

Article

Not peer-reviewed version

DualPose: Dual-Block Transformer Decoder with Contrastive DeNoising for Multi-Person Pose Estimation

[Matteo Fincato](#) and [Roberto Vezzani](#) *

Posted Date: 7 April 2025

doi: 10.20944/preprints202504.0467.v1

Keywords: Contrastive Denoising; DualPose; Human Pose Estimation; Multi-Person Pose Estimation; Transformer-Based Models



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

DualPose: Dual-Block Transformer Decoder with Contrastive DeNoising for Multi-Person Pose Estimation

Matteo Fincato  and Roberto Vezzani 

University of Modena and Reggio Emilia, DIFE, Via P. Vivarelli 10, 41125 Modena, Italy

* Correspondence: roberto.vezzani@unimore.it

Abstract: Multi-person pose estimation is the task of detecting and regressing the keypoint coordinates of multiple people in a single image. Significant progress has been achieved in recent years, especially with the introduction of transformer-based end-to-end methods. In this paper, we present DualPose, a novel framework that enhances multi-person pose estimation by leveraging a dual-block transformer decoding architecture. Class prediction and keypoint estimation are split into parallel blocks so each sub-task can be separately improved and the risk of interference is reduced. This architecture improves the precision of keypoint localization and the model's capacity to accurately classify individuals. To improve model performance, the keypoints-block uses parallel processing of self-attentions, providing a novel strategy that improves keypoint localization accuracy and precision. Additionally, DualPose incorporates a contrastive denoising (CDN) mechanism, leveraging positive and negative samples to stabilize training and improve robustness. Thanks to CDN, a variety of training samples is created by introducing controlled noise into the ground truth, improving the model's ability to discern between valid and incorrect keypoints. DualPose achieves state-of-the-art results outperforming recent end-to-end methods, as shown by extensive experiments on the MS COCO and CrowdPose datasets. The code and pretrained models are publicly available.

Keywords: contrastive denoising; DualPose; human pose estimation; multi-person pose estimation; transformer-based models

1. Introduction

The task of detecting and regressing the keypoint positions of multiple individuals in an image is known as multi-person pose estimation. Considerable progress has been made in multi-person pose estimation over the last few years. Methods are generally divided into two categories: non-end-to-end and end-to-end. Top-down [1–6], bottom-up [7–10], and one-stage model [11–14] approaches are examples of non-end-to-end methods.

Top-down methods, which are based on the highly accurate person detector, identify people in an image first and then estimate their poses. Bottom-up methods are effective in crowded environments, as they predict all keypoints in the image and then associate them into individual poses, thus removing the need for explicit detection. One-stage methods predict keypoints and their associations directly, simplifying the process by combining pose estimation and detection into a single step.

Transformers were originally introduced by Vaswani et al. [15] for machine translation. One of the main advantages of Transformers is their global computations and perfect memory, which makes them more suitable than RNNs on long sequences. Using this ability to handle sequences, DETR [16] modeled the object detection problem as generating a sequence of rectangles, one for each identified object. The same DETR paradigm was then extended to pose estimation, where the 4-dimensional representation of the detection box becomes a sequence of joint positions. PRTR [17] and TFPose [18] are two seminal examples, to which the reader can refer for further details. Inspired by the success of

DETR, end-to-end methods use transformers to improve accuracy and performance by integrating the entire process into a single framework.

By treating object detection as a direct prediction of a set of objects, these techniques transform multi-person pose estimation and eliminate the need for post-processing steps such as non-maximum suppression. GroupPose [19] optimizes the pose estimation pipeline and achieves superior performance by using group-specific attention mechanisms to refine interactions between keypoint and instance queries.

We build on this line of work and introduce a new and effective framework for end-to-end multi-person pose estimation called **DualPose**, as depicted in Figure 1. DualPose presents significant innovations that improve computational effectiveness and accuracy. First, we modify the transformer decoder architecture by employing two distinct blocks: one for processing class prediction and the other for keypoint localization. Inspired by GroupPose [19], we use $N \times K$ keypoint queries to regress $N \times K$ positions and N class queries to categorize human poses. For class queries, the class-block transformer accurately identifies and categorizes the people in the input image. The output of this class-block transformer is then added to the keypoints-block transformer, providing more comprehensive contextual information. Furthermore, in comparison to conventional sequential processing, we introduce an improved strategy within the keypoints-block by employing a novel parallel computation technique for self-attentions to enhance both efficiency and accuracy.

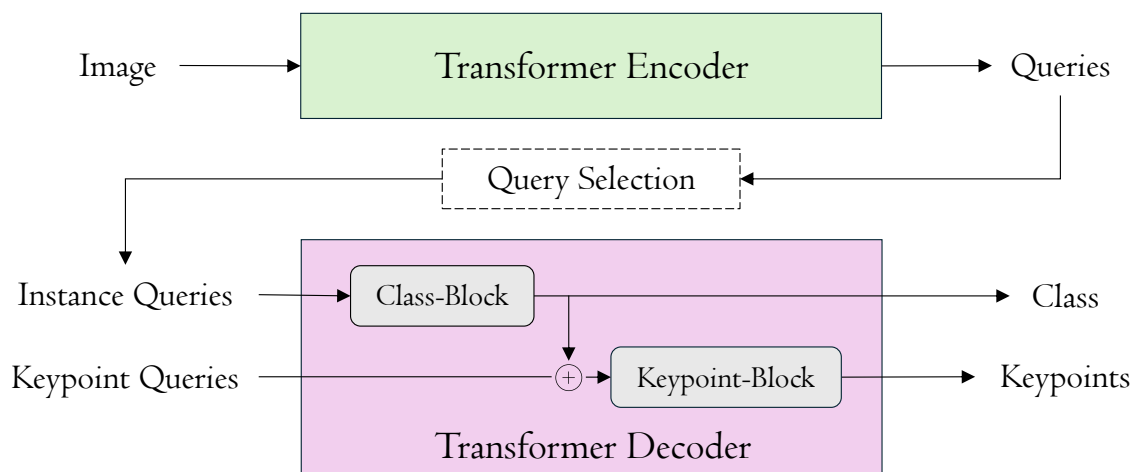


Figure 1. DualPose Scheme. DualPose operates through a two-stage strategy, where an image is processed by an Encoder to generate queries. These queries are then filtered and passed into a Decoder, which is composed of two blocks: the Class-Block, responsible for class predictions, and the Keypoint-Block, which handles keypoint localization. The final outputs include both class predictions and keypoints coordinates.

To stabilize training and improve effectiveness, DualPose also includes a contrastive denoising (CDN) mechanism [20]. This contrastive method improves the model's ability to differentiate real and false keypoints. We use two hyperparameters to control the noise scale, generating positive samples with lower noise levels (to reconstruct their corresponding ground truth keypoint positions) and negative samples with heavily distorted poses, which the model learns to predict as 'no pose'. Both types of samples are generated based on ground truth data. Furthermore, we present a novel adaptation, inherited from object detection [20], by combining both L1 and Object Keypoint Similarity (OKS) as reconstruction losses. The L1 loss ensures fine-grained accuracy in keypoint localization, while the OKS loss accounts for variations in human pose structure and aligns the predicted keypoints with ground-truth annotations.

Experimental results demonstrate that our novel approach, DualPose, outperforms recent end-to-end methods on both MS COCO [21] and CrowdPose [22] datasets. Specifically, DualPose achieves an AP of 71.2 with ResNet-50 [5] and 73.4 with Swin-Large [23] on the MS COCO val2017 dataset.

In this work, we introduce key innovations that improve the accuracy and effectiveness of multi-person pose estimation:

- DualPose uses a dual-block decoder to separate class prediction from keypoint localization, reducing task interference.
- Parallel self-attention in the keypoint block improves effectiveness and precision.
- A contrastive denoising mechanism enhances robustness by helping the model distinguish between real and noisy keypoints.
- L1 and OKS losses are applied exclusively for keypoint reconstruction, rather than for bounding boxes as in object detection.

2. Related work

Multi-person pose estimation has seen substantial advances, with methodologies commonly divided into non-end-to-end and end-to-end approaches. Non-end-to-end methods include top-down and bottom-up strategies, as well as one-stage models. End-to-end methods integrate the entire process into a single streamlined framework.

2.1. Non-End-to-End Methods

Non-end-to-end approaches for multi-person pose estimation encompass both top-down [1–4,6,24] and bottom-up [7–10] methods.

Top-down methods identify individual subjects in an image and then estimate their poses. However, the accuracy of person detection is a major factor in how effectively they perform especially in busy or obscure environments. The two-step procedure also amplifies the risk of error propagation, and increases processing complexity.

Bottom-up methods detect all keypoints first and then associate them to form poses. This approach is more efficient in crowded scenes, but adds complexity during keypoint association, often resulting in reduced accuracy in challenging scenarios.

One-stage methods [11–14,25] streamline pose estimation by combining detection and pose estimation into a single step. These models directly predict keypoints and their associations from the input image, eliminating intermediate stages such as person detection or keypoint grouping.

2.2. End-to-End Methods

End-to-end methods [19,26–28] revolutionize multi-person pose estimation by integrating the entire pipeline into a single, unified model. Inspired by Detection Transformer (DETR) [16], which pioneered the use of transformers for object detection, along with its variants [20,29–33], these methods streamline the process and improve accuracy.

PETR [26] employs a Transformer-based architecture to solve multi-person pose estimation as a hierarchical set prediction problem, first regressing poses using a pose decoder and then refining keypoints using a keypoint decoder. QueryPose [27] adopts a sparse query-based framework to perform human detection and pose estimation, eliminating the need for dense representations and complex post-processing. ED-Pose [28] integrates explicit keypoint box detection to unify global (human-level) and local (keypoint-level) information, enhancing both efficiency and accuracy through a dual-decoder architecture. GroupPose [19] uses group-specific attention to refine keypoint and instance query interactions, optimizing the process and achieving superior performance with one simple transformer decoder.

These methods have limitations: PETR struggles with keypoint precision and high computational costs, ED-Pose adds complexity with its dual-decoder, and QueryPose falters with occlusions. Our model addresses these issues with a dual-block transformer that separates class and keypoint predictions, enhancing precision and effectiveness, while the contrastive denoising mechanism improves robustness without added complexity.

3. Proposed method

We propose DualPose, a novel and effective framework for end-to-end multi-person pose estimation. Building on previous end-to-end frameworks [19,26–28], we formulate the multi-person

pose estimation task as a set prediction problem. DualPose employs a more advanced dual-block transformer decoder compared to previous approaches, which only use a basic transformer decoder [29]. This improves accuracy and computational efficiency. In the following sections, we detail the core components of DualPose.

3.1. Overall

The DualPose model consists of four primary components: a backbone [5,23], a transformer encoder [15], a transformer decoder, and task-specific prediction heads. This architecture is designed to simultaneously compute K keypoints (e.g., $K = 17$ on MS COCO) for N human instances in the input image. Refer to Figure 2 for the complete model architecture.

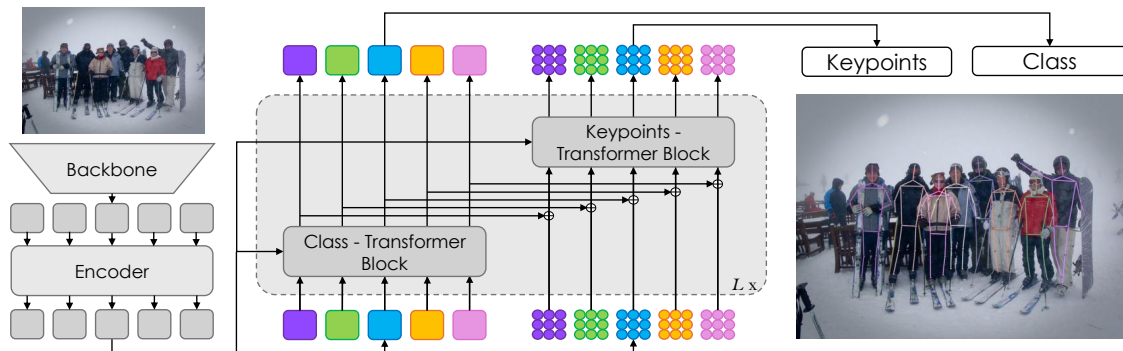


Figure 2. DualPose. Overview of the proposed system: The backbone extracts features from the input image, which are then processed through multiple encoder layers. The output is then given to the decoder. It uses two different transformer blocks: the **Keypoints-Transformer Block**, which focuses on keypoint prediction, and the **Class-Transformer Block**, which is dedicated to class prediction. To improve the keypoint predictions, the Class-Transformer Block's output is merged into the Keypoints-Transformer Block. The final output integrates class and keypoint predictions for each detected human instance.

3.1.1. Backbone and Transformer Encoder

We follow the DETR framework [29] in building the transformer encoder and backbone. The model processes an input image and generates multi-level features using 6 deformable transformer layers [29]. These features are then sent to the following transformer decoder.

3.1.2. Transformer Decoder

We introduce a dual-query system for the transformer decoder, where N instance queries and $N \times K$ keypoint queries are generated. The instance queries classify each person in the image, while the keypoint queries predict the positions of the keypoints for each detected individual. By treating each keypoint independently, these keypoint queries allow for precise regression of the $N \times K$ keypoint positions. This separation of tasks enables the model to handle instance detection and keypoint localization more effectively.

Each decoder layer follows the design of previous DETR frameworks, incorporating self-attention, cross-attention implemented with deformable attention, and a feed-forward network (FFN). However, each decoder layer architecture has been extended to include two distinct blocks: one for class prediction and one for keypoint prediction. Unlike GroupPose [19], where two group self-attentions are computed sequentially, our method computes them in parallel.

3.1.3. Prediction Heads

Following GroupPose [19], we use two prediction heads, implemented with FFNs, for human classification and human keypoints regression. In particular, DualPose predicts a classification score for N different human poses and, for each pose, computes the x, y coordinates for K keypoint locations.

3.1.4. Losses

The loss function in DualPose addresses the assignment of predicted poses to their corresponding ground-truth counterparts using the Hungarian matching algorithm [16], ensuring a one-to-one mapping between predictions and ground-truth poses. It includes classification loss (L_{cls}) and keypoint regression loss (L_{kpt}), without additional supervisions such as those in QueryPose [27] or ED-Pose [28]. The keypoint regression loss (L_{kpt}) combines ℓ_1 loss and Object Keypoint Similarity (OKS) [26]. The cost coefficients and loss weights from ED-Pose [28] are used in Hungarian matching and loss calculation.

3.2. Queries

In multi-person pose estimation, the objective is to predict N human poses, each with K keypoint positions for every input image. To achieve this, we use $N \times K$ keypoint queries, where each set of K keypoints represents a single pose. Furthermore, N instance queries are used to classify and score the predicted poses, ensuring accurate identification and assessment of each human instance.

3.3. Query Construction

The process begins by identifying human instances and predicting poses using the output memory of the transformer encoder, following previous frameworks [19,26,28]. We select N instances based on classification scores, resulting in $N \times D$ memory features.

For keypoint queries, each content part ($K \times D$) is constructed by combining randomly initialized learnable keypoint embeddings with corresponding memory features. The position part is initialized based on the predicted human poses. Instance queries use randomly initialized learnable instance embeddings ($1 \times D$) for classification tasks, without explicit position information.

Table 1. Comparative Evaluation using state-of-the-art for MS COCO **va12017.** Reference (Ref) values from previous frameworks are included in this section for context. The terms heatmap-based losses, human box regression losses, and keypoint regression losses are represented by the acronyms "HM," "BR," and "KR," in that order. The term "RLE" denotes Poseur’s residual log-likelihood estimation. † denotes the flipping test. ‡ eliminates the Poseur prediction uncertainty estimation for an equitable regression comparison. (**Bold**: Best. Underline: Second Best. * indicates retrained model on $2 \times A40$ GPU.)

Method			Ref	Backbone	Loss	AP	AP ₅₀	AP ₇₅	AP _M	AP _L
Non-End-to-End	Top-Down	Mask R-CNN [5]	CVPR 17	ResNet-50	HM	65.5	87.2	71.1	61.3	73.4
		Mask R-CNN [5]	CVPR 17	ResNet-101	HM	66.1	87.4	72.0	61.5	74.4
		PRTR [†] [17]	CVPR 21	ResNet-50	KR	68.2	88.2	75.2	63.2	76.2
		Poseur [‡] [34]	ECCV 22	ResNet-50	RLE	70.0	-	-	-	-
		Poseur [34]	ECCV 22	ResNet-50	RLE	74.2	<u>89.8</u>	81.3	71.1	<u>80.1</u>
	Bottom-Up	HrHRNet [†] [10]	CVPR 20	HRNet-w32	HM	67.1	86.2	73.0	61.5	76.1
		DEKR [†] [35]	CVPR 21	HRNet-w32	HM	68.0	86.7	74.5	62.1	77.7
		SWAHR [†] [36]	CVPR 21	HRNet-w32	HM	68.9	87.8	74.9	63.0	77.4
		LOGO-CAP [†] [37]	CVPR 22	HRNet-w32	HM	69.6	87.5	75.9	64.1	78.0
	One-Stage	DirectPose [13]	-	ResNet-50	KR	63.1	85.6	68.8	57.7	71.3
		CenterNet [†] [25]	-	Hourglass-104	KR+HM	64.0	-	-	-	-
		FCpose [11]	CVPR 21	ResNet-50	KR+HM	63.0	85.9	68.9	59.1	70.3
		InsPose [38]	ACM MM 21	ResNet-50	KR+HM	63.1	86.2	68.5	58.5	70.1
End-to-End	Previous Works	PETR* [26]	CVPR 22	ResNet-50	HM+KR	67.4	86.1	74.9	61.2	76.0
		PETR* [26]	CVPR 22	Swin-L	HM+KR	71.6	89.3	79.3	65.8	79.9
		QueryPose* [27]	NeurIPS 22	ResNet-50	BR+RLE	68.0	86.9	72.9	62.2	75.0
		QueryPose* [27]	NeurIPS 22	Swin-L	BR+RLE	71.9	89.5	77.7	66.9	79.9
		ED-Pose* [28]	ICLR 23	ResNet-50	BR+KR	69.8	87.9	76.4	64.4	78.2
		ED-Pose* [28]	ICLR 23	Swin-L	BR+KR	72.0	89.9	79.9	67.1	81.2
		GroupPose* [19]	ICCV 23	ResNet-50	KR	70.4	88.2	77.8	65.1	78.3
		GroupPose* [19]	ICCV 23	Swin-L	KR	72.8	90.0	80.6	68.0	81.6
	Ours	DualPose	-	ResNet-50	KR	71.2	88.9	77.7	66.2	79.0
		DualPose	-	Swin-T	KR	72.2	89.4	78.7	67.6	80.4
		DualPose	-	Swin-L	KR	<u>73.4</u>	91.8	<u>80.7</u>	<u>69.0</u>	82.3

Table 2. Comparisons with state-of-the-art methods on MS COCO test-dev2017 dataset. Notations are consistent with Table 1

Method			Ref	Backbone	Loss	AP	AP ₅₀	AP ₇₅	AP _M	AP _L
Non-End-to-End	Top-Down	Mask R-CNN [5]	CVPR 17	ResNet-50	HM	63.9	87.7	69.9	59.7	71.5
		Mask R-CNN [5]	CVPR 17	ResNet-101	HM	64.3	88.2	70.6	60.1	71.9
		PRTR [†] [17]	CVPR 21	ResNet-50	KR	68.8	89.9	76.9	64.7	75.8
		PRTR [†] [17]	CVPR 21	HRNet-w32	KR	<u>71.7</u>	90.6	79.6	<u>67.6</u>	78.4
	Bottom-Up	HrHRNet [†] [10]	CVPR 20	HRNet-w32	HM	66.4	87.5	72.8	61.2	74.2
		DEKR [†] [35]	CVPR 21	HRNet-w32	HM	67.3	87.9	74.1	61.5	76.1
		SWAHR [†] [36]	CVPR 21	HRNet-w32	HM	67.9	88.9	74.5	62.4	75.5
		LOGO-CAP [†] [37]	CVPR 22	HRNet-w32	HM	68.2	88.7	74.9	62.8	76.0
	One-Stage	DirectPose [13]	-	ResNet-50	KR	62.2	86.4	68.2	56.7	69.8
		CenterNet [†] [25]	-	Hourglass-104	KR+HM	63.0	86.8	69.6	58.9	70.4
		FCpose [11]	CVPR 21	ResNet-50	KR+HM	64.3	87.3	71.0	61.6	70.5
		InsPose [38]	ACM MM 21	ResNet-50	KR+HM	65.4	88.9	71.7	60.2	72.7
End-to-End	Previous Works	PETR* [26]	CVPR 22	ResNet-50	HM+KR	66.2	88.1	73.8	60.2	74.5
		PETR* [26]	CVPR 22	Swin-L	HM+KR	68.9	89.8	76.9	63.5	76.2
		QueryPose* [27]	NeurIPS 22	Swin-L	BR+RLE	70.4	90.6	77.0	65.7	77.7
		ED-Pose* [28]	ICLR 23	ResNet-50	BR+KR	68.1	88.9	75.8	62.9	75.8
		ED-Pose* [28]	ICLR 23	Swin-L	BR+KR	71.3	90.8	79.2	66.2	78.4
		GroupPose* [19]	ICCV 23	ResNet-50	KR	68.4	89.1	76.2	63.3	76.6
		GroupPose* [19]	ICCV 23	Swin-L	KR	71.5	<u>90.9</u>	<u>79.6</u>	66.3	<u>79.3</u>
	Ours	DualPose	-	ResNet-50	KR	69.5	89.4	76.6	64.4	77.4
		DualPose	-	Swin-T	KR	70.8	90.1	77.2	65.3	78.5
		DualPose	-	Swin-L	KR	71.8	92.2	79.8	67.9	80.1

3.4. Contrastive Denoising

In this work, we improve training stability and performance by integrating a contrastive denoising (CDN) mechanism into the transformer decoder. During training, CDN adds controlled noise to the ground-truth labels and poses, helping the model to better distinguish between true and false keypoints. Our approach is inspired by the work of Zhang [20], which applies CDN to object detection. However, we have adapted it to suit the context of pose estimation, allowing our model to achieve greater robustness in discerning accurate poses from noisy data. A summary schema is shown in Figure 3.

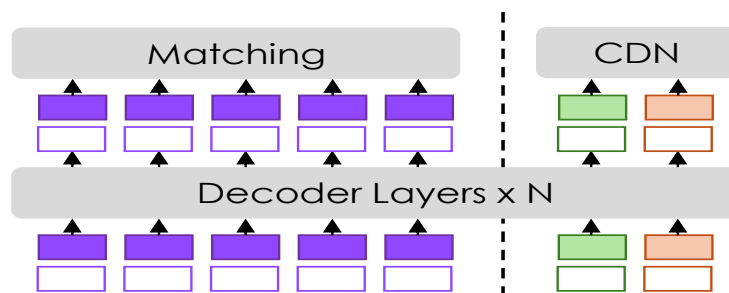


Figure 3. Contrastive DeNoising training. Additional queries are introduced, where the reference poses are derived from the ground truth (GT) combined with noise. These queries are then divided into positive and negative samples based on the amount of noise added to the GT.

To this end, we use two different hyperparameters, λ_1 and λ_2 , to control the noise scale for keypoints in pose estimation. Here, $\lambda_1 < \lambda_2$, ensuring a hierarchy in noise levels. Positive queries are designed to closely reconstruct ground truth (GT) keypoints and are generated within an inner boundary defined by λ_1 . Negative queries are meant to predict ‘no pose’, simulating difficult scenarios

that force the model to distinguish between true and false positives. They are placed between the inner and outer boundaries (λ_1, λ_2).

Positive and negative queries are organized into CDN groups. For an image containing n ground truth (GT) poses, each CDN group generates $2 \times n$ queries, with each GT pose yielding a positive and a negative query. The reconstruction losses are L_1 and OKS for keypoint regression, and focal loss for classification. An additional focal loss is used to label negative samples as background.

This contrastive denoising method greatly enhances the model’s performance in multi-person pose estimation by improving its ability to accurately identify keypoints despite varying degrees of noise.

3.5. Dual-Block Decoder

In DETR frameworks [16,29,30], the self-attention of the transformer decoder captures interactions between queries, but processing keypoint regression and class predictions simultaneously can lead to inefficiencies and interference between tasks. This often results in reduced accuracy and precision when handling complex scenes with multiple individuals or occlusions.

To address these issues, we propose a new transformer decoder architecture in DualPose that decouples the computation of keypoints and class predictions into two distinct blocks, allowing each one to process the queries independently and more effectively. This division allows a more specialized handling of the corresponding tascanhe class-block focused on class prediction and the keypoints-block on keypoint localization. This dual-block structure minimizes interference between the different query types, improving accuracy and effectiveness in multi-person pose estimation.

3.5.1. Class-Block

The class-block transformer processes class queries independently from keypoint-queries for class prediction. As a result, it can focus on accurately identifying individuals across the image. The output of the class-block is then added to the input of the keypoints-block, providing richer contextual information.

3.5.2. Keypoint-Block

The keypoints-block transformer is designed to handle the prediction of body joints. We incorporate the group self-attention mechanism from GroupPose [19], but we introduce a new enhancement by processing these self-attentions in parallel. Specifically, the queries, keys, and values are split into two halves, each processed by a separate self-attention mechanism simultaneously. This design boosts effectiveness and accuracy in predicting keypoints, while leveraging the refined contextual information from the class-block.

Table 3. Comparisons with state-of-the-art methods on CrowdPose test dataset. Notations are consistent with Tabel 1. Swin-L is used as the backbone. * indicates retrained model on $2 \times$ A40 GPU.

Method	Loss	AP	AP ₅₀	AP ₇₅	AP _E	AP _M	AP _H
PETR * [26]	HM+KR	70.0	88.7	77.0	75.8	70.5	64.5
QueryPose * [27]	BR+RLE	71.1	90.2	76.6	77.8	71.9	63.8
ED-Pose * [28]	BR+KR	71.5	88.7	78.1	79.0	72.1	62.3
GroupPose * [19]	KR	72.7	90.0	79.0	79.5	73.2	64.8
DualPose	KR	72.9	90.8	79.2	81.3	74.5	66.1

Table 4. Ablation experiments for DualPose’s queries. Evaluated on MS COCO val_{2017} . Default settings are marked in gray. (a) Designing queries for pose estimation. Both instance (inst) and keypoint (kpt) queries are crucial, with a particular emphasis on the importance of keypoint queries. (b) Number of instance queries. Testing indicates that setting the number of queries to 100 results in better performance compared to using either 50 or 200 queries. (c) Number of denoising queries. The default setting is 50 queries. Exceeding 50 queries results in similar performance but with longer execution times.

q. types	AP	AP _M	AP _L	#IQ	AP	AP _M	AP _L	#DQ	AP	AP _M	AP _L
inst & kpt	71.2	66.2	79.0	50	70.1	65.2	77.3	25	70.7	65.8	78.5
only kpt	69.5	64.0	77.9	100	71.2	66.2	79.0	50	71.2	66.2	79.0
only inst	63.4	61.5	70.0	200	70.9	66.2	78.8	100	71.0	66.2	78.8
(a)				(b)				(c)			

Table 5. Component Analysis. Performance improvements obtained with the method components.

Model	AP	AP ₅₀	AP ₇₅	AP _M	AP _L
Dual-Block	70.6	88.5	77.1	65.5	78.3
+ Denoising	70.9	88.7	77.5	65.8	78.5
+ Contrastive	71.0	88.8	77.5	66.2	79.0
+ Split	71.2	88.9	77.7	66.2	79.0

Table 6. Dual-Block Design. Both class and keypoint block are essential to achieve the best performance.

Configuration	AP	AP _M	AP _L
DualPose (Class-Block + Keypoint-Block)	71.2	66.2	79.0
Class-Block Only (Keypoint-Block disabled)	64.7	62.0	70.4
Keypoint-Block Only (Class-Block disabled)	69.9	64.9	78.3

Table 7. Comparison of FPS and Time (ms) for different methods at various input resolutions with ResNet-50 on a NVIDIA A40 GPU. (**Bold**: Best. Underline: Second Best)

Method	480 × 800 (FPS ↑/Time ↓)	800 × 1333 (FPS ↑/Time ↓)
PETR [26]	18.0 / 55 ms	10.2 / 98 ms
QueryPose [27]	16.5 / 61 ms	11.7 / 85 ms
ED-Pose [28]	33.3 / 30 ms	21.2 / 47 ms
GroupPose [19]	52.6 / 19 ms	27.7 / 36 ms
DualPose	<u>50.4 / 20 ms</u>	<u>26.0 / 38 ms</u>

4. Experiments

In this section, we describe the experimental setting, including the dataset and the training protocol used in our pipeline, and subsequently compare our results with state-of-the-art approaches.

4.1. Settings

4.1.1. Dataset

In our experiments, two popular datasets for human pose estimation are used: MS COCO [21] and CrowdPose [22]. **MS COCO** contains 250,000 person instances across 200,000 images, each with 17 keypoint annotations (we set $K = 17$ for this dataset). DualPose is evaluated using the COCO val_{2017} and $test-dev$ sets after being trained on the COCO $train_{2017}$ set.

CrowdPose includes 20,000 images with 80,000 person instances, each with 14 keypoint annotations ($K = 14$ for CrowdPose). Since crowded and occluded scenes occur frequently, CrowdPose poses additional challenges. The provided train/test split has been adopted.

4.1.2. Evaluation Metrics

The OKS-based average precision (AP) scores serve as the primary evaluation metric for both datasets. For MS COCO, we report AP scores across various thresholds and object sizes (medium and large), specifically denoted as AP , AP_{50} , AP_{75} , AP_M , and AP_L , following the standard evaluation protocol. In the case of CrowdPose, we evaluated model performance under different crowding conditions using AP scores at different thresholds and crowding levels, labeled as AP , AP_{50} , AP_{75} , AP_E , AP_M , and AP_H , which represent easy, medium, and hard crowding scenarios, respectively.

4.1.3. Implementation Details

Our training and testing procedures follow those of ED-Pose [28] and GroupPose [19]. We utilize the AdamW optimizer [39,40] with a weight decay of 1×10^{-4} . We set the base learning rate to 1×10^{-4} and the backbone's learning rate to 1×10^{-5} , consistent with DETR frameworks. A total batch size of 16 is adopted. The model is trained for 60 epochs on MS COCO [21] and 80 epochs on CrowdPose [22]. The learning rate is decayed by a factor of 0.1 at the 50th epoch for MS COCO and at the 70th epoch for CrowdPose. We use common data augmentations, such as random flips, crops, and resizing, as in DETR frameworks [16,20,28,29] during training. Images are resized so that the long side does not exceed 1333 pixels, and the short side is between 480 and 800 pixels. All experiments are conducted on $2 \times$ NVIDIA A40 GPUs.

4.2. Experimental results

Our goal is to develop an effective framework for end-to-end multi-person pose estimation. Hence, we primarily compare DualPose with previous end-to-end frameworks, such as PETR [26], QueryPose [27], ED-Pose [28], and GroupPose [19]. Additionally, to demonstrate the effectiveness of our approach, we include comparisons with non-end-to-end frameworks, encompassing top-down [17,41], bottom-up [10,35–37], and one-stage methods [11,13,25,38].

4.2.1. Comparisons with end-to-end frameworks on COCO

The comparisons on the COCO `val2017` set and `test-dev` set are shown in Tables 1 and 2. These findings confirm that DualPose performs consistently better than PETR [26] and QueryPose [27], and is on par with ED-Pose [28] and GroupPose [19]. For fairness, every model is re-trained on $2 \times$ NVIDIA A40 GPU.

DualPose, using ResNet-50 [5] as its backbone, achieves an Average Precision (AP) of 71.2 on the COCO `val2017`, outperforming models such as PETR (67.4) and QueryPose (68.0). DualPose improves further when using the Swin-Large [23] backbone, reaching an AP of 73.4, comparable to GroupPose (72.8) and ED-Pose (72.0), indicating its competitiveness with state-of-the-art models. Furthermore, the values for AP_{50} , AP_{75} , AP_M , and AP_L reflect the same trend as the overall AP, highlighting the effectiveness of the dual-block transformer decoder and contrastive denoising mechanism.

On the COCO `test-dev` set, DualPose achieves 69.5, 70.8, and 71.8 AP with ResNet-50 [5], Swin-Tiny [23], and Swin-Large [23] backbones, respectively. These results confirm DualPose's role as a top-performing framework for multi-person pose estimation.

4.2.2. Comparisons with non-end-to-end frameworks on COCO

On the COCO `val2017` set, DualPose significantly surpasses bottom-up methods [10,35–37] and one-stage methods [11,13,25,38]. One-stage approaches like FCpose [11] and InsPose [38] with ResNet-50 [5] achieve even lower APs of 63.0 and 63.1, respectively. Likewise, bottom-up methods like DEKR [35] and LOGO-CAP [37] obtain APs of 68.0 and 69.6, respectively. DualPose also outperforms previous top-down methods like PRTR [17] and shows comparable results with Poseur [34]. The results demonstrate that DualPose not only excels within top-down frameworks but also significantly outperforms all one-stage and bottom-up methods.

4.2.3. Comparisons with end-to-end frameworks on CrowdPose

We also conducted experiments on the CrowdPose [22] dataset to further validate the proposed model. Table 3 presents the performance metrics of leading competitors, including ED-Pose [28] and GroupPose [19]. The results indicate that our model achieves state-of-the-art performance, surpassing all other models across every evaluated metric. All experiments were conducted using $2 \times$ NVIDIA A40 GPU, with Swin-L [23] as the backbone.

4.2.4. Qualitative Results

Figure 4 presents the results obtained by DualPose, demonstrating its accuracy in multi-pose person estimation. Each image depicts a complex scenario, where the model successfully distinguishes between instances and accurately identifies key points. The colorful skeletons highlight the effectiveness of our approach, showing the model’s robustness in crowded or partially occluded scenes.



Figure 4. Visualization results of DualPose on MS COCO. An illustration of multi-person pose estimation in a busy environment. The model demonstrates its robustness in handling occlusions and densely populated scenarios by accurately detecting individuals and localizing keypoints.

While DualPose achieves state-of-the-art performance, challenges remain in cases of severe occlusions, where overlapping keypoints may lead to misassignments (see Figure 5). Additionally, unusual poses or complex backgrounds can cause minor localization errors.



Figure 5. Qualitative Analysis. Qualitative comparison between DualPose (top row) and GroupPose [19] (bottom row). The images illustrate the differences in multi-person pose estimation performance, highlighting improvements in keypoint localization and pose consistency achieved by our method.

4.3. Ablations

We conducted a series of ablation experiments to assess the impact of each component and set each parameter of DualPose. Using ResNet-50 as the backbone, all results are reported on the COCO val2017 dataset with 60 epochs of training, unless otherwise noted.

Query Design for Pose Estimation. In the DualPose model, accurate pose estimation relies on both instance (inst) and keypoint (kpt) queries. As Table 4(a) illustrates, the experiments underscore the crucial role of keypoint queries. In this ablation study, the dual-block architecture remains unchanged, while adjustments are made to the input queries for the decoder. When only one type of query is used, the other is omitted, and the remaining queries handle both classification and keypoint regression. The model performs best when both instance and keypoint queries are present, with AP of 71.2, AP_M of 66.2, and AP_L of 79.0. Performance decreases slightly when utilizing only keypoint queries (AP: 69.5) and drops significantly when only instance queries are used (AP: 63.4). These results clearly show that leveraging both query types is essential to achieve optimal results in human detection tasks.

Number of Instance Queries. As shown in Table 4(a), we also analyzed how different numbers of instance queries affect performance. The best results are obtained with 100 queries, and no further improvement is observed when that number is increased. For instance, using 200 queries results in a slight decrease (AP: 78.8), while reducing queries to 50 lowers the AP to 70.7. Consequently, 100 queries strike the best balance between computational efficiency and performance.

Number of Denoising Queries. The impact of the quantity of denoising queries on the model's performance is examined in Table 4(c). The optimal balance is achieved with 50 queries as the default. Increasing the number of queries to 100 does not offer additional gains and slightly reduces AP_L to 78.8, while dropping to 25 queries lowers the overall AP to 70.7, AP_M to 65.8, and AP_L to 78.5. These findings imply that the extra computational cost is not justified when denoising queries exceed 50.

Component Analysis. Table 5 provides a comprehensive overview of each novel component in the proposed model. Incorporating the denoising mechanism improves AP by 0.3 points (from 70.6 to 70.9). Adding the contrastive method further boosts the AP to 71.0. Finally, introducing the split attention mechanism raises AP to a peak of 71.2. All other metrics also improve consistently, desligntrating how each innovation enhances DualPose's state-of-the-art performance.

Dual-Block Design. In DualPose, the class and keypoint blocks are essential, each focusing on classification and keypoint localization tasks, respectively. Each block is tested individually by "disabling" the other to observe the effect on performance. As shown in Table 6, the model achieves optimal performance when both Class-Block and Keypoint-Block are active. When using only the Keypoint-Block, there is a small decrease in performance (AP: 69.5), whereas relying solely on the Class-Block leads to a more significant drop (AP: 68.1). These findings indicate that both blocks are crucial for maximizing accuracy in multi-person pose estimation.

4.3.1. Inference Speed

Table 7 compares the performance of multi-person pose estimation methods (i.e., PETR [26], QueryPose [27], ED-Pose [28], GroupPose [19], and DualPose) across two different input resolutions (480×800 and 800×1333). FPS and inference time (ms) are computed on a platform equipped with an NVIDIA A40 GPU.

DualPose exhibits comparable computational performance to GroupPose, achieving similar frame rates while offering higher accuracy. Specifically, DualPose operates at 50.4 FPS (20 ms) at a resolution of 480×800 and 26.0 FPS (38 ms) at 800×1333 , demonstrating an efficient balance between speed and accuracy.

5. Conclusion

In this paper, we present DualPose, a framework for multi-person pose estimation based on a dual-block transformer decoder architecture. The main contribution is related to the new transformer decoder architecture proposed, that decouples the computation of keypoints and class predictions

into two distinct blocks, allowing each one to process the queries independently and more effectively. DualPose enhances accuracy and precision by implementing parallel group self-attentions and distinguishing class from keypoint predictions. Furthermore, training robustness and stability are improved through a contrastive denoising mechanism. These improvements make DualPose an excellent human pose estimation system applicable in all real cases, from industrial to surveillance contexts. Comprehensive experiments on MS COCO and CrowdPose demonstrate that DualPose outperforms state-of-the-art techniques

References

1. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 483–499.
2. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 2334–2343.
3. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the Proceedings of the European Conference on Computer Vision, 2018, pp. 466–481.
4. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7103–7112.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
6. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3349–3364.
7. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.
8. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30.
9. Kreiss, S.; Bertoni, L.; Alahi, A. Pifpaf: Composite fields for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11977–11986.
10. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5386–5395.
11. Mao, W.; Tian, Z.; Wang, X.; Shen, C. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9034–9043.
12. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6951–6960.
13. Tian, Z.; Chen, H.; Shen, C. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451* **2019**.
14. Wei, F.; Sun, X.; Li, H.; Wang, J.; Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In Proceedings of the Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 527–544.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**, *30*.
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 213–229.
17. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose recognition with cascade transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1944–1953.
18. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. TFPose: Direct Human Pose Estimation with Transformers. *arXiv preprint arXiv:2103.15320* **2021**.

19. Liu, H.; Chen, Q.; Tan, Z.; Liu, J.; Wang, J.; Su, X.; Li, X.; Yao, K.; Han, J.; Ding, E.; et al. Group Pose: A Simple Baseline for End-to-End Multi-person Pose Estimation. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023.
20. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; Shum, H. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *International Conference on Learning Representations* **2022**, pp. 1–18.
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Proceedings of the European Conference on Computer Vision. Springer, 2014, pp. 740–755.
22. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.S.; Lu, C. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10863–10872.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
24. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
25. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv preprint arXiv:1904.07850* **2019**.
26. Shi, D.; Wei, X.; Li, L.; Ren, Y.; Tan, W. End-to-end multi-person pose estimation with transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11069–11078.
27. Xiao, Y.; Su, K.; Wang, X.; Yu, D.; Jin, L.; He, M.; Yuan, Z. Querypose: Sparse multi-person pose regression via spatial-aware part-level query. In Proceedings of the Advances in Neural Information Processing Systems, 2022, pp. 1–14.
28. Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; Zhang, L. Explicit box detection unifies end-to-end multi-person pose estimation. *International Conference on Learning Representations* **2023**, pp. 1–17.
29. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations* **2021**, pp. 1–16.
30. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3651–3660.
31. Chen, Q.; Chen, X.; Zeng, G.; Wang, J. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085* **2022**.
32. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329* **2022**.
33. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13619–13627.
34. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z.; van den Hengel, A. Poseur: Direct human pose regression with transformers. In Proceedings of the Proceedings of the European Conference on Computer Vision, 2022, pp. 72–88.
35. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-up human pose estimation via disentangled keypoint regression. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14676–14686.
36. Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; Zhou, E. Rethinking the heatmap regression for bottom-up human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13264–13273.
37. Xue, N.; Wu, T.; Xia, G.S.; Zhang, L. Learning local-global contextual adaptation for multi-person pose estimation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13065–13074.
38. Shi, D.; Wei, X.; Yu, X.; Tan, W.; Ren, Y.; Pu, S. Inpose: Instance-aware networks for single-stage multi-person pose estimation. In Proceedings of the Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3079–3087.
39. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.

40. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**.
41. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2961–2969.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.