Multi-view Classification Using Hybrid Fusion and Mutual Distillation

Samuel Black Richard Souvenir
Department of Computer and Information Sciences, Temple University, USA
{sam.black,souvenir}@temple.edu

Abstract

Multi-view classification problems are common in medical image analysis, forensics, and other domains where problem queries involve multi-image input. Existing multi-view classification methods are often tailored to a specific task. In this paper, we repurpose off-the-shelf Hybrid CNN-Transformer networks for multi-view classification with either structured or unstructured views. Our approach incorporates a novel fusion scheme, mutual distillation, and minimal additional parameters. We demonstrate the effectiveness and generalization capability of our approach, MV-HFMD, on multiple multi-view classification tasks and show that it outperforms other multi-view approaches, even task-specific methods. Code is available at https://github.com/vidarlab/multi-view-hybrid.

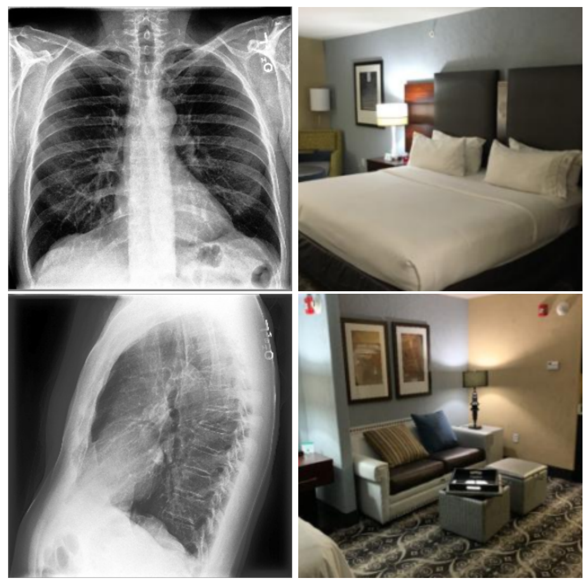


Figure 1. We present a general-purpose multi-view classifier applicable to problems in medical imaging and image forensics.

1. Introduction

In multi-view classification, the goal is to predict a target label from a *collection* of two or more images (or views). For such problems, the underlying assumption is that the component views in a collection give added context and complementary information that is useful or even necessary to make an informed prediction.

Much of the work in this area focuses on the *cross-view* setting, where each collection is comprised of a structured set of (usually two) views of the same object. Cross-view problems are prevalent in the medical image analysis domain, such as the detection of breast cancer from a pair of craniocaudal and mediolateral mammography scans [1, 3, 47, 49, 65]. These types of multi-view problems are quite structured in the sense that each view is captured from a pre-determined pose and intended to highlight a particular feature. Outside the medical domain, some less structured cross-view tasks include 3D-shape recognition [13, 38, 46, 50, 53, 58], plant species identification [8, 28, 40], and action recognition [11, 15]. Other multi-view problems are direct extensions of their single-view analogs, where the additional views are not rigidly prescribed, but may be available at inference time. Figure 1 shows image pairs from various multi-view classification problems.

Existing multi-view methods are often task-specific and not trivially transferable to other related multi-view or single-view problems. In this paper, we present a general framework that employs a novel fusion strategy, is applicable to both structured and unstructured multi-view collections, and only requires minor modifications to off-the-shelf models. We repurpose a hybrid CNN-Transformer network [10] for multi-view classification. The transformer component serves as one aspect of the multi-view fusion model by merging the learned representations from the input images. We also introduce a novel loss term where the fused prediction of the single-views and the multi-view prediction are used as sources of mutual knowledge distillation.

Notably, our approach introduces minimal extra parameters to the single-image backbone and generalizes to collections with varying number of views, including single images. In this paper, we make the following contributions:

- introduce a novel hybrid multi-view fusion strategy, which takes advantage of the hybrid CNN-Transformer architecture;
- apply mutual distillation to multi-view; and
- demonstrate the effectiveness of our approach with extensive experimentation on many multi-view tasks.

2. Related Work

The literature on multi-view methods is vast. Our approach introduces a novel method for multi-image fusion and training multi-view networks using mutual distillation. In this section, we review related methods for multi-view fusion strategies and distillation.

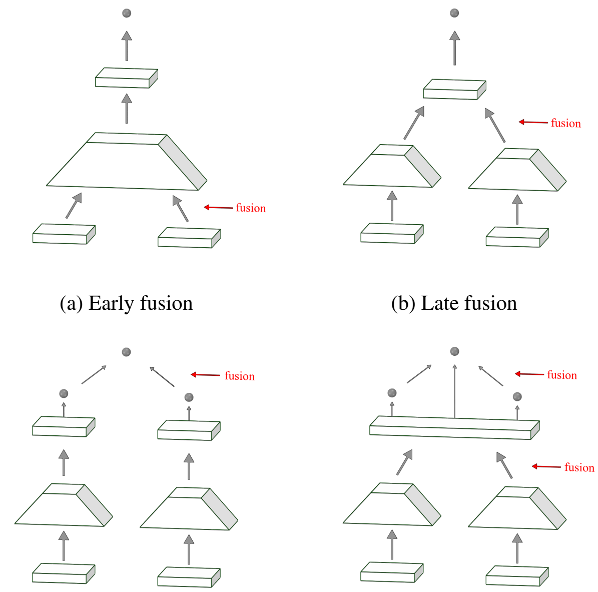
2.1. Multi-view Fusion Strategies

Multi-view methods can be broadly characterized by the stage where the information from the inputs is fused: early fusion, late fusion, and score fusion, as we move downstream the typical processing pipeline.

Early-fusion strategies involve combining low-level features from each view and continuing the training and inference processes in much the same way as the single-view case. Some methods aggregate shallow feature maps from each view before they are processed through a deep network [47, 66]. This approach is often employed in the multimodal setting, such as fusing RGB and optical flow for action recognition [15]. Cross-view Transformers [49] employ attention to transfer ResNet features across the processing streams of each view. One downside to early fusion is that task- irrelevant features may be incorporated early in the processing pipeline [32].

In late fusion, features are learned mainly independently for each input, then combined. Late fusion is a popular strategy, as evidenced by the variety of methods proposed. Some approaches simply concatenate the single-view features [3, 50] or apply pooling operations [38, 46]. Others employ additional processing between the fusion and classification stages. Group View CNN [13] for 3D object recognition uses a learned, two-stage pooling strategy. View features are first assigned to groups and pooled prior to a global pooling step. Other late-fusion strategies utilize bilinear pooling [59], graph convolutions [12, 53], recursive neural networks [33, 34], transformers [4, 57], or other specialized modules [11, 16, 18, 35, 51, 58, 65].

Score fusion can be considered as extreme late fusion where training and inference essentially follow the single-image process, and the output distributions are fused to generate a final prediction. Some methods perform element-wise pooling of the single-view class distributions to generate a multi-view prediction [1, 8, 40, 42]. Bekker et al. [1] train view-specific classifiers on cross-view mammography data, before combining the predictions. Trusted Multi-View



(a) Early fusion (b) Late fusion (c) Score fusion (d) Hybrid fusion

Figure 2. Multi-view fusion paradigms. Multi-view approaches are commonly classified as early, late, or score fusion approaches. Our hybrid fusion approach takes advantage of Hybrid CNN-Transformer architectures for multi-stage fusion.

Classification [17] aims to model prediction uncertainty by combining Dirichlet distribution estimates to generate the multi-view distribution.

Figure 2 provides a sketch of these three fusion strategies along with our proposed hybrid fusion approach, which combines late fusion with score fusion during training.

2.2. Knowledge Distillation and Mutual Learning

Knowledge distillation [2, 20] is a technique used to guide the training of a model (the *student*) using a separate, more complex model (the *teacher*). For classification, knowledge is typically transferred by modifying the student loss function to include an additional divergence term between the predicted distribution with that of the teacher [5, 9, 19, 22, 26, 31, 39, 55, 64]. It has been shown that a higher capacity teacher is not necessary, and performance gains can be achieved using equivalent teacher and student models [14, 60].

Self-knowledge distillation (self-KD) methods forgo a separate teacher model entirely. Inspired by label smoothing [48], Teacher-Free KD [60] augments cross-entropy loss with an additional KL-divergence penalty calculated between the temperature-softened class probability distribution and a uniformly smoothed target distribution. Other approaches have demonstrated the effectiveness of using previous model checkpoints as the teacher [25, 52]. Data-

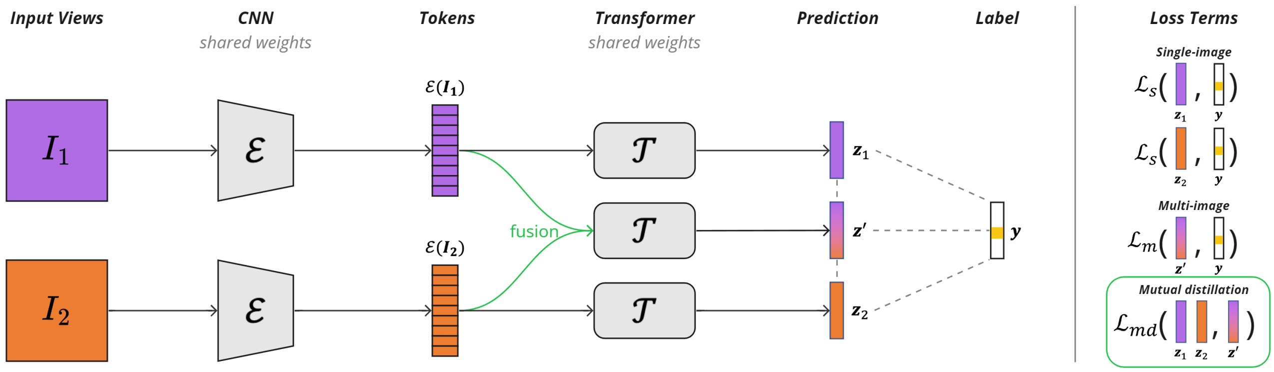


Figure 3. Illustration of our hybrid fusion model, with the two fusion points indicated in green: late fusion of the CNN feature maps and score fusion using mutual distillation of the single-view predictions.

based self-KD approaches involve minimizing the distance between intermediate features or output distributions of a given set of training examples. Augmentation methods apply data distortions to a given training example in order to generate additional model inputs to use in the calculation of the regularization term [27,56]. Other methods add regularization using pairs of different images with the same label to improve classification accuracy [43, 61].

Mutual distillation methods use a feedback loop such that the teacher generating distribution also improves over the course of training the student. For example, in Deep Mutual Learning (DML) [63], an ensemble of models is trained together, each network acting as a teacher for the others. As a given model improves, it generates a more accurate teacher distribution to help train the others. Shadow-KD [30] uses a frozen, pre-trained teacher with a learnable proxy head that facilitates mutual distillation with a student model. Other mutual learning methods utilize a single network with auxiliary output branches to use for distillation [24, 62, 67]. Various degrees of weight sharing between the branches facilitate the teacher-student feedback loop. Rather than adding additional branches, Teacher-Free Feature Distillation [29] introduces both inter and intra-layer loss terms during training.

The aforementioned methods have only been applied to single-view inputs. Only recently have distillation methods been developed for the multi-view setting. ViewsKD [37] uses a pretrained multi-view network to guide the training of a smaller student model. MVC-Net [66] introduces self-distilling mimicry loss to minimize pairwise l_2 distances between output class-probability vectors of each view.

We introduce a mutual distillation loss calculated between the multi-view and the score-fused single-view predictions. Compared to previous methods, our approach does not require a pre-trained teacher, and computation scales linearly with the number of views.

3. Preliminaries

Our main contributions, hybrid fusion and mutual distillation for multi-view classification, take advantage of the Hybrid CNN-Transformer architecture. In this section, we briefly review this model to introduce the notation and lay the foundation for our work.

CNN-Transformer hybrids combine the benefits of each component. The CNN produces a feature map, $\mathcal{C}(\boldsymbol{I}) \in \mathbb{R}^{h\times w\times c}$, where (h, w) is the downsampled resolution and c is the number of channels. This feature map is flattened along the spatial dimension and encoded into the token latent space with a linear projection matrix $\mathbf{E} \in \mathbb{R}^{c\times d}$ to produce a sequence of $S = hw$ dimensional image tokens, each $\in \mathbb{R}^{1\times d}$. A learnable positional encoding $\mathbf{E}_{\mathbf{pos}} \in \mathbb{R}^{S\times d}$ is then summed with the image tokens:

$$\mathcal{E}\left(\boldsymbol{I}\right) = \left[\mathcal{C}\left(\boldsymbol{I}\right)_1 \mathbf{E}; \ldots; \mathcal{C}\left(\boldsymbol{I}\right)_S \mathbf{E}\right] + \mathbf{E}_{\mathbf{pos}} \tag{1}$$

where $\mathcal{C}\left(\boldsymbol{I}\right)_i$ refers to the i-th spatial feature. A learnable token $\boldsymbol{x}_{\mathbf{class}} \in \mathbb{R}^{1\times d}$ is then concatenated with $\mathcal{E}\left(\boldsymbol{I}\right)$ and passed to the Transformer, which consists of a series of L encoder blocks. Information between the tokens is shared at each attention stage; the Transformer facilitates further refinement of the extracted CNN features while incorporating image-wide context. After the last encoding block, $\boldsymbol{x}^L_{\mathbf{class}}$ is passed to the classification layer to produce the class logit distribution. For brevity, we summarize the Transformer and subsequent classification layer as $\mathcal{T}$:

$$\boldsymbol{z} = \mathcal{T}\left(\left[\boldsymbol{x}_{\mathbf{class}}; \mathcal{E}\left(\boldsymbol{I}\right)\right]\right) \tag{2}$$

where $\boldsymbol{z} \in \mathbb{R}^{1\times k}$ and k is the number of classes. In the next section, we describe how this model facilitates hybrid fusion and mutual distillation.

4. Method

First, we introduce a simple modification to hybrid CNN-Transformers to classify an input set of images. Let

$\{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_N\}$ be the input collection of N views. The tokens generated for the collection are passed to the Transformer component of the hybrid model.

$$\boldsymbol{z'} = \mathcal{T}\left([\boldsymbol{x}_{\textbf{class}}; \mathcal{E}'\left(\boldsymbol{I}_1\right); \mathcal{E}'\left(\boldsymbol{I}_2\right); ...; \mathcal{E}'\left(\boldsymbol{I}_N\right)]\right) \quad (3)$$

where $\boldsymbol{z}'$ is the predicted distribution for the input collection and $\mathcal{E}'$ is $\mathcal{E}$ (Eq 1) plus another learnable encoding, $\mathbf{E}_{\textbf{img}} \in \mathbb{R}^{N \times d}$, that is shared for all tokens from a given image. These image embeddings encode the source view in the collection of each token. The number of trainable parameters only increases by Nd compared to the single-image case when the weights are shared for the CNN component of the hybrid model.[1] Figure 3 (left) illustrates our method with the token fusion highlighted.

4.1. Mutual Distillation Training

We formulate our training loss as a combination of three terms, shown visually in Figure 3 (right):

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_s + \lambda \mathcal{L}_{md} \quad (4)$$

where λ is a trade-off parameter to balance the contribution of the distillation term, $\mathcal{L}_{md}$, and classification terms. The first term, $\mathcal{L}_m$ is a classification loss between the multi-view output and the ground truth label. The second term, $\mathcal{L}_s$, is the mean classification loss between the single-view outputs and the ground truth label. For these two terms, there are many choices for loss functions. In Section 5, we show how this approach can be applied to standard loss functions, such as cross-entropy and more modern approaches that include label smoothing or regularization. We also show how both terms contribute to the performance of our method.

For the remainder of the section, we focus on the third loss term, $\mathcal{L}_{md}$, which considers the set of single-view predictions and the multi-view prediction as sources of mutual knowledge distillation [63] for each other. Our approach follows the distillation scheme of Hinton et al. [20], which minimizes the KL-divergence between temperature-softened distributions produced by a teacher and student model:

$$\mathcal{L}_{kd}\left(\boldsymbol{t}, \boldsymbol{s}; \tau\right) = \mathcal{D}_{KL}\left(\tilde{\sigma}\left(\boldsymbol{t}, \tau\right), \tilde{\sigma}\left(\boldsymbol{s}, \tau\right)\right) \quad (5)$$

where $\boldsymbol{t}$ and $\boldsymbol{s}$ are the teacher and student logits, respectively, and $\tilde{\sigma}$ denotes softmax after dividing by a temperature hyperparameter $\tau > 0$. While traditional knowledge distillation involves a one-way knowledge transfer from the teacher to the student, for $\mathcal{L}_{md}$ we compute two asymmetric distillation terms between a score-fused class distribution and $\boldsymbol{z'}$.

[1] For cross-view medical image analysis or multimodal problems, it is common for the weights associated with each view to be unshared.

$$\begin{aligned}&\mathcal{L}_{md}\left(\{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}, \boldsymbol{z'}; \tau\right) = \\ &\quad \frac{1}{2}\tau^2 \left(\mathcal{L}_{kd}\left(\hat{\bar{\boldsymbol{z}}}, \boldsymbol{z'}; \tau\right) + \mathcal{L}_{kd}\left(\hat{\boldsymbol{z}}', \bar{\boldsymbol{z}}; \tau\right)\right)\end{aligned} \quad (6)$$

where $\bar{\boldsymbol{z}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i$. Similar to other self-KD methods, the loss term, $\mathcal{L}_{md}$, uses the model predictions as sources of distillation. It is tailored to the multi-view setting by penalizing misalignment of the score-fused logits and multi-view predictions, which (as we demostrate in Section 5) improves generalization capability. Note that $\hat{\bar{\boldsymbol{z}}}$ and $\hat{\boldsymbol{z}}'$ signify gradient-detached copies of $\bar{\boldsymbol{z}}$ and $\boldsymbol{z'}$. This follows previous work in treating the teaching distributions as a constant for the purpose calculating gradients [36, 61]. Additionally, following the recommendation in [20], the distillation term is weighted by τ^2 to account for the resulting decrease in gradient magnitude when using temperature softening.

Inference While training requires computing single-view predictions from each image in the collection, inference only requires computing the multi-view prediction, $\boldsymbol{z}'$.

5. Experiments

We evaluate the effectiveness of our Multi-View Classifier with Hybrid Fusion and Mutual Distillation (MV-HFMD) on three different multi-view domains.

5.1. Experimental Setup

Unless otherwise specified, the backbone model is the ResNet26+Small ViT pre-trained on Imagnet [7] with an effective 32× CNN downsampling ratio [45]. The model is optimized using stochastic gradient descent, with a 1-cycle learning rate scheduler [44], and a batch size of 64. Experiments were conducted using NVIDIA RTX GPUs. Test set results are reported for the checkpoint that achieves the highest accuracy on the held-out validation set.

Datasets The *CheXpert* dataset [21] contains chest x-ray images collected from over 65,000 patients. Following [49], we use the subset of samples that include both a frontal and lateral scan for a given patient. Each pair is annotated for 13 different observations with one of four possible labels: "unknown" (missing), "uncertain", "negative", or "positive". There are 23,628, 3,915, and 2,802 samples in the train, validation and test splits, respectively. *Hotels-8k* [23] consists of 99,513 images of hotel rooms belonging to one of 7,774 different hotels. This dataset represents an unconstrained multi-view variant as images from the same class may contain minimal to no overlap between views. We use the designated train and test split, withholding 10% of the training data for validation. *Google LandmarksV2* [54] consists of millions of images. We use a subset of GLM that consists of

Method	CNN Arch	AUC-ROC
MVCNN [46]	ResNet26	.815 ± .004
CVT [49]*	ResNet18	.813 ± .003
MVC-NET [66]*	ResNet26	.813 ± .005
GVCNN [13]	InceptionV4	.805 ± .003
TMC [17]*	ResNet26	.802 ± .002
MVT [4]	N/A	.816 ± .003
MV-HFMD (ours)	ResNet26	.835 ± .003
MV-HFMD (ours)*	ResNet26	**.845 ± .002**

Table 1. Cross-view chest x-ray classification. AUC-ROC (mean $\pm$ SD) across 13 classification tasks, each repeated over four training runs. * indicates unshared weights for input views.

	Accuracy		Computation	
Method	T1	T5	Params	GFLOPS
MVCNN [46]	.460	.623	14.0	9.40
CVT [49]	.451	.621	12.4	14.9
GVCNN [13]	.475	.660	41.2	24.5
MVC-NET [66]	.515	.677	32.6	30.4
TMC [17]	.515	.681	14.0	9.40
MVT [4]	.597	.756	21.7	17.0
MV-HFMD (ours)	**.651**	**.807**	36.1	13.9

Table 2. Performance (Top-1 and top-5 classification accuracy) and computational efficiency (millions of parameters and GFLOPS) for multi-view classification on Hotels-8k.

104,763 images from 18,283 classes for training, 21,019 for validation and 28,098 for testing. Images in a given class include a well-defined human-made or natural landmark that is at least partially visible in each view. For Hotels-8k and Google Landmarks, images are re-sized to 224×224 prior to processing, resulting in 49 tokens per view. CheXpert images are re-sized to 384×384, resulting in 144 tokens per view. Images from Hotels-8k and Google LandmarksV2 are not naturally paired; training image pairs are dynamically generated from images of the same class. The results of the test set include all combinations for a given collection size.

Hyperparameter Tuning Hyperparameters were tuned using cross-validation with the Hotels-8k dataset. The approach is relatively insensitive to the value of the temperature hyperparameter, τ; we use $\tau = 4$, which aligns with other distillation methods [6, 26, 61]. For the weighting term, we set $\lambda = .1$.

Baselines We compare MV-HFMD to the following methods: Cross-View Transformers (CVT) [49], Multi-View Transformers (MVT) [4], Trusted Multi-View Classification (TMC) [17], Multi-View Chest Radiograph Classification Network (MVC-NET) [66], Multi-View CNN (MVCNN) [46], and Group-View CNN (GVCNN) [13]. We use the author's implementation where available and the same training and evaluation process as MV-HFMD.

5.2. Cross-view Classification

CheXpert is a benchmark dataset representative of classic cross-view classification problems common to medical image analysis. We follow the experimental protocol described in [49], which includes 13 binary classification tasks repeated over four training runs.[2]

The results, presented in Table 1, show the mean AUC-ROC score reported across all tasks. Our method outperforms all baseline methods in cross-view accuracy and in all but three of the 13 individual tasks[3]. We evaluated MV-HFMD with both shared and unshared weights for the CNN component and observed a 1% performance improvement with the latter. While common in medical image analysis, the unshared approach doubles the total CNN parameters.

5.3. Unstructured Mutli-view Classification

For the more general case of multi-view classification, we evaluate on Hotels-8k, where each class represents multiple views of hotel rooms from the same hotel. Unlike medical image analysis, the views are neither paired nor prescribed and show a much greater variance in camera pose and capture time. For all models, the CNN weights are shared due to the unstructured nature of the collections. Results are presented in Table 2, showing the Top-1 and Top-5 classification accuracy for two-image inputs.

MV-HFMD outperforms all baselines on this dataset, some by quite a wide margin. This dataset includes image pairs with very little overlap and, thus, high intracollection variability, which violates some of the assumptions of specialized cross-view methods. Figure 4 shows examples where one (or both) of the constituent views were incorrectly classified, but the multi-view collection was correctly classified.

Table 2 also includes a comparison of the model size and total computation of each method. While MV-HFMD is comparable in size to some of the larger models, the computation requirements are on par with the most efficient models in this domain. MV-HFMD can be efficiently trained using a single high-end workstation GPU.

5.4. Multi-view Training as Regularization

Similar to previous work [46], we observe that our multi-view training method acts as a regularizer for single-view classification. We follow the multi-view training pro-

[2]Methods with better CheXpert results in the literature generally use the full dataset (not just cross-view) and hi-res images.

[3]Expanded individual view and subtask results in the supplemental material

Figure 4. Examples from MV-HFMD with correct multi-view classification but one (or both) of the constituent views were incorrectly classified (correct: green, incorrect: red).

cess, but at inference evaluate single-view input. We compare this to the same hybrid CNN-Transformer architecture trained in the standard single-view manner. In these experiments, we include results from a subset of Google LandmarksV2 (GLM), which is not a dataset (or problem) typically considered in the multi-view paradigm. We do not seek to report SOTA results on GLM, but demonstrate the benefit of this training scheme for single-view inference. Table 3 shows the results for single-view classification.

Our method outperforms the single-view baselines, achieving 7% and 4% higher top-1 classification accuracy for Hotels-8k and Google Landmarks, respectively. This approach outperforms published results on Hotels-8k. Our method achieves a MAP@5 of .558. For comparison, the dataset authors report a MAP@5 of .551 [23].

	Hotels-8k		Landmarks	
Method	T1	T5	T1	T5
Baseline	.463	.633	.818	.904
MV-HFMD	**.498**	**.653**	**.851**	**.926**

Table 3. Cross-view training as regularization. Top-1 and top-5 classification accuracy for single-view classification on Hotels-8k and Google LandmarksV2. The cross-view training method acts as a regularizer and improves single-view classification performance.

#	$\mathcal{L}_s$	$\mathcal{L}_m$	$\mathcal{L}_{md}$	Multi-view	Single-view
1	✓			.562	.448
2		✓		.559	.376
3	✓	✓		.612	.458
4		✓	✓	.590	.403
5	✓		✓	.611	.490
6	✓	✓	$\rightarrow$	.628	.471
7	✓	✓	$\leftarrow$	.646	.499
8	✓	✓	✓	.651	.498

Table 4. Ablation study. Top-1 classification accuracy on Hotels-8k using different combinations of loss terms.

5.5. Ablation Study

We perform an ablation study (Table 4) on the three components of the loss function: single-image ($\mathcal{L}_s$), multi-image ($\mathcal{L}_m$) and mutual distillation ($\mathcal{L}_{md}$). For each setting, we train the model and evaluate the performance on Hotels-8k for both the multi-view and single-view predictions.

We first notice that all three components play a role in the overall performance; all subsets of the loss terms significantly underperform the full loss function. Next, we observe the significance of the single-view loss term by comparing settings 2 vs 3 and 4 vs 8. In both cases, we observe a positive contribution by including the single-view and multi-view parallel training. Our novel mutual distillation term, $\mathcal{L}_{md}$ contributes the most to the performance of the method. This can be observed by comparing settings 3 vs 8, where the improvement is roughly 6-8% depending on the classification mode. Settings 6 and 7 show the unidirectional variants of $\mathcal{L}_{md}$, which include one of the two terms of Equation 6. Both perform worse than the mutual distillation version in the multi-view setting, while using only the multi-view prediction as the teacher (setting 7) performs similarly to the full method for single-view.

5.6. Other Classification Losses

For the preceding experiments, we simply applied standard cross entropy loss for the two classification terms in our model, $\mathcal{L}_s$ and $\mathcal{L}_m$. However, more modern approaches,

Architecture	Loss	$\mathcal{L}_s + \mathcal{L}_m$	$+\mathcal{L}_{md}$	Δ
R+ViT-Ti/16	CE	.517	.550	+.033
	LS	.537	.551	+.014
	TF-KD	.529	.549	+.020
	CS-KD	.549	.567	+.018
	PS-KD	.543	.535	- .008
R26 + ViT-S/32	CE	.612	.651	+.039
	LS	.631	.663	+ .032
	TF-KD	.609	.646	+ .037
	CS-KD	.682	.692	+ .010
	PS-KD	.662	.680	+.018
R50 + ViT-B/16	CE	.664	.733	+.069
	LS	.714	.738	+.024
	TF-KD	.664	.731	+.067
	CS-KD	.738	.752	+.014
	PS-KD	.734	.736	+.002

Table 5. Multi-view accuracy using different loss functions on Hotels-8k with and without our mutual distillation loss.

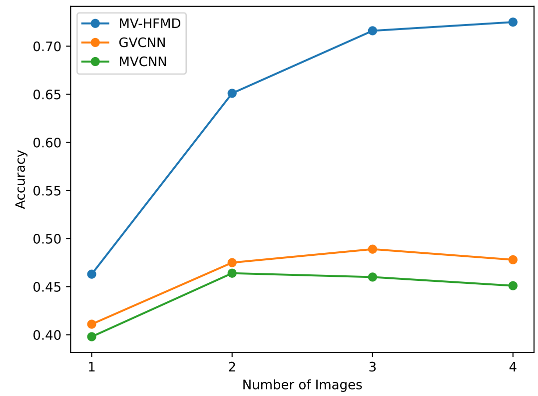


Figure 5. Top-1 multi-view classification accuracy on Hotels-8k for $N = 1, 2, 3, 4$ multi-view collections.

such as distillation, can be substituted for these loss terms. Using the Hotels-8k dataset, we evaluate other classification losses, including label smoothing (LS) [48] and three self-knowledge distillation methods: Teacher-Free KD (TF-KD) regularization [60], Classwise-KD (CS-KD) [61], and Self-Distillation with Progressive Refinement of Targets (PS-KD) [25]. For the self-KD methods, we use the implementations provided by the respective authors.[4] Table 5 shows the results for each classification loss function across three (small, medium, large) architectures. For each, we train with and without our mutual distillation term included.

In line with the single-view results presented in the respective papers, incorporating label smoothing and self-distillation improves the performance in the multi-view setting. Moreover, adding our mutual distillation term gives an extra boost in performance in all but one case. Notably, we observe the largest gains in the medium and larger sized networks, which likely benefit the most from the additional regularization that the mutual distillation term provides.

5.7. Beyond Two Views

Although we focused on the most common setting of multi-view classification with $N = 2$ images in the collection, we show that our method, MV-HFMD, continues to outperform competing approaches when more images are used in training and testing. Figure 5 shows the accuracy on Hotels-8k with collection sizes of up to 4 images for MV-HFMD and two competing methods. Although performance increases with additional views, there are diminishing returns as more are added, which is unsurprising since each additional view is more likely to contribute redundant information. Nonetheless, the marginal gains achieved with each additional view are far greater for MV-HFMD than for MVCNN and GVCNN, which both degrade at $N = 3$.

[4]For PS-KD, we compute $\mathcal{L}_s$ and $\mathcal{L}_m$ using the single and multi-view logits generated from the checkpoint from the previous epoch.

6. Discussion

We evaluated both cross-view classification and the general case of multi-image classification. Across various domains, model architectures, and other settings, our method demonstrated strong performance at a variety of multi-view tasks and also as a regularization method to improve more common single-view tasks.

Single-view vs Multi-view Activation Maps To better understand multi-view classification, we inspect the class activation maps for the same images in the single-view regime compared to the multi-view case. Each row of Figure 6 shows the activation maps generated for a pair of images with models trained for single-view classification and ($N = 2$) multi-view classification. Activation maps are computed using the method in [41] with the token embeddings that immediately precede the final transformer block.

The visualizations suggest that saliency changes significantly between these classification regimes. In the first example, the single-view maps show that the most prominent regions include the floor and curtains. For multi-view, the most activated regions are the wall in the first image, and the headboard in the second. This suggests that the model learns to associate different combinations of features with a given class, including those that span multiple views. We again observe this pattern in the second row, where the activated regions of the first image shifts to the chair, while the focus of the second image shifts away from the bed in the multi-view setting.

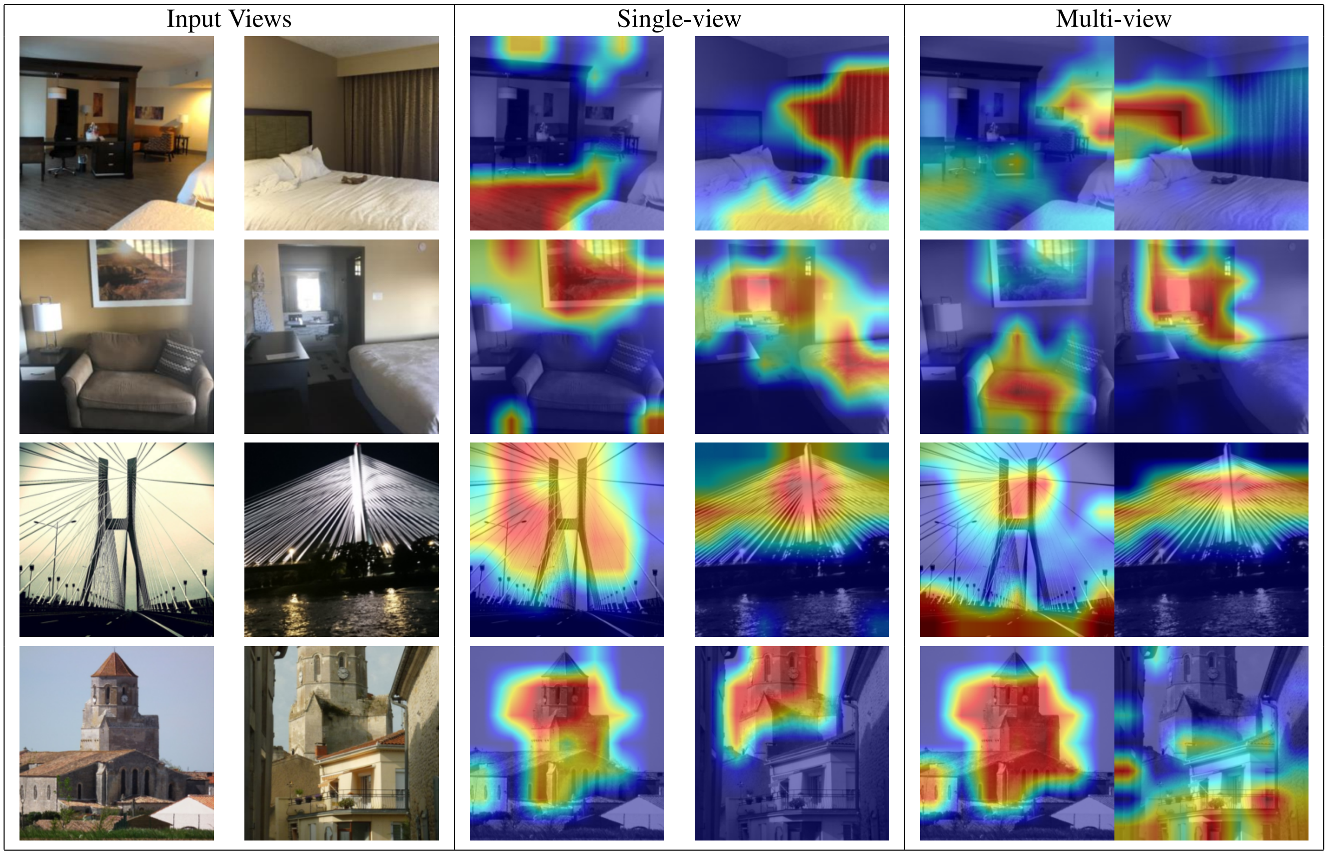


Figure 6. For each pair of input images (left), the middle columns show the class activation maps in the single-view setting and the right-most columns show the class activation maps in the multi-view setting.

Compared to the images in Hotels-8k, images in Google LandmarksV2 contain a higher degree of overlap. Consider the third example. In both single-view maps, suspension cables are highlighted as a prominent feature. However, for the multi-view case, the region for the cables remains active for the right image while the left image adds new salient regions around the road. Similarly, we see this in the fourth example with the steeple, where saliency shifts in the second image to the building facades. Typically, when there exists overlapping visual information, it activated for only one of the views in the cross-view case.

Architecture The motivations for using a hybrid CNN-Transformer for multi-view classification are twofold. First, it follows the paradigm of recent late-fusion methods, which use a CNN to extract view-specific features before aggregating them into a global collection embedding [35, 38, 46, 65]. Second, transformer-based fusion enables compatibility with structured and unstructured collections. Fusion strategies requiring knowledge of how views relate, such as graph convolutional networks [53] or sequential integration [16, 18, 33, 34], may not generalize to unstructured data.

7. Conclusion

We introduced a general-purpose approach for multi-view classification, which takes advantage of hybrid CNN-Transformer architectures and introduces hybrid fusion. Our approach outperforms baselines and specialized methods across a range of domains. For future work, we plan to investigate distillation schemes that explicitly account for the overlap between the input views. We also plan to explore multimodal applications; however, this setting would introduce non-trivial changes. Our current approach adapts an off-the-shelf model for multi-view classification. Non-image input would require separate processing and introduce additional parameters.

Acknowledgements

This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189. Special thanks to Muhsin Fatih Yorulmaz for his help with designing figures.

References

[1] Alan Joseph Bekker, Moran Shalhon, Hayit Greenspan, and Jacob Goldberger. Multi-view probabilistic classification of breast microcalcifications. *IEEE Transactions on medical imaging*, 35(2):645–653, 2015. 1, 2

[2] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, 2006. 2

[3] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. *Deep Learning for Medical Image Analysis*, pages 321–339, 2017. 1, 2

[4] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. In *Proceedings of the British Machine Vision Conference*, 2021. 2, 5

[5] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proc. International Conference on Computer Vision*, October 2019. 2

[6] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4794–4802, 2019. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 4

[8] Thanh-Binh Do, Huy-Hoang Nguyen, Hai Vu, Thi-Lan Le, et al. Plant identification using score-based fusion of multi-organ images. In *International Conference on Knowledge and Systems Engineering (KSE)*, pages 191–196. IEEE, 2017. 1, 2

[9] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11898–11908, 2023. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. OpenReview.net, 2021. 1

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016. 1, 2

[12] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proc. National Conference on Artificial Intelligence*, volume 33, pages 3558–3565, 2019. 2

[13] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018. 1, 2, 5

[14] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 2

[15] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *International Conference on Information Fusion*, pages 1–6. IEEE, 2020. 1, 2

[16] Zhizhong Han, Honglei Lu, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. 3d2seqviews: Aggregating sequential views for 3d global feature learning by cnn with hierarchical attention aggregation. *IEEE Transactions on Image Processing*, 28(8):3986–3999, 2019. 2, 8

[17] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 5

[18] Xinwei He, Tengteng Huang, Song Bai, and Xiang Bai. View n-gram network for 3d object retrieval. In *Proc. International Conference on Computer Vision*, pages 7515–7524, 2019. 2, 8

[19] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proc. International Conference on Computer Vision*, pages 1921–1930, 2019. 2

[20] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 2, 4

[21] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. National Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 4

[22] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proc. National Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021. 2

[23] Rashmi Kamath, Gregory Rolwes, Samuel Black, and Abby Stylianou. The 2021 hotel-id to combat human trafficking competition dataset. *arXiv preprint arXiv:2106.05746*, 2021. 4, 6

[24] Jangho Kim, Minsung Hyun, Inseop Chung, and Nojun Kwak. Feature fusion for online mutual knowledge distillation. In *Proc. International Conference on Pattern Recognition*, pages 4619–4625. IEEE, 2021. 3

[25] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proc. International Conference on Computer Vision*, pages 6567–6576, 2021. 2, 7

[26] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 2, 5

[27] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Self-supervised label augmentation via input transformations.

In *International Conference on Machine Learning*, pages 5714–5724. PMLR, 2020. 3

[28] Sue Han Lee, Chee Seng Chan, and Paolo Remagnino. Multi-organ plant classification based on convolutional and recurrent neural networks. *IEEE Transactions on Image Processing*, 27(9):4287–4301, 2018. 1

[29] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *Proc. European Conference on Computer Vision*, pages 347–363. Springer, 2022. 3

[30] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems*, 35:635–649, 2022. 3

[31] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 10915–10924, June 2022. 2

[32] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 2

[33] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8778–8785, 2019. 2, 8

[34] Chao Ma, Yulan Guo, Jungang Yang, and Wei An. Learning multi-view representation with lstm for 3-d shape recognition and retrieval. *IEEE Transactions on Multimedia*, 21(5):1169–1182, 2018. 2, 8

[35] Jiechao Ma, Xiang Li, Hongwei Li, Ruixuan Wang, Bjoern Menze, and Wei-Shi Zheng. Cross-view relation networks for mammogram mass detection. In *Proc. International Conference on Pattern Recognition*, pages 8632–8638. IEEE, 2021. 2, 8

[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018. 4

[37] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *Proc. European Conference on Computer Vision*, pages 93–110. Springer, 2020. 3

[38] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2016. 1, 2, 8

[39] Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montréal, Y. Bengio, Université De Montréal, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. 2

[40] Marco Seeland and Patrick Mäder. Multi-view classification with convolutional neural networks. *PLOS One*, 16(1), 2021. 1, 2

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. International Conference on Computer Vision*, pages 618–626, 2017. 7

[42] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram Van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016. 2

[43] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11943–11952, 2022. 3

[44] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 369–386. SPIE, 2019. 4

[45] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. 4

[46] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 945–953, 2015. 1, 2, 5, 8

[47] Lilei Sun, Junqian Wang, Zhijun Hu, Yong Xu, and Zhongwei Cui. Multi-view convolutional neural networks for mammographic image classification. *IEEE Access*, 7:126273–126282, 2019. 1, 2

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 2, 7

[49] Gijs van Tulder, Yao Tong, and Elena Marchiori. Multi-view analysis of unregistered medical images using cross-view transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 104–113. Springer, 2021. 1, 2, 4, 5

[50] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *Proc. International Conference on Computer Vision*, pages 1125–1133, 2015. 1, 2

[51] Hongyu Wang, Jun Feng, Zizhao Zhang, Hai Su, Lei Cui, Hua He, and Li Liu. Breast mass classification via deeply integrating the contextual information from multi-view data. *Pattern Recognition*, 80:42–52, 2018. 2

[52] Jiyue Wang, Pei Zhang, and Yanxiong Li. Memory-replay knowledge distillation. *Sensors*, 21(8):2792, 2021. 2

[53] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *Proc.*

IEEE Conference on Computer Vision and Pattern Recognition, June 2020. 1, 2, 8

[54] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[55] Liu Xiaolong, Li Lujun, Li Chao, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. In *International Conference on Learning Representations*, 2023. 2

[56] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proc. National Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019. 3

[57] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 2

[58] Ze Yang and Liwei Wang. Learning relationships for multi-view 3d object recognition. In *Proc. International Conference on Computer Vision*, pages 7505–7514, 2019. 1, 2

[59] Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 186–194, 2018. 2

[60] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 7

[61] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 3, 4, 5, 7

[62] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3713–3722, 2019. 3

[63] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 3, 4

[64] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 2

[65] Xuran Zhao, Luyang Yu, and Xun Wang. Cross-view attention network for breast cancer screening from multi-view mammograms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1050–1054. IEEE, 2020. 1, 2, 8

[66] Xiongfeng Zhu and Qianjin Feng. Mvc-net: Multi-view chest radiograph classification network with deep fusion. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 554–558. IEEE, 2021. 2, 3, 5

[67] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in Neural Information Processing Systems*, 31, 2018. 3