Florence: A New Foundation Model for Computer Vision

Lu Yuan et al.

Florence Team, Microsoft Cloud and AI

Presenter: Minho Park

A Wide Range of Downstream Tasks

- Achieve state-of-the-art in various benchmarks.

Benchmarks in 2022.01.14

Contribution

- A new paradigm of building a computer vision foundation model, Florence.

- Florence successfully extends to different tasks along space, time, and modality, with great transferability.

- Achieves new SOTA results on a wide range of vision benchmarks.

Foundation Models

- Any model that is trained from broad data at scale that is capable of being adapted (e.g., fine-tuned) to a wide range of downstream tasks.

- Foundation models become promising due to their impressive performance and generalization capabilities.
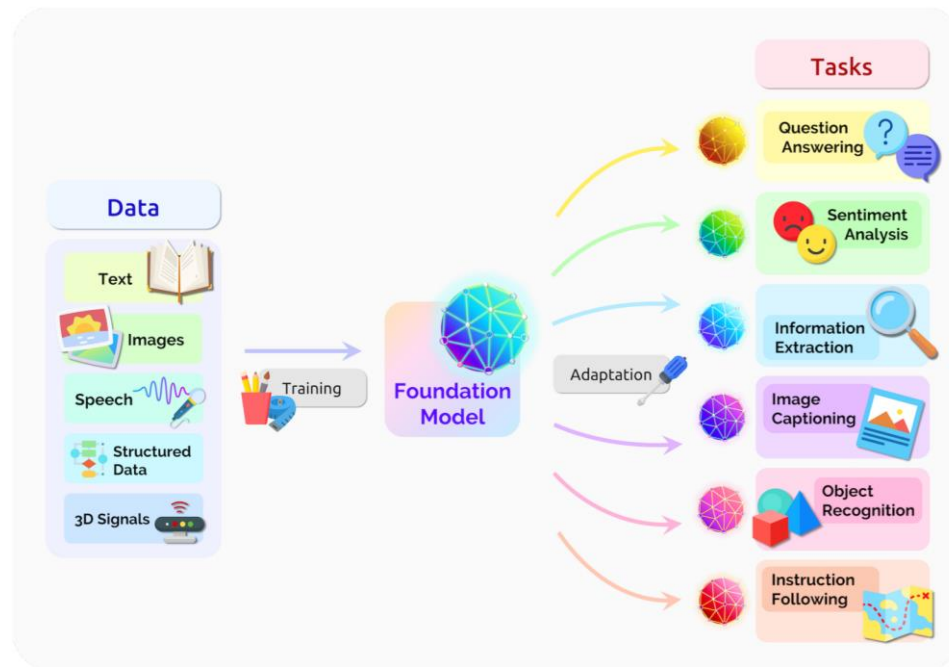
Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models

- NLP: BERT, GPT

- CV: CLIP, ALIGN, Wu Dao

- However, such models are restricted to image to text mapping.

BERT (Devlin et al.)

CLIP (Radford et al.)

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
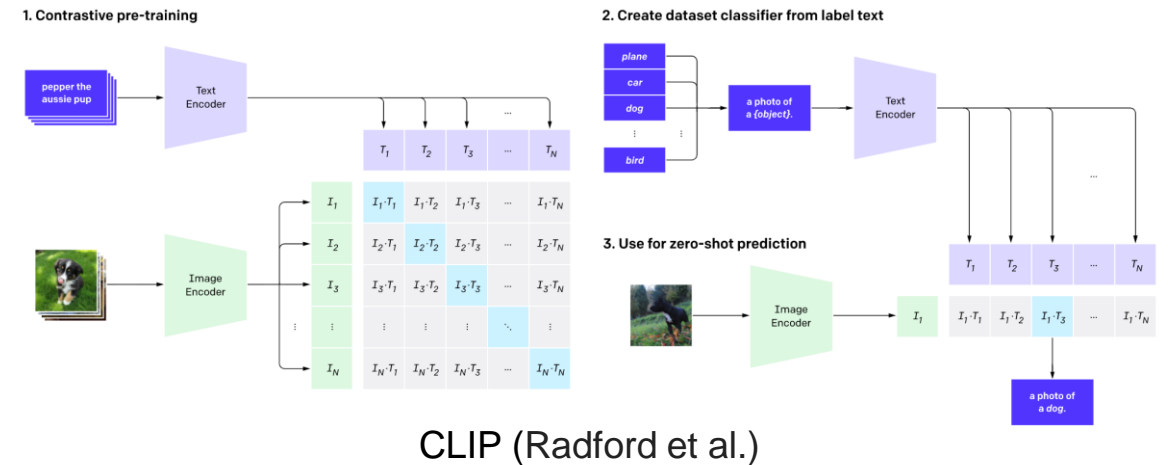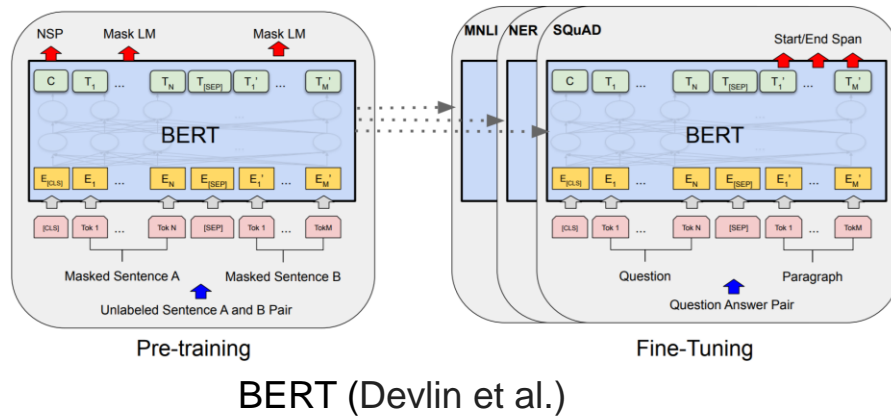
What is the foundation model for computer vision?

- Spectrum of tasks in a problem space.

- Space axis: from coarse to fine-grained.

- Time axis: from static to dynamic.

- Modality axis: from RGB only to multiple senses.

- Three axes are orthogonal.
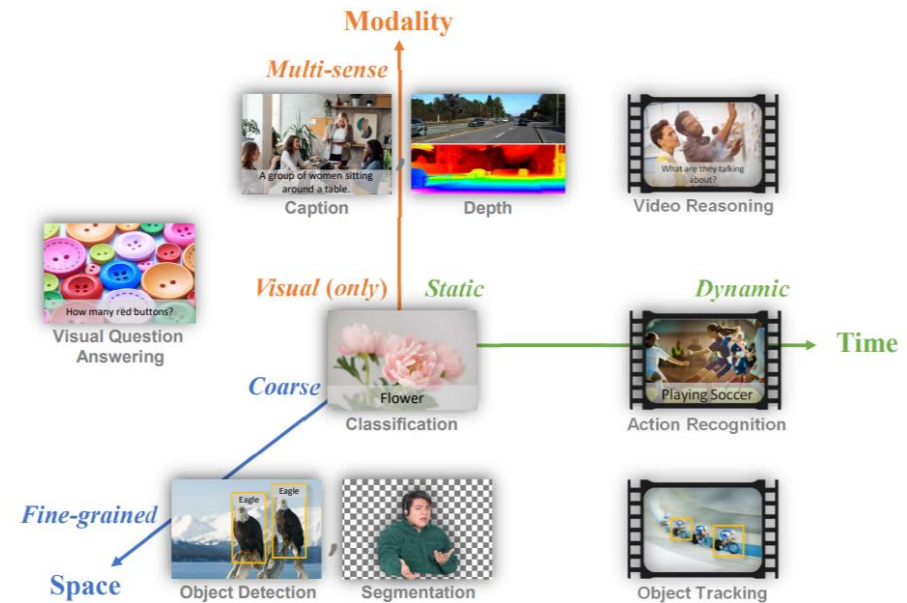
Figure 1. Common computer vision tasks are mapped to a *Space-Time-Modality* space. A computer vision foundation model should serve as general purpose vision system for all of these tasks.

Dataset Curation

- They construct FLD-900M (FLorenceDataset) with 3B image text pairs with selection and post-filtering.
  900M free-form texts (ranging from one word, phase to sentences)

  - 9.7M unique queries and 7.5B tokens in total.

- Legal and ethical constraints.

- Simple hash-based near-duplicate image removal.

- Small-size image removal.

- Image-text relevance.

- Text should include 500,000 queries (CLIP).
  Occurring at least 100 times in the English version of Wikipedia.

- Class balance the results by including up to 20,000 (image, text) pairs per query (CLIP).
  ImageNet

- Consider informativeness learnability (CLIP).

Florence

- Florence = Pretrained Models + Adaptation Models
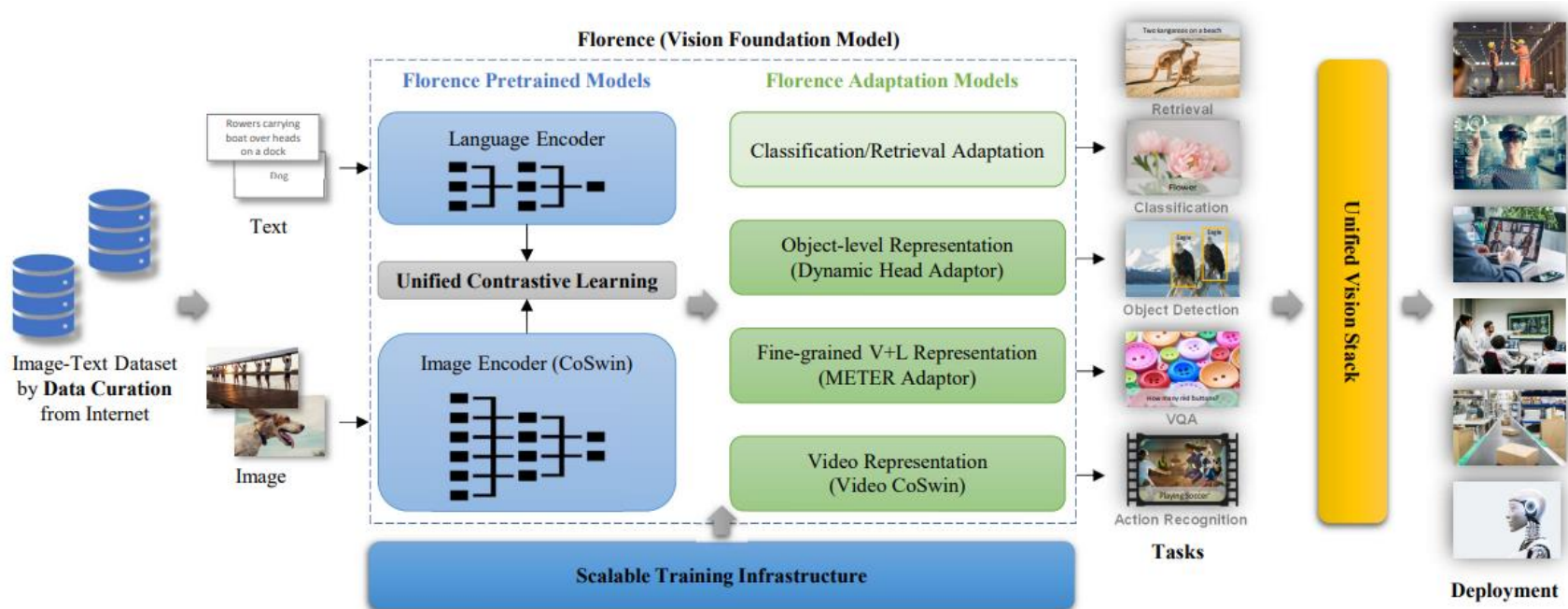
Figure 2. Overview of building Florence. Our workflow consists of data curation, unified learning, Transformer architectures and adaption. It shows the foundation model can be adapted to various downstream tasks and finally integrated into modern computer vision system to power real-world vision and multimedia applications. Compared with existing image-text pretraining models (Radford et al., 2021; Jia et al., 2021; Wud), mainly limited on cross-modal shared representation for classification and retrieval (illustrated by *light-green* adaptation module), Florence expands the representation to support object level, multiple modality, and videos respectively.

Unified Image-Text Contrastive Learning (UniCL)

- CLIP implicitly assumes that each image-text pair has its unique caption.

- However, in web-scale data, multiple images can be associated with identical captions.

- For example, in FLD-900M, there are 350M image-text pairs where there are more than one images corresponding to one identical text.

- UniCL uses image-text pair as a triplet (x, t, y).

  - x is the image, t is the language description (i.e., hash value), and y is the language label (i.e., hash key).
    generating propth templates such as "A photo of the [WORD]"

- Thus, all image-text pairs mapped to the same label y are regarded as positive in our universal image-text contrastive learning. Others are still regarded as negative.

Yang, Jianwei, et al. Unified contrastive learning in image-text-label space. In *arXiv In Preparation*, 2022.

Unified Image-Text Contrastive Learning (UniCL)

- Given a mini-batch $\mathcal{B}$, $\mathcal{L} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}$

$$\mathcal{L}_{i2t} = -\sum_{i \in \mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k \in \mathcal{P}(i)} \log \frac{\exp(\tau u_i v_k)}{\sum_{j \in \mathcal{B}} \exp(\tau u_i v_j)}$$

$$\mathcal{L}_{t2i} = -\sum_{j \in \mathcal{B}} \frac{1}{|\mathcal{Q}(j)|} \sum_{k \in \mathcal{Q}(j)} \log \frac{\exp(\tau u_k v_j)}{\sum_{i \in \mathcal{B}} \exp(\tau u_i v_j)}$$

	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\cdots$	$\mathcal{T}_N$
$\mathcal{I}_1$	1	0	1		0
$\mathcal{I}_2$	0	1	0		0
$\mathcal{I}_3$	1	0	1		0
$\vdots$				$\ddots$	
$\mathcal{I}_N$	0	0	0		1

$\mathcal{L}_{t2i}$ (column), $\mathcal{L}_{i2t}$ (row)

- where f_θ, f_ϕ are image encoder and text encoder, respectively.

- $u = \frac{f_\theta(x)}{\|f_\theta(x)\|}$, $v = \frac{f_\phi(x)}{\|f_\phi(x)\|}$ are normalized visual feature vector, and τ is a learnable temperature.

- $\mathcal{P}(i) = \{k | k \in \mathcal{B}, y_k = y_i\}$, $\mathcal{Q}(j) = \{k | k \in \mathcal{B}, y_k = y_j\}$.

Yang, Jianwei, et al. Unified contrastive learning in image-text-label space. In *arXiv In Preparation*, 2022.

Transformer-based Florence Pretrained Models

- 12-layer transformer + *CoSwin* Transformer (CvT)

Figure 1. (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers [20] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

Swin Transformer

Figure 2. Overview of building Florence. Our workflow consists of data curation, unified learning, Transformer architectures and adaption. It shows the foundation model can be adapted to various downstream tasks and finally integrated into modern computer vision system to power real-world vision and multimedia applications. Compared with existing image-text pretraining models (Radford et al., 2021; Jia et al., 2021; Wud), mainly limited on cross-modal shared representation for classification and retrieval (illustrated by *light-green* adaptation module), Florence expands the representation to support object level, multiple modality, and videos respectively.

Wu, Haiping, et al. "Cvt: Introducing convolutions to vision transformers." *arXiv preprint arXiv:2103.15808* (2021).
Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *arXiv preprint arXiv:2103.14030* (2021).

Florence Adaptation Models

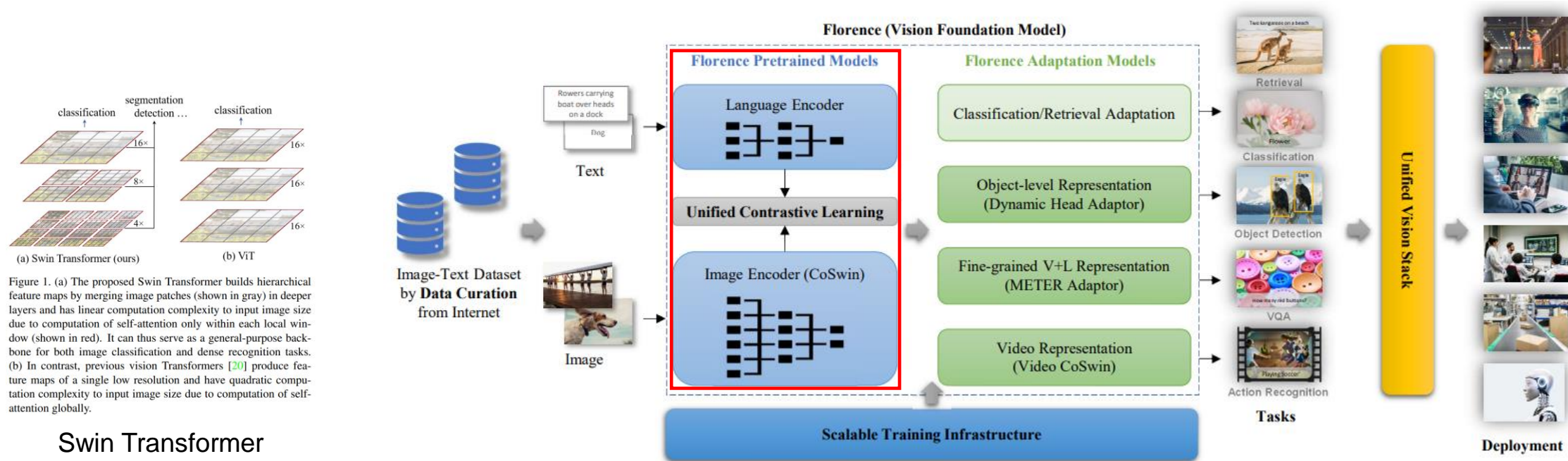- There are three additional adaptations (space, time, and modality axes).
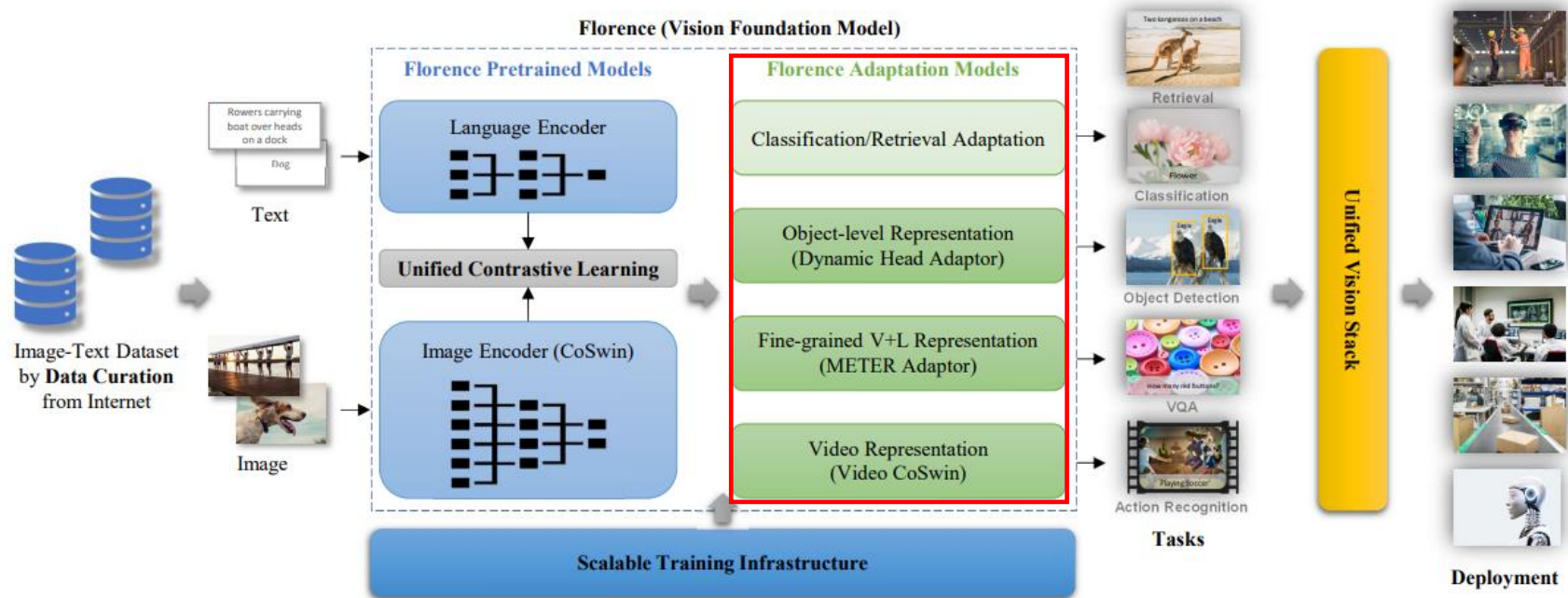
Figure 2. Overview of building Florence. Our workflow consists of data curation, unified learning, Transformer architectures and adaption. It shows the foundation model can be adapted to various downstream tasks and finally integrated into modern computer vision system to power real-world vision and multimedia applications. Compared with existing image-text pretraining models (Radford et al., 2021; Jia et al., 2021; Wud), mainly limited on cross-modal shared representation for classification and retrieval (illustrated by *light-green* adaptation module), Florence expands the representation to support object level, multiple modality, and videos respectively.

Object-level Visual Representation Learning

- They add an adaptor Dynamic Head (Dai et al., 2021a) (or Dynamic DETR (Dai et al., 2021b)), a unified attention mechanism for the detection head, to the pretrained image encoder (i.e., CoSwin).

- Deploy three attention mechanisms, each on one of the orthogonal dimensions of the tensor, i.e., level-wise, spatial-wise, and channel wise.

- FLOD-9M for object detection pre-training.
  - Merge several well-known object detection datasets, including COCO, LVIS, OpenImages, and Object365.
  - In addition, they generate pseudo bounding boxes on ImageNet-22K
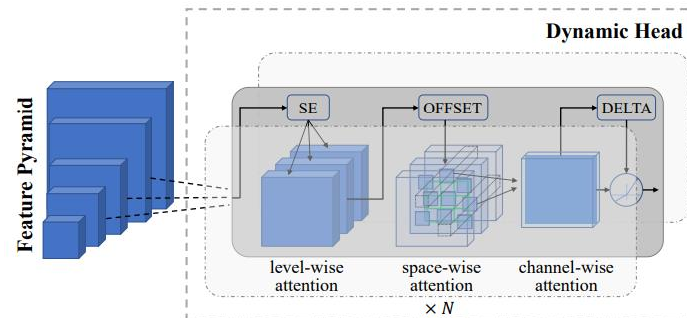
Figure 3. Dynamic Head (Dai et al., 2021a) adapter is used for object-level visual representation learning.

Dai, Xiyang, et al. "Dynamic Head: Unifying Object Detection Heads with Attentions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
Dai, Xiyang, et al. "Dynamic detr: End-to-end object detection with dynamic attention." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Fine-Grained V+L Representation Learning

- They use METER (Dou et al., 2021) adapter to expand to fine-grained vision-language representation (e.g., visual question answering (VQA), image captioning, and fine-grained representation).

- They replace the image encoder of METER with Florence pretrained model CoSwin, and use a pretrained RoBERTa (Liu et al., 2019) as the language encoder.
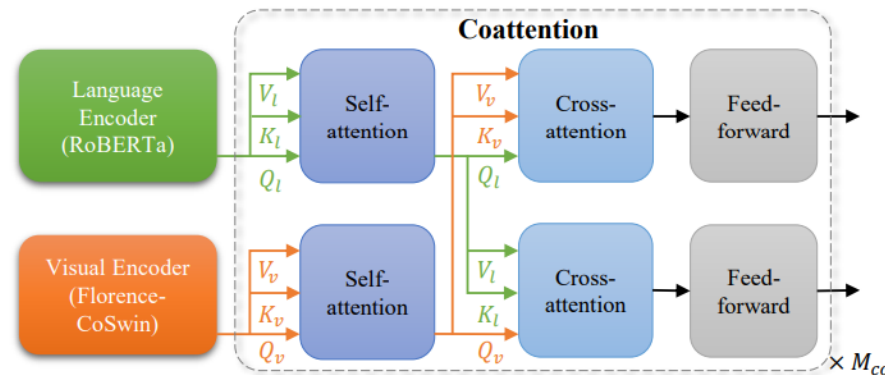
Figure 4. METER (Dou et al., 2021) is used as Florence V+L adaptation model, trained with the image-text matching (ITM) loss and the masked language modeling (MLM) loss.

Dou, Zi-Yi, et al. "An Empirical Study of Training End-to-End Vision-and-Language Transformers." *arXiv preprint arXiv:2111.02387* (2021).
Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

14

Adaptation to Video Recognition

- Video CoSwin: They use 3d convolutional layers instead of 3d patch merging and 2d patch embedding.

- As the initialization to 3D convolutional weights, the pre-trained 2D convolutional weights of CoSwin are duplicated along the temporal dimension and divided by the temporal kernel size to keep the mean and variance of the output unchanged.
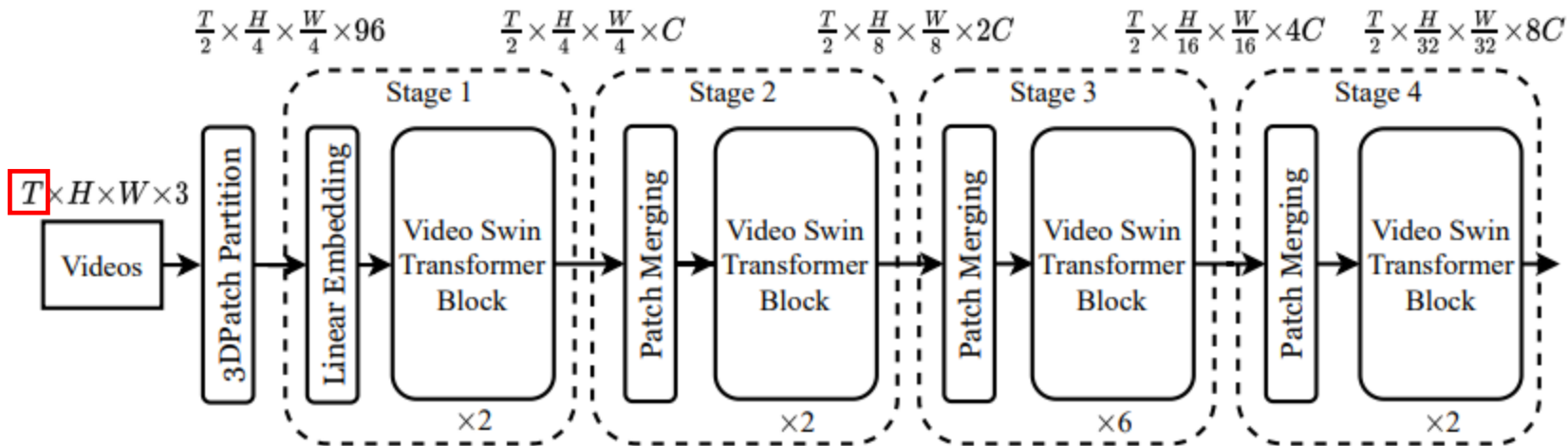
Figure 1: Overall architecture of Video Swin Transformer (tiny version, referred to as Swin-T).

Liu, Zhenhua, et al. "Post-training quantization for vision transformer." *Advances in Neural Information Processing Systems* 34 (2021).   15

Experiments

- Zero-shot Transfer in Classification

- Linear Probe in Classification

- ImageNet-1K Fine-tune Evaluation

- Few-shot Cross-domain Classification

- Image-Text Retrieval

- Object Detection and Zero-shot Transfer

- V+L Representation Learning

- Zero-shot Text-to-Video Retrieval

- Video Action Recognition

Zero-shot Transfer in Classification

- Florence can be directly used to predict if an image and a text snippet are semantically matched together in the task dataset (same method of CLIP).

	Food101	CIFAR10	CIFAR100	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	ImageNet
CLIP-ResNet-50x64	91.8	86.8	61.3	48.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	73.6
CLIP-ViT-L/14 (@336pix)	93.8	**95.7**	77.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	76.2
FLIP-ViT-L/14	92.2	**95.7**	75.3	73.1	70.8	**60.2**	-	60.7	92.0	93.0	**90.1**	78.3
Florence-CoSwin-H (@384pix)	**95.1**	94.6	**77.6**	**77.0**	**93.2**	55.5	**85.5**	**66.4**	**95.9**	**94.7**	86.2	**83.7**

Table 1. Zero-shot transfer of image classification comparisons on 12 datasets: CLIP-ResNet-50x64 (Radford et al., 2021), FLIP-ViT-L/14 (Yao et al., 2021).

Linear Probe in Classification

- Linear probe as another main metric for evaluating representation quality has been used in most recent studies

	Food101	CIFAR10	CIFAR100	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102
SimCLRv2-ResNet-152x3	83.6	96.8	84.5	69.1	68.5	63.1	86.7	80.5	92.6	94.9	96.3
ViT-L/16 (@384pix)	87.4	97.9	89.0	74.9	62.5	52.2	86.1	75.0	92.9	94.7	99.3
EfficientNet-L2 (@800pix)	92.0	**98.7**	**89.0**	75.7	75.5	68.4	89.4	82.5	95.6	94.7	97.9
CLIP-ResNet-50x64	94.8	94.1	78.6	81.1	90.5	67.7	88.9	82.0	94.5	95.4	98.9
CLIP-ViT-L/14 (@336pix)	95.9	97.9	87.4	82.2	91.5	71.6	89.9	83.0	95.1	96.0	99.2
Florence-CoSwin-H (@384pix)	**96.2**	97.6	87.1	**84.2**	**95.7**	**83.9**	**90.5**	**86.0**	**96.4**	**96.6**	**99.7**

Table 2. Comparisons of image classification linear probing on 11 datasets with existing state-of-the-art models, including Sim-CLRv2 (Chen et al., 2020c), ViT (Dosovitskiy et al., 2021a), EfficientNet (Xie et al., 2020), and CLIP (Radford et al., 2021).

ImageNet-1K Fine-tune Evaluation

- Florence can be easily adapted to support continual finetuning on target classification tasks.

- State-of-the-art on top-5 accuracy (w/o JFT dataset).

Model	Params	Data	Accuracy Top-1	Top-5
BiT-L-ResNet152x4	928M	300M	87.54	98.46
ALIGN-Efficient-L2	480M	1800M	88.64	98.67
ViT-G/14	1843M	3000M	90.45	-
CoAtNet-7	2440M	3000M	**90.88**	-
Florence-CoSwin-H	637M	900M	90.05	**99.02**

Table 3. Classification fine tuning on ImageNet-1K. Florence is compared with: BiT-L-ResNet152x4 (Kolesnikov et al., 2020), ALIGN-Efficient-L2 (Jia et al., 2021), ViT-G/14 (Zhai et al., 2021), CoAtNet-7 (Dai et al., 2021c) in terms of model scale, data scale and Top-1/Top-5 accuracy.

Few-shot Cross-domain Classification

- The Cross-Domain Few-Shot learning benchmark (Guo et al., 2020) is used to measure an algorithm's capability to adapt to downstream few-shot target tasks.

- To predict the class, we append a single linear layer as an adapter head to our image encoder CoSwin.

- Training occurs over 100 epochs per episode.
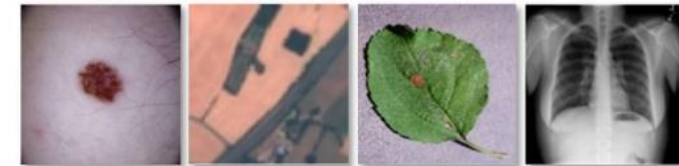
		ISIC	EuroSAT	CropDisease	ChestX	

	Model	ISIC	EuroSAT	CropD	ChestX	mean
5-shot	CW	57.4	88.1	96.6	29.7	68.0
	Florence	57.1	90.0	97.7	29.3	**68.5**
20-shot	CW	68.1	94.7	99.2	38.3	75.1
	Florence	72.9	95.8	99.3	37.5	**76.4**
50-shot	CW	74.1	96.9	99.7	44.4	78.8
	Florence	78.3	97.1	99.6	42.8	**79.5**

employs ensembes and transductive learning.

Table 4. Comparison with CW (Liu et al., 2020) (CD-FSL Challenge 2020 Winner) on CD-FSL benchmark. The average result comparison is 74.8 (Florence) vs. 73.9 (CW).

Image-Text Retrieval

- The zero-shot transfer and fine-tuning performance of Florence for both text and image retrieval.

	Method	Flickr30K (1K test set)				MSCOCO (5K test set)			
		Image → Text		Text → Image		Image → Text		Text → Image	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Zero-shot	ImageBERT (Qi et al., 2020)	70.7	90.2	54.3	79.6	44.0	71.2	32.3	59.0
	UNITER (Chen et al., 2020d)	83.6	95.7	68.7	89.2	-	-	-	-
	CLIP (Radford et al., 2021)	88.0	98.7	68.7	90.6	58.4	81.5	37.8	62.4
	ALIGN (Jia et al., 2021)	88.6	98.7	75.7	**93.8**	58.6	83.0	45.6	69.8
	FLIP (Yao et al., 2021)	89.8	**99.2**	75.0	93.4	61.3	84.3	45.9	70.6
	Florence	**90.9**	99.1	**76.7**	93.6	**64.7**	**85.9**	**47.2**	**71.4**
Fine-tuned	GPO (Chen et al., 2020a)	88.7	98.9	76.1	94.5	68.1	90.2	52.7	80.2
	UNITER (Chen et al., 2020d)	87.3	98.0	75.6	94.1	65.7	88.6	52.9	79.9
	ERNIE-ViL (Yu et al., 2020)	88.1	98.0	76.7	93.6	-	-	-	-
	VILLA (Gan et al., 2020)	87.9	97.5	76.3	94.2	-	-	-	-
	Oscar (Li et al., 2020)	-	-	-	-	73.5	92.2	57.5	82.8
	ALIGN (Jia et al., 2021)	95.3	99.8	84.9	97.4	77.0	93.5	59.9	83.3
	FLIP (Yao et al., 2021)	96.6	**100.0**	87.1	97.7	78.9	94.4	61.2	84.3
	Florence	**97.2**	**99.9**	**87.9**	**98.1**	**81.8**	**95.2**	**63.2**	**85.7**

Table 5. Image-text retrieval comparisons on Flickr30K and MSCOCO datasets (zero-shot and fine-tuned).

Object Detection and Zero-shot Transfer

- Florence is more desirable for object detection since its adaptation helps learn visual representation at the object level.

- Fine-tuning

Benchmark	Model	AP
COCO miniVal	DyHead	60.3
	Soft Teacher	60.7
	Florence	**62.0**
COCO test-Dev	DyHead	60.6
	Soft Teacher	61.3
	Florence	**62.4**
Object365	Multi-dataset Detection	33.7
	Florence	**39.3**
Visual Genome	VinVL	13.8
	Florence	**16.2**

Table 6. Object detection fine tuning comparisons with state-of-the-art methods, including DyHead (Dai et al., 2021a), Soft Teacher (Xu et al., 2021b), Multi-dataset Detection (Zhou et al., 2021), VinVL (Zhang et al., 2021b).

Object Detection and Zero-shot Transfer

- In our zero-shot transfer setting, object proposal and object classification are decoupled into two tasks.

    - Object proposal discriminates object from background, ignoring semantics of object categories.

    - Classification, on the other hand, focuses on object semantics for each bounding box proposal.

- They freeze the CoSwin backbones and pre-train the Dynamic Head on FLOD-9M by neglecting semantics from each object bounding box.

		Aquarium	BCCD	Chess Pieces	Mask Wearing	Oxford Pets	Packages	Pistols	PKLot	Pothole	Thermal	Wildfire Smoke
	Images	638	364	292	149	3680	26	2986	12416	665	203	737
	Categories	7	3	12	2	37	1	1	2	1	2	1
Fine-tuned	DyHead-Swin-L (full)	53.1	62.6	80.7	52.0	85.9	52.0	74.4	98.0	61.8	75.9	58.7
	DyHead-Swin-L (5-shot)	39.0	40.6	57.3	26.8	47.5	32.8	20.0	22.1	10.8	54.9	14.2
Zero-shot	ZSD	16.0	1.2	0.1	0.6	0.3	58.3	31.5	0.2	2.4	37.4	0.002
	Florence	43.1	15.3	13.4	15.0	68.9	79.6	41.4	31.4	53.3	46.9	48.7

Table 7. Zero-shot transfer in object detection, in comparison with previous state-of-the-art model DyHead (Dai et al., 2021a) (on COCO) fine tuning results on full-set or 5-shot respectively and zero-shot detection baseline model ZSD (Bansal et al., 2018).

V+L Representation Learning

- To evaluate the performance, we fine-tune the pre-trained model on the challenging VQA (Goyal et al., 2017) task

Model	test-dev	test-std
UNITER (Chen et al., 2020d)	73.82	74.02
Visual Parsing (Xue et al., 2021)	74.00	74.17
PixelBERT (Huang et al., 2020)	74.45	74.55
VILLA (Gan et al., 2020)	74.69	74.87
UNIMO (Li et al., 2021c)	75.06	75.27
ALBEF (Li et al., 2021a)	75.84	76.04
VinVL (Zhang et al., 2021b)	76.52	76.60
CLIP-ViL (Shen et al., 2021)	76.48	76.70
METER (Dou et al., 2021)	77.68	77.64
SimVLM (Wang et al., 2021)	80.03	80.34
Florence	**80.16**	**80.36**

Table 8. Compare our model with the existing state-of-the-art methods on VQA.

Goyal, Yash, et al. "Making the v in vqa matter: Elevating the role of image understanding in visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Zero-shot Text-to-Video Retrieval

- Although Florence is pre-trained on image-text pairs, it can be easily adapted to video tasks.

Method	Pre-training Type	Pre-training Data	R@1	R@5	R@10
MIL-NCE (Miech et al., 2020)	Video	HowTo100M	-	-	32.4
MMV (Alayrac et al., 2020)	Video	HowTo100M, AudioSet	-	-	31.1
VideoCLIP (Xu et al., 2021a)	Video*	HowTo100M	10.4	22.2	30.0
VATT (Akbari et al., 2021)	Video	HowTo100M, AudioSet	-	-	29.7
MCN (Chen et al., 2021)	Image and Video	HowTo100M	-	-	33.8
Frozen-in-Time (Bain et al., 2021)	Image and Video	ImageNet, CC, WebVid-2M	18.7	39.5	51.6
CLIP-ViT-B/16 (Radford et al., 2021)	Image	WIT400M	26.0	49.4	60.7
Florence	Image	FLD-900M	**37.6**	**63.8**	**72.6**

Table 9. Zero-shot text-to-video retrieval results on MSR-VTT 1K-A test set. (*: Feature extracted from the pre-trained model (Miech et al., 2020), followed by another stage of video-and-language pre-training) The pretraining data used in these existing methods include HowTo100M (Miech et al., 2019), AudioSet (Gemmeke et al., 2017), ImageNet (Deng et al., 2009), CC (Sharma et al., 2018), WebVid-2M (Bain et al., 2021), WIT400M (Radford et al., 2021)

Video Action Recognition

- They evaluate Florence on fine-tuned video action recognition tasks.

Method	Pretraining Data	Kinetics-400		Kinetics-600		Views	Params
		Top-1	Top-5	Top-1	Top-5		
ViViT-H/16x2	JFT-300M	84.8	95.8	85.8	96.5	4 × 3	648M
VideoSwin-L	ImageNet-22K	84.6	96.5	85.9	97.1	4 × 3	200M
VideoSwin-L	ImageNet-22K	84.9	96.7	86.1	97.3	10 × 5	200M
TokenLearner 16at18+L/10	JFT-300M	85.4	96.3	86.3	97.0	4 × 3	460M
Florence	FLD-900M	**86.5**	**97.3**	**87.8**	**97.8**	4 × 3	647M

Table 10. Comparison to state-of-the-art methods, including ViViT (Arnab et al., 2021), VideoSwin (Liu et al., 2021b), TokenLearner (Ryoo et al., 2021), on Kinetics-400 and Kinetics-600. Views indicate *#temporal clip × #spatial crop.*

Conclusion

- A new paradigm of building a computer vision foundation model.

- Although the model size is still below several other existing billion-scale models,

- Florence successfully extends to different tasks along space, time, and modality, with great transferability.

- Achieves new SOTA results on a wide range of vision benchmarks.