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

Anonymous CVPR submission

Abstract

Regression is a fundamental problem in computer vision, underpinning tasks such as gaze estimation, head pose prediction, age assessment, aesthetic quality evaluation, crowd counting and historical image dating. We introduce ***ConSel*** *(Concept-Aware Self-Supervised Regression), a unified framework that learns to predict continuous values by progressing from coarse semantic concepts to fine-grained numeric precision. ConSel follows a two-stage curriculum:* ***(1)*** *concept-aware self-supervised pretraining, which aligns visual embeddings with conceptual guidance through variance–covariance regularization without access to ground-truth labels, and* ***(2)*** *fine-tuning for precise continuous prediction.* Unlike prior approaches that are optimized only for 1D ordinal regression, ConSel generalizes to both ordinal and multi-dimensional continuous tasks. *Evaluated on 15 benchmark datasets spanning 6 domains, ConSel surpasses both domain-specialized and ordinal methods by* ***15–35%*** *while using* ***only 25% of labeled data (4× less than prior methods)****.*

1. Introduction

In computer vision, regression aims at predicting continuous numerical values directly from images, serving as a foundation for diverse computer-vision tasks such as gaze [6, 8, 89] and head-pose [84] estimation for human–computer interaction, bone-age [52] and disease-severity [81] assessment in healthcare, aesthetic-quality prediction [29] for content scoring in digital media streams, crowd-density estimation [54, 69] in video surveillance, and historical-image dating [49] in historical and archival collections. Regression approaches in computer vision can be broadly categorized into two types. **(1) Direct regression** methods and **(2) Vision–language regression** methods.

Direct regression optimizes convolutional/transformer encoders with standard losses such as Mean-Squared Error (MSE) or smooth-L1. These methods [17, 47, 56, 58, 78, 85], neglect the intermediate semantic representations underlying continuous variation. While conceptually simple and widely used, these methods are prone to overfitting with limited data and generalize poorly to unseen conditions. Recent advances in Vision–Language-Models (VLMs) (e.g., CLIP [53], BLIP [34], ALIGN [28]), which have demonstrated remarkable performance on classification [9, 14] and segmentation tasks [39, 75], have inspired an alternative paradigm to regression problem. ***Vision–language regression*** methods [13, 37, 73, 79], particularly approaches like OrdinalCLIP [37] and L2RCLIP [73], convert each label or rank into a dedicated, learnable prompt (e.g., "age 25 years", "age 26 years"). This allows VLMs to align visual features with the semantic order of labels in the text space. These approaches address the known issues in direct regression, such as overfitting and poor generalization, especially where training data is limited. However, converting each label into a textual description is not scalable, especially in dense/continuous regression tasks. For example, in tasks like eye gaze estimation or head pose prediction, where precision could matter at a scale of 0.5° or less, one would need

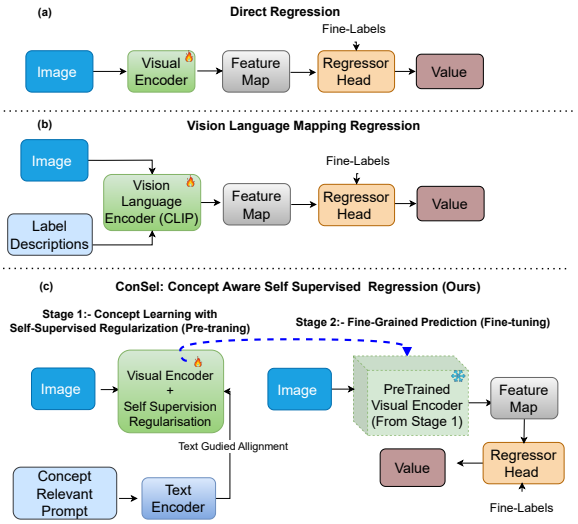

Figure 1. Comparison between (a) **Direct regression**, (b) **VLM-based regression** and (c) our proposed **(ConSel)** regression paradigm.

to generate a very large number of fine grained prompts.

Building upon this, methods such as L2RCLIP [73] and NumCLIP [13] apply the VLMs, particularly CLIP [53], to ordinal regression tasks like age estimation and historical image dating . These approaches take advantage of the natural ordering [21, 71] present in such tasks by grouping labels into broader, progressive categories, such as child, young, and adult and by using rank or order-aware losses.While, these methods excel when the target space exhibits a natural one-dimensional ranking (e.g., age groups: child $\prec$ young $\prec$ adult), the presupposition of global and natural ordinality [21, 51, 71] renders them unsuitable for regression tasks where such ranking does not exist, e.g. gaze direction, head orientation, or crowd density estimation as shown in Fig 2. Empirically, we find that in such settings, ordinal regression methods **underperform by 15–32%** compared to models designed specifically for those tasks (see results of NumClip, l2RCLIP, CLOC and Ordinal-CLIP in Table 1, 2, 5).

This performance degradation is primarily due to two reasons: 1) ordinal losses enforce discrete rank-aware margins [51, 73] that are incompatible with manifold or continuous regression. 2) Additionally, contrastive learning treats neighboring labels as negatives, for example, pushing apart representations for 24° and 25°. This directly contradicts the smoothness prior which is essential to regression [83], especially in multi-dimension and continuous tasks. Moreover, occlusion and real-world variation disrupt the progression assumed by group-aware ordinal losses, limiting their robustness to real world challenges (See Tables 6, 7).

To address the limitations stated above, we propose a novel strategy to solve regression problems by reformulating them in a concept base curriculum learning [2, 74] framework[1]. We first aim to learn high level conceptual semantics and then progressively refine the model towards fine-grained target prediction, as illustrated in Figure 1. Instead of discretizing labels into ordered bins or enumerating a prompt for each label, our approach organizes continuous targets into broader concepts, which are refined through successive curriculum stages. Specifically, our regression framework follows a two-stage learning process: (a) Concept Learning with Self-Supervised Regularization (pre-training), and (b) Fine-Grained Prediction (fine-tuning).

In **Stage 1**, we pair input images with coarse conceptual prompts (e.g., "person looking left," "person looking right," "person looking at center") that serve as soft regularization anchors during self-supervised pre-training. A customized self-supervised loss aligns these soft anchors with image embeddings while enforcing three representational constraints: (i) *consistency* to augmentations of the same sample, (ii) *variance* preservation across samples to prevent collapse, and (iii) *covariance* minimization across feature dimensions to encourage decorrelation (See Section 3.1 for details).

In **Stage 2**, the visual encoder from stage 1 is frozen, and a lightweight MLP is trained using only 25% continuous labels (e.g., age values, gaze angle, or crowd density) refining numeric precision while preserving the learned semantic structure. This version of curriculum learning mirrors human learning: understanding "left" vs "right" before predicting precise regression value such as 15.3° yaw and 17.7° pitch.

In a nutshell, our main contributions are as follows:

- We propose a conceptual curriculum learning paradigm for regression. Our two-stage approach first learns coarse semantic concepts without exposure to exact numeric labels, then fine-tunes for precise continuous values, improving both performance and generalization.
- We introduce a novel self-supervised pre-training framework for regression that learns semantic structure without imposing ordinal or rank-aware constraints. A concept-conditioned variance regularizer links semantic alignment with statistical consistency, making the approach applicable to both ordinal-style and multidimensional regression.
- We demonstrate strong performance across 6 diverse regression tasks spanning 15 benchmark datasets, surpassing previous methods and achieving consistent gains under distribution shifts and real-world challenges while using only 25% of the labeled data for fine-tuning.

2. Related work

2.1. Direct and Vision-Language Regression

Traditional regression approaches directly optimize visual encoders for continuous prediction through end-to-end learning with standard losses (MSE, MAE). For ordinal regression, methods like CORAL [3], POE [50], and SORD [12] employ ranking losses to encode ordering constraints. Task-specific architectures have been developed for age estimation [56], gaze prediction [85], and crowd counting [38]. While effective with abundant labeled data, these methods lack semantic structure and struggle with limited supervision or distribution shift. Recent CLIP-based methods for ordinal regression such as Ordinal-CLIP [37] and L2RCLIP [73] learn rank-specific prompts; NumCLIP [13] discretizes targets into 10–20 bins with linguistic descriptions; CLOC [51] employs contrastive ranking losses. While effective, they assume natural 1D ordering, making them ill-suited for multi-dimensional tasks like gaze and head pose estimation where angular relationships form 2D/3D topological structures [91] rather than linear progressions as illustrated in Figure 2.

2.2. Curriculum Learning

Curriculum learning improves efficiency by presenting training examples from easy to hard [2, 74]. Most prior

[1]Our concept-based curriculum learning is inspired by Definition 3 (Generalized Curriculum Learning) of [74].

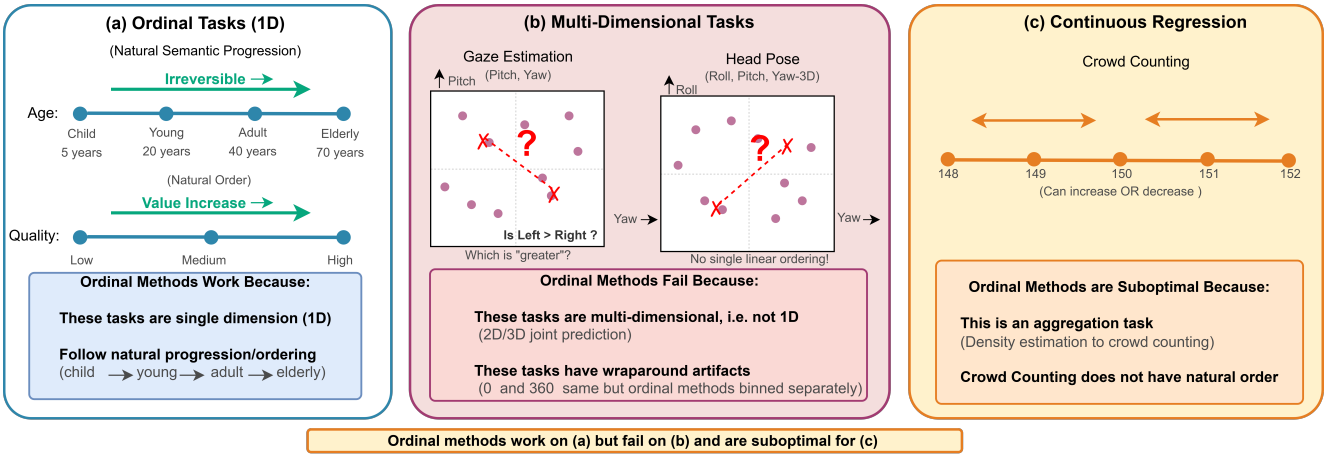

Figure 2. Taxonomy of regression tasks by dimensionality and ordering. Age and image dating are 1D with natural ordinal structure; gaze and head pose are multi-dimensional without inherent ordering. While most Ordinal methods fails on tasks (b) and (c), we show that our proposed method (ConSel) handles all these tasks very well.

work focuses on sample-level ordering [31, 77] or dynamic data selection [19], maintaining fixed loss functions while varying sample selection. We position our method within definition 3 of [74], varying the training criterion across stages, an area still under-explored for regression. Our concept-level curriculum changes the learning objective between stages: Stage-1 employs self-supervised learning with semantic language supervision for concept formation, while Stage-2 uses supervised regression for numeric precision. Unlike NumCLIP's training with discretized natural bins as target and rank aware contrastive optimisation [13], we use semantic concepts solely for self-supervised structural learning, never as classification targets, avoiding discretization artifacts while establishing concept learning representations.

3. Methodology

Our proposed paradigm teaches regression concepts through a two-stage curriculum that progresses from conceptual understanding to fine-grained prediction. *Stage-1 performs concept-aware self-supervised pretraining* by using our novel *concept-conditioned variance regularization*. These semantics act as soft anchors that organize the space into meaningful regions while preserving intra-concept diversity. *Stage-2 regression fine-tuning* specializes these representations for fine-grained continuous prediction. Figure 1(c) illustrates the overall pipeline of our framework.

3.1. Stage I: Concept Aware Self-Supervised Pretraining

Following the standard practice in non-contrastive self-supervised learning [1, 4, 20, 82], we employ a Siamese architecture with weight-shared encoders as shown in Fig 3 . Given an input image x, we generate two augmented views $x_1 = t_1(x)$ and $x_2 = t_2(x)$ using augmentations $t_1, t_2 \sim$

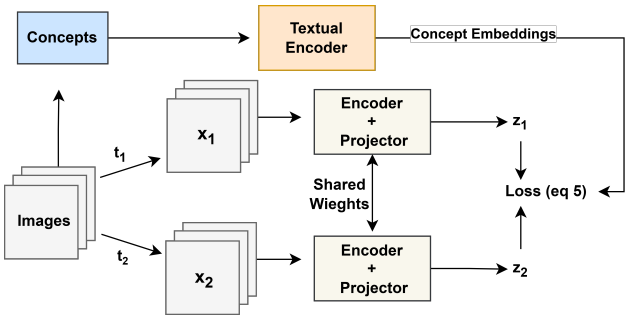

Figure 3. Overview of Stage I pretraining. Note that task relevant *Concepts* can be generated automatically or manually.

$\mathcal{T}$, where $\mathcal{T}$ includes random crops, color jittering, gaussian blur, and horizontal flips. Both views are processed through an encoder and projector $f_\theta : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ to produce embeddings $z_1 = f_\theta(x_1)$ and $z_2 = f_\theta(x_2)$, where d is the embedding dimension. We manually define a minimal, semantically complete concept set $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ tailored to each regression task. For instance, for directional regression tasks, we define the concept set $\mathcal{C} = \{Looking(\text{left}, \text{right}, \text{up}, \text{down}, \text{centre})\}$.; for age estimation, we use $\mathcal{C} = \{\text{infant}, \text{child}, \text{young}, \text{adult}\}$ (Concept usage and design details are in supplementary Section-B). These concepts are designed to capture coarse-grained semantic variations relevant to the target regression variable, acting as **soft semantic anchors** in the representation space. The concept set $\mathcal{C}$ remains fixed throughout training, providing consistent semantic structure across the dataset.

3.1.1. Loss Formulation

Our pretraining objective integrates conceptual guidance with three complementary regularization terms that prevent representational collapse:

Consistency Loss. Enforces consistent representations between augmented views:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N}\sum_{i=1}^{N} \|z_1^i - z_2^i\|_2^2 \tag{1}$$

where N is the batch size.

Variance Loss. To prevent complete collapse where all embeddings map to a constant vector, we maintain diversity across the batch by penalizing low variance for each branch (z_1, z_2) following [1]:

$$\mathcal{L}_{\text{var}} = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - \sqrt{\text{Var}(Z_j) + \epsilon}) \tag{2}$$

where Z_j denotes the j-th embedding dimension across the batch, $\gamma = 1$ is the target variance threshold, and $\epsilon = 10^{-4}$ ensures numerical stability consistent with [1].

Covariance Loss. To prevent dimensional collapse where information concentrates in few dimensions, we decorrelate embedding dimensions:

$$\mathcal{L}_{\text{cov}} = \frac{1}{d}\sum_{i\neq j} [\text{Cov}(Z)]_{ij}^2 \tag{3}$$

where $\text{Cov}(Z) \in \mathbb{R}^{d\times d}$ is the covariance matrix of batch embeddings. This term is particularly important for our curriculum based regression, as it ensures the learned representations span a rich, high-dimensional space that supports fine-grained continuous predictions.

Conceptual Alignment Loss. We align visual embeddings with semantic concept embeddings:

$$\mathcal{L}_{\text{concept}} = \frac{1}{N}\sum_{i=1}^{N} \|\bar{z}_i - s_{x_i}\|_2^2 \tag{4}$$

where $\bar{z}_i = (z_1^i + z_2^i)/2$ is the mean embedding of both augmented views. This loss acts as a soft semantic anchor, structuring the representation space according to human-interpretable concepts without requiring hard clustering or negative samples. The complete Stage I objective combines all terms:

$$\mathcal{L}_{\text{Stage I}} = \mathcal{L}_{\text{inv}} + \lambda_v\mathcal{L}_{\text{var}} + \lambda_c\mathcal{L}_{\text{cov}} + \lambda_s\mathcal{L}_{\text{concept}} \tag{5}$$

where $\lambda_v = 25$, $\lambda_c = 1.0$, and $\lambda_s = 1.0$ are hyperparameters balancing the loss components (See Supplementary Section-B for selection details).

All hyperparameters remain fixed throughout training and are not learned or adapted. The selected values transfer well across different regression tasks (see Table 1-5).

Unlike VICReg [1] which applies variance regularization uniformly, our framework achieves *concept-conditioned variance regularization* through the interplay of $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{concept}}$: maintaining high intra-concept diversity while ensuring inter-concept separability. This creates semantically structured embeddings where continuous regression attributes organize around linguistic concepts, without requiring contrastive negatives or ordinal ranking constraints.

Implicit Ordinal Structure Without Ranking Loss: Our method learns ordinal relationships implicitly through the interplay of conceptual alignment and variance-covariance regularization, without the need for explicit ranking-aware losses. First, for 1D ordinal regression (e.g., age, image aesthetic), semantic concepts from pretrained language transformer inherently encode ordering e.g., {*infant*, *child*, *young*, *adult*} which have natural sequential relationships in text embedding space. By aligning visual embeddings to these ordered concept anchors via $\mathcal{L}_{\text{concept}}$, images are pulled towards semantically ordered clusters while $\mathcal{L}_{\text{cons}}$ ensures smooth transition between concepts through augmentation invariance. Specifically, the continuous nature of the L2 alignment loss (Eq. 4) allows *representations to interpolate smoothly between concept anchors rather than forming discrete clusters*, supporting fine-grained regression within and across ordinal categories. For multi-dimensional regression (e.g., gaze with pitch-yaw coordinates), the framework's effectiveness stems from $\mathcal{L}_{\text{cov}}$ and $\mathcal{L}_{\text{concept}}$ ensuring embedding dimensions remain decorrelated [27]. Concepts like {*left*, *right*, *up*, *down*, *center*} anchor different regions of the multi-dimensional embedding space, where $\mathcal{L}_{\text{var}}$ maintains high within-concept variance to preserve fine-grained discriminability along each axis. Thus, our approach handles both ordinal 1D and multi-dimensional regression through semantic guidance and statistical regularization alone. Figure 4 depicts the learned representation spaces for age and eye gaze using UMAP. We report ablation study results comparing the role and importance of different losses with respect to conceptual loss. We also perform ablation for evaluating the effect of curriculum inspired paradigm as compared to direct and vision language regression (See Tables 8,9).

3.2. Stage II: Regression Fine-tuning

After pretraining converges, we freeze the encoder parameters θ^* and attach a lightweight regression head $h_\psi : \mathbb{R}^d \to \mathbb{R}$. The frozen encoder serves as a powerful feature extractor that has learned both semantically-structured and fine-grained visual representations during Stage I.

The regression head consists of a 3-layer MLP with hidden dimension 1024, ReLU activation, and dropout ($p = 0.2$). Given a labeled dataset $\mathcal{D}_{\text{reg}} = \{(x_i, y_i)\}_{i=1}^{M}$ where $y_i \in \mathbb{R}$ are continuous target values, we optimize:

$$\mathcal{L}_{\text{Stage II}} = \frac{1}{M}\sum_{i=1}^{M} \left(h_\psi(f_{\theta^*}(x_i)) - y_i\right)^2 \tag{6}$$

This design enables effective transfer learning even when $M \ll N$, where N is the pretraining dataset size.

4. Experiments

We have performed several experiments for evaluating of the overall performance, generalization/transferability and robustness across real world challenges like occlusion and noise. Note that our results reported in this paper were obtained using *only 25% of the labeled data during the fine-tuning stage*, while the pretraining stage was never exposed to actual labels. Despite using just a quarter of the labeled data compared to competing methods, our approach achieved stronger performance across all evaluated tasks.

4.1. Datasets

We used several benchmark datasets; specifically, ETH-XGaze [87], Gaze360 [30], MPIIGaze [86], and EYE-DIAP [18] for eye gaze estimation, while 300W-LP [93] BIWI [16] and AFLW2000 [92] were used for head pose estimation. For the latter, we trained our framework on 300W-LP and tested on BIWI and AFLW2000 following previous work [23]. **For crowd counting**, we utilized four benchmark datasets: UCF-QNRF [26], UCF-CC-50 [25], ShanghaiTech Part A [88], and JHU-CROWD++ [64]. We adopted standard train/test splits and reported Mean Absolute Error (MAE) as the primary evaluation metric following [44] . To evaluate generalization of our method, we additionally performed a cross-domain evaluation by training on UCF-QNRF and ShanghaiTech Part A and testing on the remaining datasets. For **1D Ordinal regression** tasks we used the MORPH II [55] and Adience [15] for age estimation, Crowd Beauty[60] for image aesthetic quality prediction and HCI dataset [49] for Historical Image Dating (HID). Following [35, 37, 57, 61] only 5,492 images of Caucasian descent were used from MORPH II to minimize cross-race interference. The Adience dataset contains 26,580 color images of 2,284 subjects, each assigned to one of eight age groups; we evaluated on it using 5 folds as there is no original train/test split. More details on the statistics and preprocessing of all datasets are provided in the supplementary material Section-A.

4.2. Implementation Details

For all regression tasks, we used the same two-stage training procedure with manually created task-specific concept vocabularies (e.g., "looking left/right" for gaze, "crowded/sparse" for crowd counting, "young/elderly" for age). We randomly initialized ResNet-18 [22] to avoid domain-specific biases from ImageNet pretraining, enabling task-agnostic concept learning particularly beneficial for tasks requiring spatial reasoning (crowd counting) or fine-grained analysis (age/gaze) rather than object recognition. In the first stage, we employ OREPA [24] for efficient training of ResNet-18 alongside frozen DistilBERT [59], which provides semantic concept representations. We standardized both visual and textual feature dimensions to 1024 for effective alignment and trained for 200 epochs with batch size 32 and all images were resized to 224×224. Once concept learning converges, we froze the learned ResNet-18 visual encoder and trained only a lightweight 3-layer MLP regressor head for continuous value prediction, isolating concept grounding from task-specific regression. The second stage is trained for 50 epochs and with just 25% of stratified labeled data from training set. Early stopping is used to ensure optimal generalization.

Table 1. Comparison with the state-of-the-art methods on eye gaze as well as with ordinal classification methods using domain generalization. The best results are in bold and the second-best are underlined. E:ETH-XGaze, G:Gaze360, M:MPIIGaze and D:EYE-DIAP. Results are reported as angular error (↓ lower is better). Results marked with * were obtained using authors' code.

Methods	**Train → Test**				**Avg**
	E→M	E→D	G→M	G→D	
PureGaze [7]	7.08	7.48	9.28	9.32	8.29
CLIP-Gaze[80]	<u>6.41</u>	7.51	6.89	7.06	6.97
LG-Gaze [79]	6.45	<u>7.22</u>	<u>6.83</u>	<u>6.86</u>	<u>6.84</u>
CDG [76]	6.73	7.95	7.03	7.27	7.25
CoOp [90]	7.44	7.42	7.41	7.15	7.36
CLOC [51]	9.87	11.27	9.93	10.89	10.49
MoCo-v3 [5]	9.43	10.78	9.12	10.01	9.84
DINOv2 [48]	8.02	10.14	8.17	9.36	8.92
VICReg [1]	7.93	8.02	7.45	8.54	7.99
Ordinal-CLIP* [37]	8.08	10.56	9.10	9.91	9.41
Num-CLIP* [13]	8.03	10.07	8.05	9.26	8.85
L2RCLIP* [73]	8.51	9.94	8.31	9.08	8.96
Ours	**6.19**	**6.97**	**6.41**	**6.82**	**6.59**

4.3. Comparison on Multi-Dimensional Regression

For gaze estimation, ETH-XGaze (E) and Gaze360 (G) served as training datasets, and MPIIGaze (M) and EYE-DIAP (D) were used for testing to ensure fairness with prior cross-domain evaluation protocols [7, 79, 80]. Accordingly, results were evaluated across four cross-domain settings (E→M, E→D, G→M, G→D). Table 1 reports cross-domain gaze estimation results using mean angular error in degrees (**MAE**; lower is better).

Our method achieved the lowest average error (**6.59°**) across all four train→test settings, outperforming both specialized gaze models [7, 76, 79, 80] and state-of-the-art VLM/CL based ordinal regression methods [13, 37, 51, 73, 90]. Table 1 also shows that VLM/CL-based ordinal regression networks underperform on eye-gaze estimation compared to domain-specific methods (e.g., LG-Gaze model [79]). The same holds true when compared against our method, where VLM/CL networks Ordinal-CLIP [37], Num-CLIP [13] and CLOC [51] underperform by a margin of 42.8%, 34.3%, and 59.2%, respectively. Beyond cross-dataset evaluation, we compared with PureGaze [7] and CLIPGaze [80] to assess performance under unsupervised domain adaptation (UDA), where the model is trained

Table 2. Head pose estimation on BIWI and AFLW2000. The best is in **bold** and the second-best is underlined. Results marked with * were obtained using authors' code.

Method	BIWI (MAE ↓)	AFLW2000 (MAE ↓)
Yang [78]	4.61	5.34
RankPose [10]	4.40	3.74
Valle [68]	3.66	3.83
Hempel [23]	3.78	3.61
Nguyen [46]	9.00	7.23
Musti [45]	6.72	7.82
Tourani [67]	4.47	5.28
VICReg [1]	4.21	4.30
MoCo-v3 [5]	4.62	4.27
DINOv2 [48]	4.41	4.22
CLOC[51]	4.55	4.30
Num-CLIP* [13]	6.05	7.13
L2RCLIP* [73]	7.34	7.89
Ours	**3.12**	**3.27**

on labeled source data and a small set of unlabeled target samples. Our approach outperforms all UDA methods (See Supplementary Section-C). For headpose estimation, following the cross-dataset protocol of [23], we trained on 300W-LP and evaluated on BIWI and AFLW2000 as shown in Table 2. Our method achieved 3.12° MAE on BIWI and 3.27° MAE on AFLW2000 surpassing other competitors. Valle [68] obtained 3.66° on BIWI, while Hempel [23] achieved 3.61° on AFLW2000. Self-supervised approaches including MoCo-v3 [5] and DINOv2 [48] ranged from 4.2–4.6° MAE. Vision-language methods using ordinal representations (NumCLIP [13], L2RCLIP [73]) obtained 6–7.9° MAE.

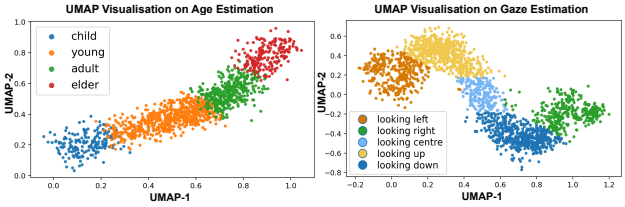

Figure 4. UMAP plot of learned representations. **Left** Age embeddings show smooth progression from young to elderly, validating implicit ordinal structure. **Right:** Gaze embeddings form a wave pattern capturing the continuous topology of directional space guided by semantic concepts. More visualizations in supplementary.

4.4. Comparison on 1D-Regression Tasks

Age-Estimation: Table 3 (col: 2, 3) show results on MORPH II and Adience dataset. Our method achieved 1.89 MAE on MORPH II (9% better than NumCLIP). On Adience, we obtained 76.0% Accuracy and 0.32 MAE, exceeding the best prior accuracy (69.4% from NumCLIP), and reducing MAE compared to the strongest prior methods (0.36 from L2RCLIP) by 0.04 (9.8% better). Notably, our method achieved these results despite being designed to handle both single-dimensional (age, aesthetics) and multi-dimensional (gaze, head pose) tasks with the same framework, whereas prior methods demonstrated their performance only on 1D ordered outputs.

Table 3. Combined results on MORPH II, Adience, and HCI. The best is in **bold**; the second-best is underlined. Blue numbers = relative (%) improvement over the second-best.

Method	MORPH II	Adience		HCI (HID)	
	MAE ↓	Acc ↑	MAE ↓	Acc ↑	MAE ↓
DRC-ORID [32]	2.26	–	–	–	–
POE [36]	2.35	60.5	0.47	54.68	0.67
PML [11]	2.31	–	–	–	–
MWR [62]	2.13	62.6	0.45	57.80	0.58
CNNPOR [42]	–	57.4	0.55	50.12	0.82
GP-DNNOR [43]	–	57.4	0.54	46.60	0.76
SORD [12]	–	59.6	0.49	53.40	0.70
GOL [33]	–	62.5	0.43	56.20	0.55
RnC [83]	–	46.8	0.86	56.64	0.72
CoOp [90]	2.39	60.6	0.50	51.90	0.76
VICReg [1]	2.26	68.3	0.46	68.53	0.38
MoCo-v3 [5]	2.29	63.2	0.51	57.53	0.72
DINOv2 [48]	2.26	65.9	0.49	62.31	0.66
OrdinalCLIP [37]	2.32	61.2	0.47	56.44	0.67
CLOC [51]	2.27	63.0	0.41	62.08	0.55
L2RCLIP [73]	2.11	66.2	0.36	67.22	0.43
NumCLIP [13]	2.08	–	–	69.61	0.35
Ours	**1.89** +9.13%	**76.0** +11.27%	**0.32** +11.11%	**78.51** +12.79%	**0.26** +25.71%

Image Quality Estimation: For image quality estimation, we evaluated on two different ordinal tasks 1)image dating (HCI dataset) and 2) aesthetics scoring (CrowdBeauty). These tasks also benefited from our semantic concept guidance, as concepts naturally encode ordering (*e.g.*, *old* vs. *recent* for dating). Our method achieved 78.51% accuracy and 0.26 MAE on HCI (See Table 3; col 4), significantly outperforming NumCLIP (69.61% accuracy, 0.35 MAE). This represents an 8.90% point accuracy gain and 25.7% MAE reduction, demonstrating that semantic concepts (*ancient*, *vintage*, *modern*, *recent*) provide ordinal structure without explicit ranking loss. On CrowdBeauty (Table 4), we achieved 79.62% overall accuracy and 0.209 MAE, improving over NumCLIP (76.53%, 0.236) by +3.09 points and 11.4% MAE reduction. Category-wise, our method excelled on Nature (83.12%), Urban (81.47%), and Animal (76.39%) categories, validating that semantic guidance transfers across diverse visual domains. Performance on People (77.52%) remains competitive, though slightly behind specialized method. Recently, Wang et al. [72] reported results on ordinal tasks (HCI, Crowd Beauty, and Adience). Despite using 100% of the training data (4 times ours), their performance on HCI and Crowd Beauty remains below ours, though they achieve a superior 0.12 MAE on the Adience age-group classification task.

4.5. Comparison on Crowd Estimation

For crowd estimation, we compared against specialized crowd density estimators (P2R [41], P2PNet [65], DM-

Table 4. Category-wise results on Image Aesthetic Quality Task. Accuracy (%) ↑ and MAE ↓.

	Accuracy (%) ↑					MAE ↓				
Method	Nature	Animal	Urban	People	Overall	Nature	Animal	Urban	People	Overall
CNNPOR [42]	71.86	69.32	69.09	69.94	70.05	0.294	0.322	0.325	0.321	0.316
SORD [12]	73.59	70.29	73.25	70.59	72.03	0.271	0.308	0.276	0.309	0.290
POE [36]	73.62	71.14	72.78	72.22	72.44	0.273	0.299	0.281	0.293	0.287
GOL [33]	73.80	72.40	74.20	69.60	72.70	0.270	0.280	0.260	0.310	0.280
CoOp [90]	72.74	71.46	72.14	69.34	71.42	0.285	0.298	0.294	0.330	0.302
OrdinalCLIP [37]	73.65	72.85	73.20	72.50	73.05	0.273	0.279	0.277	0.291	0.280
Ord2Seq [37]	78.09	75.74	72.83	69.24	74.43	0.225	0.257	0.275	0.319	0.264
L2RCLIP [73]	73.51	75.26	77.76	**78.69**	76.07	0.267	0.253	0.216	0.246	0.245
NumCLIP [13]	75.20	75.24	79.49	76.17	76.53	0.249	0.250	0.208	**0.238**	0.236
Ours	**83.12**	**76.39**	**81.47**	77.52	**79.62**	**0.187**	**0.213**	**0.199**	0.239	**0.209**

Count [70]). Table 5 shows that on four dataset, ShanghaiTech Part A (A), UCF-QNRF (B), UCF-CC-50 (C), and JHU-CROWD++ (Q), our method attained the lowest MAE (lower is better): 46.8 (A), 65.8 (B), 200.1 (C), and 51.4 (Q) respectively. While purely ordinal regression (OrdinalCLIP) methods perform poorly across splits, bin-based adaptations (Num-CLIP, CoOp) close some of the gap but still lag behind specialized methods (e.g., P2R, MPCount) and also lag behind our method. Moreover, we also performed cross-domain evaluation surpassing the prior methods (see supplementary section-C).

Table 5. Crowd counting results (MAE; ↓ lower is better) using standard splits: ShanghaiTech Part A (A), UCF-QNRF (B), UCF-CC-50 (C), and JHU-CROWD++ (Q). Best is in **bold**; second best is underlined.

Methods	A(Shang)	B(UCF)	C(50)	Q(JHU)
BL [44]	62.8	88.7	229.3	–
P2R [41]	51.02	83.2	–	58.83
P2PNet [65]	52.74	85.3	–	61.25
SASNet [66]	67.0	–	258.4	–
ChfL [63]	57.5	80.3	–	57.0
MAN [40]	56.8	77.3	–	53.4
DMCount [70]	59.7	85.6	211.0	–
Ordinal-CLIP [37]	59.9	82.8	244.4	60.5
Num-CLIP [13]	54.2	73.7	221.3	61.3
CLOC [51]	67.9	94.8	255.9	67.2
CoOp [90]	64.0	75.2	219.1	62.5
Ours	**46.8**	**65.8**	**200.1**	**51.4**

Figure 5. Grad-CAM++ visualisation on challenging samples of eye-gaze. Row 1: after Stage 1 (concept pretraining). Row 2: after Stage 2 (fine-tuning on precise labels).

4.6. Robustness to Distribution Shift

Cross-domain evaluation on eye gaze and crowd counting validates the generalizability of our method across diverse regression tasks. To further assess robustness under distribution shift, we conducted experiments on MORPHII for age estimation. Following the protocol of [37], we randomly selected several age ranges (labels) and removed a large portion of samples from those ranges, creating an imbalanced training distribution that differs significantly from the test distribution. Table 6 shows our method consistently outperformed baselines across all distribution shift scenarios. Under severe imbalance (40 classes reduced by 90%), our approach achieved 2.41 MAE compared to L2RCLIP's 2.79 and OrdinalCLIP's 3.12, representing **13.6%** and **22.8%** improvement respectively. Even under extreme shifts (10 classes, 90% reduction), we maintained 2.12 MAE versus 2.30 for L2RCLIP. Our robustness emerges from two synergistic factors: semantic anchors encode generalizable structure agnostic to training distribution, while covariance regularization maintains rich, high-dimensional representations that generalize effectively to underrepresented ranges.

Table 6. MAE↓ under distribution shift on **MORPH II**. Best is **bold**, second-best is underlined.

	number of reduced classes – %age of reduced samples/ class							
Method	**10–80**	**10–90**	**20–80**	**20–90**	**30–80**	**30–90**	**40–80**	**40–90**
CoOp [90]	2.71	2.85	2.98	3.51	3.06	3.36	2.99	3.30
OrdinalCLIP [37]	2.61	2.67	2.77	3.06	2.86	3.21	2.84	3.12
L2RCLIP	2.28	2.30	2.37	2.43	2.51	2.61	2.68	2.79
(Ours)	**2.03**	**2.12**	**2.14**	**2.19**	**2.24**	**2.28**	**2.33**	**2.41**

4.7. Robustness to Real-World Challenges

We evaluated robustness on natural variations: gaze estimation with accessories (glasses/caps) excluded from training, and age estimation with 10% facial occlusion. Table 7 demonstrates substantial advantages: 5.80° gaze error versus NumCLIP's 13.76° (57.9% improvement), and 4.45/3.38 years MAE on occluded age tasks versus 6.01/5.81 (26-42% better). Figure 5 depicts GradCAM++ attention on challenging samples with hand and glasses occlusions, showing that our two-stage curriculum enables the model to focus on discriminative regions despite these real world challenges.

Table 7. Robustness results: for gaze estimation, we evaluated on glasses/cap subset (angular MAE in degrees) after fine-tuning without those samples. For age estimation, we applied a 10% central-face occlusion on MORPH II and Adience. Results marked with * are obtained using authors' code.

Method	Gaze (deg) ↓	MORPH II (y) ↓	Adience ↓
Ours	**5.80**	**4.45**	**3.38**
CLOC [51]	16.98	6.32	5.33
NumCLIP* [13]	13.76	6.01	5.81
L2RCLIP* [73]	14.01	6.77	5.30
Ordinal-CLIP* [13]	14.54	7.33	6.11

Table 8. Ablation on loss components in the curriculum paradigm. Eye Gaze (G⇒M) is mean angular error (↓°); Age (MORPH II) and HID are MAE (↓). Blue Numbers are the relative improvement vs the row above. Results are mean±std over 3 seeds.

Configuration	Eye Gaze (↓°)	Age (↓)	HID (↓)
Consistency loss only	10.24±1.97 (—)	4.21±0.73 (—)	0.69±0.35 (—)
+ Concept loss	7.63±1.18 (+25.5%)	2.30±0.41 (+45.4%)	0.42±0.27 (+39.1%)
+ Var/Cov (no concept)	7.45±0.35 (+2.4%)	2.26±0.23 (+1.7%)	0.38±0.14 (+9.5%)
Full	**6.19±0.27** (+16.9%)	**1.89±0.15** (+16.4%)	**0.26±0.06** (+31.6%)

4.8. Ablation Studies

We conducted ablation studies to validate our design choices, analyzing both loss components (Table 8) and training paradigms (Table 9). We evaluated on three diverse tasks: multi-dimensional gaze regression, 1D ordinal age regression (MORPH II), and on HCI. We systematically added loss terms to understand their individual and synergistic effect within the curriculum learning framework.

Using only $\mathcal{L}_{\text{cons}}$ yielded the weakest performance (10.24° gaze error, 4.21 years age MAE, 0.69 HID MAE), as the model lacked both semantic structure and collapse prevention mechanisms. Without variance-covariance regularization, representations may suffer from dimensional collapse; without conceptual guidance, the learned space has no semantic organization.

Introducing $\mathcal{L}_{\text{concept}}$ produced substantial gains: **25.5%** improvement on gaze, **45.4%** on age, and **39.1%** on HID. The particularly large gain on age estimation validates our core claim that semantic concepts (*infant→adult*) provided implicit ordinal structure without explicit ranking losses. For HID image dating, temporal concepts (*ancient→modern*) similarly encode natural ordering.

Replacing conceptual alignment with $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ (Table 8 row 3) yielded comparable performance to concept-only (row 2) on age and HID, but slightly better on gaze (7.45° vs. 7.63°). This suggests that for multi-dimensional regression, decorrelating embedding dimensions is crucial for encoding independent axes (pitch vs. yaw), while for 1D ordinal tasks, semantic structure dominates.

Full Model (All Losses): Combining all four terms achieved the best performance across all tasks, with additional gains of **16.9%** (gaze), **16.4%** (age), and **31.6%** (HID) over the previous configuration. This demonstrates that conceptual alignment and variance-covariance regularization are *complementary* rather than redundant: concepts provide semantic structure while var-cov prevents collapse and maintains high-dimensional expressivity.

Table 9. Ablation across training regimes. Eye Gaze (G⇒M) is reported as mean angular error (↓°); MORPH II as MAE (↓ years). Δ_{prev} is the relative improvement vs the row above. Results are mean±std over 3 seeds.

Training	Eye Gaze (↓Err, °)		MORPHII (↓ MAE)	
	MAE Err	Δ_{prev}%	MAE	Δ_{prev}%
Direct	10.07 ± 2.58	–	5.87 ± 1.04	–
+Text (E2E)	8.04 ± 1.39	+20.16%	3.89 ± 0.95	+33.73%
+Cur (Ours)	6.19 ± 0.27	+23.01%	1.89 ± 0.15	+51.41%

Definitions: *Direct (Single Stage)* = no curriculum, no text guidance. *+Text (E2E)* = end-to-end with textual concepts and task labels. *+Cur (Ours)* = two-stage with concept guidance in Stage-1.

Training Paradigm Analysis (Table 9): Progressive enhancements demonstrate clear cumulative benefits: **Direct regression** (no text) produced scattered embeddings with no structure; **+Text guidance** improved results by 20-34%; **+Curriculum** (ours) gained additional 23-51% with the most structured embeddings; eye gaze forming clear progressive clusters and age showing smoothest ordinal progression (Fig. 4).Furthermore, the significantly lower standard deviation (±0.27° vs. ±2.58°) indicates enhanced training stability. We also conducted experiments with irrelevant concepts (semantically valid but task-unrelated) and LLM-generated concepts (automatically produced via GPT-4). Results confirmed that LLM-generated concepts also achieved good performance demonstrating automation potential (See Supplementary Section-C).

5. Conclusion

We introduced ConSel, a new framework designed to handle ordinal and non-ordinal regression through language-guided regularization. ConSel first learned conceptual representations via non-contrastive self-supervised pretraining with language alignment, allowing it to build meaningful ordinal structures from semantic language features without the need for explicit ranking losses. This approach enabled ConSel to achieve strong performance on both ordinal and non-ordinal tasks. The effectiveness of ConSel is dependent on the choice of concepts. Strong results are obtained when the concepts are semantically relevant for the task. Notably, ConSel demonstrated significant advantage in low-data regimes, outperforming its competitors while using 4x less training data. This highlights ConSel's potential for efficient learning in scenarios with limited data availability.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 3, 4, 5, 6

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009. 2

[3] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140: 325–331, 2020. 2

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 5, 6

[6] Yihua Cheng and Feng Lu. Dvgaze: Dual-view gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20632–20641, 2023. 1

[7] 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. 5

[8] Yihua Cheng, Hengfei Wang, Zhongqun Zhang, Yang Yue, Boeun Kim, Feng Lu, and Hyung Jin Chang. 3d prior is all you need: Cross-task few-shot 2d gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23891–23900, 2025. 1

[9] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 36:30662–30680, 2023. 1

[10] Donggen Dai, Wangkit Wong, and Zhuojun Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. *arXiv preprint arXiv:2005.10984*, 2020. 6

[11] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10503–10512, 2021. 6

[12] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4738–4747, 2019. 2, 6, 7

[13] 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. 1, 2, 3, 5, 6, 7, 8

[14] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024. 1

[15] 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. 5

[16] 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. 5

[17] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 1

[18] 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. 5

[19] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *International Conference on Machine Learning*, pages 1311–1320. PMLR, 2017. 3

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284, 2020. 3

[21] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015. 2

[22] 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. 5

[23] 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. 5, 6

[24] Mu Hu, Junyi Feng, Jiashen Hua, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Online convolutional re-parameterization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 568–577, 2022. 5

[25] 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. 5

[26] 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. 5

[27] Saahil Jain, Ashish Jain, and Joost van de Weijer. Barlow twins for multimodal self-supervised learning. *arXiv preprint arXiv:2112.09061*, 2021. 4

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[29] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051, 2023. 1

[30] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019. 5

[31] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010. 3

[32] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *International Conference on Learning Representations*, 2020. 6

[33] Seon-Ho Lee, Nyeong Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation. *Advances in Neural Information Processing Systems*, 35:27–39, 2022. 6, 7

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[35] 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. 5

[36] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2021. 6, 7

[37] 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. 1, 2, 5, 6, 7

[38] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 2

[39] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 1

[40] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19628–19637, 2022. 7

[41] Wei Lin, Chenyang Zhao, and Antoni B Chan. Point-to-region loss for semi-supervised point-based crowd counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29363–29373, 2025. 6, 7

[42] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 831–839, 2018. 6, 7

[43] Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5301–5309, 2019. 6

[44] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6142–6151, 2019. 5, 7

[45] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3981, 2020. 6

[46] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7588–7597, 2019. 6

[47] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016. 1

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 6

[49] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, pages 499–512. Springer, 2012. 1, 5

[50] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. 2

[51] 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. 2, 5, 6, 7, 8

[52] Monika Prokop-Piotrkowska, Kamila Marszałek-Dziuba, Elżbieta Moszczyńska, Mieczysław Szalecki, and Elżbieta

Jurkiewicz. Traditional and new methods of bone age assessment-an overview. *Journal of clinical research in pediatric endocrinology*, 13(3):251, 2021. 1

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2

[54] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Crowddiff: Multi-hypothesis crowd density estimation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12809–12819, 2024. 1

[55] 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. 5

[56] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015. 1, 2

[57] 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. 5

[58] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 1

[59] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5

[60] 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. 5

[61] 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. 5

[62] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022. 6

[63] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19618–19627, 2022. 7

[64] 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. 5

[65] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3365–3374, 2021. 6, 7

[66] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2576–2583, 2021. 7

[67] Satyajit Tourani, Siddharth Tourani, Arif Mahmood, and Muhammad Haris Khan. Unsupervised discovery of facial landmarks and head pose. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21192–21202, 2025. 6

[68] 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. 6

[69] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1130–1139, 2019. 1

[70] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33:1595–1607, 2020. 7

[71] Jinhong Wang, Jintai Chen, Jian Liu, Dongqi Tang, Danny Z Chen, and Jian Wu. A survey on ordinal regression: Applications, advances and prospects. *arXiv preprint arXiv:2503.00952*, 2025. 2

[72] Jinhong Wang, Shuo Tong, Jian Liu, Dongqi Tang, Weiqiang Wang, Wentong Li, Hongxia Xu, Danny Z Chen, Jintai Chen, and Jian Wu. Orderchain: Towards general instruct-tuning for stimulating the ordinal understanding ability of mllm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3477–3487, 2025. 6

[73] 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. 1, 2, 5, 6, 7, 8

[74] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022. 2, 3

[75] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023. 1

[76] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19376–19385, 2022. 5

[77] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR, 2018. 3

[78] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019. 1, 6

[79] 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. 1, 5

[80] 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. 5

[81] Boyang Yu, Aakash Kaku, Kangning Liu, Avinash Parnandi, Emily Fokas, Anita Venkatesan, Natasha Pandit, Rajesh Ranganath, Heidi Schambra, and Carlos Fernandez-Granda. Quantifying impairment and disease severity using ai models trained on healthy subjects. *NPJ Digital Medicine*, 7(1):180, 2024. 1

[82] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 3

[83] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023. 2, 6

[84] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8897–8906, 2023. 1

[85] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015. 1, 2

[86] 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. 5

[87] 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. 5

[88] 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. 5

[89] Wenqi Zhong, Chen Xia, Dingwen Zhang, and Junwei Han. Uncertainty modeling for gaze estimation. *IEEE Transactions on Image Processing*, 33:2851–2866, 2024. 1

[90] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5, 6, 7

[91] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 2

[92] 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. 5

[93] 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. 5