Generative Modeling for Multi-task Visual Learning

Zhipeng Bao Yu-Xiong Wang Martial Hebert
Robotics Institute, Carnegie Mellon University
{zbao, yuxiongw, hebert}@cs.cmu.edu

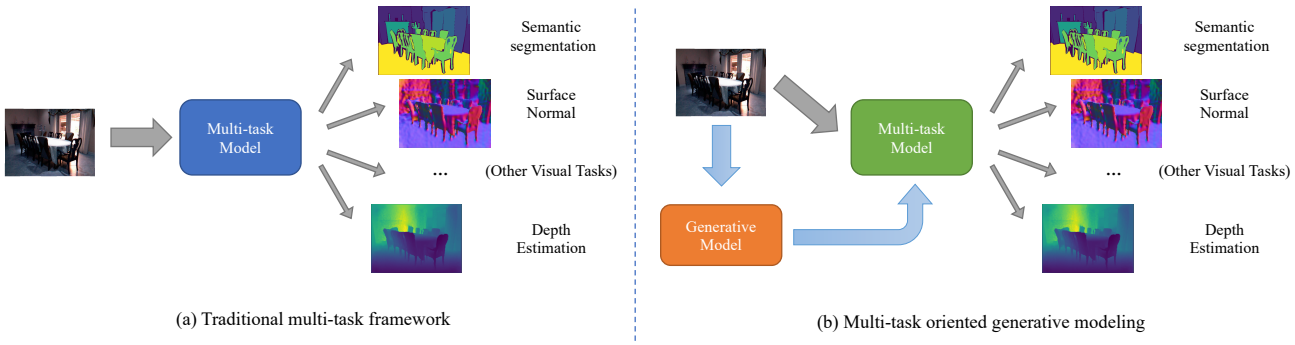

Figure 1: **Left:** Traditional multi-task learning framework (that learns shared feature representations) **v.s. Right:** our proposed multi-task oriented generative modeling (that learns a shared generative model across various visual perception tasks)

Abstract

Generative modeling has recently shown great promise in computer vision, but it has mostly focused on synthesizing visually realistic images. In this paper, motivated by multi-task learning of shareable feature representations, we consider a broader problem of learning a shared generative model that is useful across various visual perception tasks. Correspondingly, we propose a general multi-task oriented generative modeling (MGM) framework, by coupling a discriminative multi-task network with a generative network. While it is challenging to synthesize both RGB images and pixel-level annotations in multi-task scenarios, our framework enables us to use synthesized images paired with only weak annotations (i.e., image-level scene labels) to facilitate multiple visual tasks. Experimental evaluation on challenging multi-task benchmarks, including NYUv2 and Taskonomy, demonstrates that our MGM framework improves the performance of all the tasks by large margins, consistently outperforming state-of-the-art multi-task approaches.

1. Introduction

Seeing with the mind's eye — creating internal images of objects and scenes not actually present to the senses, is perhaps one of the hallmarks in human cognition [40]. For humans, this visual imagination integrates learning experience and facilities learning process in solving different problems [17, 39, 40, 16]. Inspired by such ability, there has been increasing interest in building generative models that are able to synthesize images [21]. Yet, most of the effort has focused on generating visually realistic images [6, 60], which are still far from useful for machine perception tasks [48, 5, 55]. Even though recent work has started improving the "usefulness" of synthesized images, this line of investigation is often limited to a single specific task [33, 50, 61, 51]. Could we guide generative models to benefit *multiple* visual tasks?

While similar ideas have been widely studied as multi-task learning or meta-learning of shared feature representations [19, 58], here we are taking a different perspective — *learning a shared generative model across various tasks* (their comparision is illustrated in Figure 1). Leveraging multiple tasks allows us to capture the underlying image generation mechanism for more comprehensive object and scene understanding than could be done within individual tasks. Taking simultaneous semantic segmentation, depth estimation, and surface normal prediction as an example (Figure 1), successful generative modeling requires understanding not only the semantics but also the 3D geometric structure and physical property of the input image. Meanwhile, a learned common generative model facilities knowledge to flow across tasks, thus beneficial to one another. For instance, the synthesized images provide meaningful variations in existing images and could then be used as additional training data to build better task-specific models.

This paper thus explores *multi-task oriented generative modeling* (MGM), by coupling a discriminative multi-task

network with a generative network. To make them cooperative with each other, a straightforward solution would be to synthesize both RGB images and corresponding *pixel-level annotations* (*e.g.*, pixel-wise class labels for semantic segmentation and depth map for depth estimation). For example, in the single task scenario, existing work trains a separate generative model to synthesize paired pixel-level labeled data [46, 9] as augmented set to train a better model. However, the quality and distribution of the generated annotations are not guaranteed. Moreover, these models are still highly task-dependant, and extending them to multi-task scenarios becomes difficult. A natural question then is: Do we actually need to synthesize paired image and multi-annotation data to be useful for multi-task visual learning?

Our MGM addresses this question by proposing a framework that uses synthesized images paired with *only weak annotations* (*i.e.*, image-level scene labels) to facilitate multiple visual tasks. Our key insight is to introduce *auxiliary discriminative tasks* that (i) only require image-level annotation or no annotation, and (ii) correlate with the original multiple tasks of interest. To this end, as additional components of the discriminative multi-task network, we introduce a *refinement* network and a *self-supervision* network that satisfy these properties. Through joint training, the discriminative network (the main multi-task network together with the refinement and self-supervision networks) *explicitly* guides the image synthesis process. The generative network also contributes to further refining the shared feature representation. Meanwhile, the synthesized images of the generative network are used as additional training data for the discriminative network.

In more details, the refinement network performs scene classification on the basis of the multi-task network predictions, which requires only scene labels for images, thus decoupling the use of synthesized images from ties of pixel-wise annotations. In addition, the self-supervision network can be operationalized on both real and synthesized images without reliance on annotations. With these two modules, our MGM framework learns a more effective representation that benefits multiple tasks, from both (pixel-wise) fully-annotated real images and synthesized (image-level) weakly labeled images. We instantiate MGM with the state-of-the-art encoder-decoder based multi-task network [58], self-attention GAN [60], and contrastive learning based self-supervision network [7]. Note that our framework is agnostic to the choice of these model components.

To validate our approach, we evaluate on standard multi-task benchmarks, including the NYUv2 dataset [32] and Taskonomy dataset [58]. Consistent with the previous work [53, 52], we focus on three tasks of great practical importance: semantic segmentation, depth estimation, and normal prediction. Our evaluation shows a number of interesting results. (1) Our MGM consistently outperforms both single-task and state-of-the-art multi-task approaches, *even without the contribution of the generative modeling component*. This suggests that MGM is an improved framework for multi-task learning in general. (2) When *pre-training* the generative network and using its synthesized images as additional training data for the multi-task network under our MGM framework, the performance of all the tasks gets improved. (3) *Jointly training* of the generative and multi-task networks under our MGM framework further improves the performance of all the tasks by large margins, almost reaching the *performance upper-bound* that trains with weakly annotated *real* images. (4) Finally, we demonstrate the scalability of our approach to more visual tasks.

2. Related Work

Multi-task Learning and Task Relationship: Multi-task learning (MTL) aims to leverage information coming from signals of related tasks, so that each individual task can gain benefit [14]. The authors of [45] identify that most recent works use two clusters of strategies for MTL: hard parameter sharing techniques [26, 14, 4, 41] and soft parameter sharing techniques [30, 47, 8]. These strategies have achieved a good performance for MTL with similar tasks. However, more useful but challenging explorations lie in the area of MTL with different tasks. Recent works have also carefully studied the task relationships among different tasks to make best cooperations among them. *Taskonomy* exploits the relationships among various visual tasks to benefit the transfer or multi-task learning [58]. [36] proposes a meta-learning algorithm to adapt existing models to novel zero-shot learning tasks. Task cooperation and competition are also considered, and a method is proposed for assigning tasks to a few neural networks to best balance all the tasks in [52]. Some other following works have also explored task relationship among different types of tasks [53, 57, 2]. One common point for these papers is that they are consider MTL with discriminative tasks only, but in this paper, we first introduce a generative model to multi-task visual learning.

Generative Modeling for Visual Learning: Besides the initial goal of synthesizing realistic images, some recent work has explored the potential to leverage generative models to synthesize "usefull" images for other visual tasks [49]. The most straight-forward way is to generate images and the corresponding annotations as data augmentation for the target visual task [3, 46, 9]. [54] also proposes to generate imaginary latent features rather than true images to better benefit the low-shot classification task. Another strategy to leverage generative models is through well-designed error feedback or adversarial training [29, 11, 31]. There have been a lot of works that apply generative models for different kinds of visual tasks including classification [59, 10, 20, 61], semantic segmentation [51, 29] and depth estimation [42, 1]. However, these methods are limited to a single specific task and have a relatively low generalization capability for more tasks. In comparison, the proposed MGM is applicable to various multiple visual tasks and different generative networks.

Weakly-supervised Learning: Due to the lack of strong

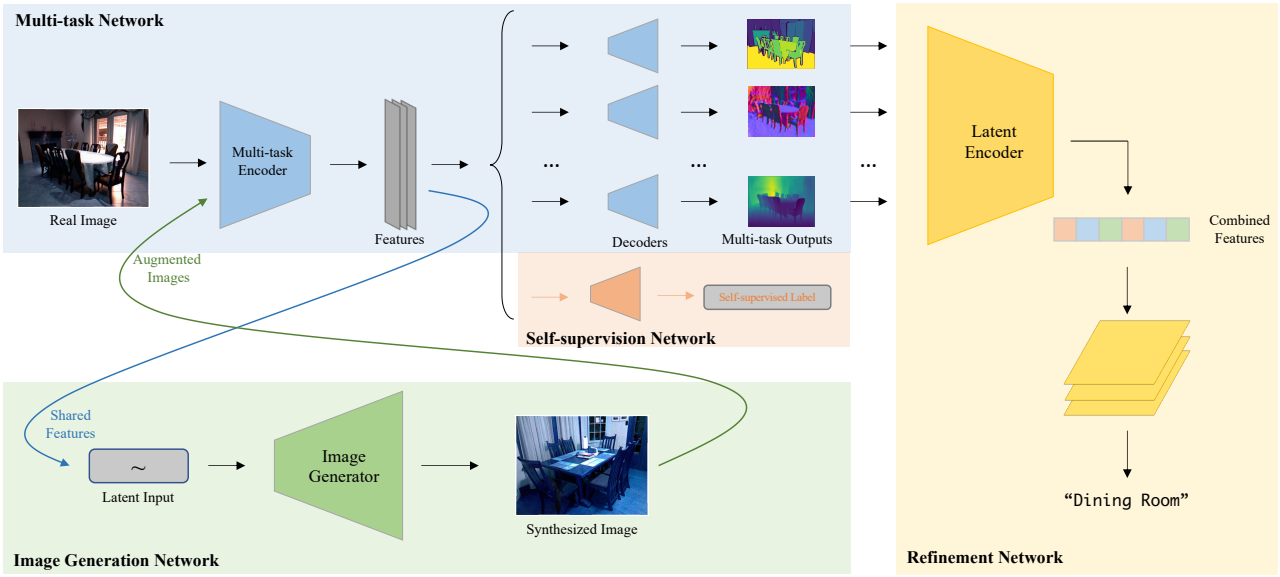

Figure 2: Architecture of our proposed multi-task oriented generative modeling (MGM) framework. There are four main components in the framework: Multi-task network to address the target multiple pixel-level prediction tasks; self-supervision network to facilitate representation learning using images without any annotation; refinement network to perform scene classification using weak annotation; image generation network to synthesize useful images that benefit multiple tasks.

pixel-level annotations of the synthesized data, weakly-supervised learning is required for the proposed MGM framework. Recent works take advantage of weakly labeled data by assigning some self-created labels (*e.g.* colorization, rotation, reconstruction) [34, 35, 7, 15, 38]. Similar self-supervised techniques have been proved useful for multi-task learning [28, 44, 14, 27]. Among these techniques, a famous one is the *Expectation-Maximization (EM)* algorithm [13], which leverages the information of weakly or unlabelled data by iteratively estimating and refining their labels. [37] further applies *EM* algorithm for semi-supervised semantic segmentation. In this work, we adapt a similar spirit and introduce a *EM*-based refinement network for MGM framework.

3. Method

In this section, we first describe the problem setting and general framework of our proposed multi-task oriented generative modeling (MGM), as summarized in Figure 2. We then explain an instantiation of the MGM model with state-of-the-art multi-task learning and image generation approaches. Finally, we discuss the detailed training strategy for the framework.

3.1. Problem Setting

Multi-task discriminative learning: Given a set of n visual tasks $\mathcal{T} = \{T_1, T_2, \cdots, T_n\}$, we aim to learn a discriminative model $\mathbf{M}$ that is able to address all of these tasks simultaneously: $\mathbf{M}(x) \rightarrow \widehat{\boldsymbol{y}} = (\widehat{y}^1, \widehat{y}^2, \cdots, \widehat{y}^n)$, where x is an input image and $\widehat{y}^i$ is the prediction for task T_i. Here we focus on the type of per-pixel level prediction tasks (*e.g.*, semantic segmentation or depth estimation). We treat image classification as a special task, which provides global semantic description (*i.e.*, scene labels) of images and only requires image-level category annotation c. Therefore, the set of fully annotated real data is denoted as $\mathcal{S}_{\text{real}} = \{(x_j, y_j^1, y_j^2, \cdots, y_j^n, c_j)\}$.

Generative learning: Meanwhile, we aim to learn a generative model $\mathbf{G}$ that produces a set of synthesized data but with only corresponding image-level scene labels (weak annotation): $\mathbf{G}(c, z) \rightarrow \widetilde{x}$, where z is a random input, and $\widetilde{x}$ is a synthesized image. The scene label of $\widetilde{x}$ is denoted as $\widetilde{c}$, and it satisfies that $\widetilde{c} = c$. We denote the set of synthesized images and their corresponding scene labels as $\mathcal{S}_{\text{syn}} = \{(\widetilde{x}_k, \widetilde{c}_k)\}$.

Cooperation between discriminative and generative learning: Our objective is that the discriminative model $\mathbf{M}$ and the generative model $\mathbf{G}$ cooperate with each other to improve the performance on the multiple visual tasks $\mathcal{T}$. So $\mathbf{M}$ effectively learns from both $\mathcal{S}_{\text{real}}$ and $\mathcal{S}_{\text{syn}}$, outperforming learning from only given $\mathcal{S}_{\text{real}}$. Note that different from most of existing work on image generation [6, 60], *here we do not focus on the visual realism of the synthesized images* $\widetilde{x}$. Instead, we hope $\mathbf{G}$ to capture the underlying image generation mechanism that benefits $\mathbf{M}$.

3.2. Framework and Architecture

Figure 2 shows the overall architecture of our proposed MGM framework with the discriminative model $\mathbf{M}$ and the generative model $\mathbf{G}$. The discriminative model $\mathbf{M}$ consists of three modules: (1) a main multi-task network for the tar-

get tasks, which is used in conventional multi-task learning, (2) a self-supervision network and (3) a refinement network that are able to leverage both real and synthesized images to facilitate the learning of latent feature representation, *without reliance on annotation or with only image-level weak annotation*. The generative model $\mathbf{G}$ is a variant of generative adversarial networks (GANs). Notice that our MGM is a *model-agnostic* framework, and here we instantiate its components with state-of-the-art models. In the ablation study (Sec. 4.3 and also in the supplementary material), we show that our framework works well with different choices of the model components.

Multi-task Network: The multi-task network takes as input an image and makes predictions for the multiple target tasks. Consistent with the most recent work on multi-task visual learning, we adopt an encoder-decoder based architecture [58, 60, 53]. Considering the trade-off between model complexity and performance, we use a shared encoder $\mathbf{E}$ to extract features from input images, and use individual decoders for each target task. We use a ResNet-18 [24] for the encoder and symmetric transposed decoders following [58]. For each task, we have its own loss function, which is used to update the corresponding decoder and the shared encoder. In principle, using a more sophisticated and powerful architecture such as [53, 52] will further improve the performance, and we leave this as future work.

Self-supervision Network: As a complementary branch in parallel with the multi-task decoders, the self-supervision network facilitates the representation learning of the encoder $\mathbf{E}$ by performing self-supervised learning tasks on images without any annotation. It can be thus operationalized on both real and synthesized images. Among the various types of self-supervised tasks [34, 35, 7, 38], we adopt an instance-level discrimination task [56, 23, 7] for its simplicity and effectiveness. We modify SimCLR [7], one of the state-of-the-art approaches to addressing this task, as our self-supervision network. This network contains an embedding network $\mathbf{E}_{\text{self}}$, working on top of the output of the multi-task encoder $\mathbf{E}$, to obtain a 1D latent feature of the input image: $\mu = \mathbf{E}_{\text{self}}(\mathbf{E}(x))$. Then, it performs contrastive learning with these latent vectors.

Specifically, given a minibatch of N images, this network first randomly sampled 2 transformed views of source images as augmented images (See supplementary material for the detailed random transformations), resulting in $2N$ augmented images in total. For each augmented image, we could find one pair of positive augmented examples from the same source image, and other $2(N-1)$ negative augmented pairs. Then the self-supervision network embeds the augmented images into latent vectors, and jointly minimizes the distance of positive pairs and maximizes the distance of negative pairs, through the normalized temperature-scaled cross-entropy (*NT-Xent*)) loss [7]:

$$\ell_{i,j} = -\log \frac{\exp\left(\text{dis}\left(\mu_i, \mu_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\text{dis}\left(\mu_i, \mu_k\right)/\tau\right)}, \quad (1)$$

where $\ell_{i,j}$ is the *NT-Xent* loss for a positive pair of examples in the latent space (μ_i, μ_j). $\mathbb{1}_{[k\neq i]} \in 0, 1$ is an indicator function evaluating to 1 iff $k \neq i$, and τ is a temperature parameter. $\text{dis}\left(\mu_i, \mu_j\right)$ is a distance function, and we use cosine distance following [7]. This loss is further back-propagated to refine the shared encoder of the multi-task network. Notice that other types of self-supervised tasks are applicable here as well. To demonstrate this, in Sec. 4.3 we report the result with another task — image reconstruction.

Refinement Network: By introducing an auxiliary classification task, the refinement network $\mathbf{R}$ is designed to further refine the shared representation using the global scene category labels. This network takes the predictions of the multi-task network as input, and predicts the category label of the input image. Importantly, because $\mathbf{R}$ only requires category labels, it can be effortlessly operationalized on the "weakly annotated" synthesized images to improve the performance of the multi-task network. Meanwhile, $\mathbf{R}$ also works to enforce the semantic consistency of the synthesized images with the image generation network.

Inspired by [37], we apply an Expectation-Maximum (EM) like algorithm to train the refinement network. We use cross-entropy as the classification loss function. We model the whole multi-task network and refinement network as a joint probability graph:

$$P(x, \boldsymbol{y}, c; \theta, \theta') = P(x)\left(\prod_{i=1}^{n} P\left(y^i \mid x; \theta\right)\right) P(c \mid \boldsymbol{y}; \theta'), \quad (2)$$

where x is an input image, $\boldsymbol{y}$ is the vector of multi-task predictions, c is the scene label, θ is the vector of parameters of the multi-task network, and θ' is the vector of parameters of the refinement network. The parameters θ and θ' are learned to maximize the joint probability. For data samples in $\mathcal{S}_{\text{real}}$, we maximize the joint probability and update θ and θ' directly with ground-truth. For data samples in $\mathcal{S}_{\text{syn}}$, we update θ in an EM-like manner. During the $\mathbf{E}$ step, we estimate the latent multi-task ground-truth by:

$$\boldsymbol{y}^{\dagger} = \underset{\boldsymbol{y}}{\text{argmax}}\, P\left(\boldsymbol{y} \mid \widetilde{x}_k; \theta\right) P(\widetilde{c}_k \mid \boldsymbol{y}; \theta'), \quad (3)$$

Then for the $\mathbf{M}$ step, we back-propagate the error between $\boldsymbol{y}^{\dagger}$ and $\widehat{\boldsymbol{y}}$ (the predicted multi-task result) to the multi-task encoder.

Image Generation Network: The image generation network $\mathbf{G}$ takes as input a latent vector z and a category label c, and synthesizes an image belonging to category c. Considering the trade-off between performance and training cost, we use self-attention generative adversarial network (SAGAN) as our $\mathbf{G}$. We achieve conditional generation by applying conditional batch normalization layers [12]:

$$\text{CBN}\left(f_{i,c,h,w} \mid \gamma_c, \beta_c\right) = \gamma_c \frac{f_{i,c,w,h} - \text{E}\left[f_{\cdot,c,\cdot,\cdot}\right]}{\sqrt{\text{Var}\left[f_{\cdot,c,\cdot,\cdot}\right] + \epsilon}} + \beta_c, \quad (4)$$

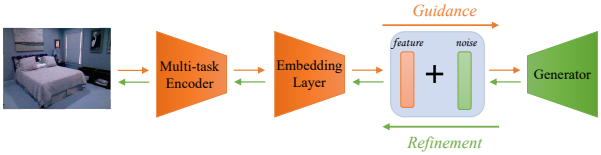

Figure 3: Joint training of the multi-task network and the image generation network. The multi-task network provides useful feature representation to guide the image generation process, while the generation network refines the shared representation through back-propagation.

where $f_{i,c,h,w}$ is an extracted c-channel 2D feature for the i-th sample, and ϵ is a small value to avoid collapse. γ_c and β_c are two parameters to control the mean and variance of the normalization, which are learned by the model for each class. We use hinge loss for the adversarial training. Notice that the proposed framework is flexible with different generative models, and we also show the effectiveness of using DCGAN [43] in the supplementary material.

3.3. Interaction Among the Networks

Cooperation Through Joint Training: We propose a simple but effective joint training algorithm shown in Figure 3. The image generation network $\mathbf{G}$ takes the transferred feature representation of the multi-task encoder $\mathbf{E}$, added with some Gaussian noise, as the latent input z to conduct conditional image generation. Hence, the generation network obtains *additional, explicit guidance* (*i.e.*, extra effective features) from the multi-task network to facilitate the generation of "better images"— images that may not look more realistic but are more useful for the multiple target tasks. Then, the generation error of $\mathbf{G}$ will be back-propagated to the multi-task encoder $\mathbf{E}$ to further refine the shared representation. This process can be also viewed as introducing image generation as an additional task in the multi-task learning framework. By doing so, the learned shared representation will be more robust and effective.

Training Procedure: Given a minibatch of data in $\mathcal{S}_{\text{real}}$, we conduct the following training procedure. (1) For the input images x, we predict $\hat{\boldsymbol{y}} = \mathbf{M}(x)$, and then use the task-specific losses between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ to update the multi-task network $\mathbf{M}$. (2) We predict the scene labels by $\hat{c} = \mathbf{R}(\hat{\boldsymbol{y}})$, and update the refinement network $\mathbf{R}$ using the cross-entropy loss between c and $\hat{c}$. (3) We randomly sample pairs of augmented images, process them with the self-supervision network, and then update the self-supervision network and the multi-task encoder $\mathbf{E}$ with the *NT-Xent* loss in Eq. (1). (4) We train the image generation network $\mathbf{G}$ through adversarial training with (x, c), and back-propagate the adversarial error and update $\mathbf{E}$ at the same time. (5) We sample another minibatch of synthesized data $(\widetilde{x}, \widetilde{c})$, and use these data to update $\mathbf{E}$ by performing both the EM-like algorithm with $\mathbf{R}$ in Eq. (3) and the self-supervised learning as in step (3).

4. Experiments

To evaluate our proposed MGM model and investigate the impact of each component, we conduct a variety of experiments on two standard multi-task learning datasets. We also perform detailed analysis and ablation studies here.

4.1. Datasets and Compared Methods

Datasets: Following the work of [53, 52], we mainly focus on three representative visual tasks in the main experiments: semantic segmentation (SS), surface normal prediction (SN), and depth estimation (DE). At the end of this section, we will show that our approach is scalable to additional number of tasks. We evaluate all the compared models on two widely-benchmarked datasets: **NYUv2** [32, 18], which contains 1,449 images with 40 types of objects [22]; **Tiny-Taskonomy**, which is the standard tiny split[1] of the Taskonomy dataset [58]. Since a certain amount of images for each category is required to train a generative network, we keep the images of the top 35 scene categories on Tiny-Taskonomy, with each one consisting of no more than 1,000 images. This resulting dataset contains 358,426 images in total. For NYUv2, we randomly select 1,200 images as the full training set and the others as the test set. For Tiny-taskonomy, we randomly pick 85% of the whole set as the full training set and the other 15% as the test set.

Compared Methods: We mainly focus on our comparison with two state-of-the-art discriminative baselines: **Single-Task (ST)** model follows the architecture of Taskonomy single task network [58], and address each task individually; **Multi-Task (MT)** model refers to the sub-network for the three tasks of interest in [52]. These two baselines can be viewed as using our multi-task network without the proposed refinement, self-supervision, and generation networks. Note that *our work is the first that introduces generative modeling for multi-task learning, and there is no existing baseline in this direction*.

Our **MGM** is the full model trained with both fully-labeled *real* data and weakly labeled *synthesized* data, which are produced by the generation network through joint training. In addition, to further validate the effectiveness of our **MGM** model, we consider its variant model $\mathbf{MGM}_{\text{r}}$ that is trained with both fully and weakly labeled *real* data. $\mathbf{MGM}_{\text{r}}$ is used to show *the performance upper bound* in the semi-supervised learning scenario, where the synthesized images are replaced by the real images in the dataset. The resolution is set to 128 for all the experiments, unless specifically mentioned. For all the compared methods, we use a ResNet-18 like architecture to build the encoder and use standard decoder architecture of Taskonomy [58].

Data Setting: We conduct experiments with three different data settings: (1) 100% data setting; (2) 50% data setting; and (3) 25% data setting. For each setting, we use 100%, 50%, or 25% of the entire labeled training set to train

[1] See https://github.com/StanfordVL/taskonomy/blob/master/data/README.md

	Data Setting	100% Data Setting			50% Data Setting				25% Data Setting			
	Models	ST	MT	MGM	ST	MT	MGM	*MGM$_r$*	ST	MT	MGM	*MGM$_r$*
NYU v2	SS-mIOU(↑)	0.253	0.257	**0.264**	0.233	0.237	**0.251**	*0.258*	0.201	0.209	**0.229**	*0.233*
	DE-mABSE(↓)	0.757	0.704	**0.698**	0.835	0.815	**0.734**	*0.720*	0.911	0.876	**0.844**	*0.820*
	SN-mAD(↓)	0.273	0.283	**0.255**	0.309	0.291	**0.273**	*0.270*	0.312	0.296	**0.277**	*0.274*
Tiny Taskonomy	SS-mLoss(↓)	0.111	0.137	**0.106**	0.120	0.138	**0.114**	*0.112*	0.119	0.141	**0.117**	*0.115*
	DE-mLoss(↓)	1.716	1.584	**1.472**	1.768	1.595	**1.499**	*1.378*	1.795	1.692	**1.585**	*1.580*
	SN-mLoss(↓)	0.155	0.153	**0.145**	0.157	0.156	**0.147**	*0.140*	0.154	0.153	**0.148**	*0.142*

Table 1: Main results on the NYUv2 and Tiny-Taskonomy datasets. SS: semantic segmentation; DE: depth estimation; SN: surface normal prediction. ↑ means higher is better; ↓ means lower is better. We use different metrics on the two datasets, following existing protocol. Our MGM consistently and significantly outperforms both single-task (ST) and multi-task (MT) baselines, *even reaching the performance upper-bound of training with weakly annotated real images* (MGM$_r$).

	Data Setting	100% Data Setting			50% Data Setting			25% Data Setting		
	Models	MGM$_{/G}$	MGM$_{/j}$	MGM	MGM$_{/G}$	MGM$_{/j}$	MGM	MGM$_{/G}$	MGM$_{/j}$	MGM
NYU v2	SS-mIOU(↑)	0.261	0.262	0.264	0.243	0.243	0.251	0.215	0.220	0.229
	DE-mABSE(↓)	0.707	0.701	0.698	0.799	0.763	0.734	0.868	0.860	0.844
	SN-mAD(↓)	0.262	0.259	0.255	0.287	0.281	0.273	0.292	0.286	0.277
Tiny Taskonomy	SS-mLoss(↓)	0.108	0.108	0.106	0.116	0.115	0.114	0.119	0.121	0.117
	DE-mLoss(↓)	1.491	1.488	1.472	1.527	1.523	1.499	1.636	1.616	1.585
	SN-mLoss(↓)	0.151	0.151	0.145	0.153	0.152	0.147	0.154	0.152	0.148

Table 2: Comparison of our MGM model with its variants. MGM$_{/G}$: *without* generating synthesized images; MGM$_{/j}$: *without* joint learning. Our MGM outperforms single-task and multi-task baselines *even without synthesized data*, showing its effectiveness as a general multi-task learning framework. The model performance further improves with joint learning.

the model. For **MGM$_r$**, we add another 50% or 25% of weakly labeled real data in the last two settings. For **MGM**, we include the same number of weakly labeled synthesized data in all the three settings.

Evaluation Metrics: For NYUv2, following the metrics in [18, 53], we measure the mean Intersection-Over-Union (mIOU) for the semantic segmentation task, the mean Absolute Error (mABSE) for the depth estimation task, and the mean Angular Distance (mAD) for the surface normal estimation task. For Tiny-taskonomy, we follow the evaluation metrics of previous work [58, 52, 53] and report the averaged loss values on the test set.

Implementation Details: We use Adam [25] optimizer for all the models. The learning rates are set to 0.001 for the multi-task, self-supervision, and refinement networks, 0.0001 for the SAGAN generator, and 0.0004 for the SAGAN discriminator. The batch size is set to 32. We use a cross-entropy loss for semantic segmentation and the scene classification task of the refinement network, and l_1 loss for surface normal and depth estimation. See the supplementary material for the detailed architecture of the entire model and the other implementation details.

4.2. Main Results

Quantitative Results: We report the main results on the two datasets in Table 1. From this table, we have the following key observations that support the effectiveness of our approach which combines generative learning with discriminative leanring. (1) Existing discriminative multi-task learning approaches may not consistently benefit all the three individual tasks. However, our MGM consistently and significantly outperforms both the single-task and multi-task baselines across all the scenarios. (2) By using weakly labeled synthesized data, the results of our model in the 50% data setting are even better than those of baselines in the 100% data setting. (3) More interesting, the performance of our MGM is close to MGM$_r$, which indicates that our synthesized images are *comparably useful* as real images for improving multiple visual perception tasks. The performance gap is especially minimal in the 25% labeled data setting, suggesting that our proposed MGM model is in particular helpful for low-data regime.

Qualitative Results: We also visualize the prediction results on the three tasks for ST, MT, and MGM in the 50% data setting in Figure 4. While obvious defects can be found for all the baselines, the results of our proposed method are quite close to the ground-truth.

How does Generative Modeling Benefit Multi-tasks? To have a better understanding of how the generative modeling and joint learning mechanism benefit multi-task visual learning, we also consider two variants of our MGM model and evaluate their performance. **MGM$_{/G}$** is the MGM model trained with $\mathcal{S}_{\text{real}}$ only (without generative modeling), which shows the performance of our proposed multi-task learning framework in general (with the help from the auxiliary refinement and self-supervision networks), and

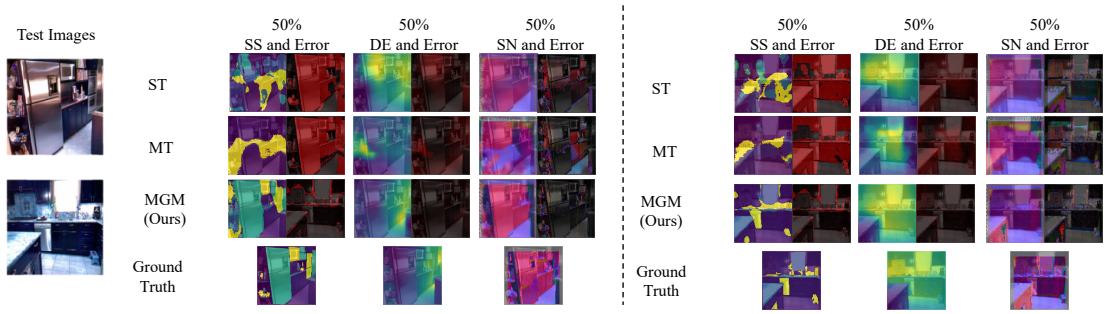

Figure 4: Visualization and error comparison of the multi-task prediction outputs in the 50% data setting. The prediction results of MGM is quite close to the ground-truth, significantly outperforming the state-of-the-art results.

helps to understand the gain of leveraging generative modeling. **MGM**$_{/\mathrm{j}}$ is trained with synthesized images produced by a pre-trained SAGAN *without* the joint training mechanism. Table 2 shows the results on the two datasets.

Combining the results of Tables 2 and 1, we find the following observations. (1) The proposed multi-task framework outperforms both single-task model and previous multi-task baseline even without generative modeling, indicating the benefit of the self-supervised task and refinement network. (2) By introducing synthesized images that are trained separately, the multi-task performance slightly improves, which shows the effectiveness of involving generative modeling into multi-task discriminative learning; but this is not enough. (3) The joint learning mechanism further improves the cooperation between generative modeling and discriminative learning, thus enabling the generative model to better facilitate multi-task visual learning. Hence, when we introduce additional synthesized images to the original labeled dataset, the performance of all the tasks consistently gets further improved.

4.3. Ablation Study

For all the experiments in this section, models are trained in the 50% data setting, unless specifically mentioned.

Impact of Self-supervision Task and Scene Classification Refinement Network: Two important components of the proposed framework are the self-supervision task and the refinement network. We evaluate their impact individually in Table 3. For the compared methods, MGM$_{/\mathrm{self}}$ is the model trained *without* the self-supervision task; MGM$_{/\mathrm{refine}}$ is the model *without* the refinement network; for MGM$_{\mathrm{recon}}$, we replace the SimCLR based self-supervision method with a weaker reconstruction task, and use Mean Square Error to update the network.

We could see that the refinement network works better for the surface normal task, and the self-supervision task works better for the semantic segmentation task; they are complementary to each other, and combining them achieves the best performance. Besides, the model could still gain some benefit even when we use some weak self-supervision task like reconstruction, which indicates the generability and robustness of our MGM model.

Model	SS-mIOU (↑)	DE-mABSE (↓)	SN-mAD (↓)
MT	0.237	0.815	0.291
MGM$_{/\mathrm{self}}$	0.239	0.776	0.279
MGM$_{/\mathrm{refine}}$	**0.254**	0.808	0.290
MGM$_{\mathrm{recon}}$	0.241	0.768	0.285
MGM	0.251	**0.734**	**0.273**

Table 3: Ablation study: (1) *without* self-supervision task, (2) *without* classification refinement network, and (3) *with* a simple reconstruction task as self-supervision. The two components are complementary and both benefit the multiple tasks. The refinement network works better for surface normal; the self-supervision network works better for semantic segmentation. Their combination achieves the best.

Number of Synthesized Images vs. Real images: From the previous results, we have found that the synthesized images could benefit the target multi-tasks in a way similar to weakly labeled real images. To further investigate the impact of the number of synthesized images, we vary if from 25% to 125% during multi-task training on NYUv2 in the 25% real data setting. Figure 5 summarizes the result. First, we can see that the performance gap between MGM$_{/\mathrm{j}}$ (without joint training) and MGM becomes larger for a higher ratio of weakly labeled data, which indicates the importance of our joint learning mechanism. *More importantly*, while the real images are constrained in number due to the human collection effort, our generation network is able to synthesize *unlimited* amounts of images. This is demonstrated in the comparison between MGM$_{\mathrm{r}}$ (with real images) and MGM: the performance of our MGM keeps improving with respect to the number of synthesized images, achieving results almost comparable to that of MGM$_{\mathrm{r}}$ when MGM$_{\mathrm{r}}$ uses all the available weakly labeled real images.

Generalization of the Shared Feature Representation: Intuitively, our MGM achieves the state-of-the-art performance by effectively learning a shared feature representation. We further show the generalization capability of this representation by designing the following experiment: for the multi-task model and our MGM model, we first learn the shared feature space with the SS and DE tasks, and we

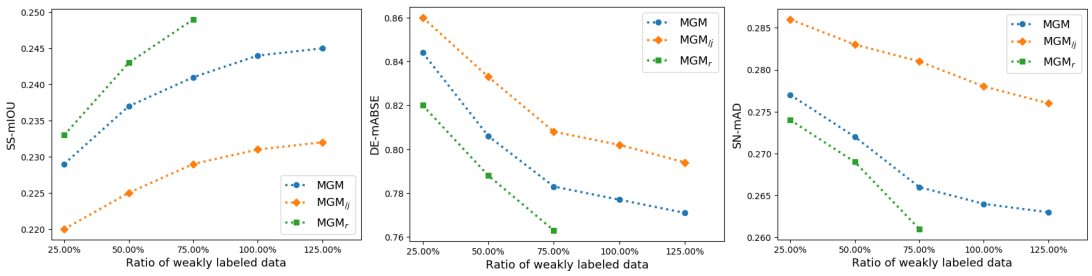

Figure 5: Performance change with different ratios of weakly labeled data. Joint learning significantly improves the performance. The performance of MGM keeps increasing with the number of the weakly labeled *synthesized* images, achieving results almost comparable to that of MGM$_r$ trained with all the available weakly labeled *real* images.

Model	mAD-100%(↓)	mAD-50%(↓)	mAD-25%(↓)
MT	0.291	0.310	0.323
MGM	**0.280**	**29.83**	**0.305**

Table 4: Results for the SN task with pre-trained feature representations by the SS and DE tasks. MGM consistently outperforms multi-task (MT), indicating that MGM learns a more effective and generalizable feature representation.

Model	SS-mIOU (↑)	DE-mABSE (↓)	SN-mAD (↓)
ST	0.239	0.849	0.282
MT	0.244	0.834	0.313
MGM	**0.257**	**0.819**	**0.275**

Table 5: Experiments with 256 image resolution on the NYUv2 dataset. Our MGM still consistently outperforms the compared baselines, showing the great robustness and flexibility of the proposed framework.

Model	SS (↓)	DE (↓)	SN (↓)	ET (↓)	Re (↓)	PC (↓)
ST	0.120	1.768	0.157	0.228	0.703	0.462
MT	0.112	1.747	0.169	0.241	0.704	0.436
MGM	**0.108**	**1.715**	**0.152**	**0.201**	**0.699**	**0.417**

Table 6: Mean test losses for six tasks on Tiny-Taskonomy. Again, our MGM outperforms the baselines, indicating its flexibility, generalizability, and scalability.

then use that learned feature space to train a new decoder for the SN task. We report the results on NYUv2 in Table 4. Our MGM outperforms the multi-task model in all the three data settings, which means that MGM indeed learns a better and robust shared feature space.

4.4. Extension

Experiments on Higher Image Resolution: In principle, the proposed framework is agnostic to the specific types of multi-task networks and image generation networks, thus flexible with image resolutions. The practical constraint lies in that it is still challenging and resource-consuming for modern generative models to synthesize very high-resolution images [60, 6], although the deep multi-task models normally work better with high-resolution images. In the main experiments, consistent with exiting image synthesis work [60], we focused on the resolution of 128×128. Here we further made an attempt to run our experiments with a higher resolution, 256×256 on NYUv2. The results of all the compared models in the 50% data setting are shown in Table 5. We could find that MGM still consistently outperforms the baselines, indicating the great robustness and flexibility of our proposed framework.

Experiments with More Tasks: The proposed framework is also flexible and scalable with different tasks. In addition to the three tasks addressed in the main experiments, here we add three extra tasks: Edge Texture (ET), Reshading (Re), and Principal Curvature (PC), leading to six tasks in total. We evaluate the performance of all the compared models on Tiny-taskonomy in the 50% data setting, and report the mean test loss for all the tasks. The result is reported in Table 6. Again, our proposed method still outperforms state-of-the-art baselines.

5. Conclusion

Motivated by multi-task learning of shareable feature representations, this paper proposes to introduce generative modeling for multi-task visual learning. The main challenge is that it is hard for generative models to synthesize both RGB images and pixel-level annotations in multi-task scenarios. We address this problem by proposing multi-task oriented generative modeling (MGM) framework equipped with the self-supervision network and the refinement network, which enable us to take advantage of synthesized images paired with image-level scene labels to facilitate multiple visual tasks. Experimental results indicate our MGM model consistently outperforms state-of-the-art multi-task approaches.

References

[1] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *Proceedings of the Euro-*

pean Conference on Computer Vision (ECCV) Workshops, September 2018. 2

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5664–5673, 2019. 2

[3] Zhipeng Bao, Yu-Xiong Wang, and Martial Hebert. Bowtie networks: Generative modeling for joint few-shot recognition and novel-view synthesis. *arXiv preprint arXiv:2008.06981*, 2020. 2

[4] Hakan Bilen and Andrea Vedaldi. Integrated perception with recurrent multi-task neural networks. In *Advances in neural information processing systems*, pages 235–243, 2016. 2

[5] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 3, 8

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 3, 4

[8] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR, 2018. 2

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 6830–6840, 2019. 2

[10] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Advances in neural information processing systems*, pages 4088–4098, 2017. 2

[11] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Monocular depth prediction using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 300–308, 2018. 2

[12] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017. 4

[13] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 3

[14] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 2, 3

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014. 3

[16] Kieran Egan. Memory, imagination, and learning: Connected by the story. *Phi Delta Kappan*, 70(6):455–459, 1989. 1

[17] Kieran Egan. *Imagination in teaching and learning: The middle school years*. University of Chicago Press, 2014. 1

[18] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 5, 6

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1

[20] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[22] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 564–571, 2013. 5

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017. 2

[27] Wonhee Lee, Joonil Na, and Gunhee Kim. Multi-task self-supervised object detection via recycling of bounding box annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4984–4993, 2019. 3

[28] Qiuhua Liu, Xuejun Liao, and Lawrence Carin. Semi-supervised multitask learning. In *Advances in Neural Information Processing Systems*, pages 937–944, 2008. 3

[29] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. 2

[30] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 2

[31] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3981, 2020. 2

[32] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5

[33] Thu H Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yongliang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *NeurIPS*, 2018. 1

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3, 4

[35] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 3, 4

[36] Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *CVPR*, 2019. 2

[37] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 3, 4

[38] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3, 4

[39] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634, 2019. 1

[40] Etienne Pelaprat and Michael Cole. "minding the gap": Imagination, creativity and human cognition. *Integrative Psychological and Behavioral Science*, 45(4):397–418, 2011. 1

[41] Anastasia Pentina and Christoph H Lampert. Multi-task learning with labeled and unlabeled tasks. In *International Conference on Machine Learning*, pages 2807–2816. PMLR, 2017. 2

[42] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2018. 2

[43] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5

[44] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018. 3

[45] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2

[46] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019. 2

[47] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018. 2

[48] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2018. 1

[49] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 2

[50] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 1

[51] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. 1, 2

[52] Trevor Standley, Amir R Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, 2020. 2, 4, 5, 6

[53] Ximeng Sun, Rameswar Panda, and Rogerio Feris. Adashare: Learning what to share for efficient deep multi-task learning. *arXiv preprint arXiv:1911.12423*, 2019. 2, 4, 5, 6

[54] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 2

[55] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016. 1

[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 4

[57] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 2

[58] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1, 2, 4, 5, 6

[59] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 2

[60] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 1, 2, 3, 4, 8

[61] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 349–360. Springer, 2018. 1, 2