2017` with Mask R-CNN 1× and RetinaNet 1×. FLOPS are evaluated with resolution 800×1280. The complete comparison table and results of 3× can be found in the Supplementary.

Simple sustainable module

Novel design for novelty

Novel design for sustainability

Appendix A

The most compared model

CycleMLP

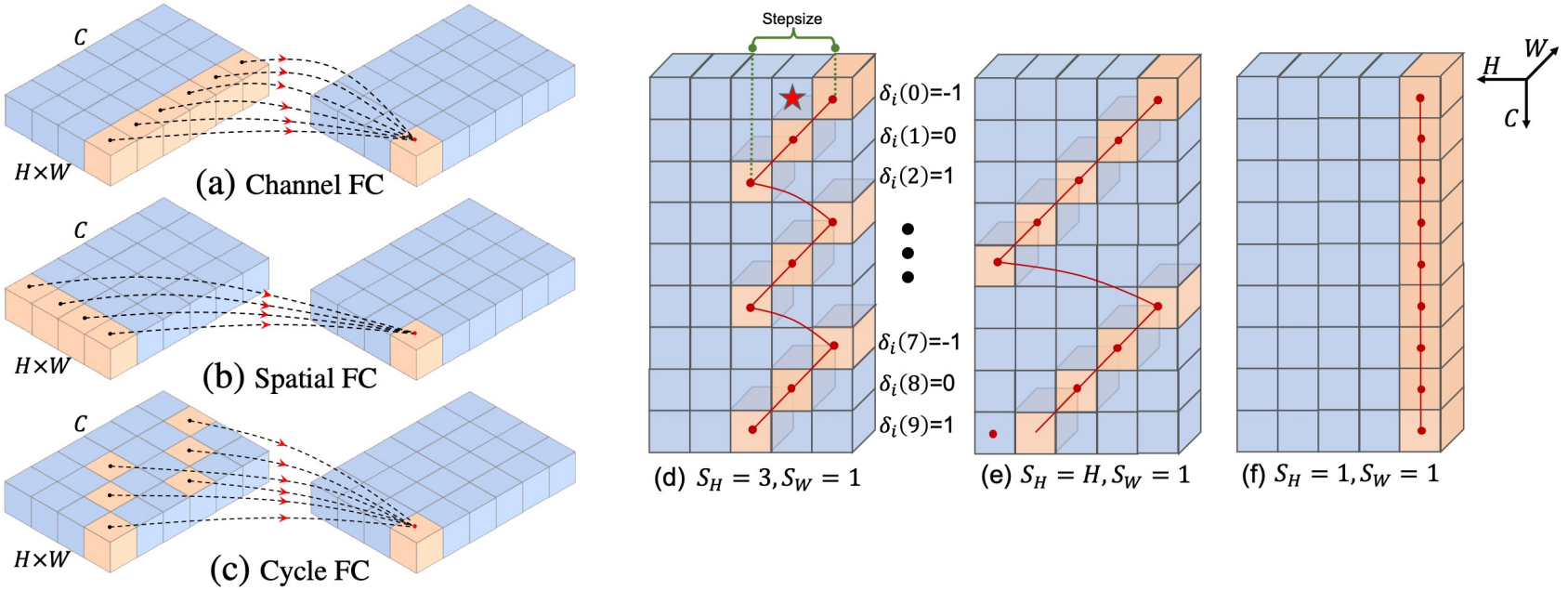


Figure 1: **(a)-(c): motivation of Cycle Fully-Connected Layer (Cycle FC)** compared to Channel FC and Spatial FC. **(a)** Channel FC aggregates features in the channel dimension with spatial size '1'. It can handle various input scales but cannot learn spatial context. **(b)** Spatial FC (Tolstikhin et al., 2021; Touvron et al., 2021a; Liu et al., 2021a) has a global receptive field in the spatial dimension. However, its parameter size is fixed and it has quadratic computational complexity to image scale. **(c)** Our proposed Cycle Fully-Connected Layer (Cycle FC) has linear complexity the same as channel FC and a larger receptive field than Channel FC. **(d)-(f): Three examples of different stepsizes.** Orange blocks denote the sampled positions. ★ denotes the output position. For simplicity, we omit batch dimension and set the feature's width to 1 here for example. Several more general cases can be found in Figure 7 (Appendix G). Best viewed in color.

Appendix B

Paper in a same way & purpose

VMamba

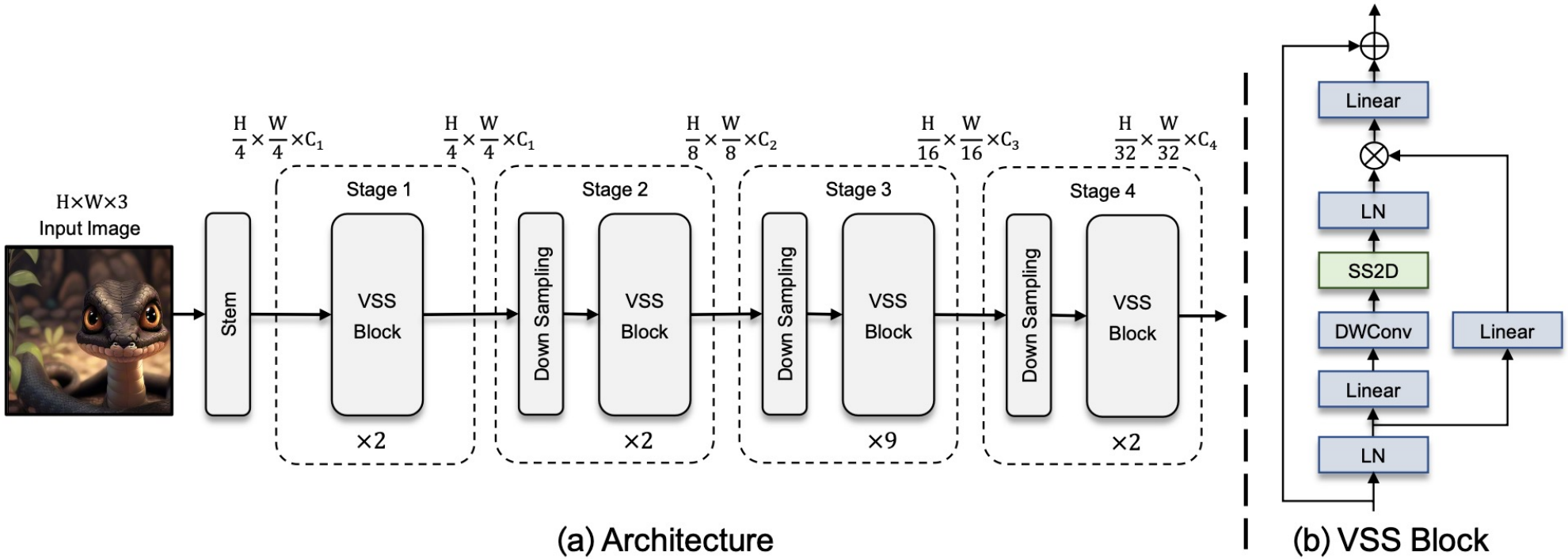


Figure 4: (a) The overall architecture of a VMamba model (VMamba-T); (b) The fundamental building block of VMamba, namely the VSS block.

VMamba

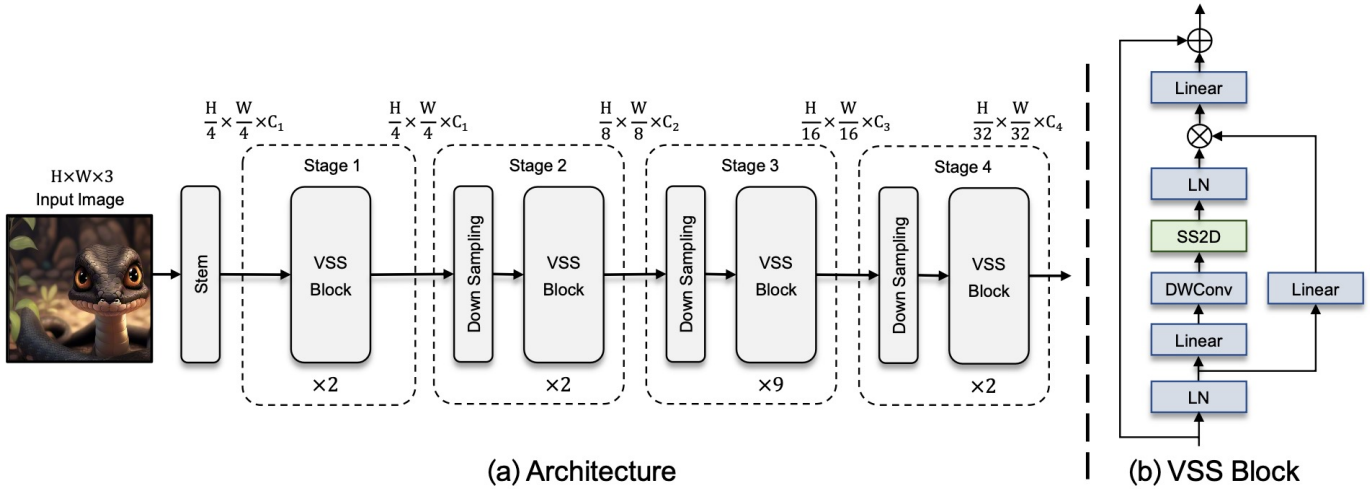


Figure 4: (a) The overall architecture of a VMamba model (VMamba-T); (b) The fundamental building block of VMamba, namely the VSS block.

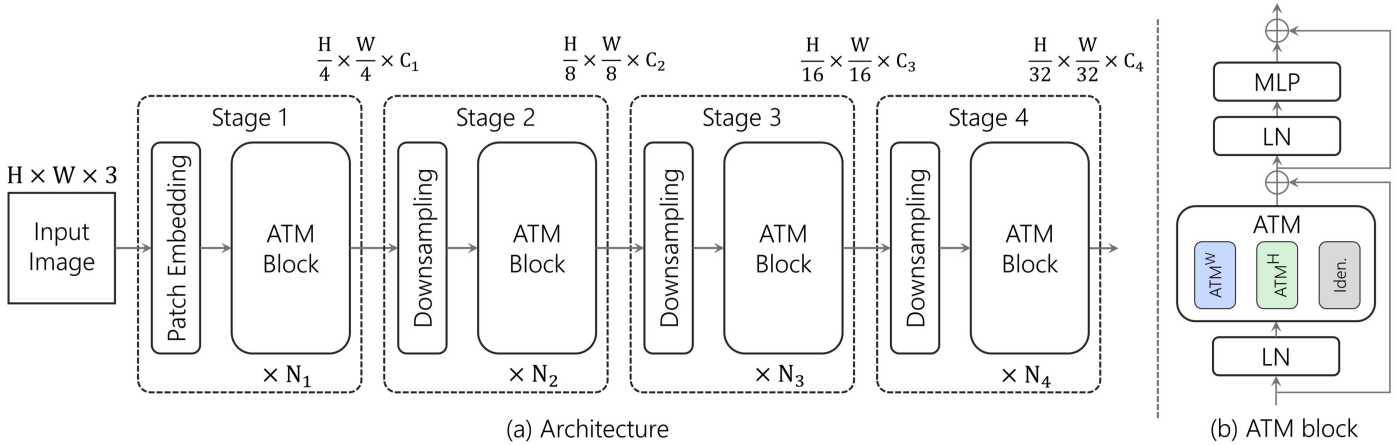


Figure 4: a) The overall architecture of ATMNet. b) The ATM block.

VMamba

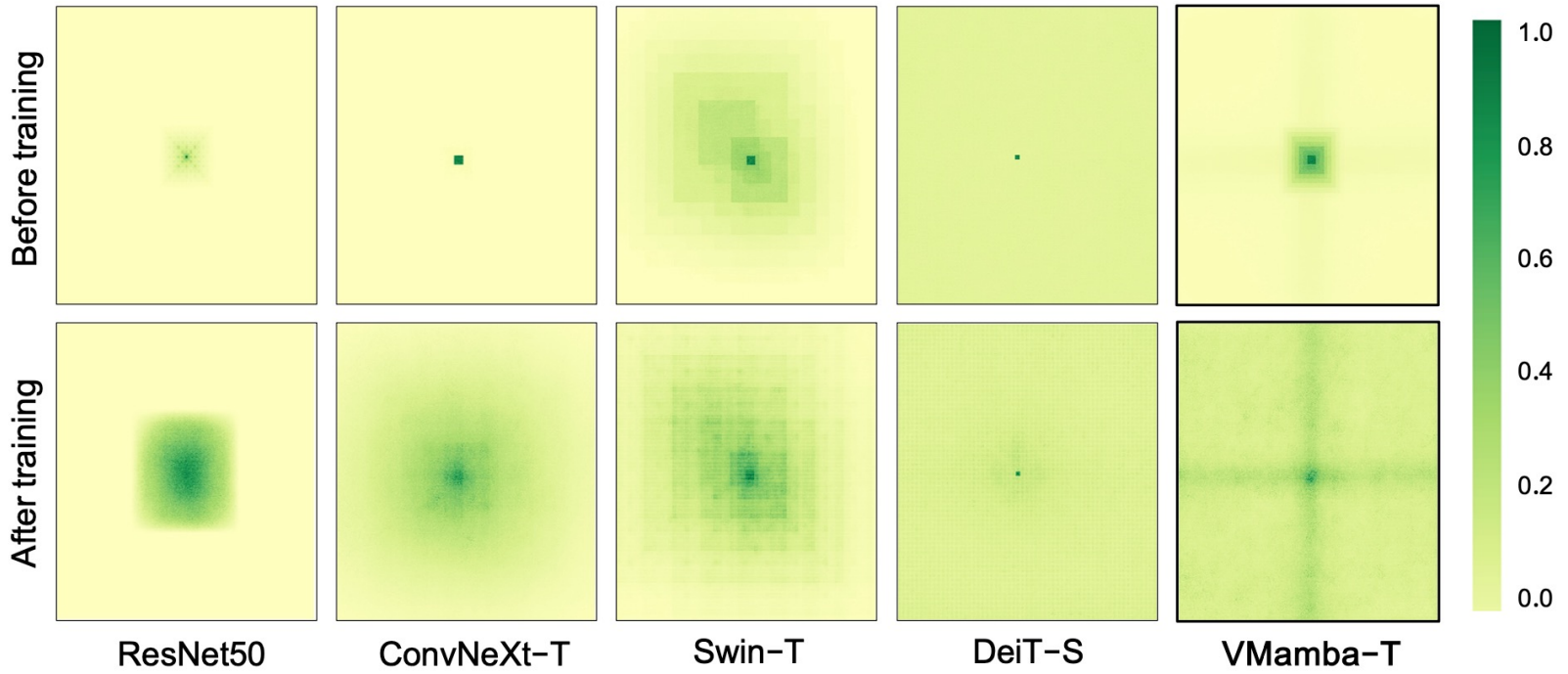


Figure 5: The **Effective Receptive Field (ERF)** is visualized for ResNet50 [19], ConvNeXt-T [29], Swin-T [28], DeiT-S [45] (ViT), and the proposed VMamba-T. A larger ERF is indicated by a more extensively distributed dark area. **Only DeiT [45] and the proposed VMamba exhibit a global ERF**. The inspiration for this visualization is drawn from [32].