End-to-End Trainable Multi-Instance Pose Estimation with Transformers

Lucas Stoffl **Maxime Vidal** **Alexander Mathis**

Swiss Federal Institute of Technology (EPFL)
alexander.mathis@epfl.ch

We propose a new end-to-end trainable approach for multi-instance pose estimation by combining a convolutional neural network with a transformer. We cast multi-instance pose estimation from images as a direct set prediction problem. Inspired by recent work on end-to-end trainable object detection with transformers, we use a transformer encoder-decoder architecture together with a bipartite matching scheme to directly regress the pose of all individuals in a given image. Our model, called POse Estimation Transformer (POET), is trained using a novel set-based global loss that consists of a keypoint loss, a keypoint visibility loss, a center loss and a class loss. POET reasons about the relations between detected humans and the full image context to directly predict the poses in parallel. We show that POET can achieve high accuracy on the challenging COCO keypoint detection task. To the best of our knowledge, this model is the first end-to-end trainable multi-instance human pose estimation method.

Introduction

Multi-human pose estimation from a single image, the task of predicting the body part locations for each individual, is an important computer vision problem. Pose estimation has wide ranging applications from measuring behavior in health care and biology to virtual reality and human-computer interactions (1–4).

Multi-human pose estimation can be thought of as a hierarchical set prediction task. An algorithm needs to predict the bodyparts of all individuals and group them correctly into humans. Due to the complexity of this process, current methods *consist of multiple steps and are not end-to-end trainable*. Fundamentally, top-down and bottom-up methods are the major approaches. Top-down methods first predict the location (bounding boxes) of all individuals based on an object detection algorithm and then predict the location of all the bodyparts per cropped individual with a separate network (5–8). Bottom-up methods first predict all the bodyparts, and then group them into individuals (9–16). However, both approaches require either post-processing or two different networks. This motivates the search for end-to-end solutions.

Inspired by DETR (17), a recent transformer-based architecture for object detection, we propose a novel end-to-end trainable method for multi-instance pose estimation. Pose estimation transformer (POET) is the first model that is trained end-to-end for multi-instance pose estimation, without the need for post-processing or two networks as typically employed in top-down approaches. POET predicts all human poses without any post-processing and is trained with a novel, yet simple loss function, which allows bipartite matching between predicted and ground-truth human poses. Our approach achieves excellent results on the difficult COCO keypoint challenge especially for large humans, and performs better than baseline models even with higher spatial resolution (Figure 1).

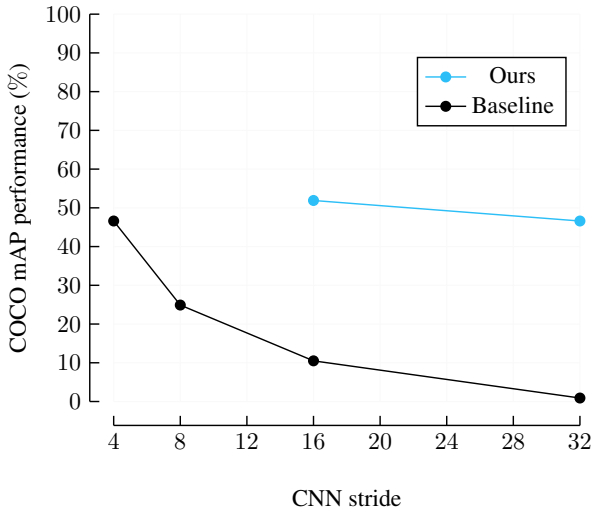

Fig. 1. POET vs. the fully convolutional associative embedding method performance as a function of the stride of the encoder on the COCO keypoint estimation task. Our method achieves strong performance despite lower resolution features.

Related work

Transformers in vision and beyond Transformers were introduced for machine translation (18), and have vastly improved the performance of deep learning models on language tasks (18–20). Their architecture inherently allows modeling and discovering long-range interactions in data. Their use has recently been extended to speech recognition (21), automated theorem proving (22), and many other tasks (23). In computer vision, transformers have been used with great effect either in combination or as an alternative to convolutional neural networks (CNNs) (23). Notably, Visual Transformer (ViT) (24) demonstrated state-of-the-art performance on image recognition tasks with pure transformer models. In other visual tasks, such as text-to-image, excellent results have been shown, e.g., by DALL-E (25).

Recently Carion et al. (17) developed a new end-to-end paradigm for visual object detection with transformers, a task which previously required either two-stage approaches or post-processing. This approach, DETR, formulated ob-

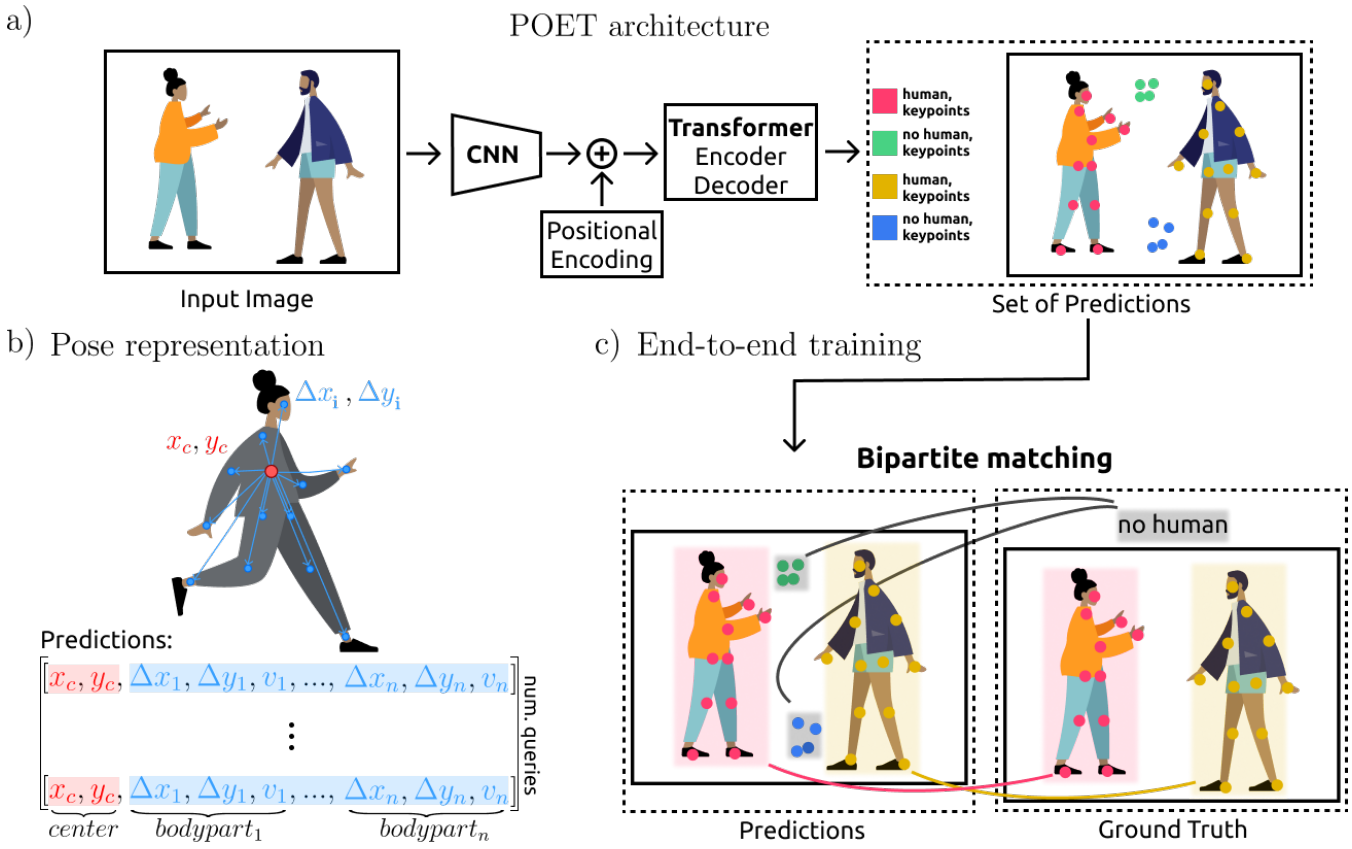

Fig. 2. Overview of our model. **a)** POET combines a CNN backbone and a transformer to directly predict the pose of multiple humans. **b)** Each pose is represented as a vector comprising the center (x_c, y_c), the relative offset $(\Delta x_i, \Delta y_i)$ of each bodypart i and its visibility v_i. **c)** POET is trained end-to-end by bipartite matching of the closest predictions to the ground truth pose, and then backpropagating the loss.

ject detection as a set prediction problem combined with a bipartite matching loss. DETR is an elegant solution, however, the model requires long training times and shows comparatively low performance on small objects (17). These problems were mitigated through further works; Deformable DETR (26) presents a multi-scale deformable attention module which only attends to a set number of points within the feature map, for different scales, and in this way reducing the training time and improving small object detection performance. Sun et al. removed the transformer decoder and fed the features coming out of the CNN backbone to a Feature Pyramid Network (27).

Importantly, end-to-end approaches were successfully applied in many complex prediction tasks such as speech recognition or machine translation (19, 20), but are still *lacking in multi-instance pose estimation*.

Pose estimation Pose estimation is a classical computer vision problem with wide ranging applications (1–3, 28). Pose estimation methods are evaluated on several benchmarks for 2D multi-human pose estimation, incl. COCO (2, 29–31).

Multi-instance pose estimation methods can be classified as top-down and bottom-up (2, 3, 28). Top-down methods predict the location of each bodypart of each *individual*, based on bounding boxes localizing individuals with a separate network (5–8, 32–34). Bottom-up methods first predict all the bodyparts, and then group them into individuals by using part affinity fields (9), pairwise predictions (10, 11, 14, 35, 36), composite fields (13, 16), or associative embeddings (12, 15). Both top-down and bottom-up approaches, require either post-processing steps (for assembly) or two different neural networks (for localization and then pose estimation).

Most recent (state-of-the-art) approaches are fully convolutional and predict keypoint heatmaps (e.g., (6, 7, 9, 12, 15)). Recently, Yang et al. proposed TransPose, a top-down method, which predicts heatmaps as well, but by using attention after a CNN encoder (8). Transformers were also used for 3D pose and mesh reconstruction on (single) humans, which achieved state-of-the-art on Human3.6M (37). Extending this work, we build on DETR (17), to propose an end-to-end trainable pose estimation method for multiple instances that directly outputs poses as vectors (without heatmaps). To cast pose estimation as a hierarchical set prediction problem, we adapt the pose representations of CenterNet (36) and Single-Stage Multi-Person Pose Machines (14).

The POET model

The overall POse Estimation Transformer (POET) model is illustrated in Figure 2. Our work is closely related to DETR (17) and fundamentally extends this object detection framework to multi-instance pose estimation. Like DETR, POET consists of two major ingredients: (1) a *transformer-*

Fig. 3. Example predictions on COCO-eval with POET-R50 model listed in Table 1. Top row: examples with good performance. Bottom row: examples with errors.

based architecture that predicts a set of human poses in parallel and (2) a *set prediction loss* that is a linear combination of simple sub-losses for classes, keypoint coordinates and visibilities. To cast multi-instance pose estimation as a set prediction problem, we represent the pose of each individual as the center (of mass) together with the relative offsets per bodypart. Each bodypart can be occluded or visible. POET is trained to directly output a vector comprising the center, relative bodyparts as well as (binary) bodypart visibility indicators (Figure 2b).

POET architecture The POET architecture contains three main elements: a CNN backbone that extracts features of the input images, an encoder-decoder transformer and a feedforward network (FFN) head that outputs the set of estimated poses.

CNN backbone. The convolutional backbone is given a batch of images, $I \in \mathbb{R}^{B\times 3\times H\times W}$, as input with batch size B, 3 color channels and image dimensions (H, W). Through several computing and downsampling steps, the CNN generates lower-resolution feature maps, $F \in \mathbb{R}^{B\times C\times H/S\times W/S}$ with stride S. Specifically, we choose different ResNets (38) with various strides S, as detailed in the experiments section.

Encoder-Decoder transformer. The encoder-decoder transformer model follows the standard architecture (17, 18). Both encoder and decoder consist of 6 layers with 8 attention heads each. The encoder takes the output features of the CNN backbone and reduces their channel dimensions by a 1×1 convolution. This downsampled tensor is then collapsed along the spatial dimension into one dimension as the multi-head mechanism expects sequential input. We add a *fixed* positional encoding to the encoder input, as the transformer architecture is (otherwise) permutation-invariant and would disregard the spatial image structure. In contrast, the input embeddings for the decoder are *learned* positional encodings, which we refer to as *object queries*. These queries must be different for the decoder to produce different results, due to the permutation-invariance. They are added to the encoder output to form the decoder input. The decoder transforms the queries into output embeddings, which are then taken by the *pose prediction head* and independently decoded into the final set of poses and class labels. Thereby, every query can search for one object/instance and predicts its pose and class. We set the number of object queries N to 25 as this is about twice higher than the maximum number of humans present in one image in the COCO dataset. With the aid of the self-attention in both the encoder and the decoder, the network is able to globally reason about all objects together using pairwise relations between them, and at the same time using the whole image as context information. The difference of this transformer decoder (and the one in DETR (17)) to the original formulation is the parallel decoding of the N objects at each layer, in contrast to the autoregressive model used by Vaswani et al. (18).

Pose prediction head. The final pose estimation is carried out by a 3-layer perceptron with ReLU activation and a linear projection layer (FFN head). This head outputs the center coordinates, the displacements to all bodyparts relative to the center and the visibility scores for every body part in a single vector (Figure 2b), and the linear layer outputs the class label using a softmax function. Thereby, we normalize the center and offsets to the image size.

Training loss In order to predict all human poses in parallel, the network is trained by the loss after optimal matching, which is calculated after finding an optimal matching between predictions and ground-truth and summing over the individuals. Therefore, our loss has to score predictions accordingly, with respect to the class, the keypoint coordinates and their visibilities, produce the matching and then optimize the multi-instance pose-specific losses.

For every instance i, in the ground truth, we compute the center as the center of mass of all visible keypoints (e.g., COCO (29) contains humans without annotations, in these cases the visibilities are set to 0). The ground truth vector for human i is

then $[x_c, y_c, \Delta x_1, \Delta y_1, v_1, \Delta x_2, \Delta y_2, v_2 \ldots, \Delta x_n, \Delta y_n, v_n]$, for center (x_c, y_c), relative offset $(\Delta x_i, \Delta y_i)$ of each body-part i and its visibility v_i. In order to make the loss functions more legible, we split this vector into $y_i = (c_i, C_i, Z_i, V_i)$, which consists of the target class label (human/non-object) c_i, the center $C_i = (x_c, y_c)$, the relative pose: $Z_i = [\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, \ldots, \Delta x_n, \Delta y_n]$ (relative joint displacements from the center C_i) and a binary visibility vector $V_i = [v_1, v_1, v_2, v_2, \ldots, v_n, v_n]$ encoding for every joint in the image, whether it is visible or not.

The prediction of the network for instance i is then defined as $\widehat{y}_i = (\widehat{p}(c_i), \widehat{C}_i, \widehat{Z}_i, \widehat{V}_i)$, where $\widehat{p}(c_i)$ is the predicted probability for class c_i, $\widehat{C}_i$ the predicted center, $\widehat{Z}_i$ the predicted pose, and $\widehat{V}_i$ the predicted visibility. Note that the network does not predict the visibility for the center.

In the following, we denote by y the ground truth set of poses, and $\widehat{y} = \{\widehat{y}_i\}_{i=1}^{N}$ the set of N predictions. Here, y is the set of humans in the image padded with non-objects. We define our pair-wise matching cost between ground truth y_i and a prediction with index $\sigma(i)$ as:

$$\begin{aligned}\mathcal{L}_{\text{match}}(y_i, \widehat{y}_{\sigma(i)})) = &-\mathbb{1}_{\{c_i \neq \varnothing\}} \widehat{p}_{\sigma(i)}(c_i) \\ &+ \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{pose}}(C_i, Z_i, V_i, \widehat{C}_\sigma(i), \widehat{Z}_\sigma(i), \widehat{V}_\sigma(i))\end{aligned} \tag{1}$$

Here, $\mathcal{L}_{\text{pose}}$ is the pose-specific cost that we will define below and involves costs for the centers, the bodyparts and their visibilities.

The optimal assignment is then found as a bipartite matching with the lowest matching cost based on the Hungarian algorithm (17, 39). This assignment is the following permutation of N elements $\sigma \in \mathfrak{S}_N$ for symmetric group $\mathfrak{S}_N$ (40):

$$\widehat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \widehat{y}_{\sigma(i)}) \tag{2}$$

Once the optimal matching is obtained, we can compute the *Hungarian loss* for all matched pairs. Like the matching cost, it contains a loss part scoring the poses, which is a linear combination of a L_1 loss to compute the differences between relative keypoint coordinates, a L_2 loss for the center coordinates and a L_2 loss for the visibilities, with hyperparameters λ_{L1}, λ_{L2} and λ_{ctr}:

$$\begin{aligned}\mathcal{L}_{\text{pose}}(C_i, Z_i, V_i, \widehat{C}_\sigma(i), \widehat{Z}_\sigma(i), \widehat{V}_\sigma(i)) = \\ \lambda_{L1} \| V_i \circ Z_i - V_i \circ \widehat{Z}_{\sigma(i)} \|_1 \\ + \lambda_{L2} \| V_i - \widehat{V}_{\sigma(i)} \|_2^2 \\ + \lambda_{ctr} \| (x_c, y_c)_i - (\widehat{x_c}, \widehat{y_c})_{\sigma(i)} \|_2^2\end{aligned} \tag{3}$$

Thereby, $\circ$ denotes point-wise multiplication. These three losses are normalized by the number of humans inside the batch.

The final loss, the *Hungarian loss*, then is a linear combination of a negative log-likelihood for class prediction and the keypoint-specific loss defined above, for all pairs from the optimal assignment $\widehat{\sigma}$:

$$\begin{aligned}\mathcal{L}_{\text{Hungarian}}(y, \widehat{y}) = \sum_{i=1}^{N} \Big[&-\log \widehat{p}_{\widehat{\sigma}(i)}(c_i) \\ &+ \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{pose}}(C_i, Z_i, V_i, \widehat{C}_{\widehat{\sigma}}(i), \widehat{Z}_{\widehat{\sigma}}(i), \widehat{V}_{\widehat{\sigma}}(i)) \Big]\end{aligned} \tag{4}$$

Most COCO images contain only few annotated humans. To account for this class imbalance, we down-weight the log-probability term by a factor of 10 for all non-objects.

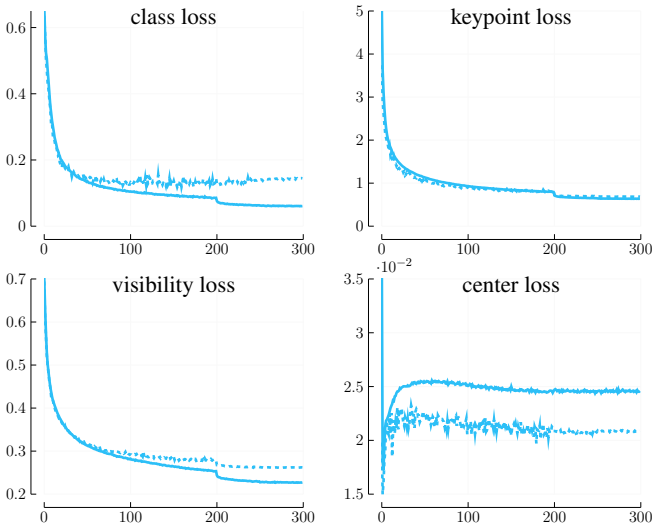

Fig. 4. Evolution of the different loss parts over training epochs (Equation 4) for POET-R50. Solid lines correspond to training losses, and dashed lines to validation losses (on COCO).

Experiments

We evaluated POET on the difficult COCO keypoint estimation challenge (29), illustrate qualitative results and show that it reaches good performance (especially for large humans). Then we show that it outperforms baseline methods that we trained based on an established bottom-up method using associative embedding with the same backbone (12, 45). Then, we analyse different aspects of the architecture and the loss. Finally, we discuss challenges and future work.

COCO keypoint detection challenge The COCO dataset (29) comprises more than $200,000$ images with more than $150,000$ people for which up to 17 keypoints are annotated. The dataset is split into train/val/test-dev sets with $57k$, $5k$ and $20k$ images, respectively. We trained on the training images (that contain humans) and report results on the validation set for our comparison study and on the test set for comparing to state-of-the-art models.

As it is customary, we assess the performance with the standard evaluation metric based on Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\delta(v_i > 0)}. \tag{5}$$

Table 1. Comparison to state-of-the-art models on COCO test-dev. Note that most models use an overall stride of 4 (or smaller), when extracting features, while POET uses stride 32 and therefore much smaller feature maps which harms performance on AP_M. However, with regard to AP_L it can compete with state-of-the-art models.

Method	Stride	#Params	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_M	AR_L
Top-Down												
Mask-RCNN (41)	4	44.4M	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
CPN (34)	-	-	72.1	91.4	80.0	68.7	77.2	78.5	95.1	85.3	74.2	84.3
HRNet-W48 (6, 42)	4	63.6M	**75.5**	92.5	83.3	71.9	81.5	80.5	-	-	-	-
TansPose-H-A6 (8)	4	17.5M	75.0	92.2	82.3	71.3	81.1	-	-	-	-	-
Bottom-Up												
OpenPose (9)	4	-	61.8	84.9	67.5	57.1	68.2	-	-	-	-	-
Hourglass (12)	4	277.8M	56.6	81.8	61.8	49.8	67.0	-	-	-	-	-
PersonLab (43)	8	68.7M	66.5	88.0	72.6	62.4	72.3	71.0	90.3	76.6	66.1	77.7
AE + R50 (44)	4	31.9M	46.6	74.2	47.9	44.6	49.3	55.2	79.7	57.5	48.1	65.1
AE + R101 (44)	4	50.9M	55.4	80.7	59.9	49.3	64.1	62.2	84.1	66.4	53.3	74.3
AE + R152 (44)	4	68.6M	59.5	82.9	64.8	51.7	71.1	65.1	85.6	69.6	55.3	78.8
PifPaf (13)	4	-	66.7	-	-	62.4	72.9	-	-	-	-	-
HigherHRNet (15)	2	63.8M	68.4	88.2	75.1	64.4	74.2	-	-	-	-	-
POET-R50 (Ours)	32	41.3M	55.4	83.3	59.8	45.4	68.8	62.5	88.3	66.8	52.9	75.4

Table 2. Comparison with the baseline method using associative embedding (AE) (12, 44) with ResNet-50 backbone *and reduced stride* on the COCO validation set. POET outperforms AE models, trained with the same backbones and reaches better performance than AE baselines utilizing an overall stride of 4.

Method	Input Size / Stride	#Params	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_M	AR_L
AE + R50	512 / 32	31.9M	0.9	3.7	0.1	0.0	2.3	6.3	18.9	2.9	0.5	14.1
AE + R50	512 / 16	31.9M	10.5	29.7	5.7	4.1	20.3	21.9	46.0	19.0	7.5	41.5
AE + R50	512 / 8	31.9M	24.9	54.0	19.8	18.7	34.4	37.6	64.6	35.9	23.4	56.9
AE + R50 (300 epochs) (44)	512 / 4	31.9M	46.6	74.2	47.9	44.6	49.3	55.2	79.7	57.5	48.1	65.1
POET-R50	512 / 32	41.3M	46.6	75.0	47.8	31.1	66.7	52.0	78.6	53.5	36.9	72.8
POET-DC5-R50	512 / 16	41.3M	**51.9**	**79.0**	**54.9**	37.7	**70.7**	**57.1**	**82.6**	**59.7**	43.3	**76.3**

Thereby, for each keypoint $i \in \{1,2,\ldots,17\}$, d_i is the Euclidean distance between the detected keypoint and its corresponding ground truth, v_i is the (boolean) visibility of the ground truth, s is the object scale, k_i is the labeling uncertainty (a COCO constant) and δ is 1 for positive visibilities and zero otherwise. We calculated the average precision and recall scores: AP_{50} (AP at OKS = 0.50), AP_{75}, AP (the mean of AP scores at OKS = 0.50, 0.55, ..., 0.90, 0.95), AP_M for medium objects, AP_L for large objects, and AR (the mean of recalls at $OKS = 0.50, 0.55, \ldots, 0.90, 0.95$), as well as AR_{50}, AR_{75}, AR_M and AR_L.

Implementation details We trained all (POET) models with the following hyperparameter setting in the keypoint loss: $\lambda_{L1} = 4$, $\lambda_{L2} = 0.2$ and $\lambda_{ctr} = 0.5$.

We set the transformer's initial learning rate to 10^{-4}, the backbone's to 10^{-5}, and weight decay to 10^{-4} (17) and train POET with AdamW (46). A dropout rate of 0.1 is applied on the transformer's weights, which are initialized with Xavier initialization (47). For the encoder, we choose ResNet50 (38) with different strides S. Accordingly, models are called POET-R50 as well as POET-DC5-R50, when using a dilated C5 stage (which decreases the stride from 32 to 16). The replacement of a stride by a dilation in the last stage of the backbone increases the feature resolution by a factor of two, but also comes with an increase in computational cost by the same factor.

During training we augment the data by applying rotation uniformly drawn from $(-25, +25)$ degrees, random cropping, horizontal flipping and coarse dropout (48) with a probability of 0.5 each. Additionally, we resize the images such that the shortest side falls in the range $[400, 800]$ and the longest side is at most $1,333$. We set the number of prediction slots N to 25, as the maximum number of keypoint annotated humans in COCO images is 13.

We carried out two different sets of experiments: (1) training POET initialized from ImageNet (49) weights to compare with current state-of-the-art models and (2) training multiple models as well as baseline models, whereby we started with COCO keypoint challenge pretrained weights from MMPose (45). For the comparison with state-of-the-art models, we train POET-R50 with a batch size of 6 on two NVIDIA V100 GPUs (hence a total batch size of 12) for 300 epochs, with a learning rate drop by a factor of 10 after 200 epochs and again after 250 epochs. One epoch takes approximately one hour in this setting. The loss curves and the evolution of mAP on COCO-val can be seen in Figures 4 and 6.

For the comparison to baseline models, we utilized the associative embedding (12, 15) implementation in MMPose (44, 45). To account for the high memory footprint and the long training times, we restrict the maximum image size to 512 during training and train POET models (POET-R50, and POET-DC5-R50) with a total batch size of 64 (50 for DC5 models) for 250 epochs, for the comparison with the baseline models. The baseline models were trained for 100 epochs with the default learning schedule for AE + ResNet models in MMPose (44), with similar augmentation methods (without coarse dropout, but with affine transformation augmentation). Despite the commonly used training schedule

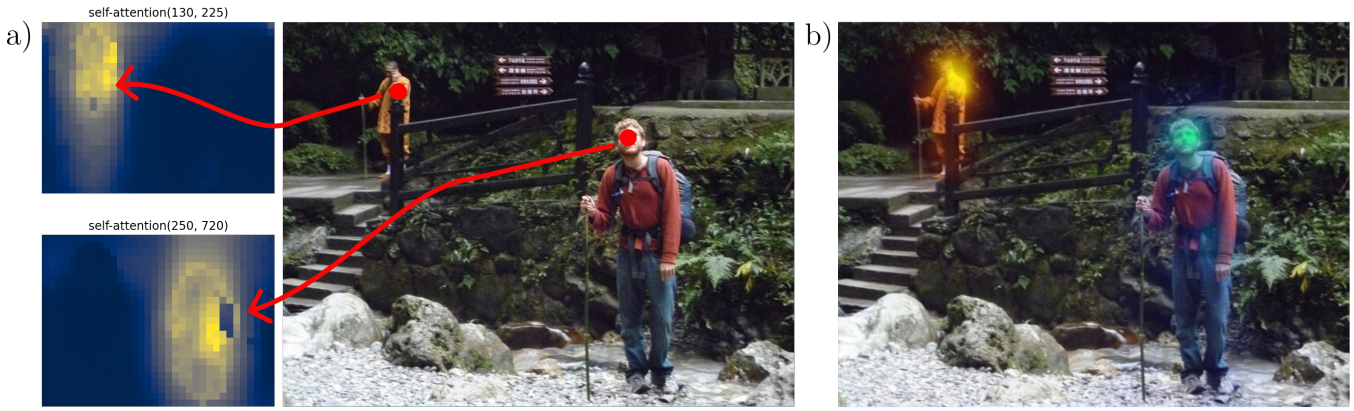

Fig. 5. a) Encoder self-attention reference points highlighted in red. The encoder attends locally to each individual. **b)** Decoder attention scores for predicted individuals. The decoder attends to a human's most distinguishable part, the face region. The image is from the COCO validation set.

with 300 epochs, we train the baseline models only for 100 epochs, as we actually initialized the the AE + ResNet models from the ones trained with stride 4 on COCO and reduced the stride by removing upsampling layers.

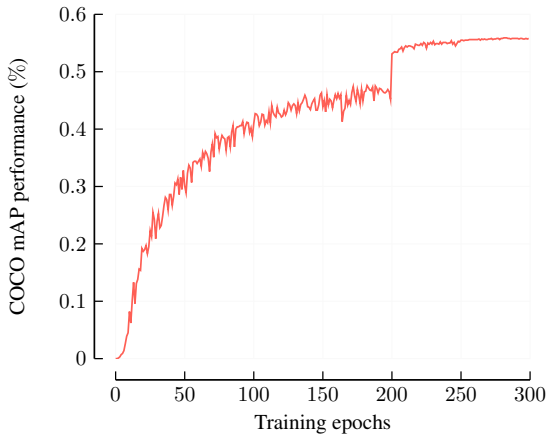

Fig. 6. Evolution of mAP on COCO validation set for POET-R50 trained for 300 epochs to compare to state-of-the-art models. Learning rate was dropped after 200 and 250 epochs.

Qualitative results When we trained POET-R50 with the loss in Equation Eq. (4) as well as the cross-validated hyperparameters, we found that class, keypoint, visibility and center loss decreased (Figure 4). We then checked if the predictions were accurate on test images. Figure 3 depicts predictions of POET-R50 on example images form COCO-val. Furthermore, we plot failure cases and conclude that POET can successfully tackle the problem of multi-instance pose estimation.

Quantitative evaluation Firstly, to quantify the performance, we calculated mAP over learning and found that it reaches high performance (Figure 6). Next, we compare our results to state-of-the-art methods on COCO test-dev (Table 1). We split the methods into top-down and bottom-up approaches and report numbers without multi-scale testing or extra training data in order to have a fair comparison. We find that POET-R50 performs competitively with other bottom-up methods for large humans (AP_L), but has lower performance for small/medium humans.

We reason that this is due to larger stride for the encoders in all strong methods (e.g. ≥ 4, see Table1), which contributes to inferior spatial resolution at the input to the transformer. Transformers scale quadratically $\mathcal{O}((H \cdot W/S^2)^2)$ in the input dimensions, and thus increasing the stride is costly. In order to demonstrate the powerful potential of our method, in comparison to previous methods, we next compare POET to baseline models with ResNet backbones (and varying strides).

We chose associative embedding (AE) (12, 45) as the model to compare to, as this method was proven to be a strong bottom-up method and currently is state-of-the-art when applied with the high resolution backbone HigherHRNet (15). We create baseline models that were trained with the same pretrained ResNet backbones, input image sizes and similar data augmentation.

We vary the overall stride of the ResNet backbones for AE from 4 to up to 32, to assure that both methods receive the same feature dimensions as input. Table 2 shows AP and AR values on COCO-val for baseline and POET models. POET outperforms the baseline methods (with the same stride) by a large margin and (for stride 16) is better than the baseline models even with stride 4, which provides evidence that the transformer head is suitable to learn multiple poses in an image, even from low-resolution feature maps (Figure 1). Future work should focus on finding hyperparameters for training POET with smaller stride, which will likely strongly improve the performance.

Analysis

Transformer attentions. In order to understand better what the roles of the encoder and decoder in the transformer architecture are, we depict the attention maps on a sample image (Figure 5). We find that the transformer's encoder attends locally to each individual while the decoder particularly puts attention on the seemingly most distinguishable part of each individual, its face.

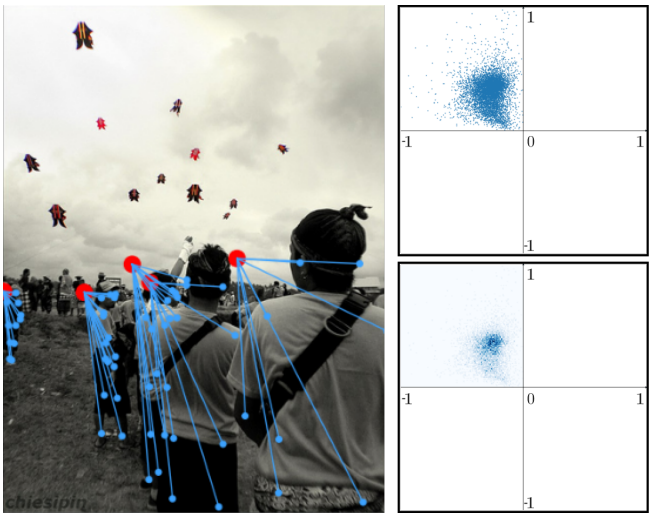

Fig. 7. Left: Figure showing the predicted center (red) and relative offset per body-part as blue vectors for multiple individuals on an example image. **Right:** Scatter plot and planar histogram of relative offset of predicted center vs. ground truth center (normalized by bounding box diagonal). Most data points fall in the IInd quadrant, demonstrating that POET-R50 indeed learns to be biased to the upper-left.

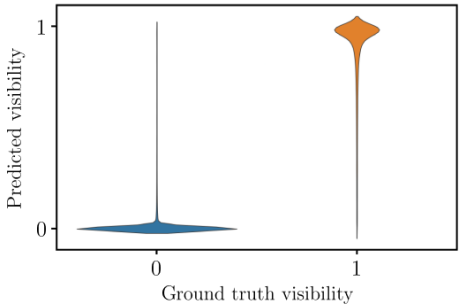

Fig. 8. Violin plot of POET-R50 predicted visibility vs. COCO ground truth visibility for all images and annotated humans in COCO-val.

Learned Centers. Interestingly, POET learns human centers to the left side of the head (Figure 7). We hypothesize that the head is the most distinguishable part of the human body and therefore putting the center next to it helps as a reference point for predicting the rest of the body. In fact, when enforcing the model to learn centers closer to the humans center of mass, by up-weighting the center loss in the the training loss (Equation 3), POET easily learns to predict the centers of mass, but fails at learning the keypoints correctly.

Visibility. The loss formulation also makes the model learn the visibility of each keypoint together with the location. However, the metrics used for the COCO keypoint detection challenge do not take into account predicted visibilities. We found that POET accurately predicts the corresponding visibility for each predicted body part (Figure 8).

Decoder Analysis. We analyze the role of the Transformer decoder and its layers by looking at the predictions at each stage of the decoding. We find that the average performance stabilizes after 3-5 decoder layers (Figure 9).

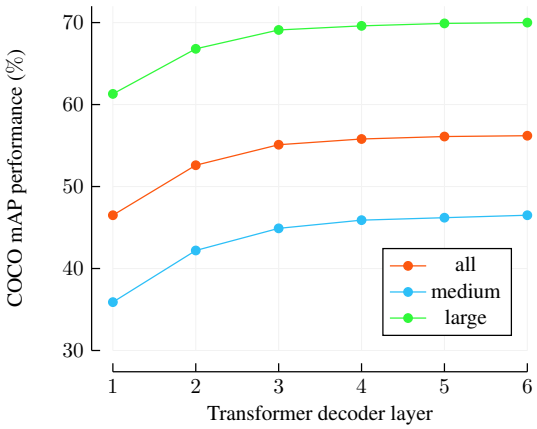

Fig. 9. Evolution of the mAP accuracy on COCO-val through decoder layers of POET-R50 illustrating that the performance saturates after four decoding layers.

Conclusions

We presented POET, a novel pose estimation method based on a convolutional encoder, transformers and bipartite matching loss for direct set prediction. Our approach achieves strong results on the difficult COCO keypoint challenge and is the first that is end-to-end trainable. POET is inspired by the recent DETR (17), which tackled object recognition and panoptic segmentation, with transformers.

Currently, POET does not achieve state-of-the-art performance, but we expect that it will inspire future research to address those challenges. Similarly, DETR performed worse for small objects and only achieved about 80% of state-of-the-art performance (17); comparable to POET (see table 1). One major limitation of POET and DETR is the slow convergence, and large memory demand which makes experimentation with high resolution backbones, which are important for accurate pose estimation, costly. Recent work directly addressed these problems for object recognition (26, 27, 50). Importantly, our method is simple and could be applied to any backbone that is trained end-to-end to do multi-instance pose estimation.

Acknowledgements

We are grateful to Steffen Schneider, Shaokai Ye, Mu Zhou, and Mackenzie Mathis for discussions as well as Axel Bisi, Alberto Chiappa, Alessandro Marin Vargas, Nicholas Robertson and Lazar Stojkovic for comments on an earlier version of this manuscript.

References

1. Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1):4 – 18, 2007. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2006.10.016. Special Issue on Vision for Human-Computer Interaction.
2. Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.
3. Alexander Mathis, Steffen Schneider, Jessy Lauer, and Mackenzie Weygandt Mathis. A primer on motion capture with deep learning: principles, pitfalls, and perspectives. *Neuron*, 108(1):44–65, 2020.
4. Thomas K Uchida and Scott L Delp. *Biomechanics of Movement: The Science of Sports, Robotics, and Rehabilitation*. MIT Press, 2021.
5. Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.
6. Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.
7. Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020.
8. Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv preprint arXiv:2012.14214*, 2020.
9. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
10. Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *CVPR'17*, 2017.
11. George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.
12. Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017.
13. Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.
14. Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6951–6960, 2019.
15. Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.
16. Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *arXiv preprint arXiv:2103.02440*, 2021.
17. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
19. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
20. Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
21. Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.
22. Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.
23. Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.
24. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
25. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
26. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.
27. Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris Kitani. Rethinking transformer-based set prediction for object detection. *arXiv preprint arXiv:2011.10881*, 2020.
28. Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392*, 2020.
29. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
30. Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.
31. Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.
32. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
33. Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision*, pages 627–642. Springer, 2016.
34. Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.
35. Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.
36. Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, Dec 2019. ISSN 1007-0214. doi: 10.26599/TST.2018.9010100.
37. Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. *arXiv preprint arXiv:2012.09760*, 2020.
38. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
39. Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
40. Bartel Leendert Van der Waerden. *Algebra*, volume 1. Springer Science & Business Media, 2003.
41. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
42. Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
43. George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.
44. MMPose Contributors. Openmmlab pose estimation toolbox and benchmark - ae-resnet. https://mmpose.readthedocs.io/en/latest/bottom_up_models.html#associative-embedding-ae-resnet, 2020.
45. MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.
46. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
47. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
48. Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
49. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
50. Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020.