Cross-Modal Generalization: Learning in Low Resource Modalities via Meta-Alignment

Paul Pu Liang[♠], Peter Wu[♠], Liu Ziyin[♡], Louis-Philippe Morency[♠], Ruslan Salakhutdinov[♠]
[♠]Carnegie Mellon University, [♡]University of Tokyo
{pliang,peterw1}@cs.cmu.edu

Abstract

The natural world is abundant with concepts expressed via visual, acoustic, tactile, and linguistic modalities. Much of the existing progress in multimodal learning, however, focuses primarily on problems where the same set of modalities are present at train and test time, which makes learning in low-resource modalities particularly difficult. In this work, we propose algorithms for cross-modal generalization: a learning paradigm to train a model that can (1) quickly perform new tasks in a target modality (i.e. meta-learning) and (2) doing so while being trained on a different source modality. We study a key research question: how can we ensure generalization across modalities despite using separate encoders for different source and target modalities? Our solution is based on meta-alignment, a novel method to align representation spaces using strongly and weakly paired cross-modal data while ensuring quick generalization to new tasks across different modalities. We study this problem on 3 classification tasks: text to image, image to audio, and text to speech. Our results demonstrate strong performance even when the new target modality has only a few (1-10) labeled samples and in the presence of noisy labels, a scenario particularly prevalent in low-resource modalities.

1 Introduction

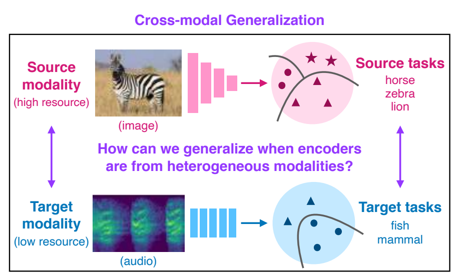

Figure 1: The *cross-modal generalization* paradigm brings discrepancies in both input and output spaces with *new tasks* expressed in *new modalities*. This raises a fundamental question: how can we ensure generalization across modalities despite using separate encoders for different source (image) and target (audio) modalities? This paper studies the minimal supervision required to perform this alignment and succeed in cross-modal generalization.

One of the hallmarks of human intelligence is the ability to generalize seamlessly across heterogeneous sensory inputs and different cognitive tasks [9]. We see objects, hear sounds, feel textures, smell odors, and taste flavors to learn underlying concepts present in our world [4]. Much of AI's existing progress in multimodal learning, however, focuses primarily on a fixed set of predefined modalities and tasks [34, 40] that are consistent between training and testing. As a result, it is unclear how to transfer knowledge from models trained for one modality (e.g. visual source modality) to another (e.g. audio target modality) at test time. This scenario is particularly important for low-resource target modalities where unlabeled data is scarce and labeled data is even harder to obtain (e.g. audio from low-resource languages [38], real-world environments [46], and medical images [14]). In the unimodal case, this is regarded as meta-learning [17] or few-shot learning [8]. In contrast, we formally define the *cross-modal generalization* setting as a learning paradigm to train a model that can (1) quickly perform new tasks in a target modality (i.e. meta-learning) and (2) doing so while being trained on a different source modality. In this paper, we study the data and algorithmic challenges for cross-modal generalization to succeed.

4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada.

As a motivating example, Figure 1 illustrates a scenario where large-scale image classification benchmarks can help audio classification, which is a less studied problem with fewer large-scale benchmarks. In this ambitious problem statement, a key research question becomes: how can we obtain generalization across modalities despite using separate encoders for different source (image) and target (audio) modalities? The technical challenge involves aligning shared knowledge learned from source image tasks with target audio tasks. Our problem statement differs from conventional meta-learning [17] and domain adaptation [59] where one can take advantage of the same source and target modality with shared encoders which helps generalization by having the same representation space. In our case, the discrepancies in modalities requires one to learn *new output concepts* expressed in *new input modalities*. As a result, cross-modal generalization requires new ideas to synchronize (align) multimodal sources and targets. What is the minimal extra supervision required to perform this alignment?

In this paper, we formalize the conditions required for successful generalization and show that another level of supervision is necessary under partial observability across modalities and tasks. Supervision comes in the form of *cross-modal meta-alignment* (Figure 2) to capture a space where representations of similar concepts in different modalities are close together while ensuring quick generalization to new tasks (i.e. with just a few labels in the target modality). We introduce a novel algorithm called CROMA (Cross-modal Meta-Alignment) that leverages readily available multimodal data from the internet (e.g. [34, 48, 67]) for meta-alignment and cross-modal generalization. Through theoretical analysis and empirical ablations, we study our proposed algorithm with both strongly and weakly paired multimodal data, showing that cross-modal generalization is possible even with limited extra supervision.

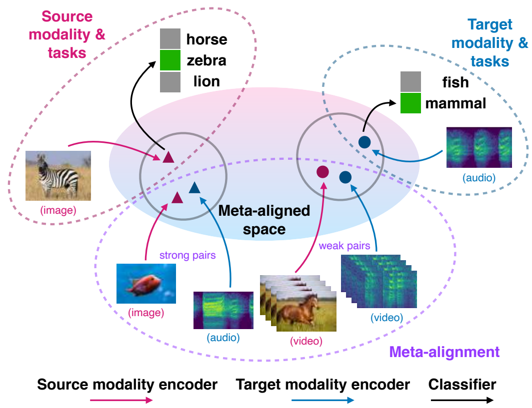

Figure 2: *Meta-alignment* aligns representation spaces while ensuring quick generalization to new tasks using strongly (image and audio) and weakly (video) paired data. This enables generalization to the target (audio) having only seen labeled data in the source (image), as assessed by few-shot classification and alignment tasks.

We present experiments on three cross-modal tasks: generalizing from (1) text to image, (2) image to audio, and (3) text to speech. In all cases, the goal is to classify data from a new target modality given only a few (1-10) labeled samples. We perform extensive experiments to compare with related approaches including target modality meta-learning that would be expected to perform well since they have seen thousands of labeled examples from the target modality during meta-training. Surprisingly, CROMA is competitive with these baselines and significantly outperforms other cross-modal approaches. In addition, we study settings where the target modality suffers from noisy or limited data, a scenario particularly prevalent in low-resource modalities [22].

2 Related Work

Few-shot learning has enabled strong performance for settings with limited labeled data [8, 20, 22] using techniques spanning data augmentation [2], metric learning [55, 61], and learning better initializations [44, 51]. In the latter, **meta-learning** has recently emerged as a popular choice due to its simplicity in combination with gradient-based methods [17].

Transfer learning focuses on transferring knowledge from external data (e.g. larger datasets [29], unlabeled data [12], and knowledge bases [32]) to downstream tasks where labeled data is expensive [57]. **Domain adaptation** similarly focuses on changing data distributions [13, 27]. However, existing works focus on data within the same modality (i.e. image domain adaptation [59], language transfer learning [12]) which simplifies the alignment problem.

Cross-modal alignment involves learning a joint space where the representations of the same concepts expressed in different modalities are close together [4]. Alignment is particularly useful for **cross-modal retrieval** (e.g. retrieving captions from images) [18] and cross-modal (or cross-lingual) representation learning [54, 63]. Several objective functions for learning aligned spaces from varying quantities of paired [6, 16, 28] and unpaired [21] data have been proposed. However, cross-modal generalization is harder since: (1) one has to learn not just the associations between modalities but also associations to labels, (2) there is weak supervision both the target modality and in the label space, (3) tasks in different modalities have different (but related) label spaces, and (4) new tasks in the target modality have to be learned using only a few samples.

Cross-modal learning: Recent work has explored more general models that enable knowledge transfer across modalities. In particular, cross-modal data programming [14] uses weak labels in a source modality to train a classifier in the target modality. Cross-modal transfer learning aims to classify the same task from different input modalities [28, 65]. Finally, few-shot learning within target modalities (e.g. images) has been shown to benefit from additional multimodal information (e.g. word embeddings [56, 58, 64] or videos [66]) during training. However, these all require labeled data from the target modality during meta-training (from a different domain). In contrast, we study *cross-modal* generalization which do not assume *any* labeled data in the target except during few-shot classification.

3 Formalizing Cross-modal Generalization

Cross-modal generalization is a learning paradigm to quickly perform new tasks in a target modality despite being trained on a different source modality. To formalize this paradigm, we build on the definition of meta-learning [25] and generalize it to study multiple input modalities. Meta-learning uses labeled data for existing source tasks to enable fast learning on new target tasks [33]. We start by defining M different heterogeneous input spaces (modalities) and N different label spaces (tasks). We denote a modality by an index $m \in \{1, ..., M\}$ and a task by $n \in \{1, ..., N\}$.

Each classification problem $\mathcal{T}(m,n)$ is defined as a triplet with a modality, task, plus a joint distribution: $\mathcal{T}(m,n) = (\mathcal{X}_m, \mathcal{Y}_n, p_{m,n}(x,y))$. $\mathcal{X}_m$ denotes the input space and $\mathcal{Y}_n$ the label space sampled from a distribution $p(m,n) := p(\mathcal{X}_m, \mathcal{Y}_n)$ given by a marginal over the entire *meta-distribution*, $p(x_1, ..., x_M, y_1, ...y_N, \mathcal{X}_{m_1}, ...\mathcal{X}_{m_M}, \mathcal{Y}_{n_1}, ...\mathcal{Y}_{n_N})$. The meta-distribution gives the underlying relationships between all modalities and tasks through a hierarchical generative process $m_i \sim p(m), n_j \sim p(n)$: first picking a modality and task (m_i, n_j) from priors $p(m)$ and $p(n)$ over input and output spaces, before drawing data x_i from $\mathcal{X}_{m_i}$ and labels y_j from $\mathcal{Y}_{n_j}$. Within each classification problem is also an underlying pairing function mapping inputs to labels through $p_{m,n}(x,y) := p(x,y|m,n)$ for all $x \in \mathcal{X}_m,\ y \in \mathcal{Y}_n$ representing the true data labeling process. To account for generalization over modalities and tasks, cross-modal generalization involves learning a single function f_w with parameters w over the meta-distribution with the following objective:

Definition 1. *The cross-modal generalization problem is*

$$\arg\max_w \mathcal{L}[f_w] := \arg\max_w \mathop{\mathbb{E}}_{\substack{m,n \sim p(m,n) \\ x,y \sim p_{m,n}(x,y)}} \log \left[\frac{f_w(x,y,m,n)}{p(x,y|m,n)} \right]. \tag{1}$$

In practice, the space between modalities and tasks is only *partially observed*: $p(x,y|m,n)$ is only observed for certain modalities and tasks (e.g. labeled classification tasks for images [11], or paired data across image, text, and audio in online videos [1]). For other modality-task pairs, we can only obtain inaccurate estimates $q(x,y|m,n)$, often due to having only *limited labeled data*. It is helpful to think about this *partial observability* as a bipartite graph $G = (V_x, V_y, E)$ between a modality set V_x and task set V_y (see Figure 3). A solid directed edge from $u \in V_x$ to $v \in V_y$ represents learning a classifier from modality u for task v given an abundance of observed labeled data, which incurs negligible generalization error. Since it is unlikely for all edges between V_x and V_y to exist, define the *low-resource subset* $\mathcal{M}$ as the complement of E in $V_x \times V_y$. $\mathcal{M}$ represents the set of low-resource modalities and tasks where it is difficult to obtain labeled data. The focus of cross-modal generalization is to learn a classifier in $\mathcal{M}$ as denoted by a dashed edge. In contrast to solid edges, the lack of data in $\mathcal{M}$ incurs large error along dashed edges.

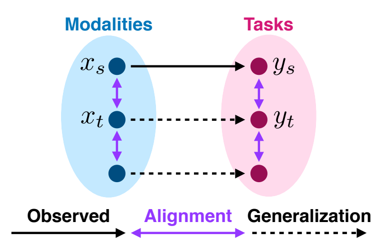

Figure 3: One obtains a subset of observed edges through labeled datasets for source modalities x_s and tasks y_s (solid edges). Generalizing to the target modalities x_t and tasks y_t (dotted edge) requires bridging modalities and tasks through alignment (purple).

Therefore, the challenge in cross-modal generalization amounts to finding the path of lowest cumulative error between an input target modality $x_t \in V_x$ and output task $y_t \in V_y$ in $\mathcal{M}$. The key insight is

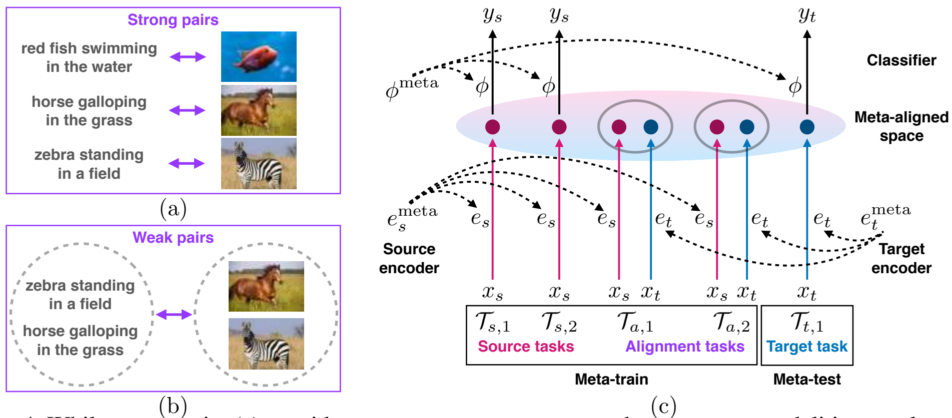

Figure 4: While strong pairs (a) provide exact, one-to-one correspondences across modalities, weak pairs (b) represent coarse semantic groupings which better reflect many-to-many cross-modal mappings and leverage weakly paired multimodal data available on the internet (e.g. videos, image captions). (c) During meta-training, meta-parameters $e_s^{\text{meta}}, e_t^{\text{meta}}, \phi^{\text{meta}}$ are trained using source modality classification tasks $\mathcal{T}_s$ and alignment tasks $\mathcal{T}_a$. Meta-testing uses trained meta-parameters for few-shot generalization to target modality tasks $\mathcal{T}_t$.

to leverage *cross-modal information* to "bridge" modalities that are each labeled for only a subset of tasks (see purple edges in Figure 3). We model cross-modal information as $p(x_s, x_t)$, i.e. *alignment* between modalities x_s and x_t, where x_s is a source modality with high-resource data and labels (x_s, y_s). When there is an abundance of paired data (x_s, x_t) (solid purple edge), we say that *strong* alignment exists; otherwise, only *weak* alignment exits. Since strong alignment incurs negligible error in estimating $p(x_s, x_t)$, the alternative *cross-modal path* $P = \{(x_t, x_s), (x_s, y_s), (y_s, y_t)\}$ might link x_t and y_t with *lower* weighted error and is preferable to direct low-resource training for the dashed edge (x_t, y_t). When only weak alignment is available, a trade-off emerges and one has to choose between the error induced by direct low-resource training and the error induced by weak alignment. (y_s, y_t) models relationships across source and target tasks using approaches such as multi-task [7] or meta-learning [17]. More formally,

Definition 2. *Let $p(x_i, x_j)$ be known for $x_i \in \mathcal{D}^x_{m_i}$, $x_j \in \mathcal{D}^x_{m_j}$ and $i \neq j$. If both $p(x_i|x_j)$ and $p(x_j|x_i)$ are delta distributions (i.e., one-to-one mapping between x_i and x_j), there is a strong alignment between modality m_i and m_j. Otherwise, there is only weak alignment.*

We now show that strong alignment can achieve optimal generalization error for tasks in the low-resource set $\mathcal{M}$.

Proposition 1. *(Benefit of strong alignment). Let all the modalities be pairwise strongly-aligned, then we can define a surrogate loss function $\tilde{\mathcal{L}}[f_w]$ such that $\mathcal{L}[\arg\max_w \tilde{\mathcal{L}}[f_w]] = 0$.*

The proof is provided in Appendix 7. On the other hand, weak alignment may not provide perfect generalization and we elucidate some conditions when it works in Section **??**.

To summarize, cross-modal information $p(x_s, x_t)$ allows us to bridge modalities that are each labeled for only a subset of tasks and achieve generalization to new modalities and tasks in $\mathcal{M}$. In the following section, we explain an algorithm based on contrastive learning [18] to estimate $p(x_s, x_t)$ from data and meta-learning to model (y_s, y_t).

4 CroMA: Cross-modal Meta-Alignment

Based on our theoretical insights, we propose a practical algorithm for cross-modal generalization involving two thrusts: (1) learning a multimodal space via meta-alignment to model (x_s, x_t) (§4.1), and (2) learning a cross-modal classifier to model (y_s, y_t) (§4.2), which jointly enable generalization to new modalities and tasks. We call our method CroMA, short for Cross-modal Meta-Alignment.

4.1 Meta-Alignment

We first simplify the problem by assuming access to strong pairs across modalities of the form (x_s, x_t) which makes it easier to learn *strong alignment* (see Figure 4(a)). At the same time, this is not an excessively strong assumption: many multimodal datasets contain paired multimodal data (e.g. activity recognition from audio and video [1] and emotion recognition from text, speech, and gestures [41, 67]).

Algorithm 1 CroMA: Cross-modal Meta-Alignment

Initialize meta-alignment encoders e_s^{meta} and e_t^{meta}, meta-classifier ϕ^{meta}.
for iteration = $1, 2, \ldots$ **do**
  Sample alignment task $\mathcal{T}_a$ with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_a}$ of pairs $\{x_s, x_t\}$.
  Initialize $e_s \coloneqq e_s^{\text{meta}}, e_t \coloneqq e_t^{\text{meta}}$ and compute alignment loss (2) on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$.
  Compute $\widetilde{e}_s$ and $\widetilde{e}_t$ after gradient updates using alignment loss wrt e_s and e_t.
  Update meta-alignment encoders $e_s^{\text{meta}} \leftarrow e_s^{\text{meta}} + \epsilon(\widetilde{e}_s - e_s^{\text{meta}})$, $e_t^{\text{meta}} \leftarrow e_t^{\text{meta}} + \epsilon(\widetilde{e}_t - e_t^{\text{meta}})$.
  Sample source modality task $\mathcal{T}_s$ with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_s}$ of pairs $\{x_s, y_s\}$.
  Initialize $\phi \coloneqq \phi^{\text{meta}}$ and compute classification loss on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$.
  Compute $\widetilde{\phi}$ after gradient updates using classification loss wrt ϕ.
  Update meta-classifier $\phi^{\text{meta}} \leftarrow \phi^{\text{meta}} + \epsilon(\widetilde{\phi} - \phi^{\text{meta}})$.

In practice, we model alignment by learning $p_\theta(x_t|x_s)$. However, directly learning a translation model $p_\theta(x_t|x_s)$ via MLE by mapping each x_s to its corresponding x_t is unlikely to work in practice since x_s and x_t are extremely high-dimensional and heterogeneous data sources which makes reconstruction difficult [37]. Instead, we use Noise Contrastive Estimation (NCE) which learns a binary classifier to distinguish paired samples $(x_s, x_t) \in \mathcal{D}$ from unpaired negative samples $x_{t,\text{neg}}$, which in the asymptotic limit is an unbiased estimator of $p(x_t|x_s)$ [15] but is much easier in practice than generating raw data.

However, the vanilla NCE objective does not handle new tasks at test time. We propose *meta-alignment* to capture an aligned space (i.e. (x_s, x_t)) while ensuring quick generalization to new tasks across different modalities (i.e. (y_s, y_t)). Meta-alignment trains encoders e_s, e_t for source and target modalities across multiple alignment tasks $\{\mathcal{T}_{a,1}, ..., \mathcal{T}_{a,T}\}$ into an aligned space. Each alignment task $\mathcal{T}_a$ consists of paired data across source and target modalities. We explicitly train for generalization to new tasks by training meta-alignment parameters e_s^{meta} and e_t^{meta} that are used to initialize instances of alignment models for new tasks [17]. When presented with a new task, we first initialize task parameters using meta parameters $e_s \coloneqq e_s^{\text{meta}}$, $e_t \coloneqq e_t^{\text{meta}}$ before training on the task by optimizing for the NCE loss:

$$\mathcal{L}_{\text{strong align}} = \sum_{(x_s, x_t) \in \mathcal{T}_a} \left(-e_s(x_s)^\top e_t(x_t) + \sum_{x_{t,\text{neg}}} e_s(x_s)^\top e_t(x_{t,\text{neg}}) \right). \tag{2}$$

where $x_{t,\text{neg}}$ denotes unpaired negative samples. The NCE objective has a nice interpretation as capturing a space where the representations of similar concepts expressed in different modalities are close together, and different concepts in different modalities are far apart [18, 47]. The meta-parameters e_s^{meta} and e_t^{meta} are updated using first-order gradient information [45] so that they gradually become better initializations for new alignment tasks spanning new concepts.

Weak pairs: We now relax the data requirements from strong to *weak pairs*. Instead of one-to-one correspondences, weak pairs represent coarse groupings of semantic correspondence (see Figure 4(b)). This better reflects real-world multimodal data since cross-modal mappings are often many-to-many (e.g. many ways of describing an image, many ways of speaking the same sentence), and are abundant on the internet such as videos constituting weak pairs of image, audio, text [48, 67]. We denote a weak pair as *sets* X_s and X_t, and define contrastive loss with an expectation over pairs across the sets (i.e. $x_s, x_t \in X_s \times X_t$) and call this *weak alignment*. We sample several $x_s \in X_s$ and $x_t \in X_t$ to treat as paired samples, and obtain negative pairs $x_{t,\text{neg}}$ by sampling outside of the paired sets.

4.2 Cross-modal Generalization

Given a well-aligned space between modalities, we now train a single classifier parametrized by a set of meta-parameters ϕ^{meta} on top of the aligned space for generalization across tasks (y_s, y_t). The joint set of classification tasks consists of tasks $\{\mathcal{T}_{s,1}, ..., \mathcal{T}_{s,T}\}$ in the source modality and tasks $\{\mathcal{T}_{t,1}, ..., \mathcal{T}_{t,T}\}$ in the target. When presented with a new task, we first initialize the classifier using meta parameters $\phi \coloneqq \phi^{\text{meta}}$ before training on the task by optimizing for the cross-entropy loss. The meta-parameters ϕ^{meta} are updated using first-order gradient information [45] towards better initialization parameters to classify new concepts. Overall, the meta-training stage consists of alignment tasks $\mathcal{T}_a$ and classification tasks in the source modality $\mathcal{T}_s$. The meta-testing stage presents tasks in the target modality $\mathcal{T}_t$. Each task consists of k labeled pairs to simulate an episode of k-shot learning. We show the full training algorithm in Algorithm 1 and a visual diagram in Figure 4(c).

During testing, a task $\mathcal{T}_t$ is sampled in the target modality. We initialize a new model with the trained meta-alignment encoder e_t^{meta} and meta-classifier ϕ^{meta}, and perform gradient updates with the k labeled samples in the target modality. Note that throughout the entire training process, only k labeled samples in the target modality are presented to CROMA, which better reflects scarce target modalities where even labeled data for different tasks is difficult to obtain.

5 Experiments

We test generalization from text to image, image to audio, and text to speech classification tasks. Anonymized code is included in the supplementary. Experimental details and additional results are included in Appendix 8 and 9.

5.1 Datasets and Tasks

Text to Image Dataset: We use the Yummly-28K dataset [43] which contains parallel text descriptions and images of recipes. We create classification labels from the metadata by concatenating the meal type and cuisine, yielding 44 distinct classes. The large number of recipes and shared concepts between text and image makes it an ideal testbed for cross-modal generalization. We used a ResNet pretrained on ImageNet [11] to encode the images, pretrained BERT encoder [12] for text, and a shared network for prediction.

Image to Audio Dataset: We combine two large unimodal classification datasets over images (CIFAR-10 and CIFAR-100 [35]) and audio made by various objects (ESC-50 [49]) with partially related label spaces. This allows us to leverage complementary information from both modalities while testing on new concepts. To obtain weak pairs, we map similar classes between the datasets using similarities from WordNet [42] and text cooccurrence. This yields 17 unique clusters of weak pairs (Appendix 8.2 lists all the clusters). We used a ResNet pretrained on ImageNet [11] to encode the images and a convolutional network pretrained on AudioSet [19] to encode audio [36, 52].

Text to Speech Dataset: We use the Wilderness dataset, a large-scale multimodal dataset composed of parallel multilingual text and speech data [5]. We use a subset of 99 languages for language classification from text (source) and speech (target) individually. The tasks are split such as there is no overlap between the text and speech samples used for classification and the pairs seen for strong alignment. We use LSTMs to encode both text and speech data.

Metrics: We report few-shot ($k = 1, 5, 10$) classification accuracy in the target modality by fixing 8 evaluation tasks, each comprised of 5 unseen target concepts during meta-test. We compute accuracy across all 8 tasks and repeat experiments 10 times to report mean and standard deviations.

Baselines: We compare with 4 broad sets of baselines:

1) **Unimodal** baselines only use unlabeled data from the target modality during meta-training following our low-resource assumption. The simplest baseline ignores meta-training and just fine-tunes on the tasks in meta-test starting from a (supervised [3] or unsupervised [12]) **pre-trained** model. To better leverage unlabeled target modality data, we also compare with **unsupervised meta-learning** [26] which performs self-supervised learning via reconstruction or weak labels during meta-training (see Appendix 8.2).

2) We modify **Domain Adaptation** (DA) methods to verify that it is necessary to use separate encoders and perform explicit alignment: a) **Shared** shares all encoder layer for both modalities except a separate linear layer that maps data from the target modality's input dimension to the source [29, 59]. b) **Shared + Align** further adds our alignment loss (contrastive loss) on top of the encoded representations, in a manner similar to [31]. c) **Shared + Domain confusion** further adds a domain confusion loss on top of the encoded representations [60]. d) **Shared + Target labels** also uses target modality labels during meta-training, similar to supervised DA [30] (details in Appendix 10.2).

3) **Cross-modal:** We adapt domain adaptation and meta-learning for cross-modal generalization under the following categories: a) **Align + Classify** which uses supervised alignment methods such as adversarial learning [59], cycle reconstruction [10, 24], or contrastive loss [62] to align input spaces from multiple domains before training a shared classifier [50]. b) **Align + Meta Classify** which learns a shared space using standard supervised alignment [18] before meta-learning a classifier [53], and

Table 1: Performance on text to image generalization on Yummly-28K (top), image to audio concept classification from CIFAR to ESC-50 (middle), and text to speech generalization on the Wilderness dataset (bottom). CROMA is on par and sometimes outperforms the oracle target modality meta-learning approach that has seen thousands of labeled target samples, and also outperforms existing unimodal, domain adaptation, and cross-modal baselines. #Target (labels) denotes the number of target modality samples and labels used during meta-training.

TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#TARGET (LABELS)
Unimodal	Pre-training [3, 12]	33.1 ± 2.8	36.4 ± 3.5	49.0 ± 3.8	0(0)
	Unsup. meta-learning [26] (reconstruct)	37.4 ± 0.6	41.7 ± 3.7	49.0 ± 1.0	5131(0)
Cross-modal	Align + Classify [10, 24, 50, 59, 62]	37.1 ± 3.0	40.0 ± 2.7	47.8 ± 6.6	5131(0)
	Align + Meta Classify [53]	39.4 ± 2.5	40.0 ± 2.3	48.8 ± 7.8	5131(0)
	CROMA (ours)	$\mathbf{39.7} \pm \mathbf{1.3}$	$\mathbf{47.1} \pm \mathbf{3.3}$	$\mathbf{51.1} \pm \mathbf{2.1}$	5131(0)
Oracle	Within modality generalization [17, 45]	38.9 ± 2.1	42.1 ± 1.4	47.9 ± 5.6	5131(5131)
Unimodal	Pre-training [3, 12]	44.2 ± 0.8	72.3 ± 0.3	77.4 ± 1.7	0(0)
	Unsup. meta-learning [26] (reconstruct)	36.3 ± 1.8	67.3 ± 0.9	76.6 ± 2.1	920(0)
	Unsup. meta-learning [26] (weak labels)	45.6 ± 1.3	74.2 ± 0.3	83.7 ± 0.1	920(0)
Cross-modal	Align + Classify [10, 24, 50, 59, 62]	45.3 ± 0.8	73.9 ± 2.1	78.8 ± 0.1	920(0)
	Align + Meta Classify [53]	47.2 ± 0.3	77.1 ± 0.7	80.4 ± 0.0	920(0)
	CROMA (ours)	$\mathbf{47.5} \pm \mathbf{0.2}$	$\mathbf{85.9} \pm \mathbf{0.7}$	$\mathbf{92.7} \pm \mathbf{0.4}$	920(0)
Oracle	Within modality generalization [17, 45]	45.9 ± 0.2	89.3 ± 0.4	94.5 ± 0.3	920(920)
Unimodal	Pre-training [3, 12]	55.2 ± 8.6	73.1 ± 3.4	84.3 ± 0.1	0(0)
	Unsup. meta-learning [26] (reconstruct)	61.5 ± 4.4	83.5 ± 4.0	88.5 ± 2.1	4395(0)
Domain Adaptation	Shared [29]	55.6 ± 10.2	75.2 ± 8.4	81.9 ± 3.9	4395(0)
	Shared + Align [31]	59.7 ± 7.6	78.4 ± 6.2	84.3 ± 1.5	4395(0)
	Shared + Domain confusion [60]	59.5 ± 7.2	76.3 ± 9.4	83.9 ± 1.8	4395(0)
	Shared + Target labels [30]	57.3 ± 9.3	76.2 ± 8.4	84.0 ± 1.9	4395(4395)
Cross-modal	Align + Classify [10, 24, 50, 59, 62]	61.1 ± 6.0	74.8 ± 2.1	86.2 ± 0.7	4395(0)
	Align + Meta Classify [53]	65.6 ± 6.1	89.9 ± 1.5	93.0 ± 0.5	4395(0)
	CROMA (ours)	$\mathbf{67.9} \pm \mathbf{6.6}$	$\mathbf{90.6} \pm \mathbf{1.5}$	$\mathbf{93.2} \pm \mathbf{0.2}$	4395(0)
Oracle	Within modality generalization [17, 45]	61.3 ± 11.2	77.0 ± 0.3	87.5 ± 0.6	4395(4395)

Table 2: Language classification predictions on low-resource speech samples after training on labeled text data. Despite seeing just 5 labeled speech samples, our method is able to accurately classify low-resource languages.

SPEECH (TEXT IN PARENTHESIS)	ORACLE	OURS
(Beda Yesus agot gu ofa oida Bua bururo Didif ojgomu)	Russian	Meax
(Ido hai Timotiu natile hampai moula Aturana Musa)	Jamaican Patois	Badaic
(Mu habotu pa kali Mataoqu osolae vekoi Rau sari Mua kana pa kauru Nenemu gua)	Avokaya	Roviana

c) **CROMA** which represents our full model of jointly training for generalization across alignment and classification tasks. Since all methods are agnostic to the specific alignment algorithm used, we use contrastive loss with negative sampling as described in Section 4.1 for fair comparison across all baselines.

4) **Oracle:** The ideal (but likely unrealistic) scenario where meta-training and meta-testing both have labeled data in the target modality. We use the Reptile algorithm [45] for target modality meta-learning. Since there is the least domain shift, we expect this method to perform best but with the requirement of large amounts of labeled target data.

5.2 Cross-modal Generalization

Comparison to oracle: For text to image (Table 1 top) and text to speech (Table 1 bottom), CROMA surprisingly outperforms even the oracle baseline, in addition to unimodal and cross-modal methods. We hypothesize this is because text data (source) is cleaner than image and speech data (target) and the community has better models for encoding text than images and speech spectrograms. Consistent with this hypothesis, we found that text classifiers performed better on Yummly-28K and Wilderness datasets than image and speech classifiers. This implies that **one can leverage abundant, cleaner, and more-predictive source modalities to improve target modality performance**. For image to audio (Table 1 middle), we observe that our cross-modal approach is on par (outperforms for $k = 1$, and within $2 - 3\%$ for $k = 5, 10$) with the oracle baseline that has seen a thousand labeled audio examples during meta-training.

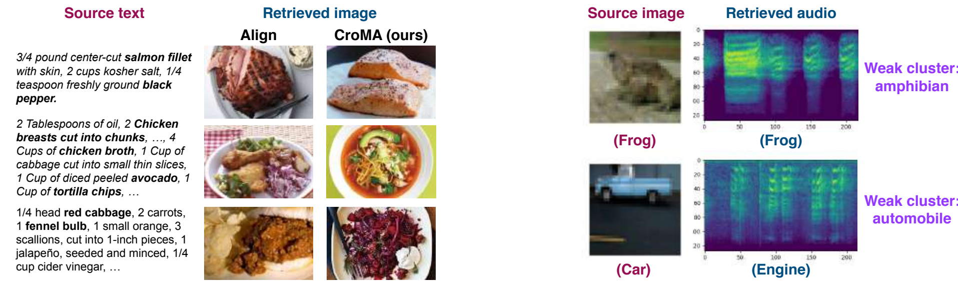

Figure 6: **Left**: samples of retrieved images given text recipes. CROMA performs few-shot retrieval of images more accurately than existing alignment approaches. **Right**: samples of retrieved audio samples given images. Despite being trained only on *weak pairs*, meta-alignment can perform few-shot cross-modal retrieval at fine granularities (e.g. amphibian, automobile).

Comparison to existing approaches: For all setups, CROMA consistently outperforms existing unimodal and cross-modal baselines. Since we use the same LSTM architecture for both text and speech, we can also apply DA approaches which share encoders. From Table 1 bottom, we see that they do not perform well on cross-modal generalization. Although domain confusion and alignment improve upon standard encoder sharing, they still fall short of our approach. Our method also outperforms the Shared + Target labels baseline which further uses target modality labels to train the shared encoder during meta-training. This serves to highlight the important differences between cross-modal generalization and domain adaptation: 1. **separate encoders** and 2. **explicit alignment are important**.

Ablation studies: Consistent across all setups in Table 1, we find that jointly meta-training across alignment and classification improves upon standard supervised alignment methods commonly used in domain adaptation [50, 53]. We find that performance improvement is greatest for the 1-shot setting, suggesting that meta-alignment is particularly suitable for low-resource target modalities.

Model predictions: We show some samples of language classification predictions on low-resource speech samples in Table 2. Despite seeing just 5 labeled speech samples, our method is able to quickly generalize and classify low-resource languages. On the text to image task (Figure 5), CROMA also quickly recognizes images from new recipes.

Figure 5: On Yummly-28K dataset, CROMA leverages source text modality to make accurate few-shot predictions on target image modality despite only seeing $1-10$ labeled image examples.

5.3 Few-shot Cross-modal Retrieval

In Figure 6, we also show samples of retrieved data in the target given input in the source modality to help us understand which source modalities the model is basing its target predictions on. Despite being trained only on weak pairs, **meta-alignment is able to perform cross-modal retrieval at fine granularities**.

5.4 Noisy Target Labels

We also evaluate the effect of noisy labels in the target modality since it is often difficult to obtain exact labels in low-resource modalities such as rare languages. To simulate label noise, we add symmetric noise [23] to all target modality labels (both meta-train and meta-test). Despite only seeing $k = 1, 5, 10$ labels in the target, CROMA is **more robust to noisy label** than the oracle baseline (see Figure 7).

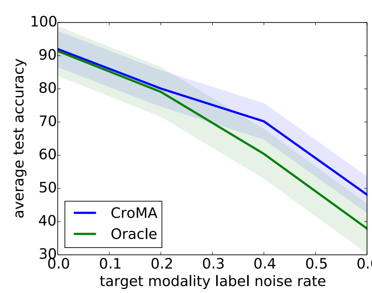

Figure 7: CROMA is robust to noisy labels in the target modality by using cross-modal information from the source, making it suitable for low-resource modalities with imperfect annotations.

6 Conclusion

In this work, we proposed *cross-modal generalization*: a learning paradigm where abundant source modalities are used to help low-resource target modalities. We showed that *meta-alignment* using cross-modal data can allow quick generalization to new concepts across different modalities. Our experiments demonstrate strong performance on classifying data from an entirely new target modality under limited samples and noisy labels, which is particularly useful for generalization to low-resource images, speech, and languages.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[3] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*, 2019.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[5] A. W. Black. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, 2019.

[6] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 81–87, 2017.

[7] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[9] François Chollet. On the measure of intelligence. *CoRR*, abs/1911.01547, 2019.

[10] Safa Cicek and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1416–1425, 2019.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Nanqing Dong and Eric P Xing. Domain adaption in one-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 573–588. Springer, 2018.

[14] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger E. Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning. *CoRR*, abs/1903.11101, 2019.

[15] Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.

[16] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[18] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[20] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *CoRR*, abs/1806.00388, 2018.

[21] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. *CoRR*, abs/1805.11222, 2018.

[22] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. Universal neural machine translation for extremely low resource languages. *CoRR*, abs/1802.05368, 2018.

[23] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-sampling: Training robust networks for extremely noisy supervision. *CoRR*, abs/1804.06872, 2018.

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *CoRR*, abs/1711.03213, 2017.

[25] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[26] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *CoRR*, abs/1810.02334, 2018.

[27] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017.

[28] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1893–1900, 2017.

[29] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614, 2016.

[30] Hal Daumé III. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815, 2009.

[31] Pratik Jawanpuria, Mayank Meghwanshi, and Bamdev Mishra. Geometry-aware domain adaptation for unsupervised alignment of word embeddings. *arXiv preprint arXiv:2004.08243*, 2020.

[32] Annervaz K M, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[33] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. *CoRR*, abs/1902.10644, 2019.

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016.

[35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[36] Anurag Kumar, Maksim Khadkevich, and Christian Fugen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330, 2018.

[37] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

[38] Xinjian Li, Zhong Zhou, Siddharth Dalmia, Alan W Black, and Florian Metze. Santlr: Speech annotation toolkit for low resource languages. *arXiv preprint arXiv:1908.01067*, 2019.

[39] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[40] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1569–1576, 2019.

[41] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 472–476, 2018.

[42] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[43] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2017.

[44] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.

[45] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2:2, 2018.

[46] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018.

[47] Akila Pemasiri, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Semantic correspondence: A hierarchical approach. *CoRR*, abs/1806.03560, 2018.

[48] Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, August 2013.

[49] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[50] Anant Raj, Vinay P. Namboodiri, and Tinne Tuytelaars. Subspace alignment based domain adaptation for RCNN detector. *CoRR*, abs/1507.05578, 2015.

[51] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[52] Tal Ridnik, Hussam Lawen, Asaf Noy, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. *arXiv preprint arXiv:2003.13630*, 2020.

[53] Doyen Sahoo, Hung Le, Chenghao Liu, and Steven C. H. Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift, 2019.

[54] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *CoRR*, abs/1902.09492, 2019.

[55] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[56] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, 2013.

[57] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *CoRR*, abs/1808.01974, 2018.

[58] Yao-Hung Hubert Tsai and Ruslan Salakhutdinov. Improving one-shot learning through fusing side information. *CoRR*, abs/1710.08347, 2017.

[59] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[60] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[61] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[62] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, page 1541–1546. AAAI Press, 2011.

[63] Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. *arXiv preprint arXiv:1910.04708*, 2019.

[64] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*. 2019.

[65] Zhenguo Yang, Min Cheng, Qing Li, Yukun Li, Zehang Lin, and Wenyin Liu. Cross-domain and cross-modality transfer learning for multi-domain and multi-modality event detection. In Athman Bouguettaya, Yunjun Gao, Andrey Klimenko, Lu Chen, Xiangliang Zhang, Fedor Dzerzhinskiy, Weijia Jia, Stanislav V. Klimenko, and Qing Li, editors, *Web Information Systems Engineering – WISE 2017*, pages 516–523, Cham, 2017. Springer International Publishing.

[66] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193, 2020.

[67] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2018.

[68] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 433–443, 2019.