ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1065

Performance Evaluation of LVFace-B with Fusion Embedding Optimization on Pose-Variant Face Recognition

Ramneet Singh Chadha[1]; Jugesh[2]; Jasmehar Singh[3]

[1,2] C-DAC, Noida, Uttar Pradesh, India
[3]Shiv Nadar University, Noida, Uttar Pradesh, India

Publication Date: 2026/03/24

Abstract: Since face recognition is so common and non-intrusive, it is used extensively in contemporary biometric systems. Recent developments in deep learning have significantly improved recognition performance on common benchmarks, particularly with Vision Transformers (ViTs). However, real-world face recognition needs to be computationally efficient and able to handle significant pose variations (such as frontal vs. profile views). A difficult pose-variant dataset (Celebrities in Frontal-Profile in the Wild) is used in this study to assess the state-of-the-art LVFace-B model, a ViT-based model with Progressive Cluster Optimization. End-to-end performance is evaluated using MediaPipe's BlazeFace, a lightning-fast face detector that operates at over 275 frames per second on a mobile CPU. Furthermore, a Fusion Embedding strategy is presented, wherein multiple embeddings from the same identity are averaged to generate a singular representative vector. Three identification scenarios are analyzed: a single embedding for each identity, multiple embeddings for each identity, and a fused mean embedding for each identity. Extensive experiments demonstrate that fusion embedding attains the highest accuracy (Rank-1 = 96.98%) while significantly decreasing computational demands. The results show that averaging embeddings makes them more robust when the pose changes and is a useful compromise for large-scale 1:N search. The suggested method is ready to be used in real time because it strikes a good balance between speed and accuracy.

Keywords: *Computer Vision, Vision Transformer, Face Recognition, LvFace, Fusion Embedding.*

How to Cite: Ramneet Singh Chadha; Jugesh; Jasmehar Singh (2026) Performance Evaluation of LVFace-B with Fusion Embedding Optimization on Pose-Variant Face Recognition. *International Journal of Innovative Science and Research Technology*, 11(3), 2006-2010. https://doi.org/10.38124/ijisrt/26mar1065

I. INTRODUCTION

Face recognition (FR) is one of the most popular and affordable ways to identify people biometrically. It has been added to security systems, mobile devices, surveillance cameras, and other uses because it's easy to take pictures of faces without touching them. In the past, FR systems used engineered features to learn face embeddings. Now, almost all FR systems use deep learning to do this. These embeddings are usually 512-dimensional vectors that group around centroids that are unique to each person. This makes it easy to compare them using cosine similarity[6]. To make it easier to tell the difference between things, well-known loss functions like ArcFace and CosFace use angular margins[15]. Convolutional neural networks (CNNs) have been successful in face recognition (FR), but Transformers have recently changed vision tasks (like classification and detection) by using self-attention. But FR still doesn't know much about Transformers. This gap exists in part because FR datasets are smaller than datasets like ImageNet, and in part because it takes a lot of work to choose the right loss function. The recent LVFace series tackles this issue through a progressive, multi-stage optimization (PCO) of the ViT model.This research specifically examines LVFace-B, which, trained on Glint360K data, attains an accuracy of 97.70% on the IJB-C benchmark (TAR@FAR=1e-4) [1], surpassing previous CNN-based systems. This shows that Transformers can work well on a large scale when trained with the right methods. This study builds on this progress by testing LVFace-B's performance on faces that have very different poses. In real-world FR systems, the pipeline includes finding faces, extracting embeddings, and matching them. Detectors with high accuracy, like RetinaFace, make well-aligned crops, but they take a lot of processing power. On the other hand, MediaPipe's BlazeFace detector is a small CNN (224 KB model) that is made for speed[3]. BlazeFace (short-range) can detect frontal faces reliably at about 275 frames per second on a single-core mobile CPU (and more than 200 FPS on a GPU). This study substitutes the conventional heavy face detector (RetinaFace) with BlazeFace to assess recognition efficacy in a CPU-efficient configuration[4].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1065

Another part of FR efficiency is keeping track of more than one picture of each person in the gallery (enrollment database). Keeping track of and comparing all of the individual embeddings can get expensive. One way to lessen this is to combine embeddings. Previous research has demonstrated that averaging several embeddings of the same identity produces a singular robust representation, enhancing identification speed without sacrificing accuracy[5]. In this research, fusion embedding is executed by calculating the element-wise mean of five embeddings for each identity (three frontal and two profile images) to create a singular final vector. This method is thought to keep the benefits of having multiple enrollment images while shortening the time it takes to search.

This paper's main contributions include: a thorough assessment of a Transformer-based FR model under pose variation; the introduction of a fusion embedding strategy to increase recognition accuracy and efficiency; a thorough pose-wise analysis with recommendations for future work; and the demonstration of an effective detection pipeline using BlazeFace.

II. LITERATURE REVIEW

The challenge of recognizing individuals from photos has long been the focus of face recognition research. Modern advancements have been fueled by deep learning, where networks are trained to map a face to a point in feature space so that the faces of the same person lie close to one another. Margin-based loss functions have been especially impactful. For instance, CosFace added an additive cosine margin and ArcFace added an additive angular margin penalty to the softmax loss [13].Variations of these ideas (SphereFace, ElasticFace, CurricularFace, etc.) all aim at tighter intra-class clustering and wider inter-class gaps. According to recent survey work, these approaches typically fall into two paradigms: (1) metric learning (triplet loss, center loss) to directly optimize distances and (2) softmax-based classification with enhanced penalties.

Interest in the application of Vision Transformers (ViTs) to FR has increased since their inception[14]. Although ViTs perform exceptionally well on general vision tasks, it is difficult to apply them directly to FR. Learning must be embedded in an open-set context for face recognition. Transformers was modified in some attempts by utilizing CNN-style augmentations or losses. For example, TransFace included patch reordering and scaling to assist Transformers in learning pose-invariant characteristics [15]. Fundamentally, the LVFace technique established a multi-stage training scheme (PCO) with three phases to stabilize ViT training for FR. This multi-stage loss scheduling, paired with a "Cosine Stage Scheduler," produced cutting-edge results: LVFace implementations set new records on IJB-B and IJB-C[1]. This suggests that transformer-based FR can outperform classic CNN algorithms when taught at scale.

On the practical side, accurate face detection is an important front-end. Highly accurate detectors (e.g., RetinaFace, MTCNN) are commonly utilized, although they need significant computation. In recent years, mobile-friendly detectors have appeared. BlazeFace is a popular example of a compact (224KB) CNN built for mobile facial identification. Bazarevsky et al. showed that BlazeFace can detect faces at over 200 frames per second on a variety of hardware. The MediaPipe library has a short-range BlazeFace variation, optimized for distances up to ~2 meters[4]. Using BlazeFace in a FR pipeline can significantly accelerate up preprocessing. Its detecting accuracy is slightly lower than heavier models, however this is a reasonable trade-off in many cases[3].

A recent study explicitly stated that averaging identity embeddings is still beneficial in many FR systems. This shows that embedding fusion can improve pose-invariant recognition without the need to store vast collections. This optimizes accuracy while increasing matching time linearly with the amount of photos. Previous research has demonstrated that merely averaging (taking the mean of) many embeddings yields a strong identity representation. Hossain et al.'s 2021 ICAECT study found that employing mean embeddings resulted in a 6.5× faster and 1.6% more accurate recognition system compared to using all embeddings as the baseline[5]. The use of multi-view information is critical in face verification techniques that incorporate cross-pose matching (such as CFP-FP). A merged template with frontal and profile features can close the pose gap[2]. This concept used for each enrolled identity, a mean embedding is calculated from numerous frontal and profile photos.

In summary, the technical foundations for this study are well-supported, including large-scale ViT training with multi-stage losses, efficient detection by BlazeFace, and identity fusion using mean embeddings. The innovative contribution is to combine these parts in a single FR pipeline and evaluate its performance under high pose fluctuation.

III. METHODOLOGY

This section describes the components of the proposed face recognition pipeline: face detection, embedding extraction, and fusion embedding.

A. Face Detection

Face detection is carried out utilizing MediaPipe BlazeFace (short-range). BlazeFace is a convolutional neural network that takes an input image and produces face bounding boxes with keypoints (left eye, right eye, nose tip, mouth, left eye tragion, right eye tragion). BlazeFace is highly efficient, processing around 275 frames per second on a single mobile CPU core. The short-range model is designed to detect faces up to around 2 meters away from the camera.

B. Cropping and Processing

If BlazeFace recognizes a face, it crops the image to the bounding box. Then cropped image is resized to 112×112 pixels, which is the expected input size for the LVFace-B model[1]. If BlazeFace fails to detect a face (as happened with a few extreme profile photos), those images are excluded from further processing[3][4].

C. Embedding Extraction (LVFace-B)

Every cropped face is fed into the LVFace-B model, which returns a 512-dimensional embedding. LVFace-B is a

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1065

Vision Transformer (ViT-B/12) trained on the Glint360K dataset with Progressive Cluster Optimization [1]. The model generates a feature vector, which is subsequently L2-normalized to fit on the unit hypersphere. Recognition is accomplished by comparing these embeddings using cosine similarity[6]. To create the gallery (enrollment dictionary), each identity's enrollment photos are processed to obtain embeddings.

D. Fusion Embedding

Fusion embedding is an approach for reducing gallery size and computing expense. Instead of storing each individual embedding for an identity, the element-wise mean is calculated to form a single fused vector. Each identity has five enrollment images: three frontal and two profile. The embeddings from these photos are averaged over each feature dimension, and the resulting 512-dimensional mean vector is normalized. This fused vector is then utilized as a template for that identity. During recognition, every probe's embedding is compared only to the fused templates. This method significantly decreases the number of comparisons (when compared to saving all five embeddings per individual) while maintaining the value of multi-view information. Prior work suggests that this mean embedding retains nearly the same discriminative power as using all embeddings[5].

In summary, for a given probe image, the recognition pipeline is as follows: face detection with BlazeFace; crop and preprocess the face; extract the embedding with LVFace-B and compute cosine similarities between the probe embedding and every identity template (fused or individual or mutliple) in the gallery; and assign the identity of the highest-scoring template. This pipeline is developed in an efficient, single-threaded CPU environment to emulate real-world deployments.

IV. EXPERIMENTAL SETUP

A. Dataset

The evaluation makes use of the CFPW (Celebrities in Frontal-Profile in the Wild) dataset. CFPW contains photos of 500 different identities, each with 10 frontal and 4 profile images. It was created primarily for testing frontal-to-profile face recognition. There are a total of 5,000 frontal and 2,000 profile photos[2]. Five photos per identity are chosen for enrollment (3 frontal and 2 profile). For consistency, the remaining photos (4th frontal and 3rd profile per identity) are used as probe images.

B. Preprocessing

BlazeFace is used to recognize faces in each image. The image is downsized to 112×112 and converted from H×W×C to CHW tensor format. Pixel values are scaled to the [-1,1] range as specified by LVFace-B. Finally, a singleton batch dimension was added to meet the input needs of the inference engine.

C. Enrollment Scenario

Three enrollment tactics were tested.

- Scenario A (Single-Image): Only the initial frontal image of each identification is utilized. The gallery has 500 embeddings (one per identity).
- Scenario B (Multi-Image): Each identity is represented by five photographs (3 frontal and two profiles). BlazeFace failed to recognize profile faces in 17 identities, resulting in 2483 embeddings (483*5+17*4) instead of the expected 2,500 (5×500).
- Scenario C (Fused-Image): The same five photos are used for each identification, but the mean embedding is stored. The gallery comprises 500 embeddings (one per identity, with an average of ~5 photos). In all cases, embeddings are matched using their cosine similarity.

D. Probs and Protocol

The fourth frontal and third profile photographs per identity (not used in enrolling) are considered probe images. This yields a total of 995 (500 frontal and 495 profile probes). BlazeFace failed to detect faces in five of the profile probe photos, leaving 495 profile probes. Each probe embedding is matched (1:N) with the gallery embeddings based on the scenario. The probe is considered correctly identified if the top match has the same identity number.

E. Metrics

The major metric is Rank-1 identification accuracy (the percentage of probes with the correct top match). Verification-style metrics are also presented by treating all pairwise comparisons: the ROC curve and its AUC (area under curve) quantify the true accept rate (TAR) at different false accept rates (FAR)[7]. The Equal Error Rate (EER) is recorded (the point at which FAR = FRR). Precision and recall at the EER level are provided, whereas TAR is reported at severe FAR criteria (10^-4, 10^-3, 10^-2) to assess performance in high-security environments. Unlike ROC and EER, which evaluate verification, Cumulative Match Characteristic(CMC) curve evaluate identification performance. It measures how well system ranks the correct identity among all enrolled identities [8][9][10]. In addition, independent accuracies for frontal and profile probes are generated to investigate pose-specific behavior.[11]

F. Implementation Detail

All experiments were performed on an Intel Core i7 CPU (no GPU). BlazeFace inference and LVFace-B embedding extraction were executed on the CPU with optimized libraries. The overall time required to analyze all probes was measured, although the stated matching durations do not include detection and feature extraction (which are common overhead). Scenario A took ~2.8 seconds to match all probes, Scenario B took ~14.02 seconds, and Scenario C took ~2.7 seconds on the same system. This demonstrates that the fused gallery (Scenario C) has essentially no additional cost over the single-image gallery, although Scenario B's gallery is five times larger.

V. RESULT

A. Overall Recognition

Table 1 highlights the important findings from each scenario. Rank-1 accuracy is broken down both by probe pose (frontal vs. profile) and overall. Two observations stick out. First, employing several enrollment photographs enhances accuracy: Scenario B (multi-image) had a greater overall

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1065

accuracy (96.8%) than Scenario A (single-image 95.9%). This is not surprising given that putting profile photos in the gallery aids in the recognition of profile probes. Second, the fusion method (Scenario C) has the highest overall accuracy (96.9%) and the best performance with profile probes.

B. Detailed Metrics

At a rigorous operating point of FAR=1e-4, Scenario C has a true accept rate (TAR) of 96.78%, compared to Scenario A's 93.16%. Scenario C likewise has the lowest EER (0.016), compared to Scenario B (0.021) and Scenario A (0.033). Precision and recall at EER also favor Scenario C. In Figure 1 shows CMC curve which lead by Scenario C on almost all rank from [1-10] while A under-performing. These benefits, are achieved without increasing the gallery size. They agree with the observation that average embeddings tend to preserve identity information.

C. Pose wise-analysis

Because frontal-frontal matches are simple, all approaches perform similarly (99% or higher rank-1). The true value of fusion emerges in profile recognition: Scenario A achieves 92.6% rank-1 on profile probes, compared to 94.3% for Scenario C. By incorporating even a modest number of profile photographs per identity (Scenarios B and C), the system can generate templates that better capture the entire view range. In essence, the fusion embedding corrects for viewpoint mismatches.

Table 1 Comparison Table of Identification and Verification Accuracy on Benchmark of CFPW Dataset on all Test Scenario (TS).

TS	Rank-1 (%)	Rank-1 @ Frontal (%)	Rank-1 @ Profile (%)	ROC_AUC [0-1]	AP (%)	EER (lower the better) [0-1]	Precision @EER [0-1]	Recall@EER (%)	Accuracy@EER (%)	TAR@FAR 1e-4 (%)	TAR@FAR 1e-3 (%)	TAR@FAR 1e-2 (%)	Average Time (s)
A	95.98	99.00	92.92	0.9891	95.01	0.0332	0.0552	96.68	96.68	93.16	94.67	96.18	2.81
B	96.88	**99.60**	94.14	0.9920	96.85	0.0219	0.0829	97.79	97.83	95.57	96.78	97.78	14.02
C	**96.98**	**99.60**	**94.34**	**0.9924**	**97.48**	**0.0162**	**0.1074**	**98.39**	**98.36**	**96.78**	**97.38**	**98.29**	**2.73**

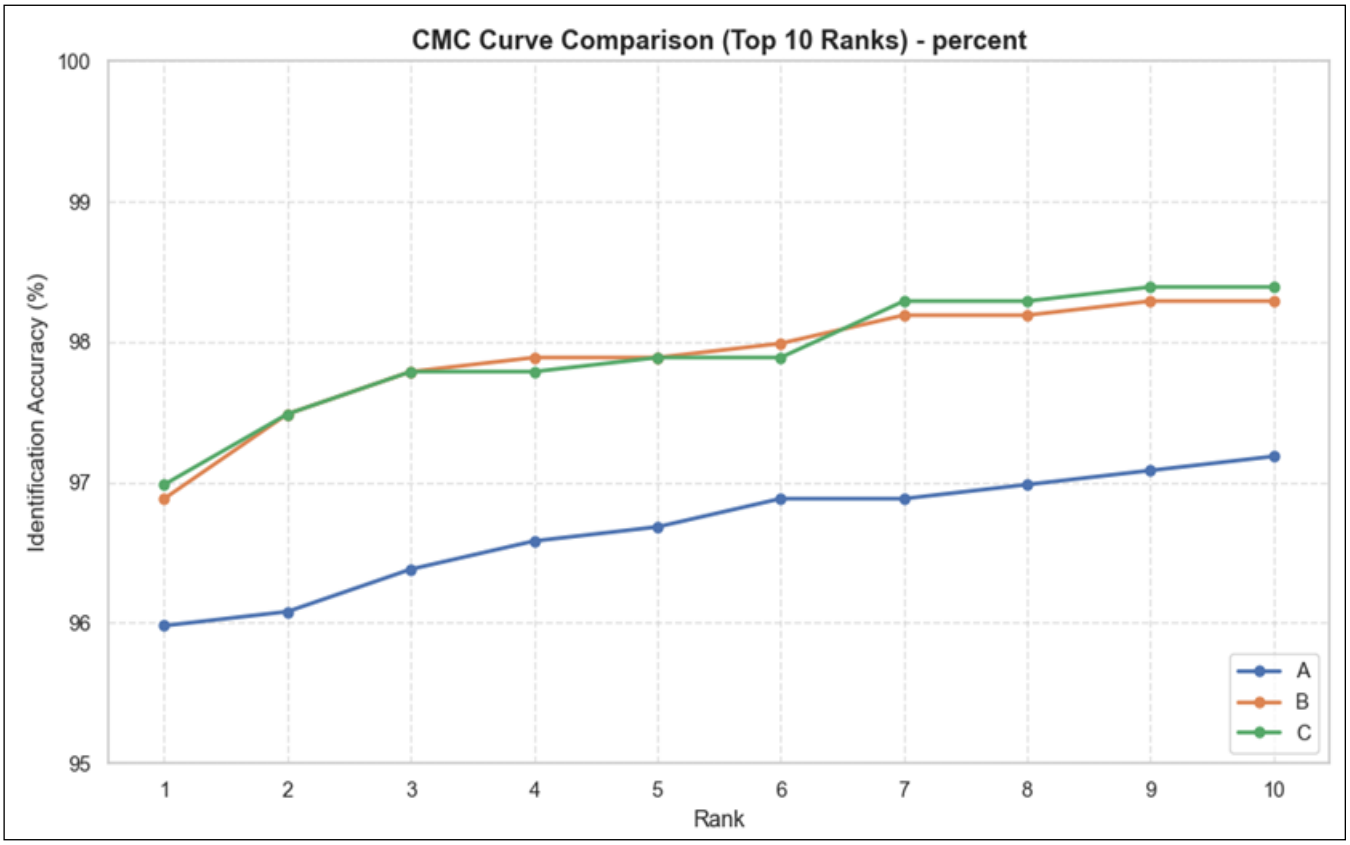

Fig 1 CMC Curve Comparison Between Scenario A, B and C.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar1065

D. Limitations

Even under the best-case scenario, a few probes were incorrectly classified. Manual examination indicates that excessive head tilts and occlusions (for example, a hand partially covering the face) resulted in inaccuracies. Furthermore, the five missing detections (blazeFace misses) show that very extreme profiles may avoid lightweight detectors. The CFPW dataset contains minimal variability in illumination and expression, therefore real-world situations may be more difficult. Finally, this evaluation assumes a fixed gallery. In dynamic systems where the gallery updates often, the fusion approach must support incremental updates.

In practice, if an application demands powerful profile face recognition, Scenario C (fusion) can significantly increase performance. If computational cost isn't an issue, Scenario B provides comparable accuracy at a greater cost. Fusion strikes the ideal balance for a variety of applications.

VI. CONCLUSION

This research offered a complete evaluation of a cutting-edge Transformer-based face recognition system with pose variation. The LVFace-B FR model, combined with a high-speed BlazeFace detector and a fusion embedding technique, produced great results on the CFPW dataset. The fusion strategy, in particular, enhanced profile-face recognition while retaining high accuracy and significantly lowering computation. These findings indicate that basic techniques such as mean embedding can successfully exploit numerous perspectives, resulting in powerful face recognition in a practical, CPU-efficient pipeline. Future advancements along these lines are planned to improve robustness to position and other variations. Overall, the study shows that careful enrollment template design and efficient detection can result in near-top-of-the-line performance in real-world face recognition tasks.

REFERENCES

[1]. J. You et al., “LVFACE: Progressive cluster optimization for large vision models in face recognition,” arXiv (Cornell University), Jan. 2025, doi: 10.48550/arxiv.2501.13420.

[2]. S. Sengupta, J. -C. Chen, C. Castillo, V. M. Patel, R. Chellappa and D. W. Jacobs, "Frontal to profile face verification in the wild," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477558.

[3]. V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, “BlazeFace: Sub-millisecond Neural face Detection on mobile GPUs,” arXiv (Cornell University), Jul. 2019, doi: 10.48550/arxiv.1907.05047.

[4]. C. Lugaresi et al., “MediaPipe: A framework for building perception Pipelines,” arXiv (Cornell University), Jun. \2019, doi: 10.48550/arxiv.1906.08172.

[5]. Md. I. Hossain, Sama-E-Shan, and H. Kabir, “An efficient way to recognize faces using mean embeddings,” 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), vol. 10, pp. 1–10, Feb. 2021, doi: 10.1109/icaect49130.2021.9392401.

[6]. Wikipedia contributors. (2025, September 17). Cosine similarity. Wikipedia. https://en.wikipedia.org/wiki/Cosine_similarity

[7]. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Dec. 2005, doi: 10.1016/j.patrec.2005.10.010.

[8]. A. K. Jain, A. A. Ross, and K. Nandakumar, Introduction to Biometrics. 2011. doi: 10.1007/978-0-387-77326-1.

[9]. B. DeCann and A. Ross, “Relating ROC and CMC curves via the biometric menagerie,” IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, 2013, pp. 1–8, Sep. 2013, doi: 10.1109/btas.2013.6712705.

[10]. N. Damer, A. Opel, and A. Nouak, “CMC curve properties and biometric source weighting in multi-biometric score-level fusion,” 17th International Conference on Information Fusion (FUSION), Salamanca, Spain, 2014, pp. 1–6, Jul. 2014, [Online]. Available: https://publica.fraunhofer.de/handle/publica/387491

[11]. “Multiclass Receiver Operating Characteristic (ROC),” Scikit-learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

[12]. A. Nemavhola, C. Chibaya, and S. Viriri, “A systematic review of CNN architectures, databases, performance metrics, and applications in face recognition,” Information, vol. 16, no. 2, p. 107, Feb. 2025, doi: 10.3390/info16020107.

[13]. Deng, Jiankang et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2018): 5962-5979.

[14]. M. Kim, A. Jain, and X. Liu, “50 years of automated face recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, pp. 1–20, Jan. 2026, doi: 10.1109/tpami.2026.3664269.

[15]. J. Dan et al., "TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 20585-20596, doi: 10.1109/ICCV51070.2023.01887.