# Uncertainty Guided Pseudo-Labeling: Estimating Uncertainty on Ambiguous Data for Escalating Image Recognition Performance

Kyung Ho Park* and HyunHee Chung*
*SOCAR, Republic of Korea*

Keywords: Label Ambiguity, Semi-Supervised Learning, Pseudo-Labeling, Uncertainty Estimation.

Abstract: Upon the dominant accomplishments of deep neural networks, recent studies have scrutinized a robust model under the inherently ambiguous samples. Prior works have achieved superior performance under these ambiguous samples through label distribution approaches, assuming the existence of multiple human annotators. However, the aforementioned problem setting is not generally feasible due to resource constraints. For a generally applicable solution to the ambiguity problem, we propose Uncertainty-Guided Pseudo-Labeling (UGPL), a proof-of-concept level framework that leverages ambiguous samples on elevating the image recognition performance. Key contributions of our study are as follows. First, we constructed synthetic ambiguous datasets as there were no public benchmark dataset that deals with ambiguity problem. Given ambiguous samples, we empirically showed that not every ambiguous sample has meaningful knowledge consistent to the obvious samples at the target classes. We then examined uncertainty can be a possible proxy for measuring the effectiveness of ambiguous sample's knowledge toward the escalation of image recognition performance. Moreover, we validated pseudo-labeled ambiguous samples with low uncertainty better contributes to the test accuracy elevation. Lastly, we validated the UGPL showed larger accuracy elevation under the small size of obvious samples; thus, general practitioners can be widely benefited. To this end, we suggest further avenues of improvement practical techniques that resolve the ambiguity problem.



## 1 INTRODUCTION

Deep neural networks have achieved significant accomplishments in various computer vision applications such as image recognition (Nath et al., 2014) and object detection (Sukanya et al., 2016) under the large-scale annotated dataset. In image classification, especially, conventional problem settings assume that each image corresponds to a particular class. However, the aforementioned assumption is not always valid in the real world. There frequently exists some samples that are inherently ambiguous to be assigned to a particular class. In Figure 1, we illustrated several ambiguous images. Considering images shown in Figure 1 (a), conventional human annotators may feel confused to annotate given images between the Bagel and the Donut as they have inherently similar characteristics to both classes. It is also challenging to classify given images at Figure 1 (b) between the Frog and the Tadpole. Considering that multiple annotators should be involved to annotate large datasets, it is difficult for them to apply a consistent standard for samples with inherent ambiguity, leading to unreliable labels. We refer to the samples with clear discrimination as *obvious samples* and the samples with inherent ambiguity as *ambiguous samples*. Following the significance of the ambiguity problem, prior studies also tackled it down to construct deep neural networks with better representation power (Rupprecht et al., 2017; Otani et al., 2020; Gao et al., 2017; Hüllermeier and Beringer, 2006). In this paper, we seek to exploit them to improve classification performance in low data regimes.

Pursuing classification models robust under ambiguous samples, previous studies have proposed approaches with label distributions, assuming the circumstance where multiple human annotators exist (Geng, 2016; Gao et al., 2017) and each annotator assigns a label to a single sample, producing multiple weak labels. Although the aforementioned label distribution approaches contribute to the robustness of the image recognition model under ambiguity, it is generally not feasible to the practitioners due to the resource constraints. Instead, it is more practical for them to separate ambiguous samples into a distinct

*Denotes equal contribution



Park, K. and Chung, H.
Uncertainty Guided Pseudo-Labeling: Estimating Uncertainty on Ambiguous Data for Escalating Image Recognition Performance.
DOI: 10.5220/0010901600003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 2*, pages 541-551
ISBN: 978-989-758-547-0; ISSN: 2184-433X

(a) Bagel v. Donut (b) Frog v. Tadpole

Figure 1: Examples of ambiguous samples in the ImageNet dataset. Due to its inherent ambiguity, human labelers' subjectiveness is highly reflected during the annotation.

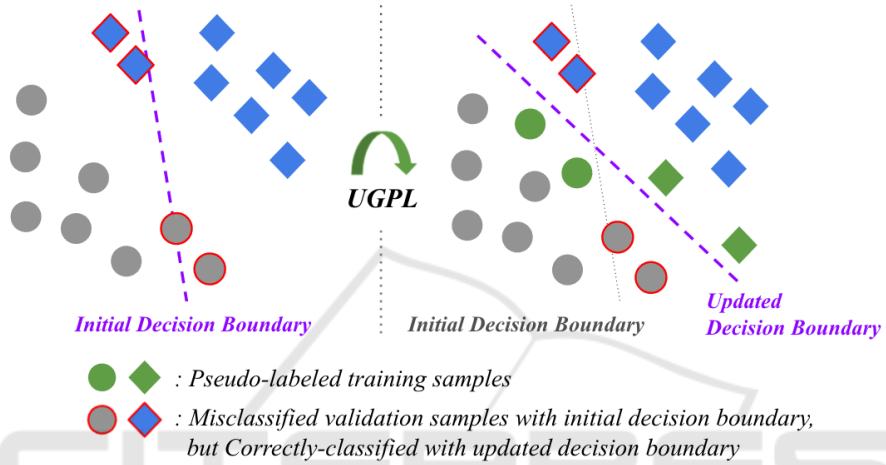



Figure 2: Simplied illustration of the proposed Uncertainty-Guided Pseudo-Labeling (UGPL). Referring to the figure at the left, an initial decision boundary (established with obvious samples only) could not classify particular validation samples precisely. The proposed UGPL acquires confident samples from the ambiguous dataset, makes pseudo-labels on them, and updates the decision boundary. As ambiguity at the embedding space is particularly unveiled, the updated decision boundary improves the misclassified results, as shown in the figure at the right.

class and simply training the model only on the obvious samples. To this end, we set a real-world circumstance where a single sample has a single label, and ambiguous samples are isolated into a separate class.

One very presumable solution is applying label propagation (Iscen et al., 2019), regarding the ambiguous samples as the samples without labels under the semi-supervised paradigm and producing pseudo-labels for them from the obvious samples. However, in our problem setting, it is not reasonable to produce pseudo-labels on every ambiguous samples due to the inherent ambiguity. Indeed, we showed that updating the training set with every pseudo-labeled ambiguous data does not contribute to the elevation of test accuracy. The pseudo-labeled ambiguous samples can escalate the classification performance if it includes knowledge consistent with the target classes. We empirically unveiled that not every ambiguous sample has consistent knowledge of the target classes. In other words, we inferred not every pseudo-labeled ambiguous samples support learning discriminative knowledge on the target classes; thus, there rises a necessity of effective pseudo-labeling designed for ambiguous samples.

To this end, we propose a **Uncertainty-Guided Pseudo-Labeling (UGPL)** to maximize the deep neural network's image recognition performance with ambiguous data. As shown in the Figure 2, our approach resolves the ambiguity problem by creating pseudo-labels on the ambiguous data and estimating uncertainty. An underlying belief of our study is that some pseudo-labels on ambiguous samples help improving the performance and others do not, and uncertainty of the pseudo-labels can be used for the distinction. Accordingly, we sort the ambiguous samples by their uncertainty in increasing order and include the samples with lowest uncertainty in the dataset to update the model. The UGPL consists of several steps. First, our approach trains an initial classifier with the obvious data to train base representations regarding the target classes. Second, we provide ambiguous data to the trained initial classifier and acquire uncertainty estimates. We expected the initial classifier to yield a low uncertainty on ambiguous samples if

it includes knowledge consistent to what the model learned from the obvious data. If the initial classifier infers a particular ambiguous sample with low uncertainty, we expect adding the ambiguous data would contribute to the representation learning. Thus, we sampled several ambiguous data with low uncertainty estimates and merged them with the initial obvious dataset. Lastly, we retrain the classifier with an updated training set and measured the performance elevation derived from the training set update.

In our study, key contributions are as follows.

- We defined a label ambiguity problem at the image classification problem, and figured out that not every ambiguous sample has consistent knowledge of the obvious samples. To this end, we designed UGPL, a proof-of-concept level framework to elevate the image classification performance leveraging ambiguous data based on estimated uncertainty.
- Leveraging well-known benchmark datasets (AFHQ (Choi et al., 2020) and CelebA-HQ (Karras et al., 2017)), we constructed two synthetic datasets named synthetic-AFHQ and synthetic-CelebA-HQ. Both synthetic datasets consist of obvious data and ambiguous data simultaneously. As there's no widely-utilized benchmark dataset that deals with ambiguity, we proposed how further researchers can create ambiguous data to examine their solution to the ambiguity problem. Specifically, we employed StyleMapGan (Kim et al., 2021) to create contextually ambiguous data given obvious images of given classes.
- We experimentally examined the UGPL framework contributes to the elevation of test accuracy. Under the synthetic-AFHQ and synthetic-CelebA-HQ datasets, our approach commonly showed the escalation of test accuracy rather than training the model only with obvious data. Based on the proof-of-concept level validation of our approach, we illustrate further room for improvement for the sake of robust image recognition research.

## 2 RELATED WORKS

### 2.1 Semi-Supervised Learning

Semi-supervised learning (SSL) is a paradigm that trains deep neural networks with both labeled and unlabeled data. The SSL aims to improve the learning performance by adding unlabeled data compared to the supervised learning approaches, which only utilize labeled data (Yang et al., 2021). There exist various studies on SSL such as consistency regularization methods (Zhou et al., 2020; Berthelot et al., 2019; Berthelot et al., 2019), graph-based approaches (Liu et al., 2019; Yang et al., 2016; Jiang et al., 2019), generative approaches (Denton et al., 2016; Odena, 2016; Rezagholiradeh and Haidar, 2018) and pseudo-labeling approaches (Arazo et al., 2020; Pham et al., 2021). Our study focused on the pseudo-labeling approaches as it is one of the presumable solutions to the ambiguity problem. The pseudo-labeling approaches create pseudo-labels on unlabeled data following the prediction of deep neural networks trained with labeled data. The recent pseudo-labeling methods add more training data by merging the labeled data and pseudo-labeled unlabeled data. Due to its simplicity and generality, numerous studies have actively scrutinized pseudo-labeling approaches on SSL.

A recent state-of-the-art study (Rizve et al., 2021) proposed an uncertainty-aware pseudo-labeling method that constructs pseudo-labels following the model prediction's confidence. The work trained a model with the labeled data and measured an uncertainty at unlabeled data to create both positive and negative pseudo-labels at each class. They dropped pseudo-labeled data if the model yielded low confidence on it, and iterated the aforementioned procedures until the number of selected pseudo-labels converges. They insisted conventional pseudo-labeling methods were ineffective due to the noises in pseudo-labels; thus, applying an uncertainty threshold could sanitize noises and elevated learning performance. While our study shares a common intuition with them, we tackle different problem: the unlabeled samples have inherent ambiguity which prevents assigning a particular label to them. The work created both positive and negative pseudo-labels simultaneously as they assumed an unlabeled sample at least has sufficient knowledge similar to one of the target classes. However, in the ambiguity problem setting, we assumed that not every unlabeled, ambiguous sample has meaningful knowledge consistent with the obvious samples. Therefore, we only acquired a positive pseudo-label on an ambiguous sample because confidence in one class does not guarantee the sample does not belong to the other class. Accordingly, we measured uncertainty with the positive pseudo-label only while the work (Rizve et al., 2021) captured uncertainties at both positive and negative pseudo-labels.

### 2.2 Uncertainty Estimation

Uncertainty estimation aims to measure how the deep neural networks made a confident decision at a given data (Abdar et al., 2021). The prior studies have actively scrutinized uncertainties in the deep neural networks through various approaches. Early studies employed a Bayesian analysis for uncertainty estimation. Given a prior distribution over the trained model's weights, Bayesian approaches estimated uncertainty capturing how much these weights vary given a particular data. (Gal and Ghahramani, 2016) proposed a Monte Carlo (MC) dropout, which employs a Dropout as a regularization term for computing an uncertainty at the model's decision. Various studies have shown the effectiveness of MC dropout in uncertainty estimation for various computer vision applications (Wang et al., 2019; Nair et al., 2020; Do et al., 2020). On the other hand, (Lakshminarayanan et al., 2016) also proposed a non-Bayesian approach to the uncertainty estimation by assembling multiple models. While the Bayesian approaches accompanied large computation overhead for uncertainty estimation, the proposed assembling approach could have efficiently yield uncertainties under the distributed computing environment. The recent progress in uncertainty estimation has been brought under the problem setting of out-of-distribution detection (OOD detection) (Bulusu et al., 2020). (Hendrycks and Gimpel, 2016) suggested a concept Maximum Softmax Probability (MSP), which is a simple but effective method for estimating uncertainty. Given a model trained with in-distribution data, the work figured out the model yielded low softmax probabilities at out-of-distribution data (which is unknown to the model); thus, the higher MSP implies higher confidence in the model's decision. Among various approaches in estimating uncertainties of deep neural networks, our study employed the MSP method considering its simplicity of implementation and lightweight resource consumption. Refer Section 4.3 for a detailed elaboration on the takeaway behind this selection.

## 3 CREATING SYNTHETIC AMBIGUOUS DATASET

The foremost challenge in examining our approach is the lack of benchmark datasets regarding the ambiguity. We tried to retrieve a particular amount of both obvious and ambiguous samples from the publicly-available natural image datasets (i.e., ImageNet (Deng et al., 2009), SUN(Xiao et al., 2010)) but failed to figure them out. Therefore, given obvious samples (which are easily acquired), we decide to create ambiguous synthetic samples with generative models (Goodfellow et al., 2014; Creswell et al., 2018; Karras et al., 2019). We fully acknowledge that it would become a more precise study to validate our approach's effectiveness with real-world samples. Note that the use of ambiguous synthetic samples is one of the improvement avenues of our work.

The ambiguous synthetic samples are designed to have attributes of multiple classes so that they have inherent ambiguity; they are somewhere in-between samples. To create the aforementioned ambiguous data, we utilize StyleMapGAN (Kim et al., 2021) which can combine multiple images with predefined masks. Given a pair of two images at the target classes, we generate a mixed image that contextually mingles characteristics of the given two images following a preset ratio 1:1 to maximize ambiguity. We primarily employed two widely utilized benchmark datasets with well-aligned images to produce realistic from StyleMapGAN: AFHQ and CelebA-HQ. We set the scope of analyses to solve two image classification tasks: classifying cat and wild animals' images from the AFHQ dataset and classifying male and female celebrities' face images from the CelebA-HQ dataset. We assumed images in the original datasets are obvious data without any ambiguity. Figure 3a,c and Figure 4a,c shows obvious images of the AFHQ dataset and CelebA-HQ dataset, respectively. In pursuit of creating ambiguous data, we randomly selected 500 obvious images from each class and provide them to the StyleMapGAN model to acquire synthetically mixed ambiguous images. For masks, we use simple half-and-half vertical or horizontal splits. The examples of ambiguous images at both the AFHQ dataset and CelebA-HQ dataset are illustrated in Figure 3b and Figure 4b, respectively. In the following sections, we use the original names to denote the synthesized datasets and present our method.

## 4 UNCERTAINTY GUIDED PSEUDO-LABELING

In this section, we illustrated UGPL, which is our approach to utilize ambiguous data to escalate the image classification performance. The overall architecture of our framework is described in Figure 5. Note that we focus on binary classification for simplicity without loss of generality.

(a) Obvious Cat (b) Ambiguous (c) Obvious Wild

Figure 3: Examples of Synthetic-AFHQ dataset, which is synthetically created examples of two obvious data and one ambiguous data from AFHQ dataset.

(a) Obvious Male (b) Ambiguous (c) Obvious Female

Figure 4: Examples of Synthetic-CelebA-HQ dataset, which is synthetically created examples of two obvious data and one ambiguous data from CelebA-HQ dataset.

### 4.1 Initial Training with Obvious Data

First, we trained an initial classifier with obvious data. We aim to let the deep neural networks learn primary representations of the obvious classes during the initial training stage. The trained initial classifier becomes a baseline for examining the effectiveness of UGPL. Our framework aims to improve image classification performance from the initial classifier. Throughout the study, we employ randomly initialized ResNet-18 (He et al., 2016) architecture and added a single fully connected layer to perform binary classification. We set the loss function as a categorical cross-entropy loss and optimized the networks' parameters with a stochastic gradient descent optimizer. We set the batch size as 64, the learning rate as 0.0001, and applied weight decay with the parameter 0.0005.

### 4.2 Pseudo-label Generation

With the initial classifier trained on the obvious samples, we created pseudo-labels on the ambiguous data. Given the initial classifier, we provided an ambiguous image $x^{(i)}$ to the classifier and extracted a logit vector $v$, a network prediction. As we set the image classification task as a binary classification, the logit vector $v$ is composed of $p_{c_1}$ and $p_{c_2}$ where $c$ describes the target class satisfying $c \subset (0,1)$, and $p_c$ implies the probability that a given data belongs to the class $c$ following the knowledge learned by the initial classifier. We created a pseudo-label $\hat{y^{(i)}}$ of ambiguous data $x^{(i)}$ following the class with higher probability. In other words, the ambiguous samples will have pseudo-labels based on the knowledge learned from the obvious data. If the initial classifier observes similar characteristics of the given ambiguous data to a particular target class, it will result in a higher probability at that class. However, it is not robust enough to use all pseudo-labels. We describe a solution below.

### 4.3 Uncertainty-guided Sampling

We additionally measured an uncertainty on the network prediction of the ambiguous sample $x^{(i)}$. While we share a common intuition with the approach (Rizve et al., 2021), we further cast doubt whether every ambiguous data with pseudo-labels contribute

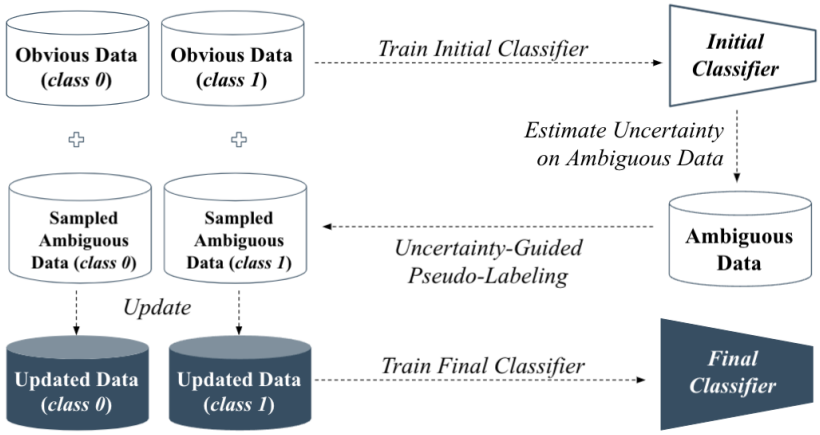



Figure 5: The overall structure of Uncertainty-Guided Pseudo-Labeling (UGPL) framework.

to improving performance. Suppose an ambiguous data inherently does not include much discriminative knowledge on the given classes. In this case, we expect the ambiguous data would not have many contributions to the representation learning. Our study exploit uncertainty as a proxy to estimate the amount of knowledge the ambiguous data contains. Specifically, we suppose that the classifier would fail to produce enough activation for a particular class and end up at high uncertainty if the input has ambiguity. On the other hand, we expect the classifier would yield a confident decision if the given ambiguous image includes any meaningful knowledge similar to the target classes.

Briefly, our framework establishes a barrier that blocks uncertain samples from participating in the training procedure. Among various approaches of uncertainty estimation on neural networks, we employed the MSP method as it is widely utilized as a baseline in various uncertainty-related domains (Hendrycks and Gimpel, 2016; Gal and Ghahramani, 2016; Harang and Rudd, 2018; Hendrycks et al., 2018; Kabir et al., 2018). Note that we did not choose the MC-Dropout and its derived versions (which are a state-of-the-art approach to the uncertainty estimation) as they require comparatively-heavy computation resources during the uncertainty estimation. Our approach measures uncertainty following the MSP method, selects ambiguous data with low uncertainty order, and merges them with obvious data following their pseudo-labels. Note that a high MSP value at a particular sample implies that the neural network confidently made a prediction with low uncertainty on the sample. Lastly, we established a final classifier by training the network with the updated training set. Note that we employed the same architecture and other configurations with section 4.1 to eliminate the effect of other factors on the image recognition performance. Throughout experiments illustrated in Section 5, we empirically examined how uncertainty can be a proxy for measuring the amount of knowledge that ambiguous data has.

# 5 EXPERIMENTS

## 5.1 Experiment Setup

In this section, we aimed to examine UGPL's effectiveness in the elevation of image classification performance. Throughout the experiments, we scrutinized answers to the following research questions described below sections.

### 5.1.1 Does Every Ambiguous Sample Have Knowledge Consistent to the Obvious Samples?

First and foremost, we tried to examine that not every ambiguous sample has meaningful knowledge to discriminate the target classes. For the experiment setups, we trained the initial classifiers with 400 obvious images per target class at both the synthetic-AFHQ dataset and synthetic-CelebA-HQ dataset. We then prepared three types of validation subsets: Ambiguous set and two obvious sets from each class used to generate the ambiguous subsets. We extracted the distribution of uncertainty estimates at each set and examined whether the ambiguous set shows a particular amount of uncertainty compared to the obvious sets. Note that we constructed validation sets with 500 samples. Suppose that the ambiguous set shows higher uncertainty compared to obvious sets. In this case, we would make a naive discovery that ambigu-

ous data do actually show higher uncertainty rather than obvious data. Therefore, we would experimentally examine that not every ambiguous sample has consistent knowledge of the obvious samples. The experiment results are shown in Figure 6 and Figure 7.

#### 5.1.2 Does UGPL Achieves Better Image Classification Performance?

We then examined whether UGPL contributes to the escalation of image recognition performance. We employed an evaluation metric as Top-1 Accuracy and scrutinized the relationship between a performance escalation and selected ambiguous samples' uncertainties. During the uncertainty-guided sampling illustrated in Section 4.3, we extracted ambiguous samples' MSP values and sorted them in increasing order. We divided uncertainties into four windows based on quartile values as follows: Q1 (0% to 25%), Q2 (25% to 50%), Q3 (50% to 75%), Q4 (75% to 100%). A sampling strategy Q1 implies that we selected ambiguous samples where their MSP values lie lower than 25% at total MSP distribution; thus, it selects ambiguous samples which is the most ambiguous with high uncertainty. Note the lower MSP denotes less confidence in the network's prediction, which describes a high uncertainty. On the other hand, the sampling strategy Q4 denotes that selected ambiguous samples yield low uncertainty estimates as the initial classifier confidently decided. To ensure that the experiment result is not affected by the number of ambiguous data at each class in the training set, we sampled 100 ambiguous samples for each target class (200 images in total) to be updated into the training set. Lastly, we set the upper bound on the elevation of the test accuracy by adding the same number of obvious samples to each target class. As obvious samples presumably have consistent knowledge with the trained samples, we evaluated it as an upper bound for the proposed UGPL framework. Throughout experiments, we tried to examine whether ambiguous samples with low uncertainties contributed more to the test accuracy escalation.

#### 5.1.3 Is UPGL Still Valid under the Small Size of Obvious Samples?

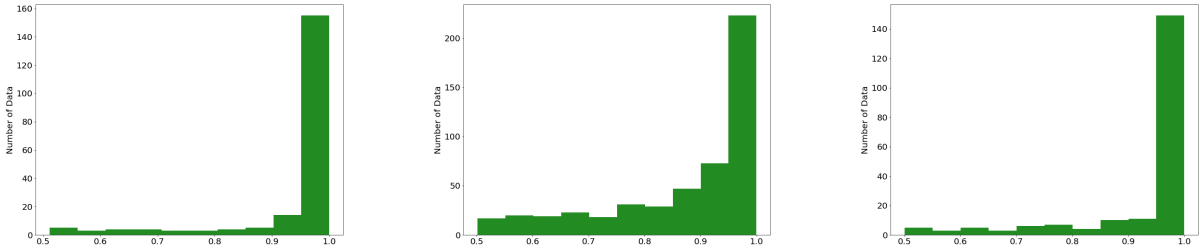
Lastly, we aimed to validate whether UGPL is valid under the different sizes of obvious samples. If our approach achieves meaningful test accuracy escalation under the small size of obvious samples, we expect UGPL can benefit general practitioners under resource constraints. We set the number of obvious samples per target class at Synthetic-AFHQ and Synthetic-CelebA-HQ as (100, 400) and (150, 600), respectively. Note each target class has the same number of obvious data to maintain a class balance at the initial training set. We set the former experiment setup to illustrate the circumstance where the initial classifier did not learn sufficient characteristics of the target classes. On the other hand, the latter setup implies the model learned more about target classes compared to the former setup.

### 5.2 Analogy 1: Not Every Ambiguous Samples Bear Consistent Knowledge to the Obvious Samples

Referring the experiment results described in Figure 6 and Figure 7, we discovered estimated uncertainties vary among ambiguous samples. Following Figure 6a, 6c and 7a, 7c, the classifier resulted in low uncertainty in most of validation obvious sets. We reconfirmed a common notion that obvious samples share a consistent knowledge to understand the target class. However, the initial classifier showed different uncertainty distribution in ambiguous validation sets as shown in Figure 6b and 7b. Compared to obvious validation samples, there existed more uncertain samples in ambiguous validation sets. The initial classifier cast doubt on particular ambiguous samples as they do not seem to have similar knowledge to the obvious data. In a nutshell, we could have discovered not every ambiguous sample has meaningful knowledge similar to the obvious data by unveiling uncertainty distribution at ambiguous sets.

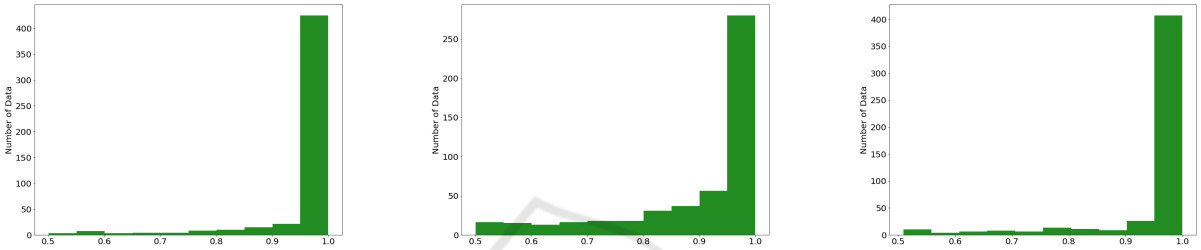
### 5.3 Analogy 2: Ambiguous Samples with Low Uncertainty Accomplish Higher Test Accuracy

Following the experiment results shown in Figure 8 and Figure 9, we examined UGPL achieves better image recognition performance rather than utilizing obvious data only. Regardless of datasets and the size of the initial training set, our approach (**UGPL-4Q**) achieved the most similar test accuracy to the upper bound. We also discovered a positive relationship between ambiguous samples' uncertainty and accuracy elevation; the lower uncertainty in ambiguous samples correlates to the higher test accuracy escalation. From the experiment results, we could examine our proof-of-concept level statements regarding the ambiguity. As ambiguous samples with low uncertainty escalated the test accuracy at most, uncertainty can be a possible proxy for measuring the ambiguous sample's knowledge. Specifically, low uncertainty im-




(a) Obvious Cat (b) Ambiguous (c) Obvious Wild

Figure 6: The distribution of uncertainty at validation sets of obvious cat images, ambiguous images, and obvious wild images on the synthetic-AFHQ dataset. The ambiguous set included particularly many uncertain data compared to others; thus, we resulted in ambiguous samples do not always bear the knowledge similar to the obvious samples.




(a) Obvious Male (b) Ambiguous (c) Obvious Female

Figure 7: The distribution of uncertainty at validation sets of obvious male images, ambiguous images, and obvious female images on the synthetic-CelebA-HQ dataset. The ambiguous set included particularly many uncertain data compared to others; thus, we resulted in ambiguous samples do not always bear the knowledge similar to the obvious samples.

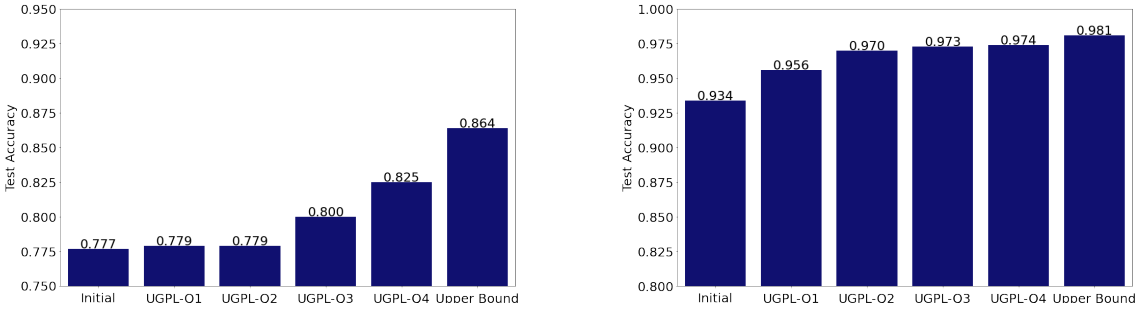
plies an ambiguous sample includes similar knowledge to the obvious samples. Last but not least, we validated UGPL can be an effective method to figure out meaningful ambiguous samples and elevate the test accuracy without further resource consumption on acquiring additional obvious data.

### 5.4 Analogy 3: UGPL Is More Effective When There Exists Small Obvious Samples

Lastly, we figured out UGPL elevated larger test accuracy under the small initial training set. In Figure 8a, our approach **UGPL-Q4** escalated the test accuracy for 0.479 (+6% from the initial classifier's test accuracy) at the synthetic-AFHQ dataset. Our approach also elevated the test accuracy for 0.469 (+5.8%) at the synthetic-CelebA-HQ dataset as shown in Figure 9a. On the other hand, UGPL improved the test accuracy at both datasets, but the impact was comparatively small. Our approach contributed to the accuracy escalation of 0.039 (+4.28%) at the synthetic-AFHQ dataset, and increased the performance by 0.028 (+3.1%) at the synthetic-CelebA-HQ dataset. Considering the experiment results, we interpret the UGPL's effectiveness is not depreciated under the small size of obvious samples at the initial training set. We discovered the UPGL even achieved a larger escalation of test accuracy in a small size of the initial training set. Thus, general practitioners can utilize our approach even if they do not have a large number of obvious samples at very first. Still, we could not propose a clear, definite analysis of the discovery, and it remains room for improvement for a deeper understanding of our approach.

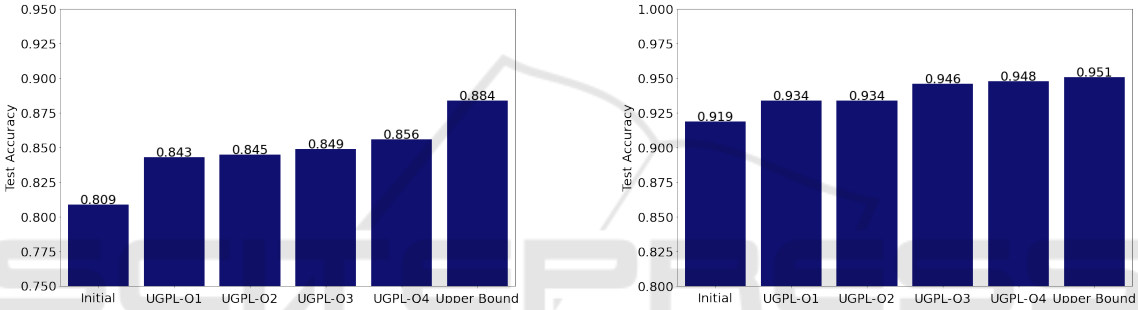
## 6 DISCUSSIONS AND CONCLUSION

Throughout the study, we described the UGPL framework and examined that our approach escalates image classification performance with ambiguous samples. Although we validated the UGPL's effectiveness at a proof-of-concept level, we fully acknowledge that there shall be more strict and various experiments to unveil the mechanism of our approach's effectiveness. Our approach should be examined with more diverse experiment setups such as real-world datasets, the size of initial training size, the number of sampled ambiguous data, and various uncertainty estimation methods. While our baseline work limited the scope of analysis into binary image classification, further studies shall tackle the ambiguity problem on multi-




(a) Small Initial Training Set (b) Large Initial Training Set

Figure 8: Experiment result on the synthetic-AFHQ dataset. **Initial** implies the test accuracy with the classifier trained by obvious samples only. **UGPL-Q1** to **UGPL-Q4** describes the test accuracy from the classifier trained with ambiguous samples as well as obvious samples. UGPL-Q1 option added the most uncertain ambiguous samples, and the UGPL-Q4 option, which is our approach, merged the least uncertain ambiguous samples to the updated training set. **Upper Bound** represents the test accuracy with the classifier trained by an updated training set where obvious samples are additionally merged into the **initial** option. Following the result, our approach **UGPL-Q4** escalated the test accuracy at most from the **Initial**.




(a) Small Initial Training Set (b) Large Initial Training Set

Figure 9: Experiment result on the synthetic-CelebA-HQ dataset. **Initial** implies the test accuracy with the classifier trained by obvious samples only. **UGPL-Q1** to **UGPL-Q4** describes the test accuracy from the classifier trained with ambiguous samples as well as obvious samples. UGPL-Q1 option added the most uncertain ambiguous samples, and the UGPL-Q4 option, which is our approach, merged the least uncertain ambiguous samples to the updated training set. **Upper Bound** represents the test accuracy with the classifier trained by an updated training set where obvious samples are additionally merged into the **initial** option. Following the result, our approach **UGPL-Q4** escalated the test accuracy at most from the **Initial**.

class image classification or object detection tasks. If further studies revisit the aforementioned points, it will become a meaningful breakthrough to elevate the image recognition performance with ambiguous samples in the computer vision domain.

Throughout the study, we proposed an ambiguity problem in the computer vision domain that inquires an answer to the following question: How can we leverage the ambiguous samples to escalate the image recognition performance rather than solely using obvious samples? One of the presumable approaches to this ambiguity problem is pseudo-labeling under the semi-supervised learning paradigm. However, our study proposes a doubt on pseudo-labeling at every ambiguous sample because applying confident pseudo-labels at inherently ambiguous, which might not have sufficient knowledge on the target class, is not reasonable. To this end, our study illustrated a series of analyses to establish a practical pseudo-labeling framework under the ambiguous samples and suggest several proof-of-concept level validations. First, we utilized a generative neural network (StyleMapGAN) to create synthetic ambiguous data based on two widely-used datasets (AFHQ, CelebA-HQ) due to the lack of a benchmark dataset that deals with ambiguity. Second, we empirically showed that not every ambiguous sample bears similar knowledge to obvious samples. We further illustrated the uncertainty can be a possible proxy for describing the effectiveness of ambiguous sample's knowledge toward the escalation of image classification performance. Third, we designed the UGPL framework, which selects ambiguous samples with low uncertainty to update the training set. We experimentally examined UGPL contributes to the elevation of image classification performance. Lastly, the UGPL accomplished higher performance elevation under the small size of the initial training set.

Therefore, general practitioners can be widely benefited from our approach, although they did not acquire a particular amount of obvious samples as an initial training set. Based on our study's proposition and further revisits on the aforementioned avenues of improvement, we expect upcoming studies can provide an effective solution to leverage ambiguous samples on escalating the image recognition performance.

## REFERENCES


Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*.

Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., and Song, D. (2020). Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Denton, E., Gross, S., and Fergus, R. (2016). Semi-supervised learning with context-conditional generative adversarial networks. *arXiv preprint arXiv:1611.06430*.

Do, H. P., Guo, Y., Yoon, A. J., and Nayak, K. S. (2020). Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn. *Magnetic resonance in medicine*, 83(5):1863–1874.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., and Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838.

Geng, X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Harang, R. and Rudd, E. M. (2018). Towards principled uncertainty estimation for deep neural networks. *arXiv preprint arXiv:1810.12278*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D., Mazeika, M., and Dietterich, T. (2018). Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hüllermeier, E. and Beringer, J. (2006). Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439.

Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. (2019). Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079.

Jiang, B., Zhang, Z., Lin, D., Tang, J., and Luo, B. (2019). Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11313–11320.

Kabir, H. D., Khosravi, A., Hosen, M. A., and Nahavandi, S. (2018). Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE access*, 6:36218–36234.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Kim, H., Choi, Y., Kim, J., Yoo, S., and Uh, Y. (2021). Stylemapgan: Exploiting spatial dimensions of latent in gan for real-time image editing. *arXiv preprint arXiv:2104.14754*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Liu, B., Wu, Z., Hu, H., and Lin, S. (2019). Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for

multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557.

Nath, S. S., Mishra, G., Kar, J., Chakraborty, S., and Dey, N. (2014). A survey of image classification methods and techniques. In *2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT)*, pages 554–557. IEEE.

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

Otani, N., Otsubo, Y., Koike, T., and Sugiyama, M. (2020). Binary classification with ambiguous training data. *Machine Learning*, 109(12):2369–2388.

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. (2021). Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568.

Rezagholiradeh, M. and Haidar, M. A. (2018). Reg-gan: Semi-supervised learning based on generative adversarial networks for regression. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2806–2810. IEEE.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. (2017). Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600.

Sukanya, C., Gokul, R., and Paul, V. (2016). A survey on object recognition methods. *International Journal of Science, Engineering and Computer Technology*, 6(1):48.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Yang, X., Song, Z., King, I., and Xu, Z. (2021). A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*.

Yang, Z., Cohen, W., and Salakhudinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR.

Zhou, T., Wang, S., and Bilmes, J. (2020). Time-consistent self-supervision for semi-supervised learning. In *International Conference on Machine Learning*, pages 11523–11533. PMLR.