ConSel: *Con*cept-Aware *Sel*f-supervised Learning for Regression Beyond Ordinal Tasks

Supplementary Material

A. Dataset Statistics

Eye Gaze Estimation: For the eye-gaze experiments, we use four public datasets: ETH-XGaze [30], Gaze360 [27], MPIIFaceGaze (MP2) [29], and EyeDiap [8]. ETH-XGaze [30] comprises more than 1 million images, with data collected from 80 subjects who vary in age, gender, glasses-wearing, lighting conditions, and other factors. Gaze360 [27] covers a broad range of participant ages and genders (58% female, 42% male), and includes 238 participants, of which 53 are indoor (5 scene types) and 185 are outdoor (2 scene types); we use the official training set (129K images) and testing set (26K images), and only retain samples with valid face-detection annotations while strictly following the original split. Images of MPIIFaceGaze come from 15 subjects; after the pre-processing described in [4], each participant contributes roughly 3k images. Following [27, 28], we crop the original images to face patches and resize them to 224×224, use only samples with valid face-detection annotations, and keep the training/test split consistent with the original protocol. EyeDiap offers a total of 94 video clips from 16 subjects, who perform different eye movements while looking at a screen or a physical target. Following [27], we sample images os of the screen-target sessions by selecting one frame every 15 frames, resulting in a processed set of 16,674 images. In our setup, ETH-XGaze and Gaze360 are used exclusively for training, while MPIIFaceGaze and EyeDiap are held out for cross-dataset evaluation without any fine-tuning on these target domains.

Head-Pose Estimation: For head-pose estimation, we use three standard benchmarks dataset i.e., 300W-LP [33], AFLW2000 [32], and the BIWI Kinect Head Pose dataset [7]. 300W-LP is a large-pose extension of the 300W collection, it aggregates several in-the-wild face alignment datasets (AFW [2], LFPW[1], HELEN[12], IBUG [20]) into a unified set of ≈122k images with 68-point landmarks and head poses spanning roughly -90° to 90° in yaw. AFLW2000 consists of the first 2,000 images from the AFLW dataset, each annotated with 68 facial landmarks; the images exhibit large pose variation, occlusions, and diverse illumination, and are widely used as a challenging evaluation benchmark. The BIWI Kinect Head Pose dataset contains ≈15k RGB–D frames of 20 subjects (6 female, 14 male, with 4 subjects recorded twice) captured at a resolution of 640×480, with ground-truth head poses covering about $\pm 75^\circ$ yaw and $\pm 60^\circ$ pitch. In our experiments, we follow common practice [9, 24] and use 300W-LP as the primary supervised training source, while AFLW2000 and BIWI are held out for cross-dataset evaluation. For all three datasets, we convert the provided annotations to a common yaw–pitch–roll representation (in degrees) following [9, 24]. All the images were resized into 224×224.

Crowd Counting: For crowd counting, we evaluate on four benchmark datasets: UCF-QNRF [11], UCF-CC-50 [10], ShanghaiTech Part A [31], and JHU-CROWD++ [23]. UCF-QNRF contains 1,535 high-resolution images with 1,251,642 head annotations; the official split has 1,201 training and 334 test images, with per-image counts ranging from 49 to 12,865 (mean $\approx$ 815). UCF-CC-50 is a small but extremely dense dataset of 50 web images with 63,974 annotated heads; crowd counts range from 94 to 4,543 (mean $\approx$ 1,279), and we follow the standard 5-fold cross-validation protocol. ShanghaiTech Part A comprises 482 congested crowd images (300 train, 182 test) collected from the Internet, with 241,677 head annotations in total, an average count of about 501 people per image, and a maximum count of 3,139. JHU-CROWD++ is a large-scale unconstrained dataset with 4,372 images and approximately 1.51M head annotations; counts span from empty scenes to extremely dense crowds with up to 25,791 people (average $\approx$ 346), and the images cover a wide range of scenes and weather conditions (rain, snow, haze).

1D Ordinal Tasks: we use four benchmarks: MORPH II [18] and Adience [6] for age estimation, CrowdBeauty [21] for image aesthetic quality prediction, and the Historical Color Image (HCI) dataset [16] for historical image dating. MORPH II contains 55,134 mugshot-style face images from 13,618 individuals with ages from 16 to 77 years; following common practice [14, 15, 19, 22], we use only 5,492 images of Caucasian descent from 2,193 individuals to minimise cross-race interference, and perform experiments under a standard 80/20 random split with five-fold cross-validation. The Adience dataset comprises 26,580 colour images of 2,284 subjects captured "in the wild", each assigned to one of eight age groups (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60+); we follow the recent works [17, 26] report results under five-fold, cross-validation as there is no fixed train/test split. CrowdBeauty consists of 13,929 Flickr photos across four categories (nature, animal, urban, people), each annotated with an ordinal aesthetic score on a five-level scale: unacceptable", flawed", ordinary", professional", and "exceptional"; we treat these as ordered labels, use the median score per image as the ground truth, and adopt the common 80%/20% train/test split with five-fold cross-validation fol-

lowing [5, 25]. The HCI dataset contains historical colour photographs from five decades (1930s–1970s), with 265 images per decade (1,325 images in total); each image is labelled by its decade index (1–5). Following recent ordinal-regression works [5, 17] on HCI, we randomly split the 265 images of each decade into 210 training, 5 validation, and 50 test samples, and perform 5-fold cross-validation, reporting the mean performance across folds.

B. Concepts Utilization & Hyperparameter

For concept utilisation, we manually define a small set of high-level semantic concepts for each task. The design goal is to cover all plausible coarse categories for that task with as few concepts as possible, so that every training sample can be associated with one meaningful concept without requiring fine-grained manual labelling. We map each sample to exactly one concept using the rules in Tables Sup1–Sup6. There is no standard protocol for defining such concept ranges in these tasks, so we adopt simple hand-crafted thresholds that reflect intuitive coarse categories (e.g., near-frontal vs peripheral gaze, low vs high crowd density). These ranges were chosen once using dataset statistics to avoid extreme class imbalance and were kept fixed across all experiments, without any task-specific tuning.

For **eye gaze**, we use five directional concepts that jointly cover both yaw and pitch variation: "a person looking left", "a person looking right", "a person looking centre", "a person looking up", and "a person looking down". For **head pose**, we use seven concepts that capture both in-plane and out-of-plane rotations: "face directly facing camera", "face turned left", "face turned right", "head tilted upward", "head tilted downward", "head rolled left", and "head rolled right".

Table Sup1. Concept definitions for eye gaze. Pitch and yaw are in degrees.

Concept	Pitch condition	Yaw condition
Center	$-8 \leq \text{pitch} \leq 8$	$-8 \leq \text{yaw} \leq 8$
Left	$-8 \leq \text{pitch} \leq 8$	$\text{yaw} < -8$
Right	$-8 \leq \text{pitch} \leq 8$	$\text{yaw} > 8$
Up	$\text{pitch} > 8$	$-8 \leq \text{yaw} \leq 8$
Down	$\text{pitch} < -8$	$-8 \leq \text{yaw} \leq 8$

For **crowd estimation**, we encode scene-level density with five ordinal concepts: "image of a very sparse crowd", "image of a sparse crowd", "image of a moderate crowd", "image of a dense crowd", and "image of a very dense crowd". For **age estimation**, we use four broad age-related concepts: "child", "young", "adult", and "elderly". For **historical image dating**, we employ three temporal concepts that describe the overall visual style of the photograph: "vintage", "old", and "recent".

Table Sup2. Concept definitions for head pose. Yaw, pitch, and roll are in degrees.

Concept	Yaw condition	Pitch condition	Roll condition
Face facing camera	$\lvert\text{yaw}\rvert \leq 8$	$\lvert\text{pitch}\rvert \leq 8$	$\lvert\text{roll}\rvert \leq 5$
Face turned left	$\text{yaw} < -8$	$\lvert\text{pitch}\rvert \leq 25$	$\lvert\text{roll}\rvert \leq 10$
Face turned right	$\text{yaw} > 8$	$\lvert\text{pitch}\rvert \leq 25$	$\lvert\text{roll}\rvert \leq 10$
Head tilted upward	$\lvert\text{yaw}\rvert \leq 25$	$\text{pitch} > 8$	$\lvert\text{roll}\rvert \leq 10$
Head tilted downward	$\lvert\text{yaw}\rvert \leq 25$	$\text{pitch} < -8$	$\lvert\text{roll}\rvert \leq 10$
Head rolled left	any	any	$\text{roll} > 5$
Head rolled right	any	any	$\text{roll} < -5$

Table Sup3. Concept definitions for crowd density. c denotes the ground-truth person count for an image.

Concept	Count condition
Very sparse crowd	$c < 300$
Sparse crowd	$300 \leq c < 1000$
Moderate crowd	$1000 \leq c < 3500$
Dense crowd	$3500 \leq c < 10500$
Very dense crowd	$c \geq 10500$

Table Sup4. Concept definitions for age estimation. a is age in years (MORPH II) and g is the Adience age-group label.

Concept	Condition
Child	MORPH II: $a < 15$; Adience: $g \in \{0\text{–}2,\ 4\text{–}6,\ 8\text{–}13\}$
Young	MORPH II: $15 \leq a < 30$; Adience: $g \in \{15\text{–}20,\ 25\text{–}32\}$
Adult	MORPH II: $30 \leq a < 53$; Adience: $g \in \{38\text{–}43,\ 48\text{–}53\}$
Elderly	MORPH II: $a \geq 53$; Adience: $g = 60+$

Table Sup5. Concept definitions for historical image dating (HCI). d denotes the decade label (1–5).

Concept	Decade condition	
Vintage	$d \in \{1, 2\}$	(1930s–1940s)
Old	$d \in \{3, 4\}$	(1950s–1960s)
Recent	$d = 5$	(1970s)

For **image aesthetic quality**, we directly reuse the five ordinal labels provided in the CrowdBeauty dataset as aesthetic concepts, namely "unacceptable", "flawed", "ordinary", "professional", and "exceptional". These concepts are applied across the different semantic categories present

Table Sup6. Concept definitions for image aesthetic quality in CrowdBeauty. $r \in \{1, \ldots, 5\}$ denotes the original ordinal rating.

Concept	Rating condition
Unacceptable	$r = 1$
Flawed	$r = 2$
Ordinary	$r = 3$
Professional	$r = 4$
Exceptional	$r = 5$

in the dataset (for example, nature images, people images, urban scenes), and are used as high-level aesthetic descriptors in our concept-guided pre-training stage.

B.1. Hyperparameters Selection:

We empirically selected the hyperparameter values. Table Sup7 reveals synergy between variance and semantic alignment. At low variance ($\lambda_v = 1$), concepts provide minimal benefit (13.69° vs 14.45°), as collapsed embeddings lack diversity for semantic organization. As variance increases (λ_v: 5→15→25), semantic alignment becomes progressively effective: at $\lambda_v = 15$, doubling λ_s yields 13-15% gains; at $\lambda_v = 25$, progressive λ_s increases (0.25→1.0) provide 29% cumulative improvement (9.03°→6.59°). This interplay reflects complementary roles: variance creates high-dimensional diverse space (preventing collapse), concepts organize that space using pretrained DistilBERT structure (requiring only $\lambda_s = 1$ due to pretrained semantic knowledge). Small λ_s values (0.25-1.0) yield 29% gains because pretrained DistilBERT already encodes rich semantic relationships. Notably, variance regularization alone ($\lambda_v = 25, \lambda_s = 0$) achieves only 9.03°, and progressively increasing semantic alignment (λ_s: 0→0.25→0.5→0.75→1.0) yields consistent gains (9.03°→8.44°→7.39°→6.63°→6.59°), demonstrating that neither component suffices independently but together they create the structured diversity essential for regression as shown in Figure 1.

C. Additional Results

Results on UDA Settings for Gaze: Following prior work [13, 28], we additionally evaluate our method under an unsupervised domain adaptation (UDA) protocol, where a small number of unlabeled samples from the target domain are incorporated during fine-tuning (as indicated by Dt). As shown in Table Sup8, our method consistently surpasses existing approaches. In the zero-shot setting (Dt = 0), our approach already achieves noticeable improvements over both PureGaze and CLIP-Gaze, demonstrating stronger generalisation under domain shift. More significantly, when 100 unlabeled target samples are introduced during fine-tuning,

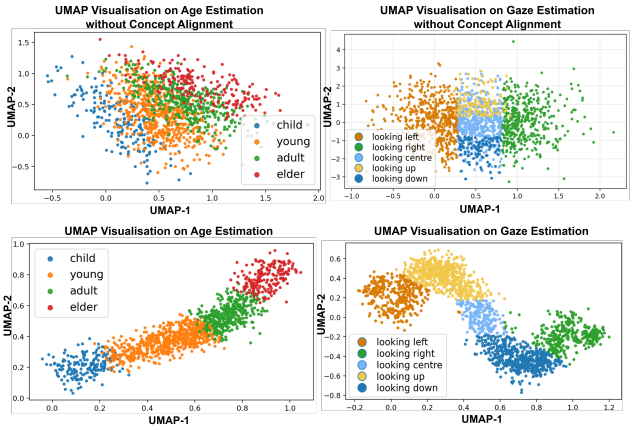

Figure 1. UMAP plot of learned representations of Gaze Estimation and age with or without concept alignment.

Table Sup7. Ablation of loss-weight hyperparameters on performance. Results are reported as average MAE for Eye Gaze (angular error) and average MAE for Age Estimation. Best results are obtained with $\lambda_v = 25, \lambda_c = 1, \lambda_s = 1$, which are therefore adopted for all remaining experiments.

λ_v	λ_c	λ_s	Eye Gaze (↓ MAE°)	MORPH II (↓ MAE)
1	1	0.5	13.69	4.57
1	5	1	14.45	4.18
1	5	0.5	14.63	4.24
5	1	0.5	10.32	3.85
5	1	1	10.29	3.32
15	1	0.5	9.81	2.98
15	1	1	8.51	2.53
25	1	0	9.03	2.92
25	1	0.25	8.44	2.42
25	1	0.5	7.39	2.18
25	1	0.75	6.63	1.91
25	**1**	**1**	**6.59**	**1.89**

our method yields a substantial performance gain, achieving an average MAE of 4.52, outperforming the (CLIP-Gaze) by a clear margin. This consistent advantage across source to target pairs highlights the robustness of our representation and its effectiveness in leveraging limited unlabeled target data, confirming the superior adaptability of our model under cross-domain scenarios.

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[2] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016.

Table Sup8. Performance comparison under unsupervised domain adaptation settings (lower is better). Dt denotes the number of samples taken from the testing set during fine-tuning.

Method	Dt	E→M	E→D	G→M	G→D	Avg
PureGaze [3]	0	7.08	7.48	9.28	9.32	8.29
CLIP-Gaze [28]	0	6.41	7.51	6.89	7.06	6.97
Ours	0	**6.19**	**6.97**	**6.41**	**6.82**	6.59
LatentGaze [13]	100	5.21	7.81	–	–	6.51
CLIP-Gaze	100	4.45	5.27	4.94	5.60	5.07
Ours	**100**	**4.01**	**4.63**	**4.49**	**4.95**	**4.52**

[3] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 436–443, 2022.

[4] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7509–7528, 2024.

[5] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach clip to develop a number sense for ordinal regression. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[6] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

[7] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101(3):437–458, 2013.

[8] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pages 255–258, 2014.

[9] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. Toward robust and unconstrained full range of rotation head pose estimation. *IEEE Transactions on Image Processing*, 33:2377–2387, 2024.

[10] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[11] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018.

[12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[13] Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *Proceedings of the asian conference on computer vision*, pages 3379–3395, 2022.

[14] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1145–1154, 2019.

[15] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *Advances in Neural Information Processing Systems*, pages 35313–35325, 2022.

[16] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, pages 499–512. Springer, 2012.

[17] Dileepa Pitawela, Gustavo Carneiro, and Hsiang-Ting Chen. Cloc: Contrastive learning for ordinal classification with multi-margin n-pair loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15538–15548, 2025.

[18] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006.

[19] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

[20] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.

[21] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of the international AAAI conference on web and social media*, pages 397–406, 2015.

[22] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2304–2313, 2018.

[23] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020.

[24] Roberto Valle, José M Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2874–2881, 2020.

[25] Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, and Jian Wu. Ord2seq: Regarding ordinal regression as label sequence prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5875, 2023.

[26] Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *Advances in Neural Information Processing Systems*, 2023.

[27] Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[28] Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: Towards general gaze estimation via visual-linguistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6729–6737, 2024.

[29] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[30] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European conference on computer vision*, pages 365–381. Springer, 2020.

[31] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016.

[32] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015.

[33] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.