Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.

Vision Transformers (ViT) adapt the Transformer model for computer vision tasks by converting images into sequences of patches.

Each patch is flattened and embedded before being passed to a standard Transformer encoder.

This approach contrasts with CNNs, which process images using localized filters and hierarchical representations.

ViTs have shown that, given sufficient data and compute, they can outperform traditional CNNs.

Variants like DeiT, Swin Transformer, and PiT have improved the efficiency and performance of ViTs.

Applications include image classification, object detection, segmentation, and video understanding.