MUSIQ: Multi-scale Image Quality Transformer

Junjie Ke[1], Qifei Wang[1], Yilin Wang[2], Peyman Milanfar[1], Feng Yang[1]
[1]Google Research, [2]Google
{junjiek, qfwang, yilin, milanfar, fengyang}@google.com

ICCV 2021

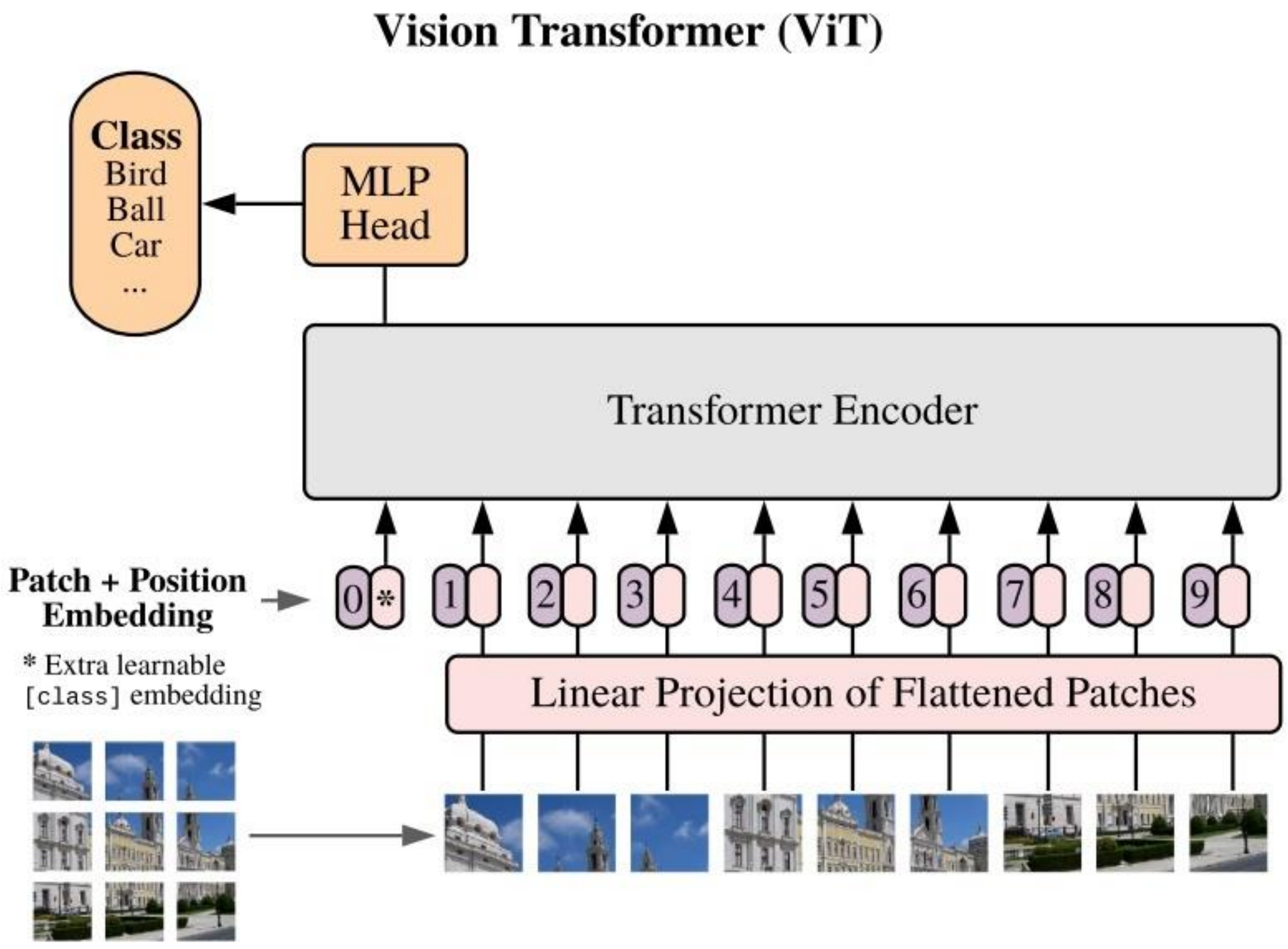
Transformer

- image patches->embeddings
- encoder: self-attention

-> 计算任意两个位置之间关联，所需计算量都相同

-> 这与CNN不同

- 在许多视觉任务上，基于transformer的性能已超过CNN

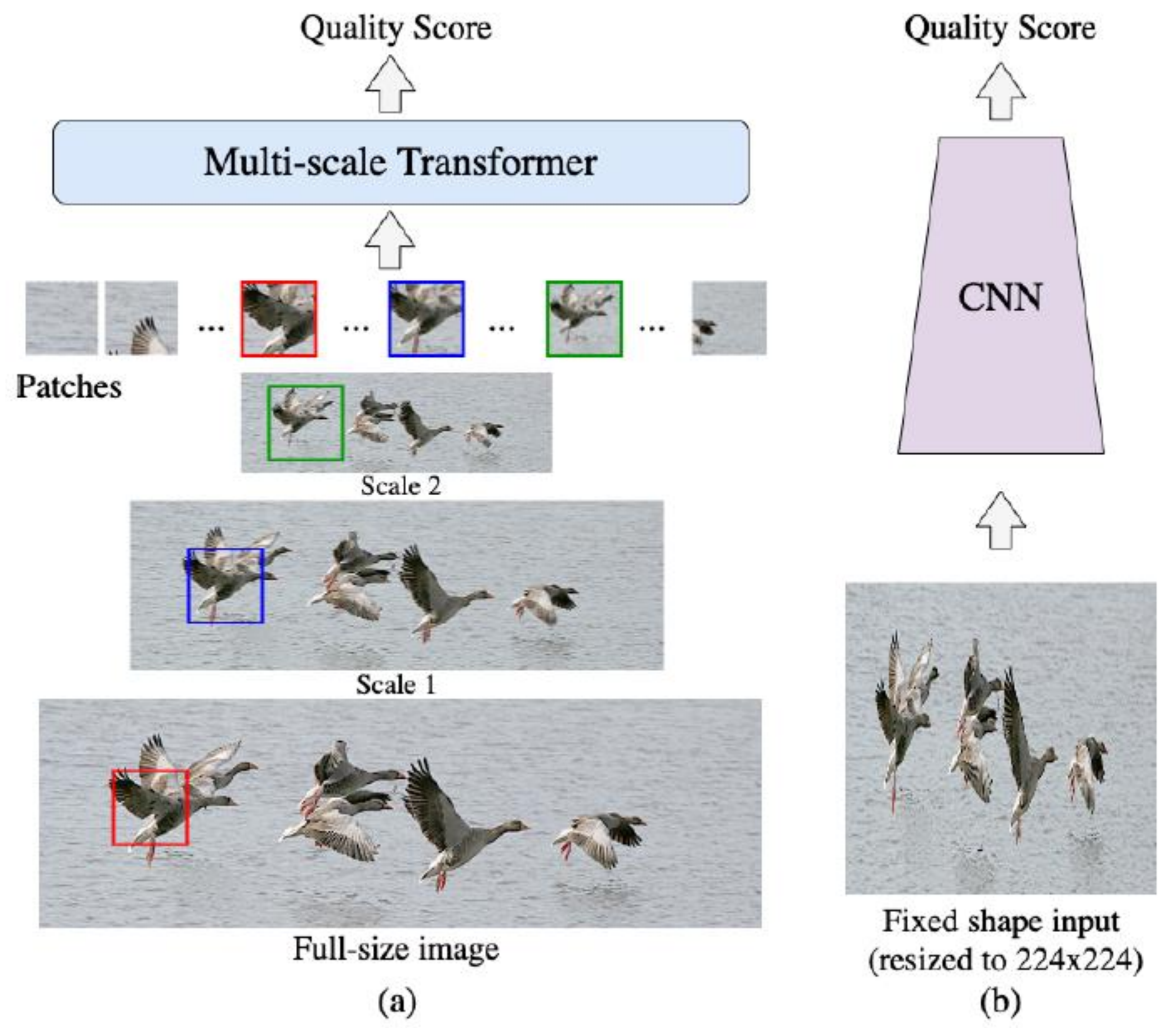


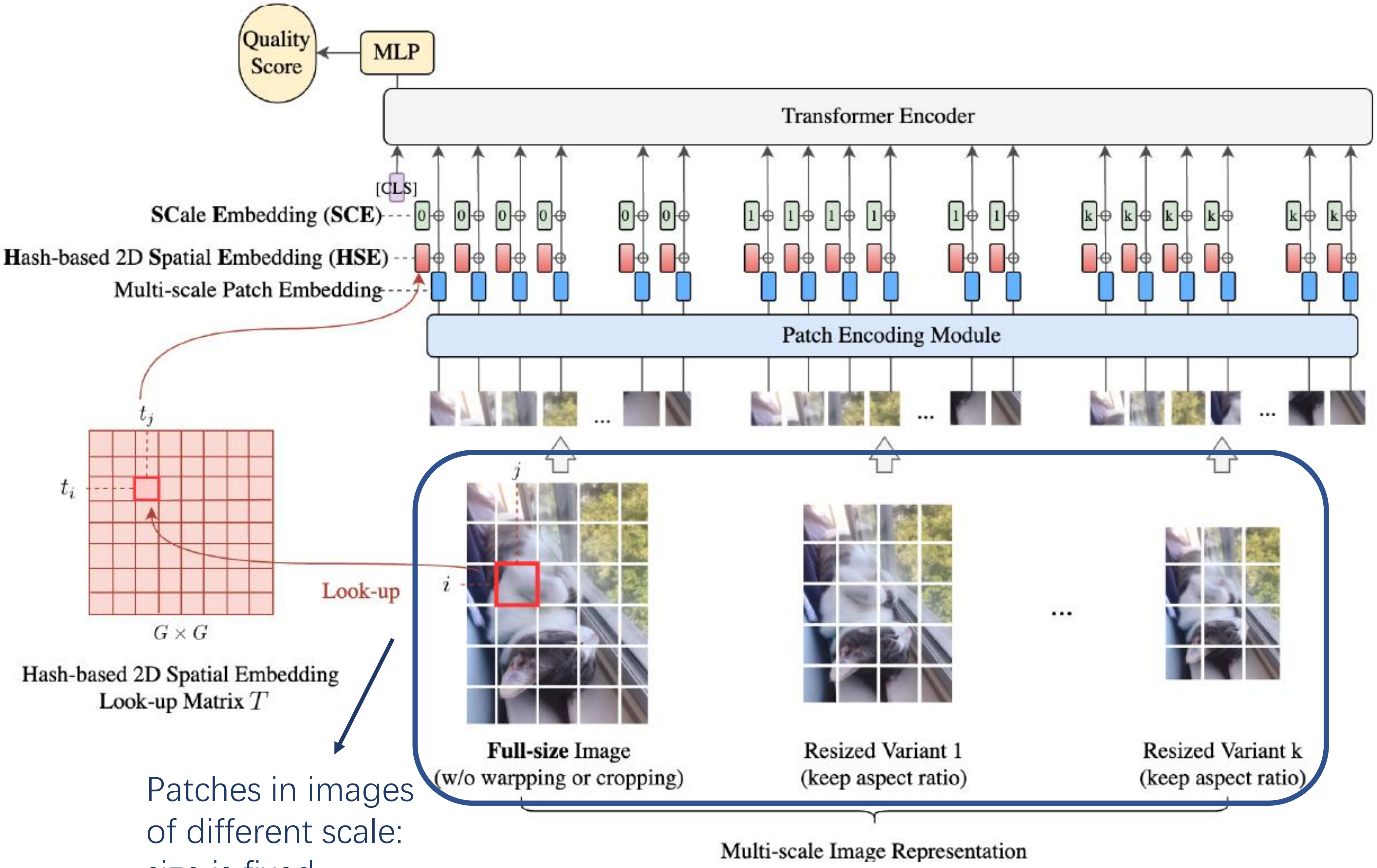
Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale.

Motivation

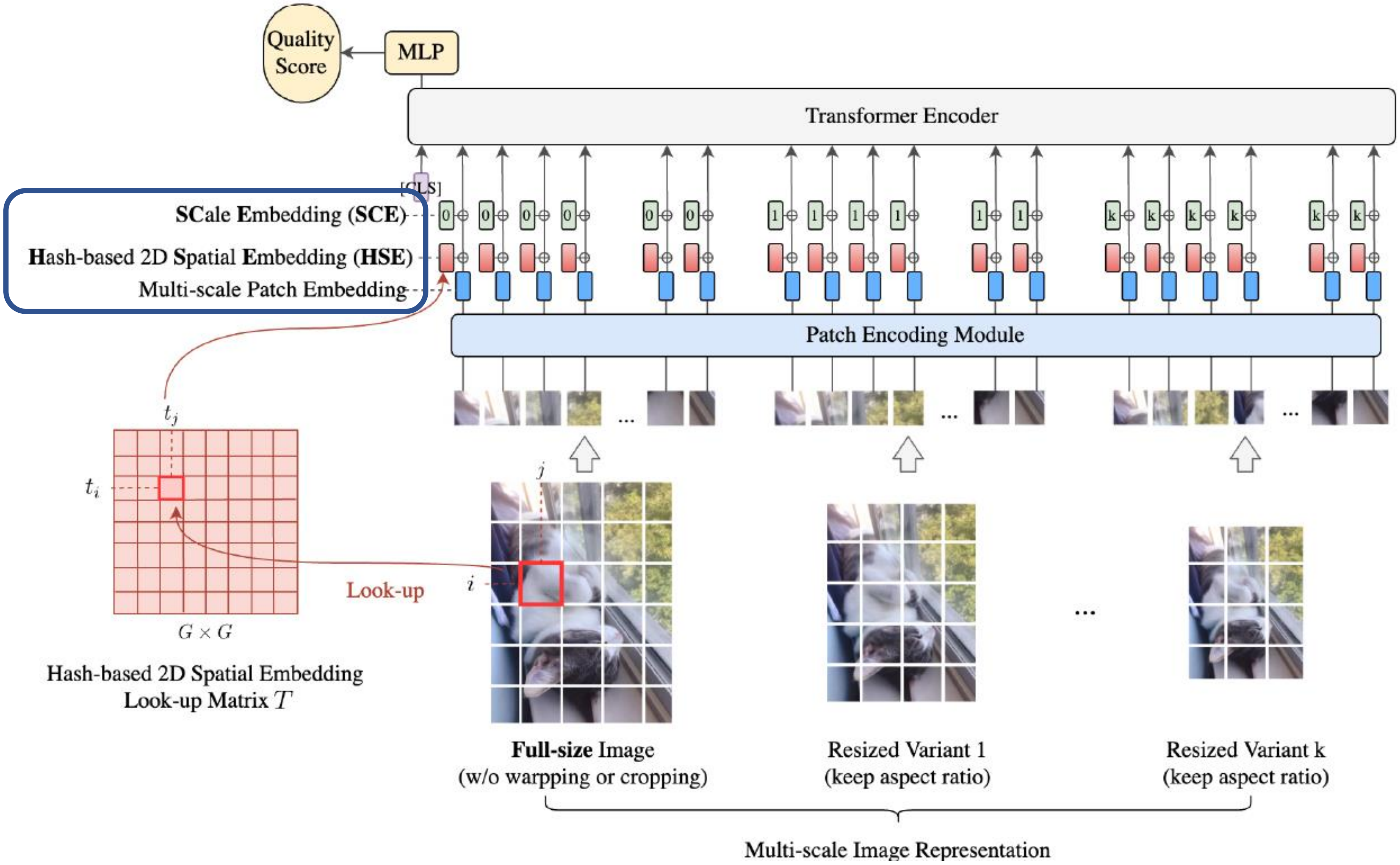
- IQA with CNN: input images are usually resized and cropped to a fixed shape
- Use a multi-scale image representation to capture image quality at different granularities

- a hash-based 2D spatial embedding
- a scale embedding





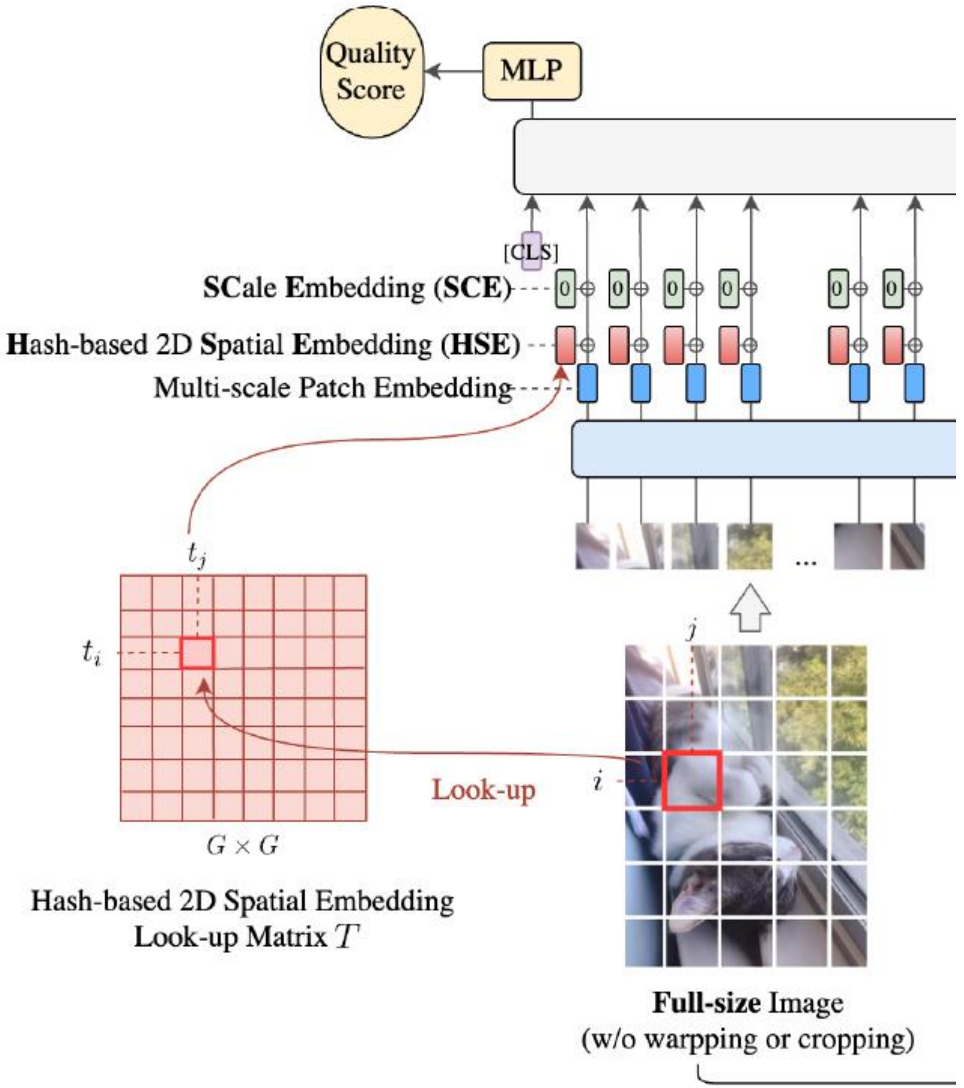
Patches in images of different scale: size is fixed



Quality Score
MLP
Transformer Encoder
[CLS]
SCale Embedding (SCE)
Hash-based 2D Spatial Embedding (HSE)
Multi-scale Patch Embedding
Patch Encoding Module
t_j
t_i
j
i
Look-up
$G \times G$
Hash-based 2D Spatial Embedding Look-up Matrix T
Full-size Image (w/o warpping or cropping)
Resized Variant 1 (keep aspect ratio)
Resized Variant k (keep aspect ratio)
Multi-scale Image Representation

Hash-based 2D Spatial Embedding

$$t_i = \frac{i \times G}{H/P}, \; t_j = \frac{j \times G}{W/P}$$



Experiment

- Pre-train: ImageNet
  - resize some images
  - augmentation: random cropping
  - 300 epochs, batch size=4096
- Fine-tune: IQA dataset
  - no resize or crop of input images

	Validation Set		Test Set	
method	SRCC	PLCC	SRCC	PLCC
BRISQUE [26]	0.303	0.341	0.288	0.373
NIQE [27]	0.094	0.131	0.211	0.288
CNNIQA [18]	0.259	0.242	0.266	0.223
NIMA [36]	0.521	0.609	0.583	0.639
Ying *et al.* [43]	0.562	0.649	0.601	0.685
MUSIQ-single	**0.563**	**0.651**	**0.640**	**0.721**
MUSIQ (Ours)	**0.566**	**0.661**	**0.646**	**0.739**
std	±0.002	±0.003	±0.005	±0.006

Table 1. Results on PaQ-2-PiQ full-size validation and test sets.

method	SRCC	PLCC
BRISQUE [26]	0.665	0.681
ILNIQE [47]	0.507	0.523
HOSA [39]	0.671	0.694
BIECON [19]	0.618	0.651
WaDIQaM [3]	0.797	0.805
PQR [44]	0.880	0.884
SFA [21]	0.856	0.872
DBCNN [49]	0.875	0.884
MetaIQA [50]	0.850	0.887
BIQA [34] (**25** crops)	**0.906**	0.917
MUSIQ-single	0.905	**0.919**
MUSIQ (Ours)	**0.916**	**0.928**
std	±0.002	±0.003

Table 2. Results on KonIQ-10k dataset. Blue and black numbers in bold represent the best and second best respectively. We take numbers from [34, 50] for results of the reference methods.

Experiment

method	SRCC	PLCC
DIIVINE [28]	0.599	0.600
BRISQUE [26]	0.809	0.817
CORNIA [42]	0.709	0.725
QAC [40]	0.092	0.497
ILNIQE [47]	0.713	0.721
FRIQUEE [14]	0.819	0.830
DBCNN [49]	0.911	0.915
Fang *et al.* [12] (w/o extra info)	0.908	0.909
MUSIQ-single	**0.917**	**0.920**
MUSIQ (Ours)	**0.917**	**0.921**
std	± 0.002	± 0.002

Table 3. Results on SPAQ dataset. Blue and black numbers in bold represent the best and second best respectively. We take numbers from [12] for results of the reference methods.

method	cls. acc.	MSE $\downarrow$	SRCC	PLCC
MNA-CNN-Scene [25]	0.765	-	-	-
Kong *et al.* [20]	0.773	-	0.558	-
AMP [29]	0.803	0.279	0.709	-
A-Lamp [24] (**50** crops)	0.825	-	-	-
NIMA (VGG16) [36]	0.806	-	0.592	0.610
NIMA (Inception-v2) [36]	0.815	-	0.612	0.636
MP_{ada} [33] ($\geq$ **32** crops)	0.830	-	-	-
Zeng *et al.* (ResNet101) [45]	0.808	0.275	0.719	0.720
Hosu *et al.* [16] (**20** crops)	0.817	-	**0.756**	**0.757**
AFDC + SPP (single warp) [7]	**0.830**	0.273	0.648	-
AFDC + SPP (**4** warps) [7]	**0.832**	0.271	0.649	0.671
MUSIQ-single	0.814	**0.247**	0.719	0.731
MUSIQ (Ours)	0.815	**0.242**	**0.726**	**0.738**
std	± 0.121	± 0.001	± 0.001	± 0.001

Table 4. Results on AVA dataset. Blue and black numbers in bold represent the best and second best respectively. cls. acc. stands for classification accuracy. MSE stands for mean square error. We take numbers from [7] for results of the reference methods.

Ablation Studies

- Importance of Aspect-Ratio-Preserving (ARP)

method	# Params	SRCC	PLCC
NIMA(Inception-v2) [36] (224 square input)	56M	0.612	0.636
NIMA(ResNet50)* (384 square input)	24M	0.624	0.632
ViT-Base 32* (384 square input) [11]	88M	0.654	0.664
ViT-Small 32* (384 square input) [11]	22M	0.656	0.665
MUSIQ w/ square resizing (512, 384, 224)	27M	0.706	0.720
MUSIQ w/ ARP resizing (512, 384, 224)	27M	0.712	0.726
MUSIQ w/ ARP resizing (full, 384, 224)	27M	**0.726**	**0.738**

Table 5. Comparison of ARP resizing and square resizing on AVA dataset. * means our implementation. ViT-Small* is constructed by replacing the Transformer backbone in ViT with our 384-dim lightweight Transformer. The last group of rows show our method with different resizing methods. Numbers in the bracket show the resolution used in the multi-scale representation.

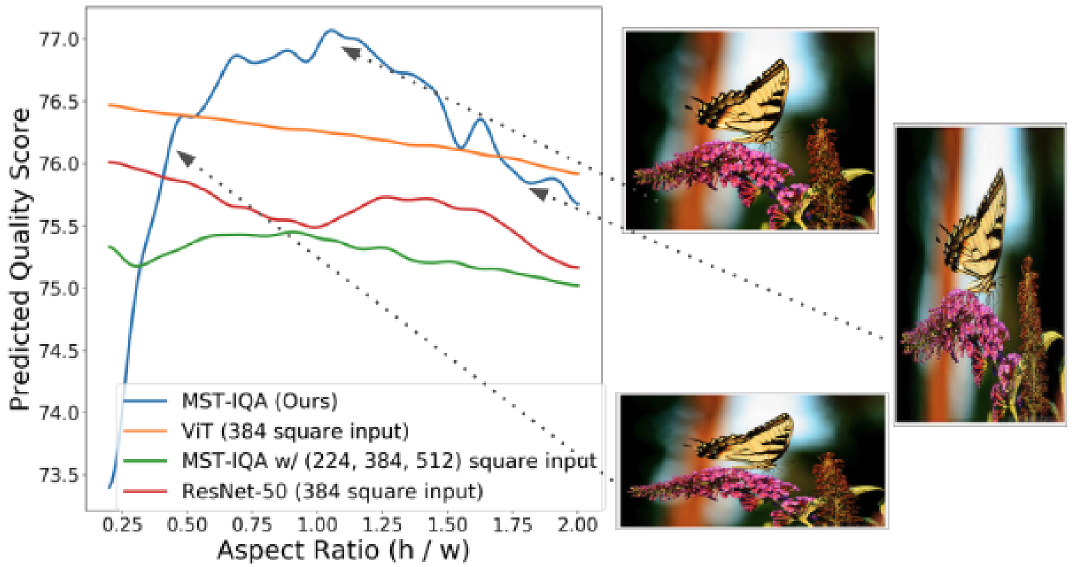


Figure 3. Model predictions for an image resized to different aspect ratios. The blue curve shows MUSIQ with ARP resizing. The green curve shows our model trained and evaluated with square input. Orange and red curves show the ViT and ResNet-50 with square input. MUSIQ can detect quality degradation due to unnatural resizing while other methods are not sensitive.

Ablation Studies

- Effect of Full-size Input and the Multi-scale Input Composition

Multi-scale Composition	SRCC	PLCC
(224)	0.600	0.667
(384)	0.618	0.695
(512)	0.620	0.691
(384, 224)	0.620	0.707
(512, 384, 224)	0.629	0.718
(full)	0.640	0.721
(full, 224)	0.643	0.726
(full, 384)	0.642	0.730
(full, 384, 224)	**0.646**	**0.739**
Average ensemble of (full), (224), (384)	0.640	0.710

Table 6. Comparison of multi-scale representation composition on PaQ-2-PiQ full-size test set. The multi-scale representation is composed of the resolutions shown in the brackets. Numbers in brackets indicate the longer side length L for ARP resizing. "full" means full-size input image.

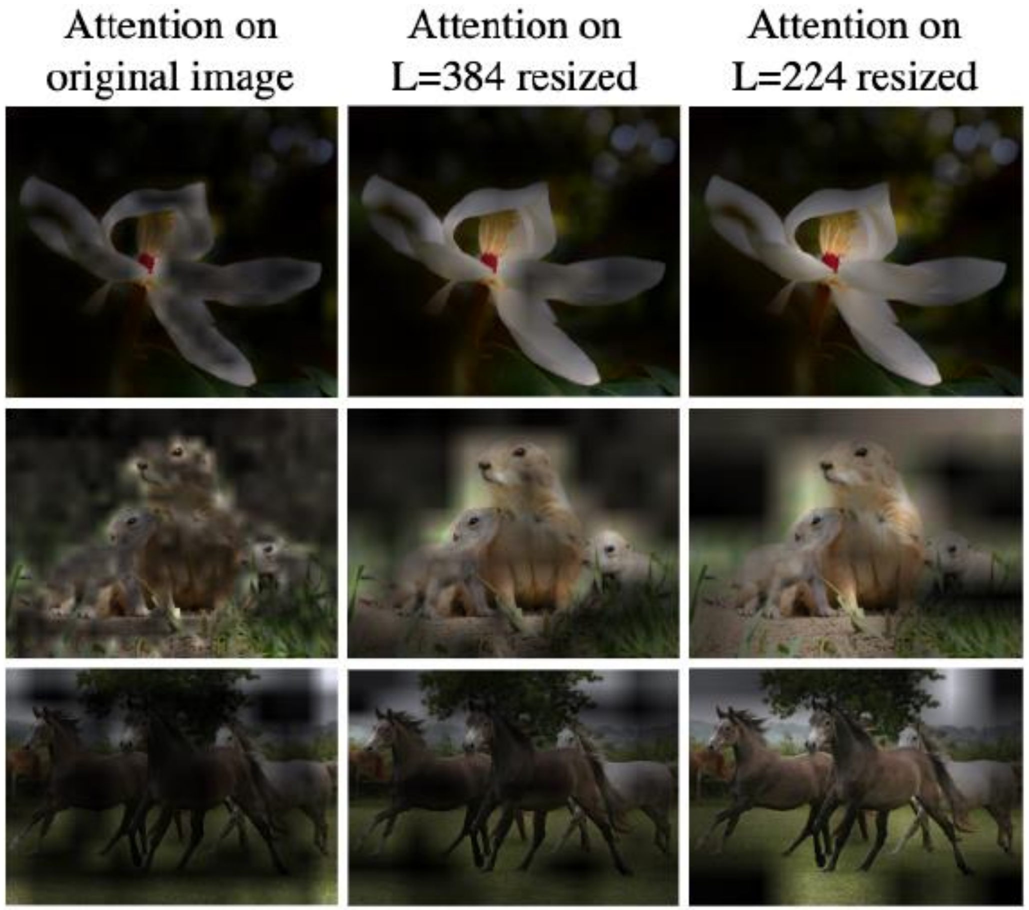


Figure 4. Visualization of attention from the output tokens to the multi-scale representation (original resolution image and two ARP resized variants). Note that images here are resized to fit the grid, the model inputs are 3 different resolutions. The model is focusing on details in higher resolution image and on global area in lower resolution ones.

Spatial Embedding	SRCC	LCC
HPE ($G = 5$)	0.720	0.733
HPE ($G = 8$)	0.723	0.734
HPE ($G = 10$)	**0.726**	**0.738**
HPE ($G = 12$)	0.722	0.736
HPE ($G = 15$)	0.724	0.735
HPE ($G = 20$)	0.722	0.734

Table 12. Ablation study for different grid size G in HSE on AVA dataset.

Patch Size	16	32	48	64
SRCC	0.715	**0.726**	0.713	0.705
PLCC	0.729	**0.738**	0.727	0.719

Table 14. Comparison of different patch size on AVA dataset.

Summary

- 本文在transformer中加入多尺度的图像输入，以及对应的位置编码方式
- 方法较为简单，写作清晰，在附录中提供了较为详细的说明
- 给出了代码链接，但还未放出代码

- transfomer是一种可以代替CNN的框架
  - 视觉任务上效果较好
  - 对长距离的图像块的联系更为方便
  - 为了实现不同的功能需要在embedding上加入需要的信息，或是对head进行不同的约束
- 好训练吗？
  - 训练所需数据量大
  - 直接从头在ImageNet上训练似乎有些难
  - [1]将几个IQA数据集合并进行了从头训练，发现比使用预训练后再微调的效果好

[1] You J, Korhonen J. Transformer for image quality assessment[C]//2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 1389-1393.