## **Related Work**

### **Perception in Bird’s-Eye-Space**

#### Lift-Splat-Shoot: lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D

The goal of perception for autonomous vehicles is to extract semantic representations from multiple sensors and fuse these representations into a single “bird’s-eye-view” coordinate frame for consumption by motion planning. We propose a new end-to-end architecture that directly extracts a bird’s-eye-view representation of a scene given image data from an arbitrary number of cameras. The core idea behind our approach is to “lift” each image individually into a frustum of features for each camera, then “splat” all frustums into a rasterized bird’s-eyeview grid. By training on the entire camera rig, we provide evidence that our model is able to learn not only how to represent images but how to fuse predictions from all cameras into a single cohesive representation of the scene while being robust to calibration error. On standard bird’seye-view tasks such as object segmentation and map segmentation, our model outperforms all baselines and prior work. In pursuit of the goal of learning dense representations for motion planning, we show that the representations inferred by our model enable interpretable end-to-end motion planning by “shooting” template trajectories into a bird’s-eyeview cost map output by our network. We benchmark our approach against models that use oracle depth from lidar.

#### BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers

3D visual perception tasks, including 3D detection and map segmentation based on multi-camera images, are essential for autonomous driving systems. In this work, we present a new framework termed BEVFormer, which learns unified BEV representations with spatiotemporal transformers to support multiple autonomous driving perception tasks. In a nutshell, BEVFormer exploits both spatial and temporal information by interacting with spatial and temporal space through predefined grid-shaped BEV queries. To aggregate spatial information, we design a spatial cross-attention that each BEV query extracts the spatial features from the regions of interest across camera views. For temporal information, we propose a temporal self-attention to recurrently fuse the history BEV information. Our approach achieves the new state-of-the-art 56.9% in terms of NDS metric on the nuScenes test set, which is 9.0 points higher than previous best arts and on par with the performance of LiDAR-based baselines. We further show that BEVFormer remarkably improves the accuracy of velocity estimation and recall of objects under low visibility conditions.

#### BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View

Abstract

Autonomous driving perceives its surroundings for decision making, which is one of the most complex scenarios in visual perception. The success of paradigm innovation in solving the 2D object detection task inspires us to seek an elegant, feasible, and scalable paradigm for fundamentally pushing the performance boundary in this area. To this end, we contribute the BEVDet paradigm in this paper. BEVDet performs 3D object detection in Bird-Eye-View (BEV), where most target values are defined and route planning can be handily performed. We merely reuse existing modules to build its framework but substantially develop its performance by constructing an exclusive data augmentation strategy and upgrading the Non-Maximum Suppression strategy. In the experiment, BEVDet offers an excellent trade-off between accuracy and time-efficiency. As a fast version, BEVDet-Tiny scores 31.2% mAP and 39.2% NDS on the nuScenes val set. It is comparable with FCOS3D, but requires just 11% computational budget of 215.3 GFLOPs and runs 9.2 times faster at 15.6 FPS. Another high-precision version dubbed BEVDet-Base scores 39.3% mAP and 47.2% NDS, significantly exceeding all published results. With a comparable inference speed, it surpasses FCOS3D by a large margin of +9.8% mAP and +10.0% NDS.

#### BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection

Abstract:

Single frame data contains finite information which limits the performance of the existing vision-based multicamera 3D object detection paradigms. For fundamentally pushing the performance boundary in this area, a novel paradigm dubbed BEVDet4D is proposed to lift the scalable BEVDet paradigm from the spatial-only 3D working space into the spatial-temporal 4D working space. We upgrade the naive BEVDet framework with a few modifications just for fusing the feature from the previous frame with the corresponding one in the current frame. In this way, with negligible additional computing budget, we enable BEVDet4D to access the temporal cues by querying and comparing the two candidate features. Beyond this, we simplify the task of velocity prediction by degenerating it into the positional offset prediction in the two adjacent features. As a result, BEVDet4D with robust generalization performance reduces the velocity error by up to -62.9%. This makes the vision-based methods, for the first time, become comparable with those relied on LiDAR or radar in this aspect. On challenge benchmark nuScenes, we report a new record of 54.5% NDS with the high-performance configuration dubbed BEVDet4D-Base. At the same inference speed, this notably surpasses the previous leading method BEVDet-Base by +7.3% NDS.

#### BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection

Abstract

In this research, we propose a new 3D object detector with a trustworthy depth estimation, dubbed BEVDepth, for camera-based Bird’s-Eye-View (BEV) 3D object detection. By a thorough analysis of recent approaches, we discover that the depth estimation is implicitly learned without camera information, making it the de-facto fake-depth for creating the following pseudo point cloud. BEVDepth gets explicit depth supervision utilizing encoded intrinsic and extrinsic parameters. A depth correction sub-network is further introduced to counteract projecting-induced disturbances in depth ground truth. To reduce the speed bottleneck while projecting features from image-view into BEV using estimated depth, a quick view-transform operation is also proposed. Besides, our BEVDepth can be easily extended with input from multi-frame. Without any bells and whistles, BEVDepth achieves the new state-of-the-art 60.0% NDS on the challenging nuScenes test set while maintaining high efficiency. For the first time, the performance gap between the camera and LiDAR is largely reduced within 10% NDS.

.

#### FIERY: Future Instance Segmentation in Bird's-Eye view from Surround Monocular Cameras

Driving requires interacting with road agents and predicting their future behaviour in order to navigate safely. We present FIERY: a probabilistic future prediction model in bird’s-eye view from monocular cameras. Our model predicts future instance segmentation and motion of dynamic agents that can be transformed into non-parametric future trajectories. Our approach combines the perception, sensor fusion and prediction components of a traditional autonomous driving stack by estimating bird’s-eye-view prediction directly from surround RGB monocular camera inputs. FIERY learns to model the inherent stochastic nature of the future solely from camera driving data in an end-toend manner, without relying on HD maps, and predicts multimodal future trajectories. We show that our model outperforms previous prediction baselines on the NuScenes and Lyft datasets.

#### BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving

Abstract

In this paper, we present BEVerse, a unified framework for 3D perception and prediction based on multicamera systems. Unlike existing studies focusing on the improvement of single-task approaches, BEVerse features in producing spatio-temporal Birds-Eye-View (BEV) representations from multi-camera videos and jointly reasoning about multiple tasks for vision-centric autonomous driving. Specifically, BEVerse first performs shared feature extraction and lifting to generate 4D BEV representations from multi-timestamp and multi-view images. After the egomotion alignment, the spatio-temporal encoder is utilized for further feature extraction in BEV. Finally, multiple task decoders are attached for joint reasoning and prediction. Within the decoders, we propose the grid sampler to generate BEV features with different ranges and granularities for different tasks. Also, we design the method of iterative flow for memory-efficient future prediction. We show that the temporal information improves 3D object detection and semantic map construction, while the multi-task learning can implicitly benefit motion prediction. With extensive experiments on the nuScenes dataset, we show that the multi-task BEVerse outperforms existing single-task methods on 3D object detection, semantic map construction, and motion prediction. Compared with the sequential paradigm, BEVerse also favors in significantly improved efficiency.

#### M2BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Bird’s-Eye View Representation

Abstract:

In this paper, we propose M2BEV, a unified framework that jointly performs 3D object detection and map segmentation in the Bird’s Eye View (BEV) space with multi-camera image inputs. Unlike the majority of previous works which separately process detection and segmentation, M2BEV infers both tasks with a unified model and improves efficiency. M2BEV efficiently transforms multi-view 2D image features into the 3D BEV feature in ego-car coordinates. Such BEV representation is important as it enables different tasks to share a single encoder. Our framework further contains four important designs that benefit both accuracy and efficiency: (1) An efficient BEV encoder design that reduces the spatial dimension of a voxel feature map. (2) A dynamic box assignment strategy that uses learning-to-match to assign ground-truth 3D boxes with anchors. (3) A BEV centerness re-weighting that reinforces with larger weights for more distant predictions, and (4) Large-scale 2D detection pre-training and auxiliary supervision. We show that these designs significantly benefit the ill-posed camera-based 3D perception tasks where depth information is missing. M2BEV is memory efficient, allowing significantly higher resolution images as input, with faster inference speed. Experiments on nuScenes show that M2BEV achieves state-of-the-art results in both 3D object detection and BEV segmentation, with the best single model achieving 42.5 mAP and 57.0 mIoU in these two tasks, respectively.

#### HDMapNet: An Online HD Map Construction and Evaluation Framework

Abstract:

Constructing HD semantic maps is a central component of autonomous driving. However, traditional pipelines require a vast amount of human efforts and resources in annotating and maintaining the semantics in the map, which limits its scalability. In this paper, we introduce the problem of HD semantic map learning, which dynamically constructs the local semantics based on onboard sensor observations. Meanwhile, we introduce a semantic map learning method, dubbed HDMapNet. HDMapNet encodes image features from surrounding cameras and/or point clouds from LiDAR, and predicts vectorized map elements in the bird’s-eye view. We benchmark HDMapNet on nuScenes dataset and show that in all settings, it performs better than baseline methods. Of note, our camera-LiDAR fusion-based HDMapNet outperforms existing methods by more than 50% in all metrics. In addition, we develop semantic-level and instance-level metrics to evaluate the map learning performance. Finally, we showcase our method is capable of predicting a locally consistent map. By introducing the method and metrics, we invite the community to study this novel map learning problem.

#### VECTORMAPNET: END-TO-END VECTORIZED HD MAP LEARNING

Autonomous driving systems require a good understanding of surrounding environments, including moving obstacles and static High-Definition (HD) semantic map elements. Existing methods approach the semantic map problem by offline manual annotation, which suffers from serious scalability issues. Recent learning-based methods produce dense rasterized segmentation predictions to construct maps. However, these predictions do not include instance information of individual map elements and require heuristic post-processing to obtain vectorized maps. To tackle these challenges, we introduce an end-to-end vectorized HD map learning pipeline, termed VectorMapNet. VectorMapNet takes onboard sensor observations and predicts a sparse set of polylines in the bird’s-eye view. This pipeline can explicitly model the spatial relation between map elements and generate vectorized maps that are friendly to downstream autonomous driving tasks. Extensive experiments show that VectorMapNet achieve strong map learning performance on both nuScenes and Argoverse2 dataset, surpassing previous state-of-the-art methods by 14.2 mAP and 14.6mAP. Qualitatively, we also show that VectorMapNet is capable of generating comprehensive maps and capturing more fine-grained details of road geometry. To the best of our knowledge, VectorMapNet is the first work designed towards end-to-end vectorized map learning from onboard observations.

#### BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird’s-Eye View Representation

Multi-sensor fusion is essential for an accurate and reliable autonomous driving system. Recent approaches are based on point-level fusion: augmenting the LiDAR point cloud with camera features. However, the camera-to-LiDAR projection throws away the semantic density of camera features, hindering the effectiveness of such methods, especially for semantic-oriented tasks (such as 3D scene segmentation). In this paper, we break this deeply-rooted convention with BEVFusion, an efficient and generic multi-task multi-sensor fusion framework. It unifies multimodal features in the shared bird’s-eye view (BEV) representation space, which nicely preserves both geometric and semantic information. To achieve this, we diagnose and lift key efficiency bottlenecks in the view transformation with optimized BEV pooling, reducing latency by more than 40×. BEVFusion is fundamentally task-agnostic and seamlessly supports different 3D perception tasks with almost no architectural changes. It establishes the new state of the art on nuScenes, achieving 1.3% higher mAP and NDS on 3D object detection and 13.6% higher mIoU on BEV map segmentation, with 1.9× lower computation cost.

Simple-BEV: What Really Matters for Multi-Sensor BEV Perception?

Lidar and camera are both suseptible to occulusion/bad weathers.

Abstract— Building 3D perception systems for autonomous vehicles that do not rely on high-density LiDAR is a critical research problem because of the expense of LiDAR systems compared to cameras and other sensors. Recent research has developed a variety of camera-only methods, where features are differentiably “lifted” from the multi-camera images onto the 2D ground plane, yielding a “bird’s eye view” (BEV) feature representation of the 3D space around the vehicle. This line of work has produced a variety of novel “lifting” methods, but we observe that other details in the training setups have shifted at the same time, making it unclear what really matters in top-performing methods. We also observe that using cameras alone is not a real-world constraint, considering that additional sensors like radar have been integrated into real vehicles for years already. In this paper, we first of all attempt to elucidate the high-impact factors in the design and training protocol of BEV perception models. We find that batch size and input resolution greatly affect performance, while lifting strategies have a more modest effect—even a simple parameter-free lifter works well. Second, we demonstrate that radar data can provide a substantial boost to performance, helping to close the gap between camera-only and LiDAR-enabled systems. We analyze the radar usage details that lead to good performance, and invite the community to re-consider this commonly-neglected part of the sensor platform.

#### Fast-BEV: Towards Real-time On-vehicle Bird’s-Eye View Perception

Data augmentation, pre-computation of 2d-3d transformation matrix, dense voxel projection.

Recently, the pure camera-based Bird’s-Eye-View (BEV) perception removes expensive Lidar sensors, making it a feasible solution for economical autonomous driving. However, most existing BEV solutions either suffer from modest performance or require considerable resources to execute on-vehicle inference. This paper proposes a simple yet effective framework, termed Fast-BEV, which is capable of performing real-time BEV perception on the on-vehicle chips. Towards this goal, we first empirically find that the BEV representation can be sufficiently powerful without expensive view transformation or depth representation. Starting from M2BEV [19] baseline, we further introduce (1) a strong data augmentation strategy for both image and BEV space to avoid over-fitting (2) a multi-frame feature fusion mechanism to leverage the temporal information (3) an optimized deployment-friendly view transformation to speed up the inference. Through experiments, we show Fast-BEV model family achieves considerable accuracy and efficiency on edge. In particular, our M1 model (R18@256×704) can run over 50FPS on the Tesla T4 platform, with 47.0% NDS on the nuScenes validation set. Our largest model (R101@900x1600) establishes a new stateof-the-art 53.5% NDS on the nuScenes validation set.

### **Token quantization techniques**

#### Learned Token Pruning for Transformers

Efficient deployment of transformer models in practice is challenging due to their inference cost including memory footprint, latency, and power consumption, which scales quadratically with input sequence length. To address this, we present a novel token reduction method dubbed Learned Token Pruning (LTP) which adaptively removes unimportant tokens as an input sequence passes through transformer layers. In particular, LTP prunes tokens with an attention score below a threshold, whose value is learned for each layer during training. Our threshold-based method allows the length of the pruned sequence to vary adaptively based on the input sequence, and avoids algorithmically expensive operations such as top-𝑘 token selection. We extensively test the performance of LTP on GLUE and SQuAD tasks and show that our method outperforms the prior state-of-the-art token pruning methods by up to ∼2.5% higher accuracy with the same amount of FLOPs. In particular, LTP achieves up to 2.1× FLOPs reduction with less than 1% accuracy drop, which results in up to 1.9× and 2.0× throughput improvement on Intel Haswell CPUs and NVIDIA V100 GPUs. Furthermore, we demonstrate that LTP is more robust than prior methods to variations in input sequence lengths. Our code has been developed in PyTorch and open-sourced1 .

Unstructured pruning allows arbitrary patterns of sparsification for parameters and feature maps and can, in theory, produce significant computational savings while preserving accuracy. However, commodity DNN accelerators cannot efficiently exploit unstructured sparsity patterns. Thus, structured pruning methods are typically preferred in practice due to their relative ease of deployment to hardware.

Another orthogonal approach that we consider in this paper is token pruning, which reduces computation by progressively removing unimportant tokens in the sequence during inference. For NLP tasks such as sentence classification, token pruning is an attractive approach to consider as it exploits the intuitive observation that not all tokens (i.e., words) in an input sentence are necessarily required to make a successful inference.

Recently, there has been work on pruning input sentences to transformers, rather than model parameters. This is referred to as token pruning, where less important tokens are progressively removed during inference.

#### Token Merging: Your ViT But Faster

We introduce Token Merging (ToMe), a simple method to increase the throughput of existing ViT models without needing to train. ToMe gradually combines similar tokens in a transformer using a general and light-weight matching algorithm that is as fast as pruning while being more accurate. Off-the-shelf, ToMe can 2× the throughput of state-of-the-art ViT-L @ 512 and ViT-H @ 518 models on images and 2.2× the throughput of ViT-L on video with only a 0.2-0.3% accuracy drop in each case. ToMe can also easily be applied during training, improving in practice training speed up to 2× for MAE fine-tuning on video. Training with ToMe further minimizes accuracy drop, leading to 2× the throughput of ViT-B on audio for only a 0.4% mAP drop. Qualitatively, we find that ToMe merges object parts into one token, even over multiple frames of video. Overall, ToMe’s accuracy and speed are competitive with state-of-the-art on images, video, and audio.

#### MiniViT: Compressing Vision Transformers with Weight Multiplexing

Vision Transformer (ViT) models have recently drawn much attention in computer vision due to their high model capability. However, ViT models suffer from huge number of parameters, restricting their applicability on devices with limited memory. To alleviate this problem, we propose MiniViT, a new compression framework, which achieves parameter reduction in vision transformers while retaining the same performance. The central idea of MiniViT is to multiplex the weights of consecutive transformer blocks. More specifically, we make the weights shared across layers, while imposing a transformation on the weights to increase diversity. Weight distillation over self-attention is also applied to transfer knowledge from large-scale ViT models to weight-multiplexed compact models. Comprehensive experiments demonstrate the efficacy of MiniViT, showing that it can reduce the size of the pre-trained Swin-B transformer by 48%, while achieving an increase of 1.0% in Top-1 accuracy on ImageNet. Moreover, using a single-layer of parameters, MiniViT is able to compress DeiT-B by 9.7 times from 86M to 9M parameters, without seriously compromising the performance. Finally, we verify the transferability of MiniViT by reporting its performance on downstream benchmarks.

#### Token Pooling in Vision Transformers

Despite the recent success in many applications, the high computational requirements of vision transformers limit their use in resource-constrained settings. While many existing methods improve the quadratic complexity of attention, in most vision transformers, self-attention is not the major computation bottleneck, e.g., more than 80% of the computation is spent on fully-connected layers. To improve the computational complexity of all layers, we propose a novel token downsampling method, called Token Pooling, efficiently exploiting redundancies in the images and intermediate token representations. We show that, under mild assumptions, softmax-attention acts as a high-dimensional low-pass (smoothing) filter. Thus, its output contains redundancy that can be pruned to achieve a better trade-off between the computational cost and accuracy. Our new technique accurately approximates a set of tokens by minimizing the reconstruction error caused by downsampling. We solve this optimization problem via cost-efficient clustering. We rigorously analyze and compare to prior downsampling methods. Our experiments show that Token Pooling significantly improves the cost-accuracy trade-off over the state-of-the-art downsampling. Token Pooling is a simple and effective operator that can benefit many architectures. Applied to DeiT, it achieves the same ImageNet top-1 accuracy using 42% fewer computations.

Vision transformers (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021; Heo et al., 2021; Zheng et al., 2021) have demonstrated state-of-the-art results in many vision applications, from image classification to segmentation. However, the high computational cost limits their use in resource-restricted, real-time, or low-powered applications. While most prior work in Natural Language Processing (NLP) improve the time-complexity of attention (Tay et al., 2020b; Ilharco et al., 2020), in vision transformers the main computation bottleneck is the fully-connected layers, as we show in §3.1. The computational complexity of these layers is determined by the number of tokens and their feature dimensionality. While reducing the dimensionality improves computational cost, it sacrifices model capacity and often significantly deteriorates the accuracy of the model. On the other hand, since images often contain mostly smooth surfaces with sparsely located edges and corners, they contain similar (and thus redundant) features. Moreover, we show that, under mild assumptions, softmax-attention is equivalent to low-pass filtering of tokens and thereby produces tokens with similar features, as empirically observed by Goyal et al. (2020) and Rao et al. (2021). This redundancy in representations suggests that we can reduce the number of tokens, i.e., downsampling, without a significant impact to the accuracy, achieving a better cost-accuracy trade-off than reducing feature dimensionality alone.

Downsampling is widely used in Convolutional Neural Network (CNN) architectures to improve computational efficiency, among other purposes. Given a grid of pixels or features, downsampling gradually reduces the grid dimensions via combining neighboring vertices on the grid. The prevailing max/average pooling and sub-sampling are examples of (spatially uniform) grid-downsampling that only uses locations on the grid to decide which vertices to combine. Such methods do not efficiently address non-uniformly distributed redundancy in images and features (Recasens et al., 2018; Marin et al., 2019). Unlike CNNs that require grid preservation, transformers allow a wider range of nonuniform data-aware downsampling layers, where a better operator can be designed.

We propose Token Pooling, a novel nonuniform data-aware downsampling operator for transformers efficiently exploiting redundancy in features. See the illustration and performance metric in Figures 1a & 1b. Motivated by nonuniform sampling and image compression (Marvasti, 2012; Unat et al., 2009; Belfor et al., 1994; Rabbani, 2002), we formulate token downsampling as an optimization problem that minimizes the reconstruction error caused by downsampling. We show that clustering algorithms, K-Means and K-Medoids, efficiently solve this problem, see illustration in Figure 1c. To the best of our knowledge, we are the first to use this formulation and simple clustering analysis for token donwsampling in transformers. We also compare with various prior downsampling techniques, including grid-downsampling (Pan et al., 2021) and token pruning (Goyal et al., 2020; Rao et al., 2021). Our results show that the proposed Token Pooling outperforms existing methods and provides the best trade-off between computational cost and classification accuracy.

#### TokenLearner: Adaptive Space-Time Tokenization for Videos

In this paper, we introduce a novel visual representation learning which relies on a handful of adaptively learned tokens, and which is applicable to both image and video understanding tasks. Instead of relying on hand-designed splitting strategies to obtain visual tokens and processing a large number of densely sampled patches for attention, our approach learns to mine important tokens in visual data. This results in efficiently and effectively finding a few important visual tokens and enables modeling of pairwise attention between such tokens, over a longer temporal horizon for videos, or the spatial content in image frames. Our experiments demonstrate strong performance on several challenging benchmarks for video recognition tasks. Importantly, due to our tokens being adaptive, we accomplish competitive results at significantly reduced computational cost. We establish new state-of-the-arts on multiple video datasets, including Kinetics-400, Kinetics-600, Charades, and AViD.

Videos provide an abundance of visual information. Video understanding particularly requires employing effective spatial-temporal processing of frames to capture long-range interactions [5, 37, 21, 17, 24, 12, 34, 20, 25, 1]. An important aspect of this understanding is how to quickly learn which parts of the input video stream are important, both spatially and temporally, and to focus computational resources on them. But what basic processing mechanism are able to do so successfully?

Recent advancements in Transformers demonstrate improved accuracy on vision classification tasks. For example, departing from standard convolutional approaches, the Vision Transformer (ViT) [9] treats the image as a sequence of patches, utilizing the Transformer architecture [39] similar to text understanding. Standard approaches for video recognition take videos as stacked images (i.e., a spacetime volume) and tend to extend 2D neural architectures to 3D (e.g., 3D-ResNets [17, 5, 38, 11]). Motivated by ViT, recent approaches [2, 3] also extend Transformers for videos by creating 3D ‘tubelet’ video tokens with regular 3D-grids, which often result in computationally heavy models. There are often too many tokens to process, especially for longer videos.

The main question addressed in this work is how to adaptively learn the representation from visual inputs to most effectively capture the spatial information for image frames and spatio-temporal interactions for videos. Here are our main ideas:

The first key observation is we are able to learn to represent visual data by learning to ‘tokenize’ the representations. This is in contrast to previous approaches which used densely sampled tokens e.g., 16x16 or 32x32 over a series of attention layers [9, 3].

Specifically, we can learn to compute important regions in the input image/video, making the tokens adapt to the input data. We compute multiple spatial weight maps per frame with a spatial attention mechanism, and use it for the tokenization. The goal of these maps is to learn which areas are of importance. Here, each spatial weight map is multiplied with the input to form a ‘token’, to be processed by the subsequent learning modules.

Furthermore, we find that very few tokens may be sufficient for a visual understanding task. More specifically, we show that one can significantly reduce the computational budget of video Transformers, by utilizing 8-16 tokens as an intermediate frame representation (instead of keeping 200∼500). Our TokenLearner is able to reduce the number of total FLOPS by half, while maintaining or even increasing the classification accuracy.

The approach is simple, efficient, and, as shown by the results, outperforms methods including both convolutional methods and previous space-time Transformer ones from prior art. In video understanding tasks, we establish new state-of-the-art numbers on Kinetics-400, Kinetics-600, Charades, and AViD datasets by outperforming prior models.

In visual Transformer architectures such as ViT [9], an input image is first tokenized by splitting it into small (e.g., 16x16) spatial patches, which are used as input to the model. Similarly, in recent video Transformer architectures, such as ViViT [2] and TimeSformer [3], the video is tokenized by cutting the video into 2d spatial or 3d spatio-temporal cubes on a regular grid.

Instead of processing fixed, tokenized inputs, our attention module learns the tokens that are to be used for the recognition task. We gain several important properties by doing so: (1) We enable the adaptive tokenization so that the tokens can be dynamically selected conditioned on the input. (2) This also effectively reduces the total number of tokens for the transformer, which is particularly beneficial considering that there are many tokens in videos (e.g., 14 × 14 × 64) and the computation is quadratic to the number of tokens. (3) Finally, we provide an ability for each subsequent layer to learn to rely on different space-time tokenizations, potentially allowing different layers to capture different aspects of the video. These dynamically and adaptively generated tokens can be used in standard transformer architectures such as ViT for images and ViViT for videos

We have presented TokenLearner, a novel approach for visual representation learning, which adaptively tokenizes the representations. The goal is to learn to extract important tokens in image frames and videos for the recognition tasks at hand. Our approach is more efficient, than contemporary work, by finding few important space-time tokens which can model visual representations of images and videos. We observe improved accuracies across challenging video understanding tasks, and outperformed prior approaches in many datasets. One of the remaining challenges is in learning full spatio-temporal tokens. The current TokenLearner focuses on finding spatial tokens over a sequence of frames, and it could be extended to directly mine tokens over space-time volumes.

#### 

#### Masked Autoencoders As Spatiotemporal Learners

This paper studies a conceptually simple extension of Masked Autoencoders (MAE) [31] to spatiotemporal representation learning from videos. We randomly mask out spacetime patches in videos and learn an autoencoder to reconstruct them in pixels. Interestingly, we show that our MAE method can learn strong representations with almost no inductive bias on spacetime (only except for patch and positional embeddings), and spacetime-agnostic random masking performs the best. We observe that the optimal masking ratio is as high as 90% (vs. 75% on images [31]), supporting the hypothesis that this ratio is related to information redundancy of the data. A high masking ratio leads to a large speedup, e.g., > 4× in wall-clock time or even more. We report competitive results on several challenging video datasets using vanilla Vision Transformers [18]. We observe that MAE can outperform supervised pre-training by large margins. We further report encouraging results of training on real-world, uncurated Instagram data. Our study suggests that the general framework of masked autoencoding (BERT [15], MAE [31], etc.) can be a unified methodology for representation learning with minimal domain knowledge.

Our observations on video data support this hypothesis. We find that the optimal masking ratio of MAE is 90% for videos (Fig. 2), higher than the masking ratio of 75% for its image counterpart [31]

#### 

#### SPVit: Pruning Self-attentions into Convolutional Layers in Single Path

Vision Transformers (ViTs) have achieved impressive performance over various computer vision tasks. However, modelling global correlations with multi-head self-attention (MSA) layers leads to two widely recognized issues: the massive computational resource consumption and the lack of intrinsic inductive bias for modelling local visual patterns. To solve both issues, we devise a simple yet effective method named Single-Path Vision Transformer pruning (SPViT), to efficiently and automatically compress the pretrained ViTs into compact models with proper locality added. Specifically, we first propose a novel weight-sharing scheme between MSA and convolutional operations, delivering a single-path space to encode all candidate operations. In this way, we cast the operation search problem as finding which subset of parameters to use in each MSA layer, which significantly reduces the computational cost and optimization difficulty, and the convolution kernels can be well initialized using pre-trained MSA parameters. Relying on the single-path space, we further introduce learnable binary gates to encode the operation choices, which are jointly optimized with network parameters to automatically determine the configuration of each layer. We conduct extensive experiments on two representative ViTs showing that our SPViT achieves a new SOTA for pruning on ImageNet-1k. For example, our SPViT can trim 52.0% FLOPs for DeiT-B and get an impressive 0.6% top-1 accuracy gain simultaneously.

#### EfficientFormer: Vision Transformers at MobileNet Speed

Vision Transformers (ViT) have shown rapid progress in computer vision tasks, achieving promising results on various benchmarks. However, due to the massive number of parameters and model design, e.g., attention mechanism, ViT-based models are generally times slower than lightweight convolutional networks. Therefore, the deployment of ViT for real-time applications is particularly challenging, especially on resource-constrained hardware such as mobile devices. Recent efforts try to reduce the computation complexity of ViT through network architecture search or hybrid design with MobileNet block, yet the inference speed is still unsatisfactory. This leads to an important question: can transformers run as fast as MobileNet while obtaining high performance? To answer this, we first revisit the network architecture and operators used in ViT-based models and identify inefficient designs. Then we introduce a dimension-consistent pure transformer (without MobileNet blocks) as a design paradigm. Finally, we perform latencydriven slimming to get a series of final models dubbed EfficientFormer. Extensive experiments show the superiority of EfficientFormer in performance and speed on mobile devices. Our fastest model, EfficientFormer-L1, achieves 79.2% top-1 accuracy on ImageNet-1K with only 1.6 ms inference latency on iPhone 12 (compiled with CoreML), which runs as fast as MobileNetV2×1.4 (1.6 ms, 74.7% top-1), and our largest model, EfficientFormer-L7, obtains 83.3% accuracy with only 7.0 ms latency. Our work proves that properly designed transformers can reach extremely low latency on mobile devices while maintaining high performance.

#### Rethinking Vision Transformers for MobileNet Size and Speed

With the success of Vision Transformers (ViTs) in computer vision tasks, recent arts try to optimize the performance and complexity of ViTs to enable efficient deployment on mobile devices. Multiple approaches are proposed to accelerate attention mechanism, improve inefficient designs, or incorporate mobile-friendly lightweight convolutions to form hybrid architectures. However, ViT and its variants still have higher latency or considerably more parameters than lightweight CNNs, even true for the years-old MobileNet. In practice, latency and size are both crucial for efficient deployment on resource-constraint hardware. In this work, we investigate a central question, can transformer models run as fast as MobileNet and maintain a similar size? We revisit the design choices of ViTs and propose an improved supernet with low latency and high parameter efficiency. We further introduce a fine-grained joint search strategy that can find efficient architectures by optimizing latency and number of parameters simultaneously. The proposed models, EfficientFormerV2, achieve about 4% higher top-1 accuracy than MobileNetV2 and MobileNetV2×1.4 on ImageNet-1K with similar latency and parameters. We demonstrate that properly designed and optimized vision transformers can achieve high performance with MobileNet-level size and speed1 .

#### Visual Transformer Pruning

Visual transformer has achieved competitive performance on a variety of computer vision applications. However, their storage, run-time memory, and computational demands are hindering the deployment on mobile devices. Here we present an visual transformer pruning approach, which identifies the impacts of channels in each layer and then executes pruning accordingly. By encouraging channelwise sparsity in the Transformer, important channels automatically emerge. A great number of channels with small coefficients can be discarded to achieve a high pruning ratio without significantly compromising accuracy. The pipeline for visual transformer pruning is as follows: 1) training with sparsity regularization; 2) pruning channels; 3) finetuning. The reduced parameters and FLOPs ratios of the proposed algorithm are well evaluated and analyzed on ImageNet dataset to demonstrate its effectiveness.

#### NVIT: VISION TRANSFORMER COMPRESSION AND PARAMETER REDISTRIBUTION

Transformers yield state-of-the-art results across many tasks. However, they impose huge computational costs during inference. We apply global structural pruning with latency-aware regularization on all parameters of the Vision Transformer (ViT) model for latency reduction. Furthermore, we analyze the pruned architectures and find interesting regularities in the final weight structure. Our discovered insights lead to a new architecture called NViT (Novel ViT), with a redistribution of where parameters are used. This architecture utilizes parameters more efficiently and enables control of the latency-accuracy trade-off. On ImageNet-1K, we prune the DEIT-Base (Touvron et al., 2021) model to a 2.6× FLOPs reduction, 5.1× parameter reduction, and 1.9× run-time speedup with merely 0.07% loss in accuracy. We achieve more than 1% accuracy gain when compressing the base model to the throughput of the Small/Tiny variants. NViT gains 0.1-1.1% accuracy over the hand-designed DEIT family when trained from scratch, while being faster.

Self-attention based transformer models demonstrate high model capacity, easy scalability, and superior ability in capturing long-range dependency (Vaswani et al., 2017; Devlin et al., 2018; Radford et al., 2018; Jiao et al., 2019; Brown et al., 2020). They have thus been widely applied to natural language processing (NLP) tasks, and recently received growing attention for computer vision tasks. Vision Transformer, i.e., the ViT (Dosovitskiy et al., 2020), shows that embedding image patches into tokens and passing them through a sequence of transformer blocks can lead to higher accuracy compared to state-of-the-art CNN models. DEIT, recent work by Touvron et al. (2021), further presents a data-efficient training method such that acceptable accuracy can be achieved without extensive pretraining. Offering competitive performance to CNNs under similar training regimes, transformers now point to the appealing perspective of solving both NLP and vision tasks with the same architecture.

Unlike CNNs built with convolutional layers that are mainly parameterized by few dimensions like the kernel size and the number of filters, the ViT has multiple distinct components, i.e., QKV projection, multi-head attention, multi-layer perceptron, etc. (Vaswani et al., 2017), each defined by independent dimensions. As a result, the dimensionality of each component in each ViT block needs to be carefully designed to achieve a decent trade-off between efficiency and accuracy. However, this is typically not the case for state-of-the-art models. Models such as ViT (Dosovitskiy et al., 2020) and DEIT (Touvron et al., 2021) mainly inherit the design heuristics from NLP tasks, e.g., use MLP expansion ratio 4, fix QKV per head, all the blocks having the same dimensions, etc., which may not be optimal for computer vision (Chen et al., 2021a), causing significant redundancy in the base model and a worse efficiency-accuracy trade-off upon scaling, as we show extensively in our experiments.

This work targets efficient ViTs by exploring latency-aware global structural pruning, leveraging the insights to redistribute parameters for enhanced accuracy-efficiency trade-off. Our approach, as visualized in Figure 1, starts from analyzing the blocks in the computation graph of ViT to identify all the dimensions that can be independently controlled. We apply global structural pruning over all the components in all blocks. This offers complete flexibility to explore their combinations towards an optimal architecture in a complicated design space. Our global pruning utilizes an importance score based on the first-order Taylor expansion of the pruning loss, offering comparability among all prunable components from all layers. Furthermore, we incorporate the estimated latency reduction of each neuron into its importance score. This guides the final pruned architecture to be faster on target devices, as we show in experiments. The pruned models are distilled utilizing the information of the ground truth labels, a pretrained CNN teacher akin to DEIT (Touvron et al., 2021), and the original full model. On the ImageNet-1K benchmark (Russakovsky et al., 2015), structural pruning enables a nearly lossless 5.14× parameter reduction, 2.57× FLOPs reduction and 1.86× speed up on V100 GPU over the DEIT-Base model. An 1% and 1.7% accuracy gain is observed over DEIT-Small and DEIT-Tiny models when we compress the base model to a similar latency.

Using structural pruning for architectural guidance, we further make an important observation that the popular uniform distribution of parameters across all layers is, in fact, not optimal. A simple redistribution of parameters can already provide stronger architectural alternatives, as we show in our experiments for both pretraining and downstream tasks. To this end, we present a new parameter distribution rule to scale ViT architectures, enabling a breed of models named as NViT. When scaling to similar FLOPs and latency, NViT architectures achieve 0.1%, 0.2% and 1.1% accuracy gains over the DEIT-Base, Small, and Tiny models respectively when trained from scratch on ImageNet-1K

Our main contributions are as follows:

* Provide a systematic analysis on the prunable components in the ViT model. We identify the ability to perform structural pruning on the embedding dimension, number of heads, MLP hidden dimension, QK dimension and V dimension of each head separately;
* Propose a latency-aware, importance-based criteria that enables hardware-friendly global structural pruning of all the components, achieving a nearly lossless 1.9× speedup;
* Present a new architecture scaling rule that enables NViT, a new family of efficient vision transformer architectures that redistributes the dimensions of DEIT models to outperform them under similar FLOPs and latency. This is the first work showing the potential of discovering novel scalable architectures by pruning vision transformers;
* Demonstrate that the high performance achieved by the pruned models and NViT models transfer effectively to downstream tasks

#### A-ViT: Adaptive Tokens for Efficient Vision Transformer

We introduce A-ViT, a method that adaptively adjusts the inference cost of vision transformer (ViT) for images of different complexity. A-ViT achieves this by automatically reducing the number of tokens in vision transformers that are processed in the network as inference proceeds. We reformulate Adaptive Computation Time (ACT [17]) for this task, extending halting to discard redundant spatial tokens. The appealing architectural properties of vision transformers enables our adaptive token reduction mechanism to speed up inference without modifying the network architecture or inference hardware. We demonstrate that A-ViT requires no extra parameters or sub-network for halting, as we base the learning of adaptive halting on the original network parameters. We further introduce distributional prior regularization that stabilizes training compared to prior ACT approaches. On the image classification task (ImageNet1K), we show that our proposed A-ViT yields high efficacy in filtering informative spatial features and cutting down on the overall compute. The proposed method improves the throughput of DeiT-Tiny by 62% and DeiT-Small by 38% with only 0.3% accuracy drop, outperforming prior art by a large margin.

#### Pruning Attention Heads of Transformer Models Using A\* Search

#### A Novel Approach to Compress Big NLP Architectures

Recent years have seen a growing adoption of Transformer models such as BERT in Natural Language Processing and even in Computer Vision. However, due to their size, there has been limited adoption of such models within resource-constrained computing environments This paper proposes novel pruning algorithm to compress transformer models by eliminating redundant Attention Heads. We apply the A\* search algorithm to obtain a pruned model with strict accuracy guarantees. Our results indicate that the method could eliminate as much as 40% of the attention heads in the BERT transformer model with no loss in accuracy.

#### DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification

Attention is sparse in vision transformers. We observe the final prediction in vision transformers is only based on a subset of most informative tokens, which is sufficient for accurate image recognition. Based on this observation, we propose a dynamic token sparsification framework to prune redundant tokens progressively and dynamically based on the input. Specifically, we devise a lightweight prediction module to estimate the importance score of each token given the current features. The module is added to different layers to prune redundant tokens hierarchically. To optimize the prediction module in an end-to-end manner, we propose an attention masking strategy to differentiably prune a token by blocking its interactions with other tokens. Benefiting from the nature of self-attention, the unstructured sparse tokens are still hardware friendly, which makes our framework easy to achieve actual speed-up. By hierarchically pruning 66% of the input tokens, our method greatly reduces 31% ∼ 37% FLOPs and improves the throughput by over 40% while the drop of accuracy is within 0.5% for various vision transformers. Equipped with the dynamic token sparsification framework, DynamicViT models can achieve very competitive complexity/accuracy trade-offs compared to state-of-the-art CNNs and vision transformers on ImageNet.

#### AdaViT: Adaptive Vision Transformers for Efficient Image Recognition

Built on top of self-attention mechanisms, vision transformers have demonstrated remarkable performance on a variety of vision tasks recently. While achieving excellent performance, they still require relatively intensive computational cost that scales up drastically as the numbers of patches, self-attention heads and transformer blocks increase. In this paper, we argue that due to the large variations among images, their need for modeling long-range dependencies between patches differ. To this end, we introduce AdaViT, an adaptive computation framework that learns to derive usage policies on which patches, self-attention heads and transformer blocks to use throughout the backbone on a per-input basis, aiming to improve inference efficiency of vision transformers with a minimal drop of accuracy for image recognition. Optimized jointly with a transformer backbone in an end-to-end manner, a light-weight decision network is attached to the backbone to produce decisions on-the-fly. Extensive experiments on ImageNet demonstrate that our method obtains more than 2× improvement on efficiency compared to state-of-the-art vision transformers with only 0.8% drop of accuracy, achieving good efficiency/accuracy trade-offs conditioned on different computational budgets. We further conduct quantitative and qualitative analysis on learned usage polices and provide more insights on the redundancy in vision transformers.

**In particular, we insert a light-weight multihead subnetwork (i.e. a decision network) to each transformer block of the backbone network, which learns to predict binary decisions on the usage of patch embeddings, self-attention heads and blocks throughout the network. Since binary decisions are non-differentiable, we resort to Gumbel-Softmax [25] during training to make the whole framework end-to-end trainable. The decision network is jointly optimized with the transformer backbone with a usage loss that measures the computational cost of the produced usage policies and a normal cross-entropy loss, which incentivizes the network to produce policies that reduce the computational cost while maintaining classification accuracy. The overall target computational cost can be controlled by hyperparameter γ ∈ (0, 1] corresponding to the percentage of computational cost of the full model with all patches as input during training, making the framework flexible to suit the need of different computational budgets。**

#### NOT ALL PATCHES ARE WHAT YOU NEED: EXPEDITING VISION TRANSFORMERS VIA TOKEN REORGANIZATIONS

Vision Transformers (ViTs) take all the image patches as tokens and construct multi-head self-attention (MHSA) among them. Complete leverage of these image tokens brings redundant computations since not all the tokens are attentive in MHSA. Examples include that tokens containing semantically meaningless or distractive image backgrounds do not positively contribute to the ViT predictions. In this work, we propose to reorganize image tokens during the feed-forward process of ViT models, which is integrated into ViT during training. For each forward inference, we identify the attentive image tokens between MHSA and FFN (i.e., feed-forward network) modules, which is guided by the corresponding class token attention. Then, we reorganize image tokens by preserving attentive image tokens and fusing inattentive ones to expedite subsequent MHSA and FFN computations. To this end, our method EViT improves ViTs from two perspectives. First, under the same amount of input image tokens, our method reduces MHSA and FFN computation for efficient inference. For instance, the inference speed of DeiT-S is increased by 50% while its recognition accuracy is decreased by only 0.3% for ImageNet classification. Second, by maintaining the same computational cost, our method empowers ViTs to take more image tokens as input for recognition accuracy improvement, where the image tokens are from higher resolution images. An example is that we improve the recognition accuracy of DeiT-S by 1% for ImageNet classification at the same computational cost of a vanilla DeiT-S. Meanwhile, our method does not introduce more parameters to ViTs. Experiments on the standard benchmarks show the effectiveness of our method. The code is available at <https://github.com/youweiliang/evit>.

#### SPViT: Enabling Faster Vision Transformers via Latency-aware Soft Token Pruning

Abstract. Recently, Vision Transformer (ViT) has continuously established new milestones in the computer vision field, while the high computation and memory cost makes its propagation in industrial production difficult. Considering the computation complexity, the internal data pattern of ViTs, and the edge device deployment, we propose a latencyaware soft token pruning framework, SPViT, which can be set up on vanilla Transformers of both flatten and hierarchical structures, such as DeiTs and Swin-Transformers (Swin). More concretely, we design a dynamic attention-based multi-head token selector, which is a lightweight module for adaptive instance-wise token selection. We further introduce a soft pruning technique, which integrates the less informative tokens chosen by the selector module into a package token rather than discarding them completely. SPViT is bound to the trade-off between accuracy and latency requirements of specific edge devices through our proposed latency-aware training strategy. Experiment results show that SPViT significantly reduces the computation cost of ViTs with comparable performance on image classification. Moreover, SPViT can guarantee the identified model meets the latency specifications of mobile devices and FPGA, and even achieve the real-time execution of DeiT-T on mobile devices. For example, SPViT reduces the latency of DeiT-T to 26 ms (26% ∼41% superior to existing works) on the mobile device with 0.25% ∼4% higher top-1 accuracy on ImageNet. Our code is released at https://github.com/PeiyanFlying/SPViT

#### IA-RED2 : Interpretability-Aware Redundancy Reduction for Vision Transformers

The self-attention-based model, transformer, is recently becoming the leading backbone in the field of computer vision. In spite of the impressive success made by transformers in a variety of vision tasks, it still suffers from heavy computation and intensive memory costs. To address this limitation, this paper presents an Interpretability-Aware REDundancy REDuction framework (IA-RED2 ). We start by observing a large amount of redundant computation, mainly spent on uncorrelated input patches, and then introduce an interpretable module to dynamically and gracefully drop these redundant patches. This novel framework is then extended to a hierarchical structure, where uncorrelated tokens at different stages are gradually removed, resulting in a considerable shrinkage of computational cost. We include extensive experiments on both image and video tasks, where our method could deliver up to 1.4× speed-up for state-of-the-art models like DeiT [53] and TimeSformer [3], by only sacrificing less than 0.7% accuracy. More importantly, contrary to other acceleration approaches, our method is inherently interpretable with substantial visual evidence, making vision transformer closer to a more human-understandable architecture while being lighter. We demonstrate that the interpretability that naturally emerged in our framework can outperform the raw attention learned by the original visual transformer, as well as those generated by off-the-shelf interpretation methods, with both qualitative and quantitative results. Project Page: http://people.csail.mit.edu/bpan/ia-red/.

## 