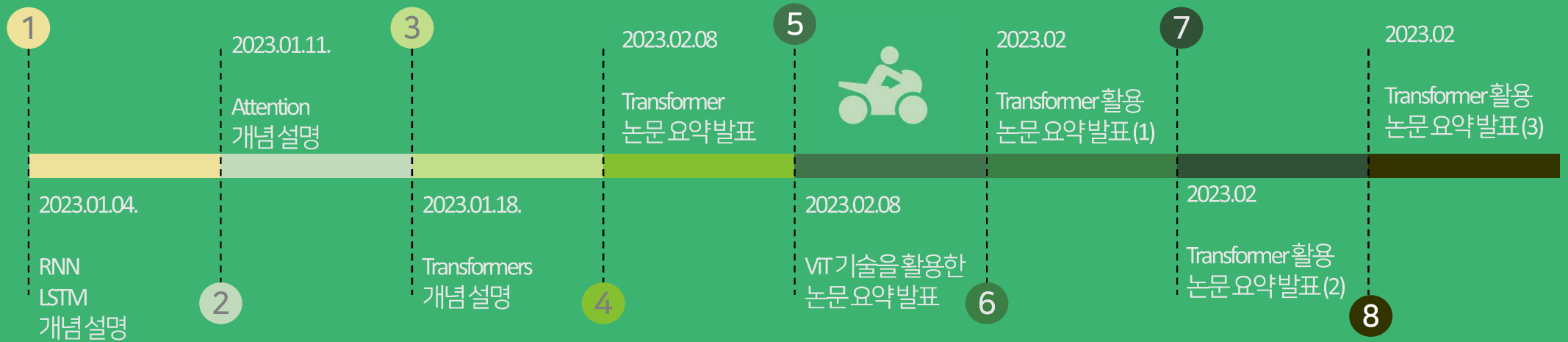
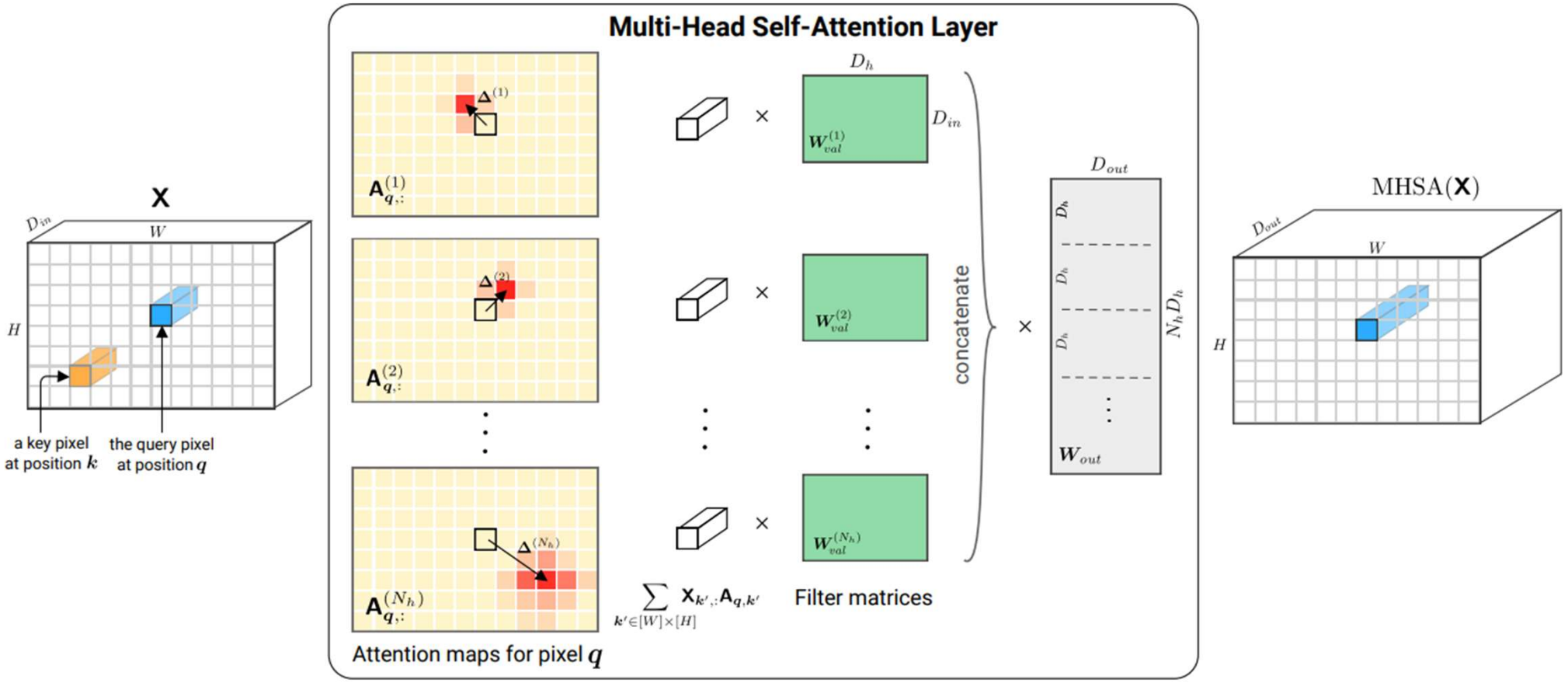
VIT Papers Rough Summary

CGVM Transformer study in 2023 winter

>> Transformer 주차별 계획



01 On the relationship between self-attention and convolutional layers



Cordonnier, J. B., Loukas, A., and Jaggi, M. (2019). On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.

02 An image is worth 16x16 words: Transformers for image recognition at scale

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

02 An image is worth 16x16 words: Transformers for image recognition at scale

Key points

- CNN에서 해방된 Computer vision

- Image를 일정한 크기의 Patch로 나누어서 이를 NLP 분야의 Token(words)처럼 처리

- VIT의 한계 : Inductive Bias (주어지지 않은 입력을 예측하는 능력) 의 부재

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

02 An image is worth 16x16 words: Transformers for image recognition at scale

Generalization Problem

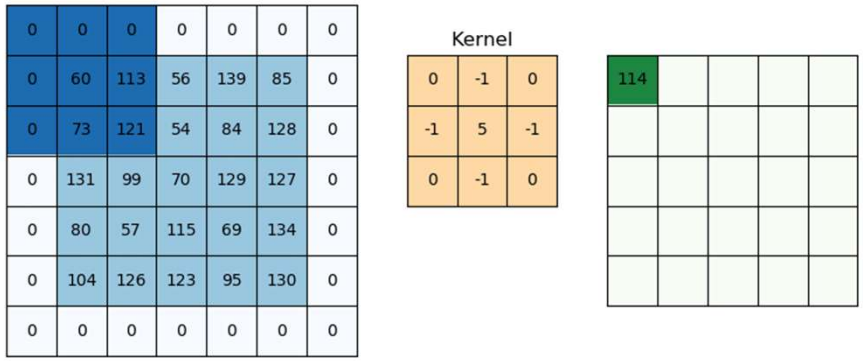
- Models are Brittle : 아무리 같은 의미의 데이터라도 조금만 바뀌면 모델이 망가진다.

- Models are Spurious : 데이터의 진정한 의미를 파악하지 못하고 결과(Arifacts)와 편향(Bias)를 암기한다.

Inductive Bias

학습 시에는 만나보지 않았던 상황에 대하여 정확한 예측을 하기 위해 사용하는 추가적인 가정

Inductive Bias



- Locality(spatial) Inductive bias

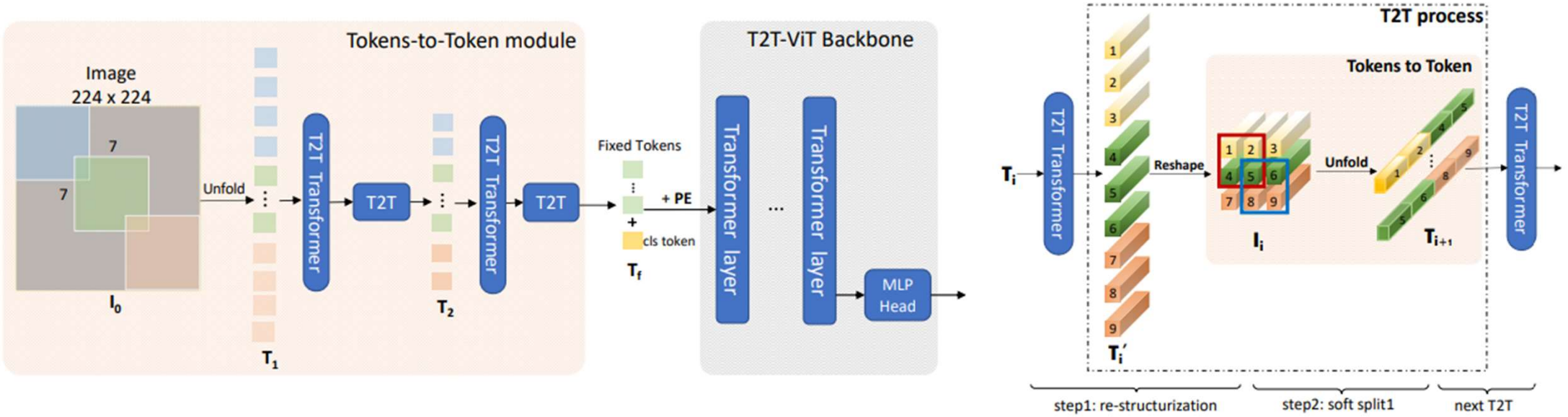
- Translation invariance Inductive bias

Battaglia, Peter W., et al. "Relational inductive biases, deep learning, and graph networks." *arXiv preprint arXiv:1806.01261* (2018).

02 An image is worth 16x16 words: Transformers for image recognition at scale

목적	Transformer를 Vision task(image recognition)에 적용하자
방법	ViT(Vision Transformer) - Transformer 를 사용 - image를 일정한 크기의 patch로 나누어 이를 NLP 분야의 Token처럼 처리
장점	Transformer의 특징(연산 효율성과 확장성) ㄴ 데이터셋과 모델 크기가 계속 커져도 모델 성능이 포화되지 않고 지속적으로 개선
단점	Inductive Bias의 부재 1. 같은 parameter의 CNN 모델보다 성능 저하 2. 재성능을 내려면 방대한 데이터 필요
검증방법	accuracy
코드	https://github.com/lucidrains/vit-pytorch

03 Tokens-to-token vit: Training vision transformers from scratch on imagenet



Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., and Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 558-567).

03 Tokens-to-token vit: Training vision transformers from scratch on imagenet.

Key points

- Tokens-to-Tokens(T2T) Module

- 패치 단위로 나뉘어진 토큰들을 다시 이미지 형태로 합쳐주는 Re-structurization

- 합친 이미지를 Local structure를 잘 포착하도록 다시 패치로 나눠 주는 Soft Split

- Deep Narrow Backbone(dense connection, GhostNet, MSA 헤드 개수 증가, Deep-narrow)

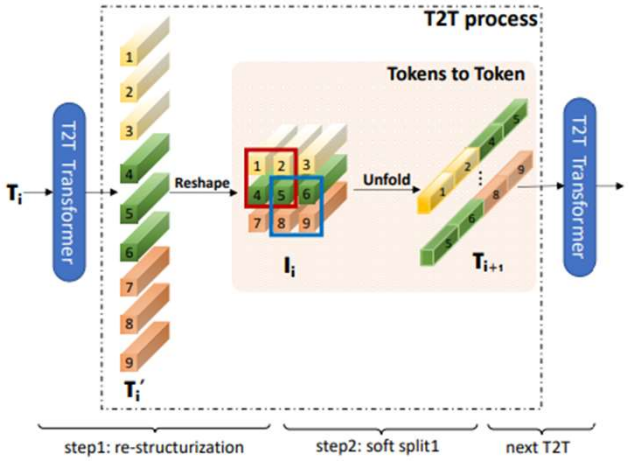


Table 2. Comparison between T2T-ViT and ViT by training from scratch on ImageNet.

Models	Top1-Acc (%)	Params (M)	MACs (G)
ViT-S/16 [12]	78.1	48.6	10.1
DeiT-small [36]	79.9	22.1	4.6
DeiT-small-Distilled [36]	81.2	22.1	4.7
T2T-ViT-14	**81.5**	21.5	4.8
T2T-ViT-14↑384	**83.3**	21.5	17.1
ViT-B/16 [12]	79.8	86.4	17.6
ViT-L/16 [12]	81.1	304.3	63.6
T2T-ViT-24	**82.3**	**64.1**	**13.8**

03 Tokens-to-token vit: Training vision transformers from scratch on imagenet.

목적	- Vit가 NLP와 Vision의 차이를 충분히 고려하지 않아 성능이 떨어진다는 것을 개선 1. ViT는 Edge나 Line 같은 Local structure를 잘 포착하지 못한다. 2. 파라미터를 효과적으로 활용하지 못해 불필요한 Feature를 너무 많이 만든다.
방법	- Tokens-to-Tokens(T2T) Module 1. 패치 단위로 나뉘어진 토큰들을 다시 이미지 형태로 합쳐주는 Re-stucturization 2. 합친 이미지를 Local structure를 잘 포착하도록 다시 패치로 나누어 주는 soft split - 다양한 선행 연구의 모델을 참조하여 vision task에 적합한 backbone을 찾음 ㄴ Deep Narrow Backbone(**Hidden 384, depth 16**)
장점	Vit 보다 적은 파라미터로 보다 좋은 성능
검증방법	Recognition accuracy
코드	https://github.com/yitu-opensource/T2T-ViT

04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

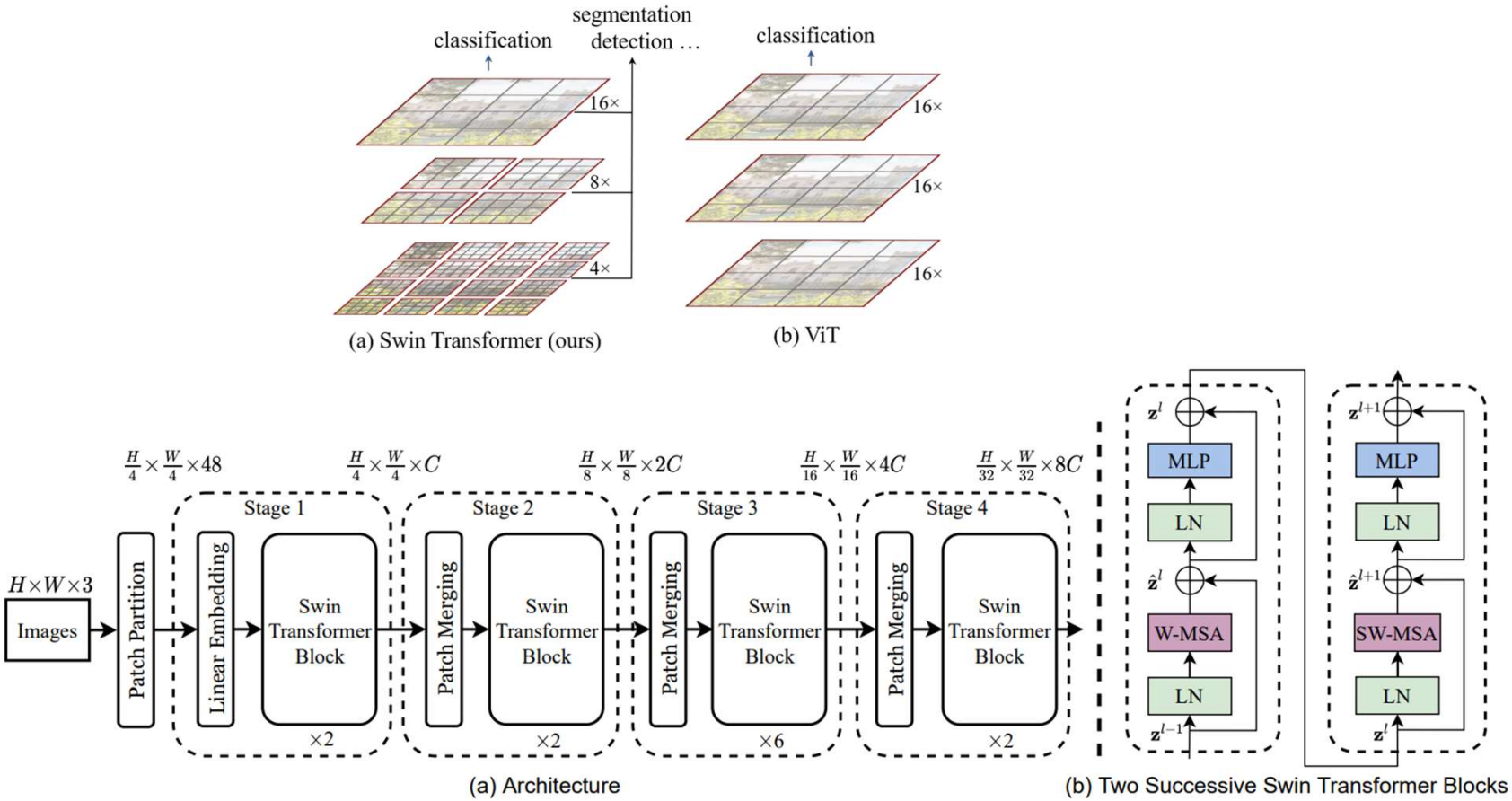


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z.,and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Key points

- 영상에 사이즈에 따라 quadratic 하게 증가하는 ViT의 연산량을 개선

- 계층적인 feature map을 이용하고 feature map의 window 내에서만 self-attention을 적용

- 계층적인 feature map 에서 다른 window 에 속한 patch 간의 연관성을 파악할 수 없다는 점을 shifted window로 해결

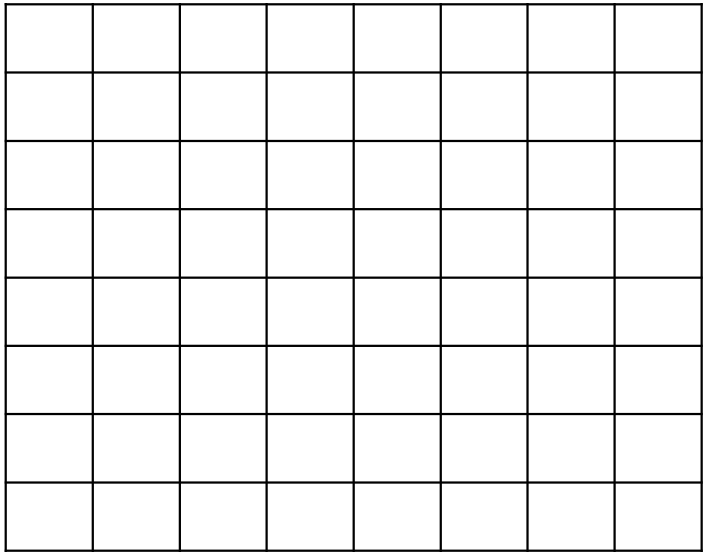
Method	mini-val AP^{box}	mini-val AP^{mask}	test-dev AP^{box}	test-dev AP^{mask}	#param.	FLOPs
RepPointsV2* [12]	-	-	52.1	-	-	-
GCNet* [7]	51.8	44.7	52.3	45.4	-	1041G
RelationNet++* [13]	-	-	52.7	-	-	-
SpineNet-190 [21]	52.6	-	52.8	-	164M	1885G
ResNeSt-200* [78]	52.5	-	53.3	47.1	-	-
EfficientDet-D7 [59]	54.4	-	55.1	-	77M	410G
DetectoRS* [46]	-	-	55.7	48.5	-	-
YOLOv4 P7* [4]	-	-	55.8	-	-	-
Copy-paste [26]	55.9	47.2	56.0	47.4	185M	1440G
X101-64 (HTC++)	52.3	46.0	-	-	155M	1033G
Swin-B (HTC++)	56.4	49.1	-	-	160M	1043G
Swin-L (HTC++)	57.1	49.5	57.7	50.2	284M	1470G
Swin-L (HTC++)*	**58.0**	**50.4**	**58.7**	**51.1**	284M	-

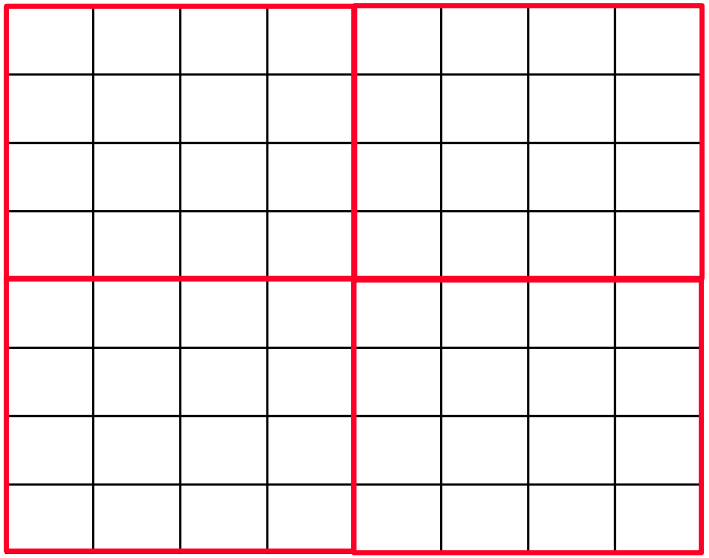
Table 2. Results on COCO object detection and instance segmentation. †denotes that additional decovolution layers are used to produce hierarchical feature maps. * indicates multi-scale testing.

Method	Backbone	val mIoU	test score	#param.	FLOPs	FPS
DANet [23]	ResNet-101	45.2	-	69M	1119G	15.2
DLab.v3+ [11]	ResNet-101	44.1	-	63M	1021G	16.0
ACNet [24]	ResNet-101	45.9	38.5	-		
DNL [71]	ResNet-101	46.0	56.2	69M	1249G	14.8
OCRNet [73]	ResNet-101	45.3	56.0	56M	923G	19.3
UperNet [69]	ResNet-101	44.9	-	86M	1029G	20.1
OCRNet [73]	HRNet-w48	45.7	-	71M	664G	12.5
DLab.v3+ [11]	ResNeSt-101	46.9	55.1	66M	1051G	11.9
DLab.v3+ [11]	ResNeSt-200	48.4	-	88M	1381G	8.1
SETR [81]	T-Large‡	50.3	61.7	308M	-	-
UperNet	DeiT-S†	44.0	-	52M	1099G	16.2
UperNet	Swin-T	46.1	-	60M	945G	18.5
UperNet	Swin-S	49.3	-	81M	1038G	15.2
UperNet	Swin-B‡	51.6	-	121M	1841G	8.7
UperNet	Swin-L‡	**53.5**	**62.8**	234M	3230G	6.2

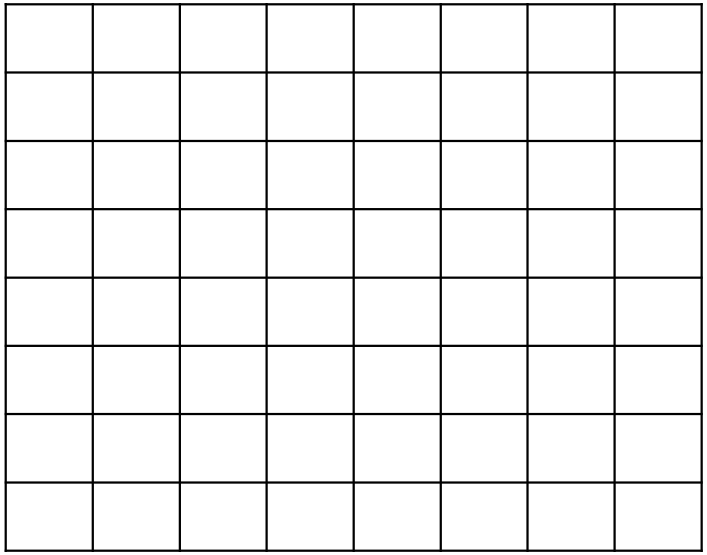
Table 3. Results of semantic segmentation on the ADE20K val and test set. † indicates additional deconvolution layers are used to produce hierarchical feature maps. ‡ indicates that the model is pre-trained on ImageNet-22K.

04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

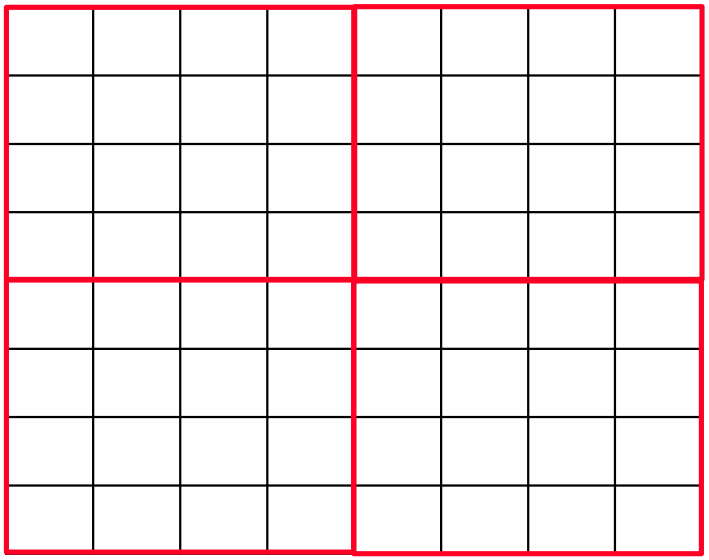




04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

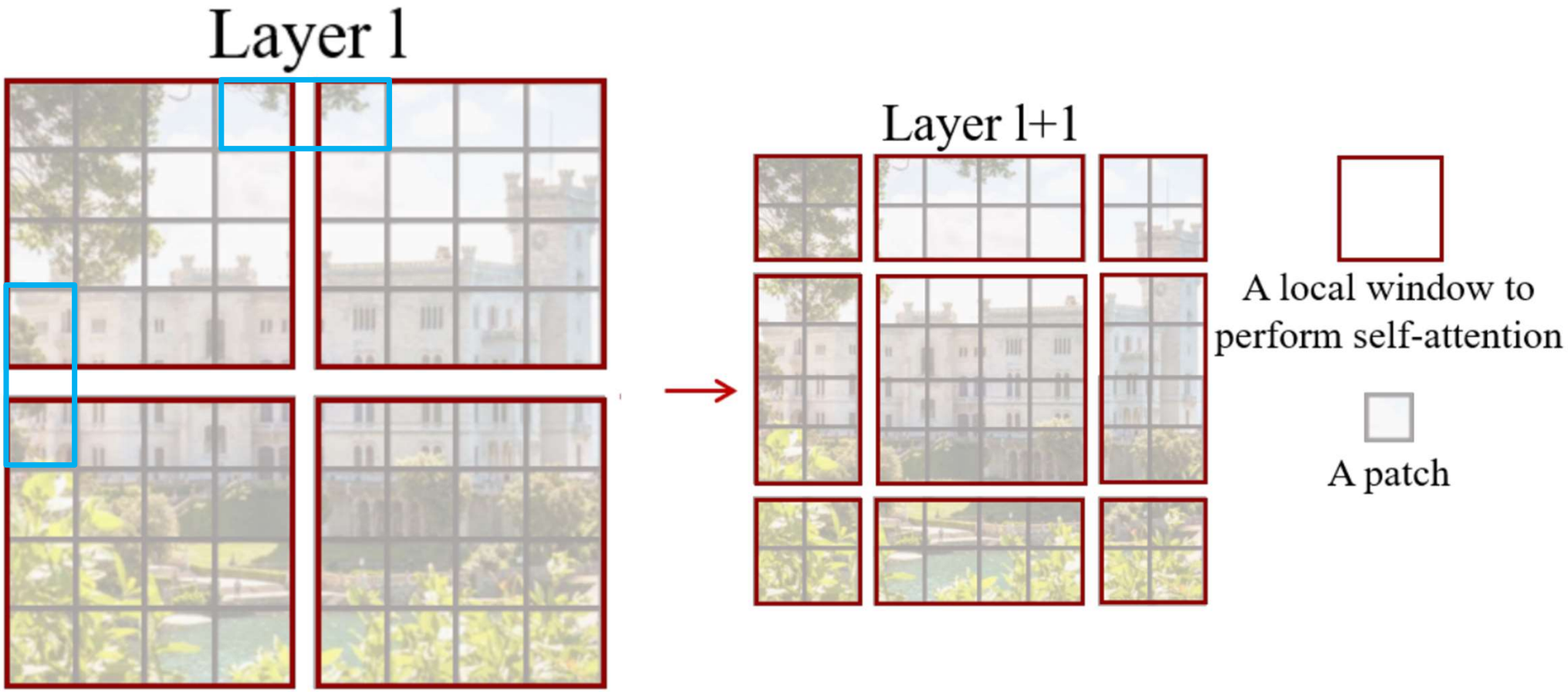


64 x 64 = 4096
Self Attention



16 x 16 x 4 = 1024
Self Attention

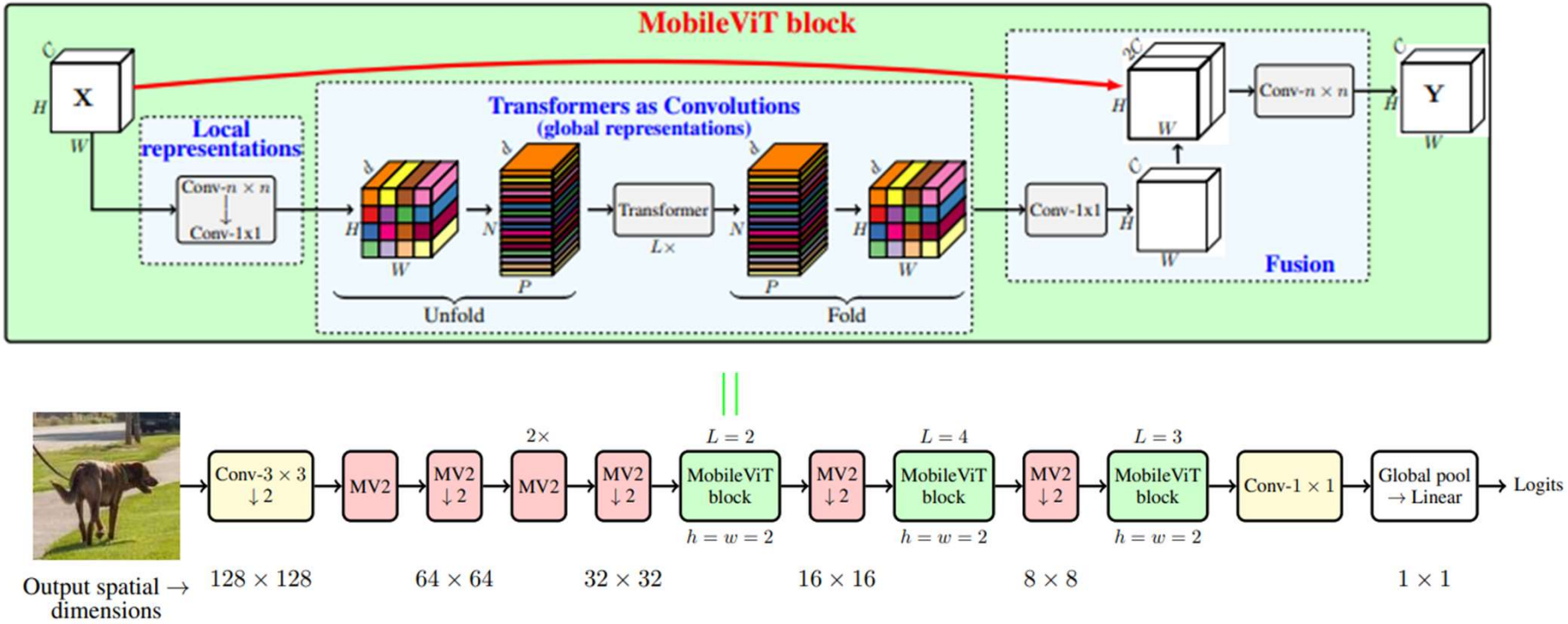
04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows



04 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

목적	- Vit의 연산량이 영상의 사이즈에 quadratic하게 증가한다는 문제를 해결
방법	- hierarchical feature map ㄴ 계층적인 feature map의 window 내에서만 self-attention을 적용하여 연산량 감소 - shifted window ㄴ 다른 window에 속한 patch 간의 연관성을 파악할 수 없다는 점을 해결 ㄴ Cyclic shift로 window 개수와 사이즈 유지
장점	Vit 보다 적은 파라미터로 CNN기반 모델보다 좋은 성능 연산량 감소로 object detection, Semantic segmentation이 가능
검증방법	Detection AP, Segmentation mIoU
코드	https://github.com/microsoft/Swin-Transformer

05 MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer



Mobile과 같은 하드웨어의 자원이 제한된 곳에서 범용적으로(General purpose) 사용할 수 있도록 만든 작고(Light-weight) 빠른(Low-latency) ViT 모델

Mehta, S., and Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

05 MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer

Key points

- CNN과 transformer의 강점을 결합

- 적은 parameter로 input의 local information과 global information을 모두 학습하는 것이 목표

- Inductive Bias의 부재를 CNN과의 결합으로 개선

- General purpose, Light-weight, Low-latency

Feature backbone	# Params. ⇓	mAP ⇑
MobileNetv3	4.9 M	22.0
MobileNetv2	4.3 M	22.1
MobileNetv1	5.1 M	22.2
MixNet	4.5 M	22.3
MNASNet	4.9 M	23.0
MobileViT-XS (Ours)	**2.7 M**	24.8
MobileViT-S (Ours)	5.7 M	**27.7**

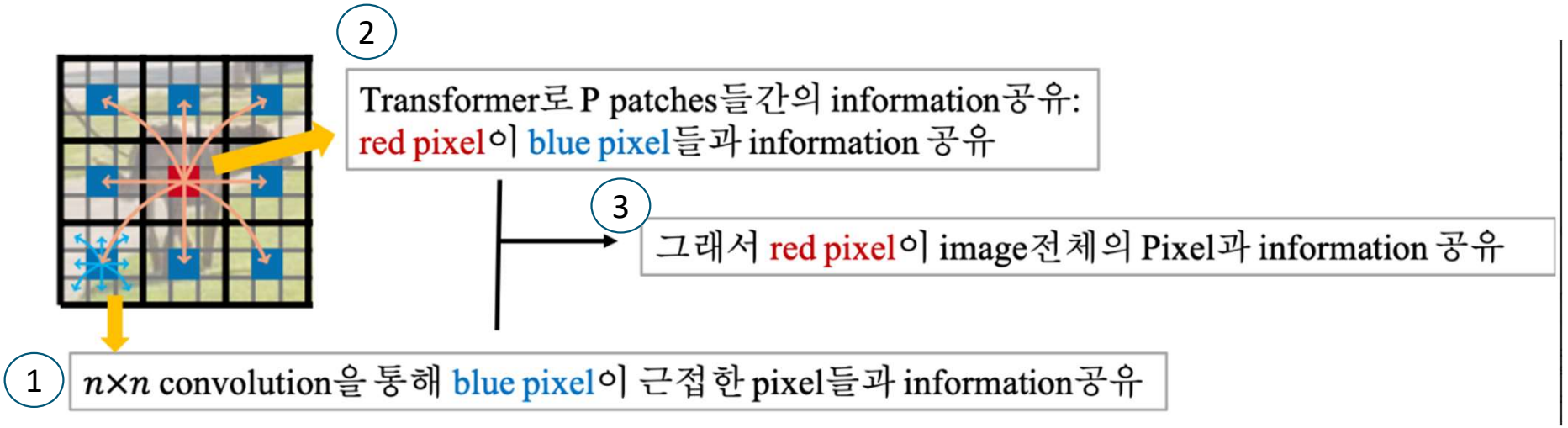
(a) Comparison w/ light-weight CNNs

Feature backbone	# Params. ⇓	mAP ⇑
VGG	35.6 M	25.1
ResNet50	22.9 M	25.2
MobileViT-S (Ours)	**5.7 M**	**27.7**

(b) Comparison w/ heavy-weight CNNs

Feature backbone	# Params. ⇓	mIOU ⇑
MobileNetv1	11.2 M	75.3
MobileNetv2	4.5 M	75.7
MobileViT-XXS (Ours)	1.9 M	73.6
MobileViT-XS (Ours)	2.9 M	**77.1**
ResNet-101	58.2 M	**80.5**
MobileViT-S (Ours)	6.4 M	79.1

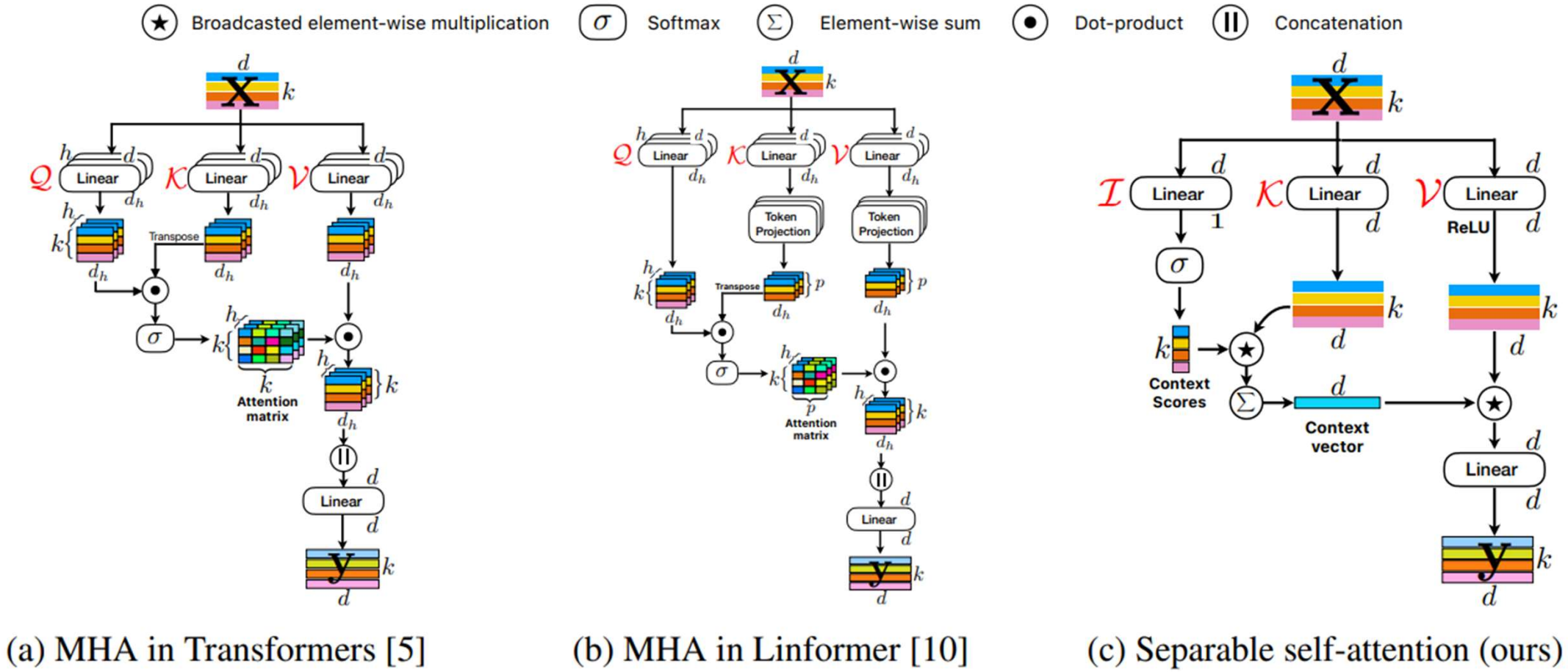
05 MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer



05 MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer

목적	- ViT가 Inductive bias의 부재로 생기는 문제점을 해결 1. 모델이 무겁다 2. 최적화하기 어렵다 3. 성능을 위해 large data가 필요하다 - 적은 parameter로 local information과 global information을 모두 학습하는 것이 목표
방법	- CNN과 ViT의 결합 - Mobile ViT Block
장점	- General purpose, Light-weight, Low-latency - CNN과 결합하여 local information과 global information를 효율적으로 학습
검증방법	Detection AP, Segmentation mIoU
코드	

06 Separable Self-attention for Mobile Vision Transformers



(a) MHA in Transformers [5] (b) MHA in Linformer [10] (c) Separable self-attention (ours)

Mehta, S., & Rastegari, M. (2022). Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*.

06 Separable Self-attention for Mobile Vision Transformers

Key points

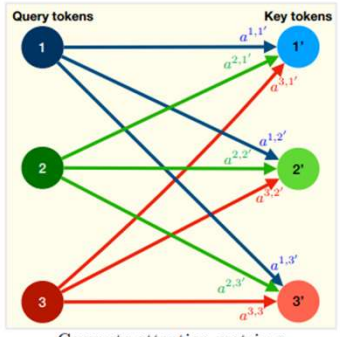
- 기존 VIT 는 파라미터는 적지만 high latency

- multi head attention 의 시간 복잡도는 $O(k^2)$

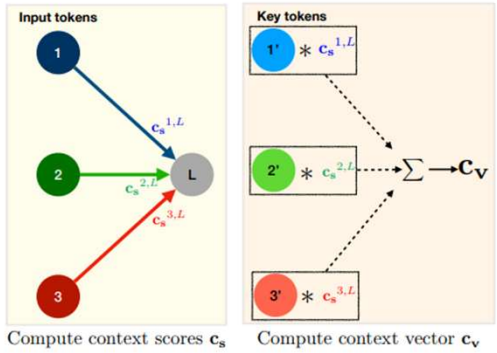
- 이를 위해 Multi head attention 연산을 개선하여 O(k)로 개선

- MobileViTv2

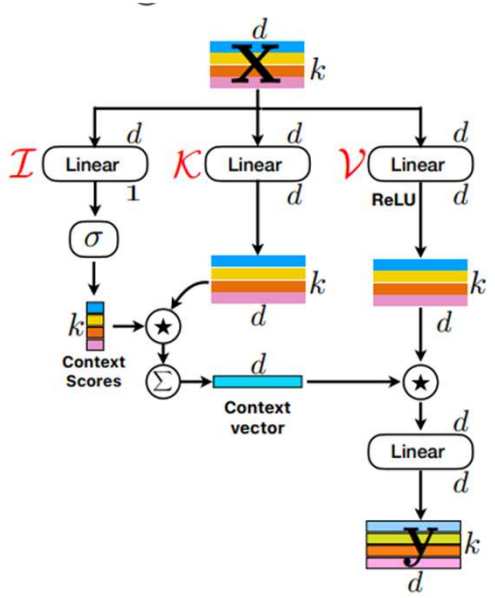
- Latent token L 에대한 context score를 계산하는 것이 주요 아이디어



(a) Self-attention in transformers



(b) Proposed separable self-attention method



(c) Separable self-attention (ours)

Attention unit	Latency ↓	Top-1 ↑
Self-attention in Transformer	9.9 ms	**78.4**
Self-attention in Linformer	10.2 ms	78.2
Separable self-attention	**3.4** ms	78.1

06 Separable Self-attention for Mobile Vision Transformers

목적	- Mobile ViT의 문제점을 개선한 모델 ㄴ Mobile ViT(Multi-head attention)의 high latency, $O(k^2)$의 시간복잡도
방법	- Multi-head attention을 개선한 Separable self-attention ㄴ Latent token L에 대한 context score를 계산 ㄴ 각 토큰을 내적 연산을 통해 스칼라에 매핑한 후 latent token 과의 거리를 계산(context score) ㄴ fusion block은 성능향상에 도움이 되지 않는다 판단하여 제거
장점	- 다른 self-attention 방법과 비교했을 때 훨씬 빠른 추론 속도
검증방법	Recognition의 accuracy, Latency 측정, 비교
코드	

07 Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features

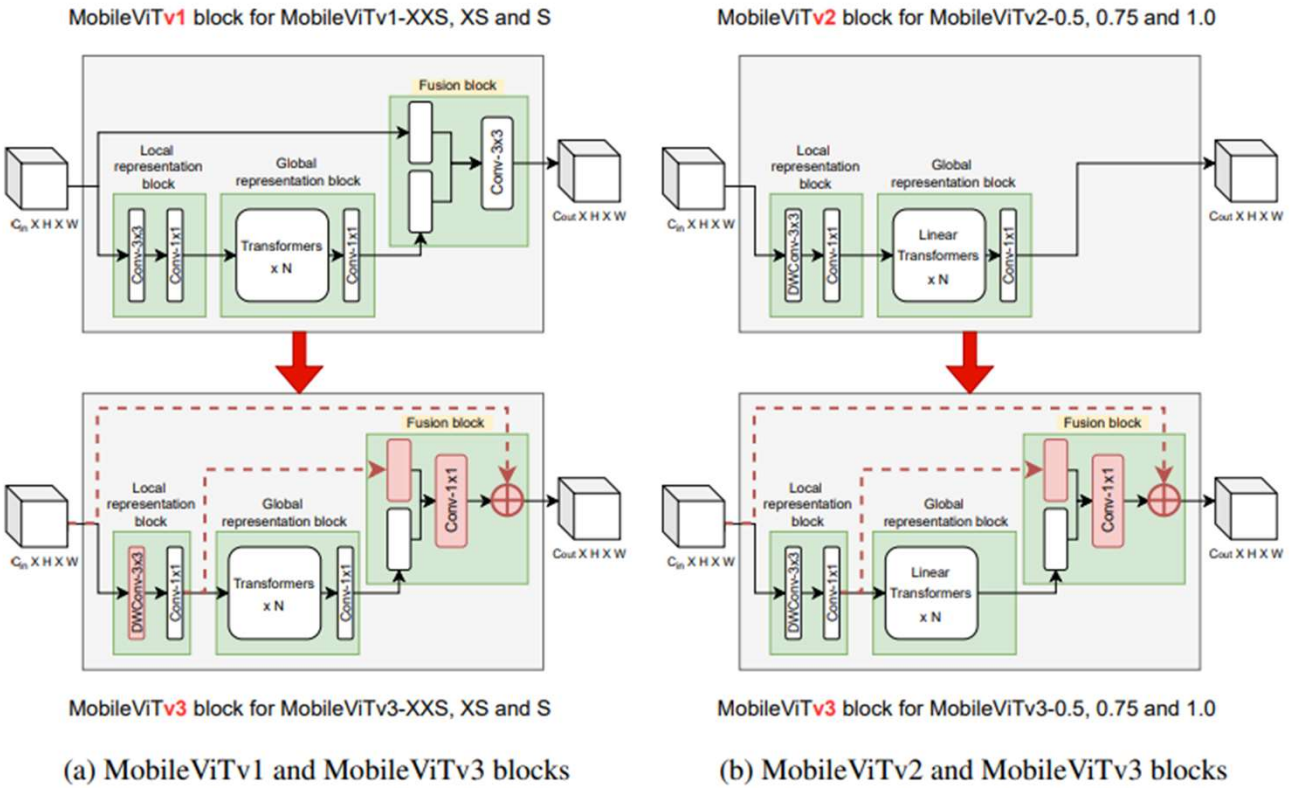


Figure 2: A comparison between: (a) MobileViTv1 and MobileViTv3 modules, and (b) MobileViTv2 and MobileViTv3 modules. The proposed architectural changes are highlighted in red.

Wadekar, S. N., and Chaurasia, A. (2022). Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*.

07 Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features

Key points

- Mobile ViT의 fusion block을 개선하자.

- scaling challenge, complex learning task

- Conv 3x3 → Conv 1x1

- Residual connection

Backbone	# Params (M)↓	mAP(%) ↑
MobileViTv1-XXS	1.50	18.5
MobileViTv3-XXS	1.53	**19.3** (↑0.8%)
MobileViTv2-0.5	2	21.2
MobileViTv3-0.5	2	**21.8** (↑0.6%)
MobileViTv2-0.75	3.6	24.6
MobileViTv3-0.75	3.7	**25.0** (↑0.4%)
MobileNetv3	4.9	22.0
MobileNetv2	4.3	22.1
MobileNetv1	5.1	22.2
MixNet	4.5	22.3
MNASNet	4.9	23.0 (↑0.0%)
MobileViTv1-XS	2.7	24.8 (↑1.8%)
MobileViTv3-XS	2.7	**25.6** (↑2.6%)

(a) Comparison w/light-weight CNNs

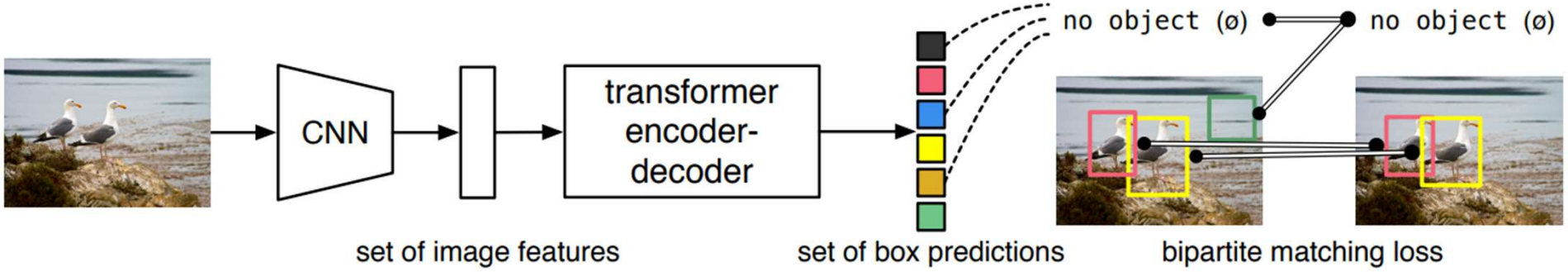
Backbone	# Params. (M) ↓	mIOU(%) ↑
MobileViTv1-XXS	1.90	73.60
MobileViTv3-XXS	1.96	**74.04** (+0.44%)
MobileViTv2-0.5	6.2	75.07
MobileViTv3-0.5	6.3	**76.48** (+1.41%)
MobileViTv1-XS	2.9	77.10
MobileViTv3-XS	3.3	**78.77** (+1.60%)
MobileViTv1-S	6.4	79.10
MobileViTv3-S	7.2	**79.59** (+0.49%)
MobileViTv2-1.0	13.32	78.94
MobileViTv3-1.0	13.56	**80.04** (+1.10%)

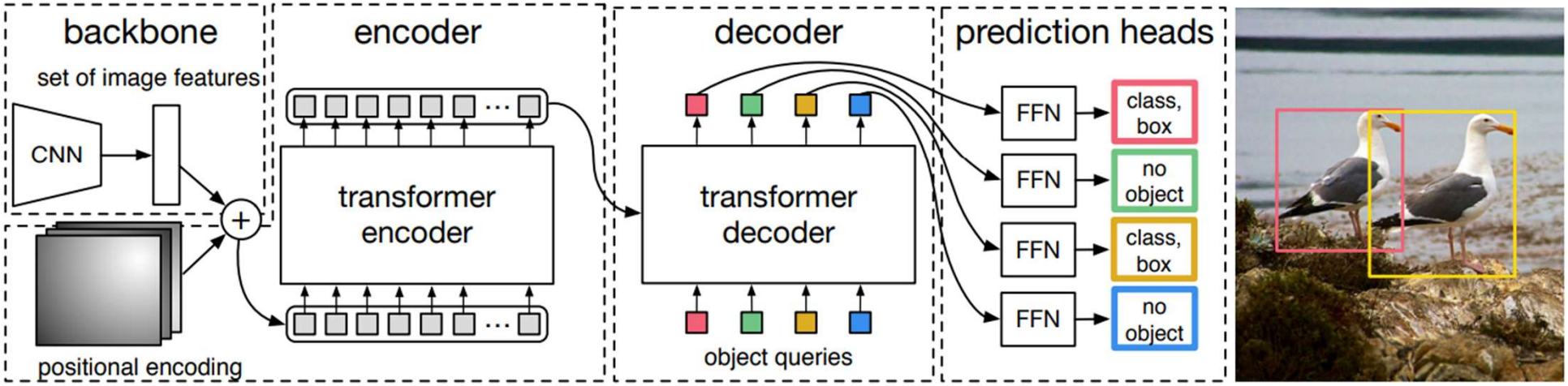
(a) Segmentation on PASCAL VOC 2012 dataset

07 Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features

목적	- Mobile ViT v1, v2의 문제점을 개선한 모델 ㄴ Mobile ViT의 fusion block은 vision task에 적합하지 않다.
방법	- 개선된 Mobile ViT의 fusion block ㄴ fusion block의 목표는 input과 global feature를 합치는 것이므로 Conv 1x1로 변경 ㄴ Residual connection을 사용하여 더 깊은 layer를 최적화
장점	- Output channel의 채널수가 증가하는 것을 해결해 scaling 문제 해결 - detection, segmentation 두 task에서 v1,v2 보다 좋은 성능
검증방법	Detection AP, Segmentation mIoU
코드	https://github.com/micronDLA/MobileViTv3 (저자x)

08 End-to-end object detection with transformers





Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (pp. 213-229). Springer International Publishing.

08 End-to-end object detection with transformers

Key points

- Object detection을 direct set predicition problem으로 간주하고 해결

- Prior knowledge (NMS(Non-Maximum Suppression, anchor genetration)를 사용하지 않아 detection pipeline을 간소화

- DEtection TRansformer (DETR)

- Partite matching (이분매칭) & Transformer encoder-decoder architecture

- Hungarian Algorithm

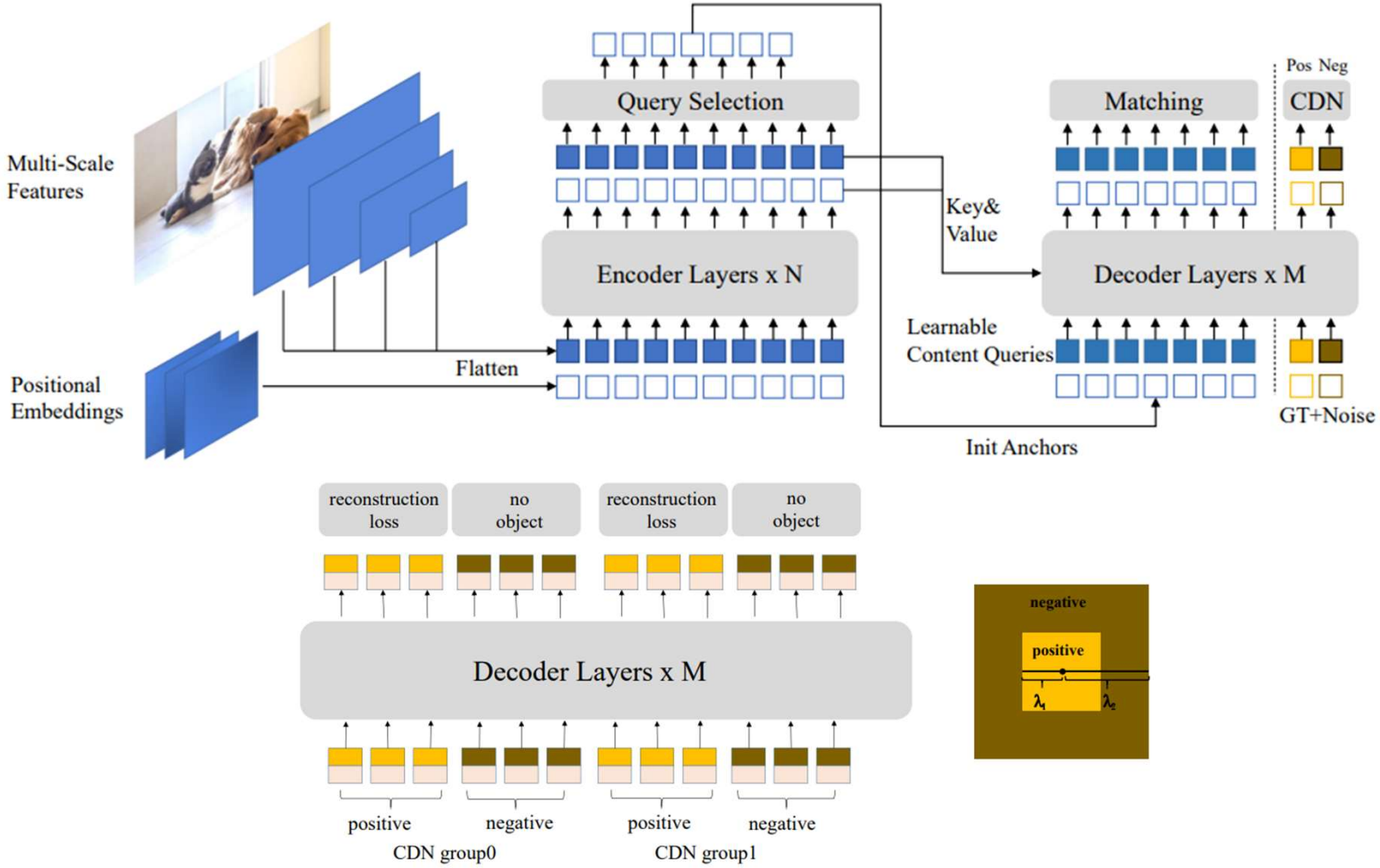
- SOTA baseline 모델 중 하나인 Faster R-CNN과 견주어도 부족함 없음

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

08 End-to-end object detection with transformers

목적	Object detection task에 Transfoermer를 사용
방법	DETR - Object detection을 direct set predicition problem으로 - Prior knowledge를 사용하지 않아 detection pipeline을 간소화 - Partite matching (이분매칭) & Transformer encoder-decoder architecture - Hungarian Algorithm
장점	SOTA baseline 모델 중 하나인 Faster R-CNN과 견주어도 부족함 없는 성능
단점	General purpose, Light-weight, Low-latency
검증방법	Detection AP
코드	https://github.com/facebookresearch/detr

09 DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection



Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Shum, H. Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

09 DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

Key points

- DETR을 개선한 모델

- DETR이 객체가 없을 때 Negative로 분류하는 능력이 떨어지는 것을 해결

- CDN(Contrastive DeNoising) → positive query, negative query

- Mixed Query Selection → 동적 쿼리, 정적 쿼리 selection 의 문제를 개선

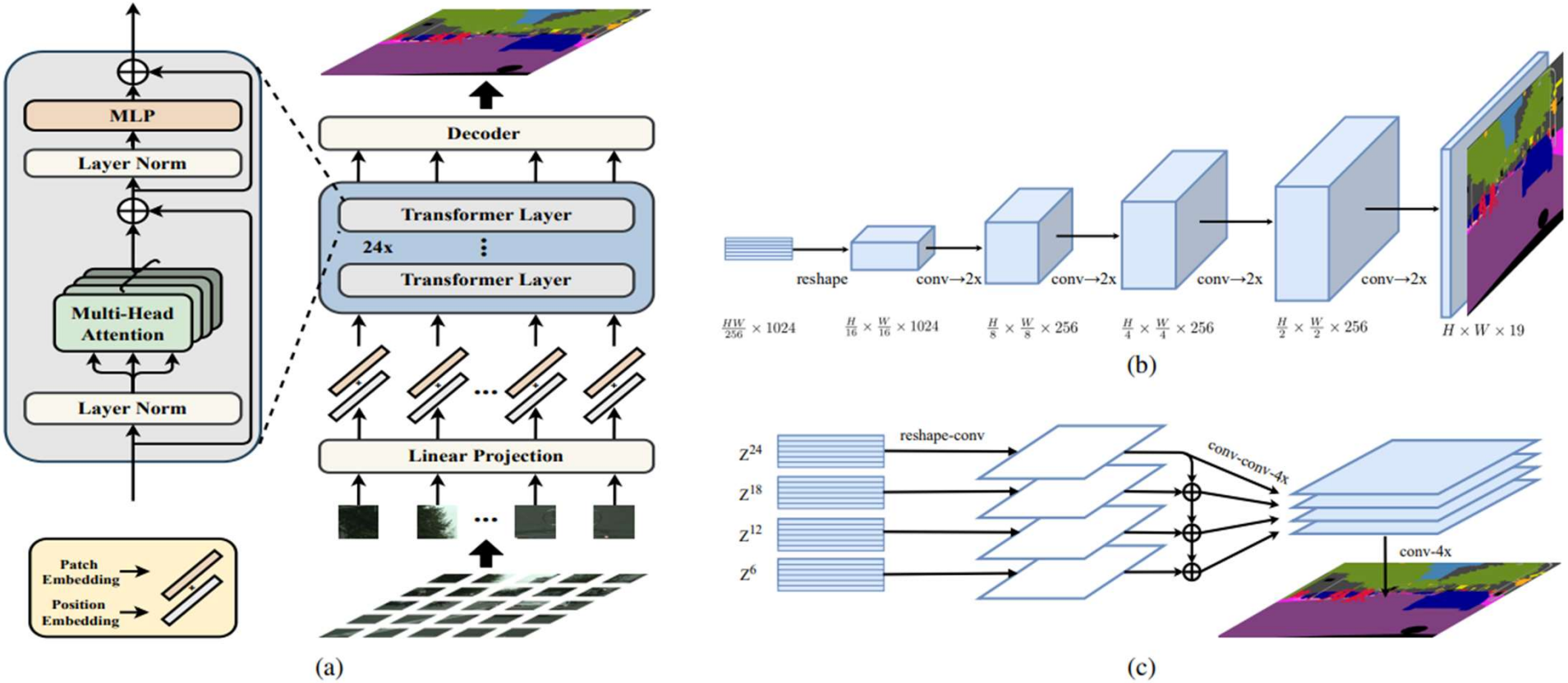
- Look forward Twice → 레이어의 loss가 다음 레이어 까지 전달되어 학습 수렴속도를 가속화

#Row	QS	CDN	LFT	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1. DN-DETR [17]	No			43.4	61.9	47.2	24.8	46.8	59.4
2. Optimized DN-DETR	No			44.9	62.8	48.6	26.9	48.2	60.0
3. Strong baseline (Row2+pure query selection)	Pure			46.5	64.2	50.4	29.6	49.8	61.0
4. Row3+mixed query selection	Mixed			47.0	64.2	51.0	31.1	50.1	61.5
5. Row4+look forward twice	Mixed		✓	47.4	64.8	51.6	29.9	50.8	61.9
6. DINO (ours, Row5+contrastive DN)	Mixed	✓	✓	**47.9**	**65.3**	**52.1**	**31.2**	**50.9**	**61.9**

09 DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

목적	DETR이 객체가 없을 때 Negative로 분류하는 능력이 떨어지는 것을 해결
방법	- CDN(Contrastive DeNosing) → Positive query와 negative query 를 생성해서 디코더에 입력 - Mixed Query Selection →동적 쿼리에서는 이미지 특성 정보가 소실될 수 있다는 점을 해결 - Look forward Twice → 레이어의 loss가 다음 레이어 까지 전달되어 학습 수렴 속도를 가속화
장점	DETR 보다 좋은 성능
단점	General purpose, Light-weight, Low-latency
검증방법	Detection AP
코드	https://github.com/IDEA-Research/DINO

10 Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers



b : Progressive Upsampling(PUP) c : Multi-Level feature Aggregation(MLA)

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881-6890).

10 Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers

Key points

- Segmentation task의 encoder-decoder 구조에 ViT 사용

- ViT 를 encoder로 사용

- Naïve upsampling(Naïve)

- Progressive UPsampling(PUP)

- Multi-level feature Aggregation(MLA)

Method	Pre	Backbone	mIoU
FCN (80k, SS) [39]	1K	ResNet-101	44.47
FCN (80k, MS) [39]	1K	ResNet-101	45.74
DANet [18]	1K	ResNet-101	52.60
EMANet [31]	1K	ResNet-101	53.10
SVCNet [16]	1K	ResNet-101	53.20
Strip pooling [23]	1K	ResNet-101	54.50
GFFNet [30]	1K	ResNet-101	54.20
APCNet [19]	1K	ResNet-101	54.70
SETR-*Naïve* (80k, SS)	21K	T-Large	52.89
SETR-*Naïve* (80k, MS)	21K	T-Large	53.61
SETR-*PUP* (80k, SS)	21K	T-Large	54.40
SETR-*PUP* (80k, MS)	21K	T-Large	55.27
SETR-*MLA* (80k, SS)	21K	T-Large	54.87
SETR-*MLA* (80k, MS)	21K	T-Large	**55.83**
SETR-*PUP-DeiT* (80k, SS)	1K	T-Base	52.71
SETR-*PUP-DeiT* (80k, MS)	1K	T-Base	53.71
SETR-*MLA-DeiT* (80k, SS)	1K	T-Base	52.91
SETR-*MLA-DeiT* (80k, MS)	1K	T-Base	53.74

Table 5. **State-of-the-art comparison on the Pascal Context dataset.** Performances of different model variants are reported. SS: Single-scale inference. MS: Multi-scale inference.

10 Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers

목적	Transformer를 semantic Segmentation에 적용
방법	- Segmentation task의 encoder-decoder 구조에 ViT 사용 - Vit를 Encoder로 사용 - 3가지 Decoder를 사용하는 3가지의 submodel 1. Naïve upsampling(Naïve) → bilinearly unsampling 2. Progressive UPsampling(PUP) → adversarial effect 방지 3. Multi-level feature Aggregation(MLA) → feature pyramid network과 유사한 구조를 사용하여 속도 향상 메모리 사용량 감소
장점	ResNet 기반의 모델보다 근소하게 좋은 성능
검증방법	Segmantation mIoU
코드	https://github.com/fudan-zvg/SETR

Thank you

Hyoungbum Kim