

CARE: A Coverage-Aware Retrieval Ensemble for Isolated Sign Language Recognition

Nguyen Tu Nam and Hiroki Takahashi

The University of Electro-Communications, Tokyo, Japan
g2340024@gl.cc.uec.ac.jp, rocky@inf.uec.ac.jp

Abstract

Isolated sign language recognition is commonly formulated as supervised classification, but this formulation is challenging when nearly every target gloss has only one labeled video. We study the Isolated Sign Language Recognition track of the WSLP 2026 Shared Task on Indian Sign Language Processing, a predominantly one-shot target setting with 4,398 training videos spanning 4,361 glosses, over 99% of which are singletons. We introduce CARE, a Coverage-Aware Retrieval Ensemble that augments a complete local visual dictionary with a CISLR reference bank covering a subset of the target vocabulary. A frozen Video-Swin encoder ranks all target labels locally, while frozen I3D and SignRep encoders provide external retrieval evidence for lexically matched labels. CARE estimates orthogonal Procrustes mappings from shared-class centroids, retains both raw and aligned representation geometries, and preserves external videos as individual nearest-neighbor references. Query-wise score standardization and a deterministic coverage mask restrict external evidence to supported labels. The method requires neither target-specific backbone fine-tuning nor a large-vocabulary classification head. CARE obtains 74.31%/76.15%/76.15% Top-1/5/10 accuracy on the WSLP 2026 validation split and 65.59%/67.11%/68.25% on the official test split. The 8.72% Top-1 gap warrants caution when interpreting performance on the small 109-query validation split used for model selection.

1 Introduction

Isolated sign language recognition (ISLR) maps a sign video to its corresponding lexical label, or gloss. ISLR can support dictionary lookup and corpus annotation, but achieving robust recognition under unconstrained conditions remains a challenge. In this paper, we address the Isolated Sign Language Recognition track of the WSLP 2026

Shared Task on Indian Sign Language Processing.¹ We refer to the task’s word-recognition dataset as WR.

WR presents a predominantly one-shot regime in which nearly all target glosses occur exactly once in the training set. A standard classifier would therefore have almost no within-class evidence for estimating most decision boundaries. An alternative formulation interprets the labeled target set as a visual dictionary: each training video acts as a reference entry, reducing recognition to a query-to-entry ranking problem. This retrieval formulation also permits vocabulary expansion without retraining a classification head, a property explored by prior work on one-shot and dictionary-based sign recognition (Joshi et al., 2022; De Coster and Dambre, 2023; Desai et al., 2023; Vandendriessche et al., 2025). CARE builds on this formulation by augmenting the complete target dictionary with a cross-dataset reference bank that supports only a known subset of labels. Its coverage-aware fusion prevents labels without external references from receiving external scores.

Public labeled corpora can enlarge the reference dictionary, but their use introduces three coupled problems. First, dataset-specific differences in signers, temporal boundaries, and recording conditions can shift feature distributions even when the same encoder is used. Second, the external lexicon only partially overlaps the target vocabulary. Normalized exact matching with the public CISLR corpus (Joshi et al., 2022) supports only part of the target vocabulary; all other labels must still compete in the same global ranking. Third, distances produced by different representations and geometries can have different distributions. Assigning large external distances to unsupported labels would conflate absence of external support with negative vi-

¹<https://exploration-lab.github.io/WSLP-2026/task/>

sual evidence.

We address this setting with CARE. A frozen Video-Swin (Liu et al., 2022) encoder supplies a complete local dictionary score, while frozen WLASL-adapted I3D (Carreira and Zisserman, 2017; Li et al., 2020) and SignRep (Wong et al., 2025) encoders provide external CISLR retrieval scores for the lexically covered labels. We estimate orthogonal Procrustes mappings from shared-gloss class centroids and retain both raw and aligned representations. The resulting scores are standardized per query and fused through a deterministic coverage mask, so unsupported labels retain only the local score. CARE thus combines local retrieval, cross-dataset alignment, and coverage-aware fusion without target-specific backbone fine-tuning.

2 Related Work

2.1 Large-Vocabulary and Low-Resource ISLR

Large-scale datasets have shifted ISLR from small controlled vocabularies toward recognition under substantial variation in signer and recording conditions. MS-ASL introduced a signer-independent benchmark with 1,000 American Sign Language labels (Vaezi Joze and Koller, 2019), while WLASL provides videos covering 2,000 word classes and compares appearance- and pose-based models (Li et al., 2020). These resources established strong supervised baselines and encouraged the use of spatio-temporal networks such as I3D (Carreira and Zisserman, 2017) and hierarchical video transformers such as Video Swin (Liu et al., 2022).

These benchmarks mostly support fixed classifiers trained with repeated examples per class. One approach to low-resource recognition transfers pre-trained representations across datasets or languages. OpenHands standardizes pose-based representations across six sign languages (Selvaraj et al., 2022). Most relevantly, CISLR proposes prototype-based one-shot recognition for Indian Sign Language using features transferred from resource-rich American Sign Language data (Joshi et al., 2022). SignBERT adds a prediction head and fine-tunes its pretrained encoder for supervised isolated-sign classification (Hu et al., 2021). This setup differs from our WR setup, where nearly all target classes have only one labeled example and we formulate recognition as retrieval. SignRep, by contrast, demonstrates that a frozen encoder can support sign-dictionary retrieval (Wong et al., 2025).

We therefore use it in CARE’s external retrieval branch.

2.2 Dictionary Retrieval

Several works formulate sign recognition as query-by-example search. De Coster and Dambre retrieve dictionary entries with dense vector search (De Coster and Dambre, 2023), while Boháček and Hruží study few-shot recognition using videos collected from online dictionaries (Boháček and Hruží, 2023). ASL Citizen frames large-vocabulary ISLR as dictionary retrieval (Desai et al., 2023), and Vandendriessche et al. (Vandendriessche et al., 2025) learn representations for one-shot retrieval across changing dictionaries and sign languages. Metric-based few-shot learning provides a related comparison: Matching Networks predict from attention-weighted query-support similarities (Vinyals et al., 2016), whereas Prototypical Networks compare each query with the mean embedding of a class’s support examples (Snell et al., 2017). CARE is closer to instance-level matching: although we use shared-class centroids to estimate the alignment maps, retrieval retains every external video rather than replacing each class with a single prototype. It additionally combines a complete local dictionary with an external bank that covers only a known subset of target labels.

2.3 Cross-Dataset Alignment and Partial Coverage

Cross-dataset retrieval must account for both representation shift and incomplete label overlap. CORAL aligns second-order feature statistics (Sun et al., 2016), and Procrustes analysis has been used for manifold alignment (Wang and Mahadevan, 2008). Universal domain adaptation considers shared and private source and target classes (You et al., 2019), but must typically infer their relationship. In CARE, normalized lexical matching directly identifies the externally supported target labels. We estimate an orthogonal map from shared-gloss centroids, retain the raw geometry as a complementary channel, and mask external scores for unsupported labels.

3 Methodology

3.1 Problem Formulation and Overview

Let $\mathcal{G}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ be the labeled local dictionary with target vocabulary $\mathcal{Y}$, where x_i is a reference video, $y_i \in \mathcal{Y}$ is its gloss, and N_L is the

number of local references. Given a query video q , the task is to rank every $y \in \mathcal{Y}$. Because almost every target label is a singleton, CARE retains the training videos as reference instances instead of learning a separate classifier output for every target label.

CARE uses three retrieval branches that produce five score channels: one local Video-Swin channel, raw and aligned I3D channels, and raw and aligned SignRep channels. The external channels compare the query with a CISLR reference bank, whereas the local channel scores every target label. Since the external lexicon covers only part of $\mathcal{Y}$, these scores are fused through an explicit label-coverage mask. We precompute the representations and fit closed-form orthogonal alignment maps without updating the encoders. Fusion weights are selected using validation scores. Figure 1 summarizes the retrieval and fusion pipeline.

3.2 Representations and External Coverage

For Video-Swin (Liu et al., 2022) and WLASL-adapted I3D (Carreira and Zisserman, 2017; Li et al., 2020), each WR video is represented by a uniform temporal view and a motion-quantile view. The latter samples according to smoothed frame-difference energy, emphasizing frames in higher-motion regions. For SignRep (Wong et al., 2025), we process overlapping windows and pool their embeddings using predicted hand activity to downweight windows with low predicted activity. All resulting embeddings are ℓ_2 -normalized and all encoders remain frozen. For a WR video z , let $h_b^v(z)$ denote its normalized embedding from encoder $b \in \{S, I, R\}$ under view v , where S , I , and R denote Video-Swin, I3D, and SignRep, respectively. For Video-Swin and I3D, $v \in \{u, m\}$ denotes the uniform or motion-quantile view; for SignRep, $v = \text{act}$ denotes its single activity-pooled view. Accordingly, $\mathcal{V}_I = \{u, m\}$ and $\mathcal{V}_R = \{\text{act}\}$. For a retained CISLR video p , let $h_b^E(p)$ denote its normalized external embedding for $b \in \{I, R\}$. We use $\text{norm}(a) = a/\|a\|_2$.

We construct the external bank by normalizing WR and CISLR glosses through Unicode decomposition, lowercasing, and non-alphanumeric replacement, followed by unambiguous exact matching. Let $\mathcal{E}_y$ be the retained CISLR videos for target label y , and let $\mathcal{C} \subseteq \mathcal{Y}$ contain the labels with both I3D and SignRep external features. Every matched video remains an individual prototype; no class averaging is used during retrieval.

3.3 Local Prototype Retrieval

Video-Swin defines the complete local expert. For each label, CARE selects the closest local instance after taking the strongest agreement among the four uniform/motion view pairs:

$$D_{S,L}(q, y) = \min_{i: y_i=y} \left[1 - \max_{v, w \in \{u, m\}} \langle h_S^v(q), h_S^w(x_i) \rangle \right]. \quad (1)$$

This reduces to nearest-neighbor retrieval for singleton labels and remains valid for the few labels with multiple references. Here, L denotes the local bank and $\langle \cdot, \cdot \rangle$ is the inner product, which equals cosine similarity because the embeddings are normalized. The final system uses no additional local SignRep score.

3.4 Cross-Dataset Alignment and External Retrieval

WR and CISLR can occupy different regions of the same pretrained feature space because of differences in signers, framing, temporal boundaries, and recording conditions. For encoder $b \in \{I, R\}$, we compute normalized local and external centroids for every shared gloss $y \in \mathcal{C}$. The matrices $X_{b,v}$ and Y_b stack the paired local-view and external centroids, respectively, and $\mu_{b,v}^L$ and μ_b^E denote their row means. Here, $v \in \{u, m\}$ for I3D, while SignRep has one activity-pooled channel. Let $R_{b,v}^*$ be the standard orthogonal Procrustes solution obtained by singular value decomposition after centering the two matrices (Schönemann, 1966). A local query is mapped into the external coordinate system as

$$\tilde{h}_b^v(q) = \text{norm}((h_b^v(q) - \mu_{b,v}^L)R_{b,v}^* + \mu_b^E). \quad (2)$$

CARE retains both the original and aligned geometries. Let $h_{b,\text{raw}}^v(q) = h_b^v(q)$ and $h_{b,\text{align}}^v(q) = \tilde{h}_b^v(q)$. External instance similarities are reduced to label distances by nearest-reference matching for $r \in \{\text{raw}, \text{align}\}$:

$$D_{b,E}^r(q, y) = 1 - \max_{p \in \mathcal{E}_y} \max_{v \in \mathcal{V}_b} \langle h_{b,r}^v(q), h_b^E(p) \rangle, \quad y \in \mathcal{C}. \quad (3)$$

Here, E denotes the external bank. Thus, class centroids determine only the alignment map, while retrieval retains individual external instances rather than class averages.

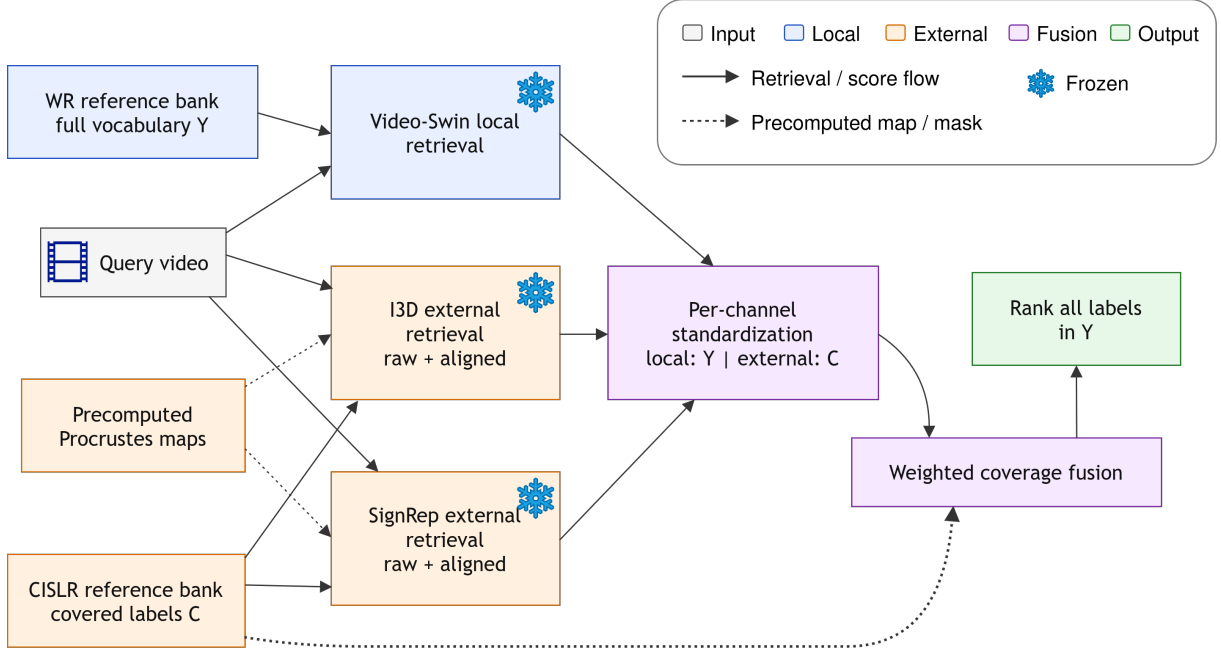


Figure 1: Overview of CARE.

3.5 Coverage-Aware Retrieval Ensemble

Because the retrieval channels can have different distance distributions, each is standardized independently for every query. For a distance function F with finite label support $\mathcal{A}$, $Z_{\mathcal{A}}[F](q, y)$ subtracts $\text{median}_{y' \in \mathcal{A}} F(q, y')$ from $F(q, y)$ and divides by $\text{std}_{y' \in \mathcal{A}} F(q, y') + \varepsilon$, where $\varepsilon > 0$ prevents division by zero. Local scores are standardized over $\mathcal{Y}$, whereas each external channel is standardized separately over $\mathcal{C}$. The complete fusion rule is

$$D(q, y) = Z_{\mathcal{Y}}[D_{S,L}](q, y) + \mathbb{I}[y \in \mathcal{C}] \sum_{b \in \{I, R\}} \sum_{r \in \{\text{raw}, \text{align}\}} \lambda_b^r Z_{\mathcal{C}}[D_{b,E}^r](q, y). \quad (4)$$

Here, λ_b^r is the weight of external channel (b, r) and $\mathbb{I}[\cdot]$ is the indicator function. The indicator is the coverage-aware component: external evidence is added only to labels with valid CISLR support, while uncovered labels retain their complete local score. The term ensemble denotes the late fusion of these normalized score channels. At inference, labels are ranked in ascending order of $D(q, y)$; no backbone update or target-vocabulary classifier is required.

4 Experiments

4.1 Datasets

WSLP 2026 ISLR (WR). The dataset was released for the Isolated Sign Language Recognition track of the WSLP 2026 Shared Task on Indian Sign Language Processing. It contains 5,033 videos: 4,398 labeled training examples, 109 validation queries, and 526 official test queries. Its training vocabulary contains 4,361 glosses, of which 4,336 (99.43%) are singletons. Validation and test labels are withheld, and predictions are evaluated by the official WSLP 2026 task server.

CISLR. We use the public CISLR corpus (Joshi et al., 2022) only as labeled external data. The corpus contains 7,050 videos and 4,765 glosses. The lexical alignment defined in the Methodology retains 3,739 videos covering 2,016 target labels (46.23%). This bank comprises 2,017 videos from the CISLR prototype partition and 1,722 from its released test partition; we do not evaluate performance on the original CISLR task.

4.2 Evaluation Metrics

Following the official WSLP 2026 evaluation metrics, performance is measured using Top-1, Top-5, and Top-10 accuracy. Top- K accuracy is the percentage of videos for which the ground-truth gloss appears among the first K predictions. Top-1 therefore measures exact recognition, whereas Top-5

and Top-10 measure the quality of the short ranked candidate list returned by the system.

4.3 Implementation Details

Video preprocessing and local retrieval. All RGB frames are resized to 224×224 pixels. For Video-Swin, each uniform and motion-quantile view contains 32 frames. We use a frozen Video-Swin-S model initialized with its default pretrained weights. We remove the classification head and apply the normalization associated with those weights.

I3D features. The I3D branch uses the public RGB checkpoint fine-tuned on WLASL-2000. Each WR view contains 96 frames. We extract the final spatio-temporal feature sequence, normalize each temporal vector, average over time, and normalize the result again. External CISLR I3D sequences are pooled using the same normalize-average-normalize rule.

SignRep features. We use the released SignRep checkpoint with its backbone frozen. Videos are divided into 16-frame windows with a stride of 8 frames. Window embeddings are pooled using predicted hand activity to downweight windows with low predicted activity. CARE uses the resulting activity-weighted representation.

Alignment and score fusion. We fit one Procrustes map for each I3D view and one for SignRep using the complete shared-label set. We set $\varepsilon = 10^{-6}$ and use $\lambda_I^{\text{raw}} = 3.25$, $\lambda_I^{\text{align}} = 3.50$, $\lambda_R^{\text{raw}} = 1.00$, and $\lambda_R^{\text{align}} = 6.00$. We assign the local Video-Swin score a weight of 1.00.

5 Results

5.1 Validation and Test Performance

Table 1 reports the performance of the final CARE configuration. Official test accuracy is lower than validation accuracy by 8.72%, 9.04%, and 7.90% at Top-1, Top-5, and Top-10, respectively. Although all backbones remain frozen, repeated validation-based selection of the fusion weights and system configuration may have introduced selection bias. Sampling variability and differences between the split distributions may also contribute to the gap; the aggregate scores alone do not establish their relative contributions. On the test split, Top-10 accuracy exceeds Top-1 accuracy by 2.66%, corresponding to 14 additional queries with the correct label at ranks 2–10.

5.2 Archived External Encoder Comparison

Table 2 compares the two external configurations; the second row corresponds to the final CARE system in Table 1. The SignRep-based configuration scores 2.75% higher at Top-1 and 1.84% higher at both Top-5 and Top-10 than the Video-Swin-based configuration. Because the fusion weights of the two configurations were selected separately on the same validation split, this is a descriptive comparison rather than a controlled ablation.

6 Conclusion

We presented CARE, which combines a complete local dictionary with raw and aligned external retrieval under a coverage mask. Without target-specific backbone fine-tuning, it achieves 65.59% Top-1 accuracy on the official 526-query WSLP 2026 test split. The validation–test gap warrants caution when interpreting performance on the small validation split used for model selection. The reported comparison does not isolate the contributions of individual components; a controlled ablation remains necessary to assess them.

References

- Matyáš Boháček and Marek Hruš. 2023. Learning from what is already out there: Few-shot sign language recognition with online dictionaries. In *2023 17th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6.
- Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4724–4733. IEEE.
- Mathieu De Coster and Joni Dambre. 2023. [Querying a sign language dictionary with videos using dense vector search](#). In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, pages 1–5.
- Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2023. [ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition](#). In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *2021 IEEE/CVF Interna-*

Table 1: CARE results on the WSLP 2026 validation and official test splits. Values in parentheses give the number of queries for which the correct gloss appears within the first K positions.

Split	Queries	Top-1 (%)	Top-5 (%)	Top-10 (%)
Validation	109	74.31 (81)	76.15 (83)	76.15 (83)
Official test	526	65.59 (345)	67.11 (353)	68.25 (359)

Table 2: Validation performance of two archived external-encoder configurations. Values in parentheses give the number of queries whose correct gloss appears within the first K positions.

External branches	Top-1 (%)	Top-5 (%)	Top-10 (%)
I3D + Video-Swin	71.56 (78)	74.31 (81)	74.31 (81)
I3D + SignRep	74.31 (81)	76.15 (83)	76.15 (83)

tional Conference on Computer Vision (ICCV), pages 11067–11076. IEEE.

Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. [CISLR: Corpus for Indian Sign Language recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3192–3201. IEEE.

Peter H. Schönemann. 1966. [A generalized solution of the orthogonal procrustes problem](#). *Psychometrika*, 31(1):1–10.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. [OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. [Return of frustratingly easy domain adaptation](#). *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence, 30(1):2058–2065.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *Proceedings of the British Machine Vision Conference*.

Toon Vandendriessche, Mathieu De Coster, Annelies Lejon, and Joni Dambre. 2025. [Representing signs as signs: One-shot ISLR to facilitate functional sign language technologies](#). *arXiv preprint arXiv:2502.20171*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638.

Chang Wang and Sridhar Mahadevan. 2008. [Manifold alignment using procrustes analysis](#). In *Proceedings of the 25th International Conference on Machine Learning*, pages 1120–1127.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. SignRep: Enhancing self-supervised sign representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22804–22814.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2019. Universal domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2715–2724. IEEE.