MixMatch

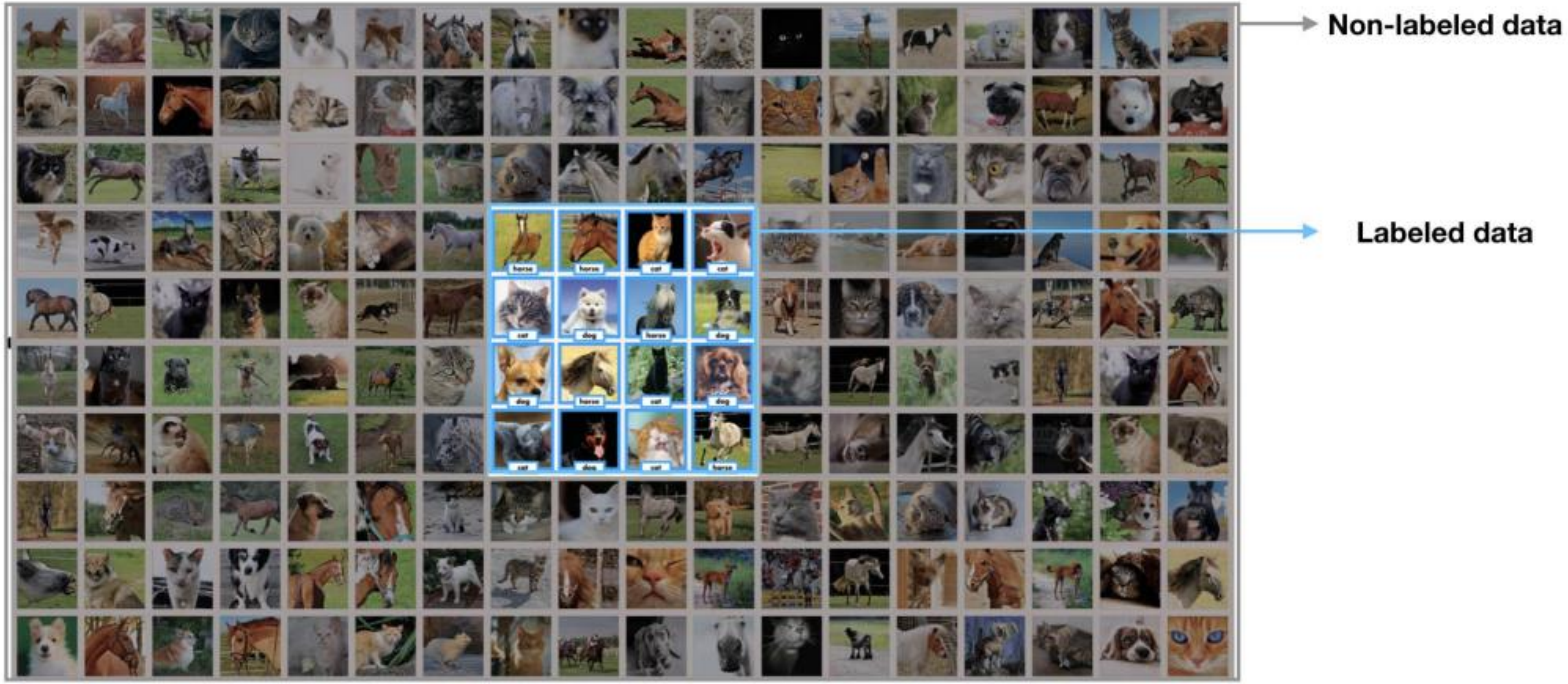
A Holistic Approach to Semi-Supervised Learning

index

1. Introduction

▪ Semi-Supervised Learning

In situations where labeled data is not abundant, let's get help from unlabeled data!



1. Introduction

- Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

1. Introduction

- Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

Mainly used to generalize about data that has never been seen before.

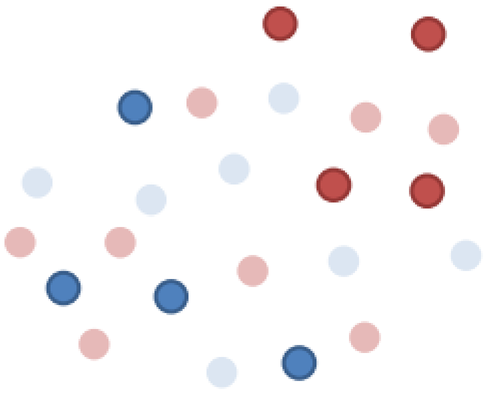
1. Introduction

▪ Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

1. Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.



Labeled & Unlabeled data

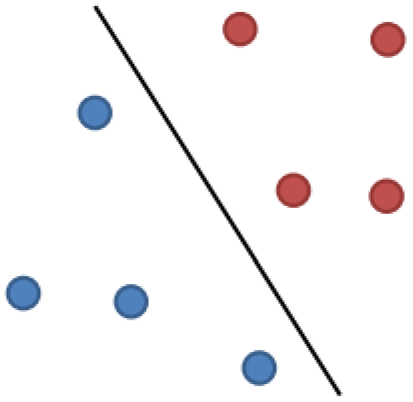
1. Introduction

▪ Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

1. Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.



Supervised Learning

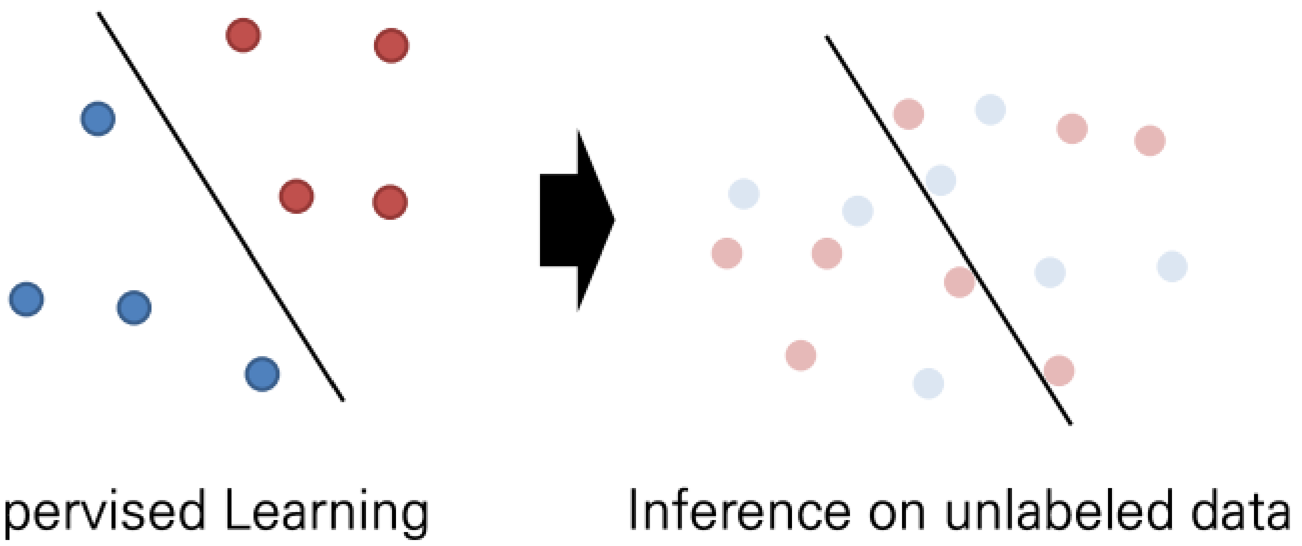
1. Introduction

▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

1. Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.



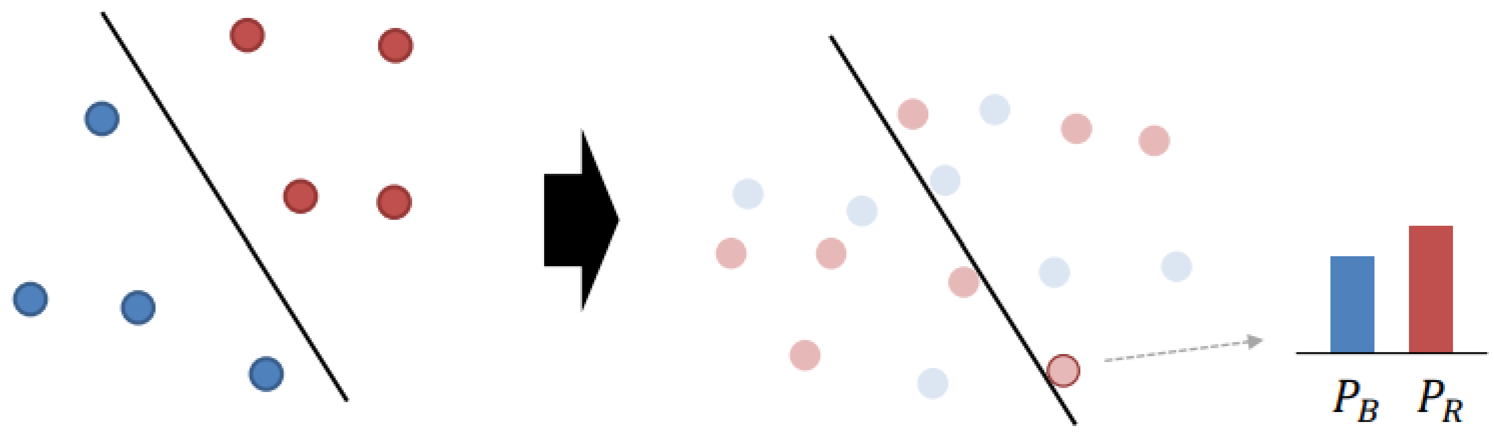
1. Introduction

▪ Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

1) Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.



Supervised Learning

Inference on unlabeled data

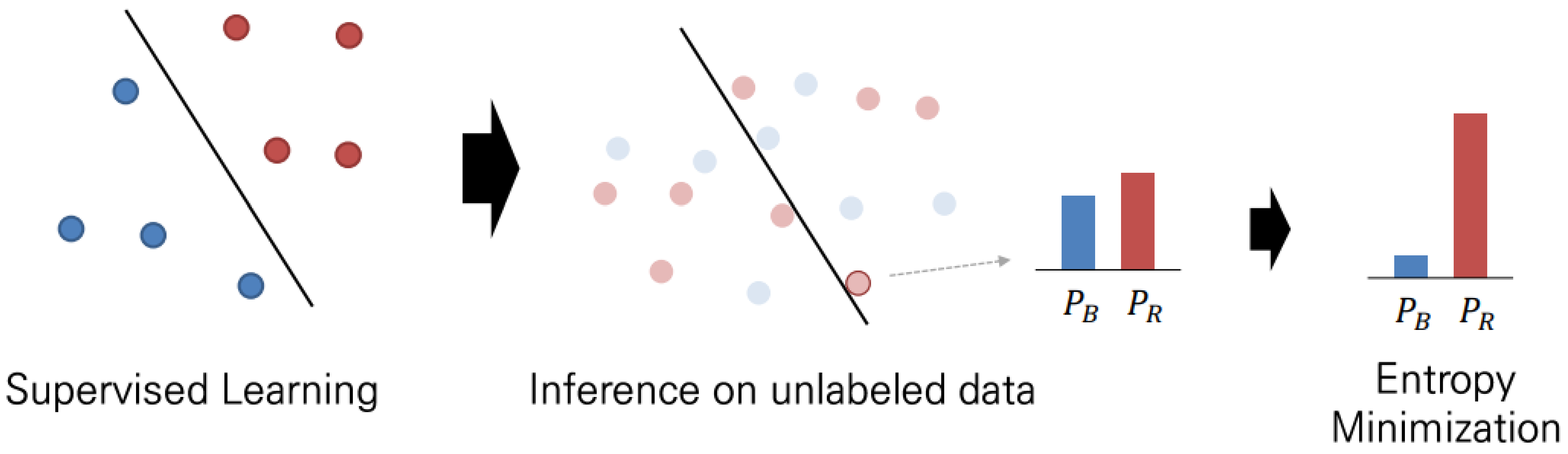
1. Introduction

▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

1) Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.



1. Introduction

▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

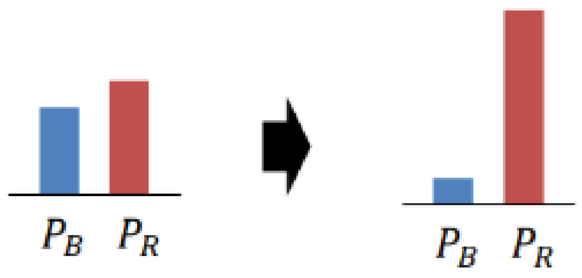
1) Entropy Minimization

The purpose is to increase the confidence of prediction values for unlabeled data.

Softmax Temperature

$$\frac{p_i}{\sum p_j} \Rightarrow \frac{p_i^{\frac{1}{T}}}{\sum p_j^{\frac{1}{T}}}$$

Low entropy
for low temperature (T → 0)



Entropy
Minimization

1. Introduction

▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

2) Consistency Regularization

Data Augmentation

- Supervised: Class information will not be affected even if slight modifications are made to the data.

Label : Smile

1. Introduction

- Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

2) Consistency Regularization

Data Augmentation

- Supervised: Class information will not be affected even if slight modifications are made to the data.

Label : Smile　　Smile　　Smile　　Smile

1. Introduction

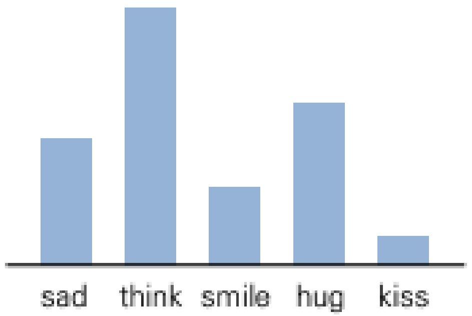
▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

2) Consistency Regularization

Data Augmentation

- Semi-Supervised: Augmenting unlabeled data changes the predicted distribution of classes.



Unlabeled data

1. Introduction

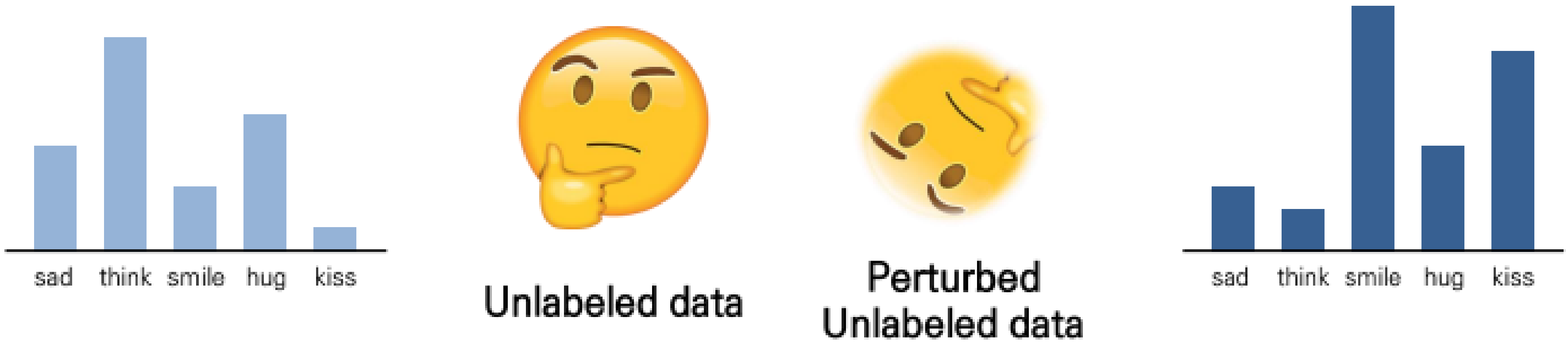
▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

2) Consistency Regularization

Data Augmentation

- Semi-Supervised: Augmenting unlabeled data changes the predicted distribution of classes.



1. Introduction

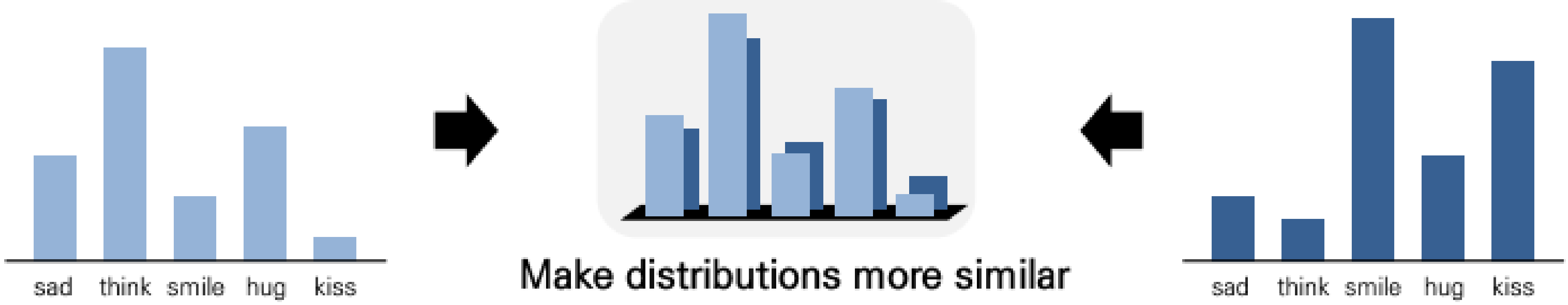
▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

2) Consistency Regularization

Learn to predict the same class distribution even when performing augmentation on unlabeled data.

Increase the similarity of distributions by using Squared Loss Term, etc.



1. Introduction

▪ Semi-Supervised Learning – Recent Trends

$$Loss = L_S + L_U$$

3) Traditional Regularization – MixUp

- Supervised: Create new data through convex combination of each data and label.

  Adapts well to unseen data and prevents overfitting

[1 , 0] [0 , 1] [0.7 , 0.3]

1. Introduction

▪ Semi-Supervised Learning - Recent Trends

$$Loss = L_S + L_U$$

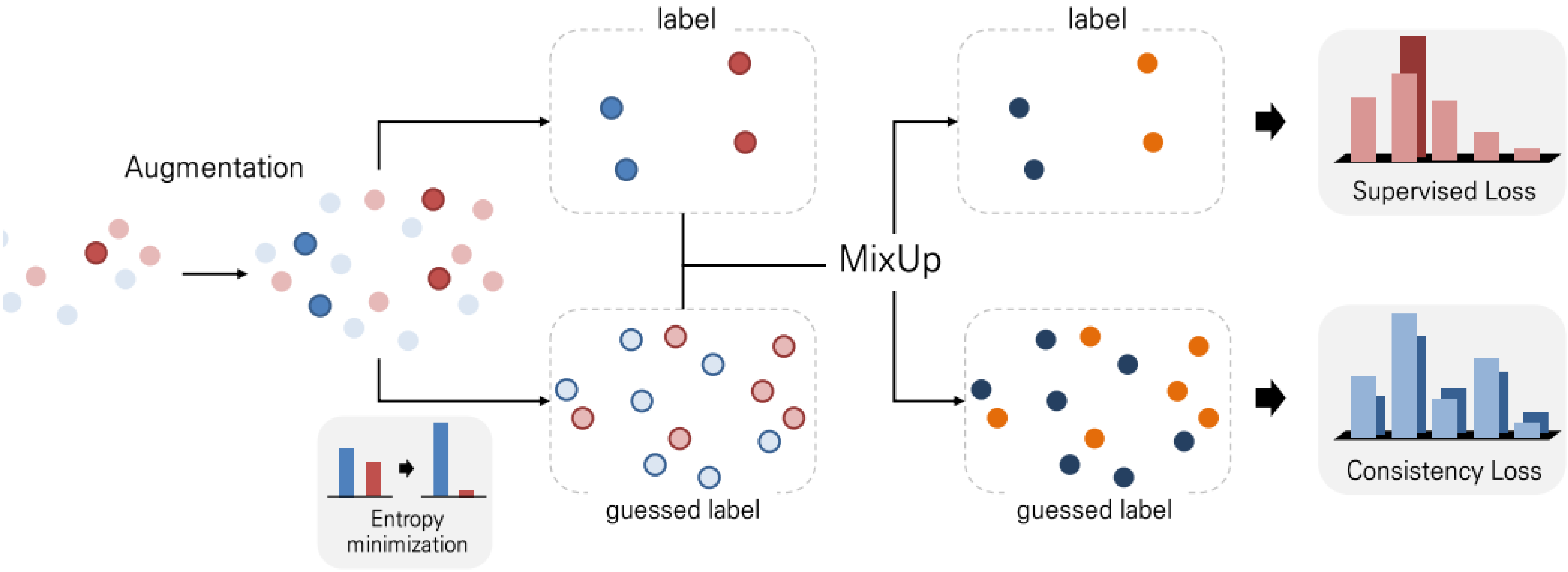
3) Traditional Regularization – MixUp

- Semi-Supervised: The model uses fake labels generated for unlabeled data.

1. Introduction

- MixMatch : A Holistic Approach to Semi-Supervised Learning

MixMatch, a new methodology that encompasses all of the previously introduced SSL methodologies, is presented.



2. MixMatch

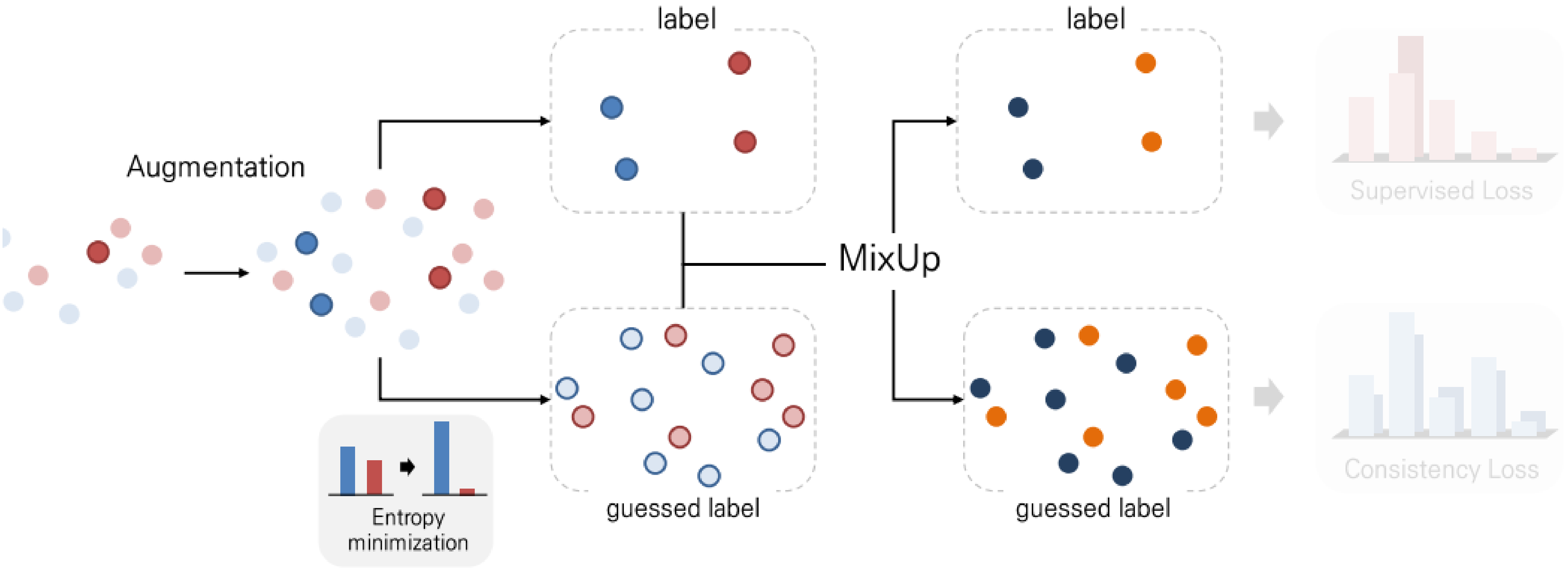
New Labeled and Unlabeled data are generated through MixMatch for each batch.

$$\mathcal{X}', \mathcal{U}' = \mathrm{MixMatch}(\mathcal{X}, \mathcal{U}, T, K, \alpha)$$

X : Labeled Examples

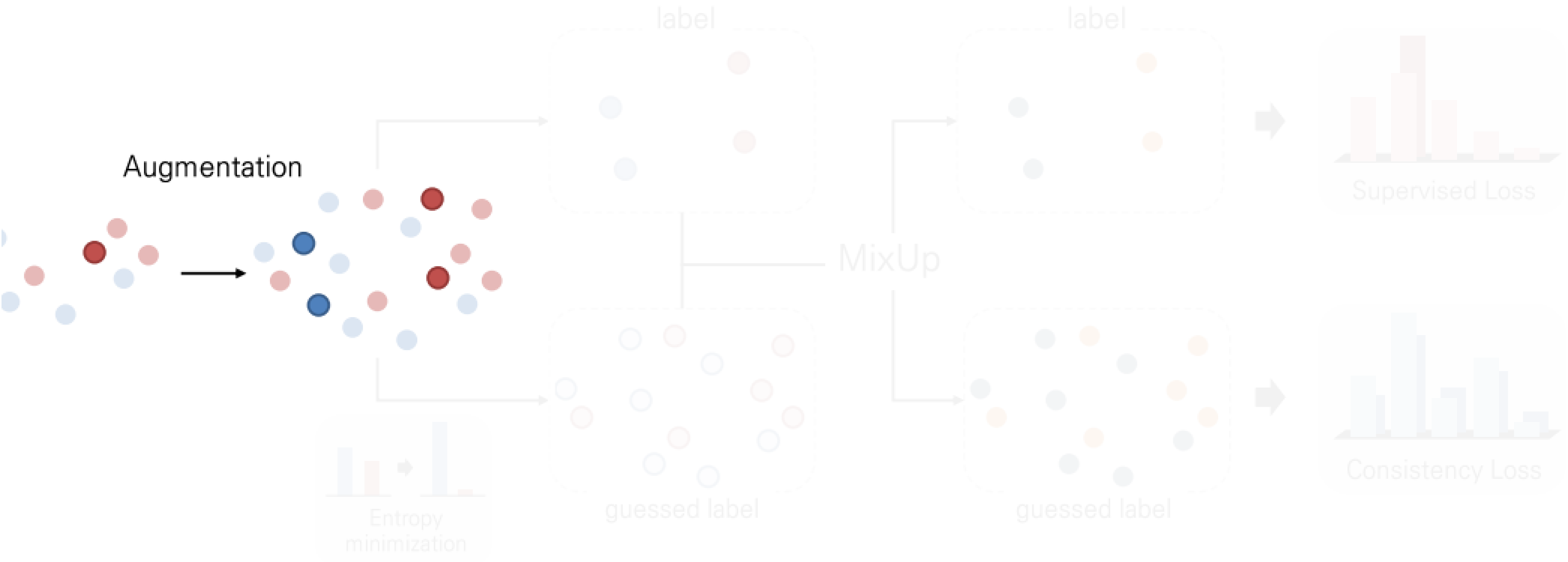
U : Unlabeled Examples

T, K, α : Hyperparameters



2. MixMatch

- Data Augmentation

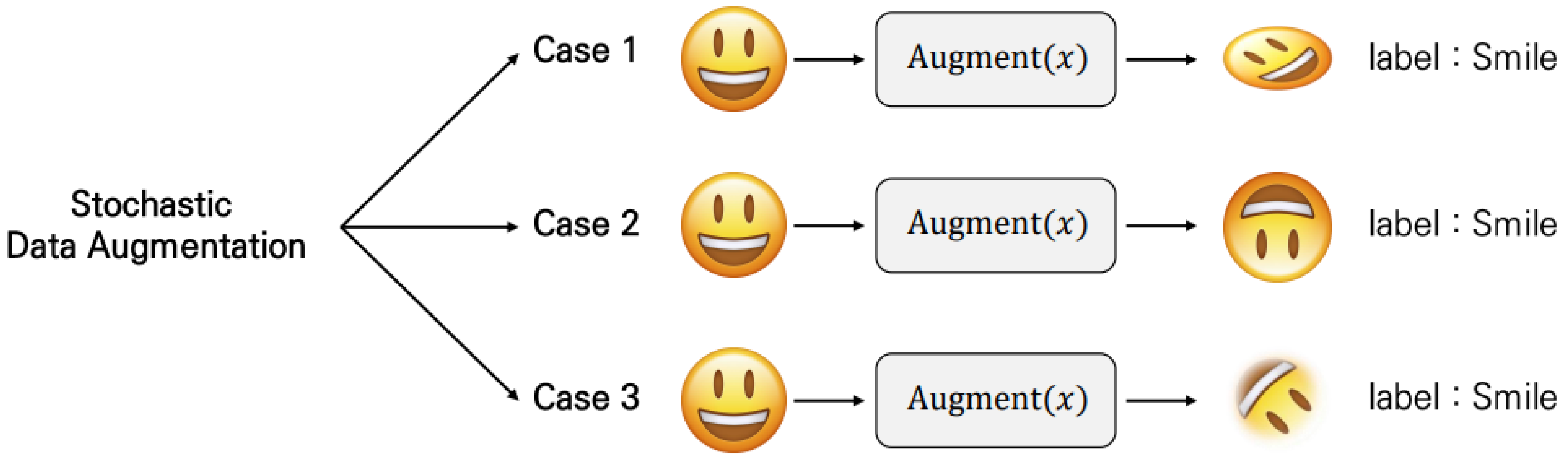


2. MixMatch

- **Data Augmentation**

$\hat{x}_b = \mathrm{Augment}(x_b)$: Stochastic Data Augmentation

Randomly apply one of the predefined Image Augmentation techniques to the labeled data.



2. MixMatch

- Data Augmentation

$\hat{x}_b = \text{Augment}(x_b)$: Stochastic Data Augmentation

Randomly apply one of the predefined Image Augmentation techniques to the labeled data.

ûb,k = Augment(up): Stochastic Data Augmentation on Unlabeled Data

k∈ (1, ..., K)

Applying Stochastic Data Augmentation K times to unlabeled data

2. MixMatch

- Data Augmentation

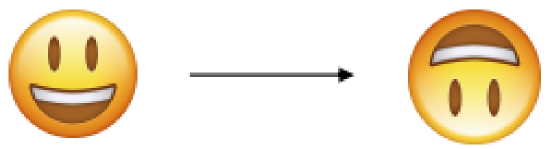
$\hat{x}_b = \text{Augment}(x_b)$: Stochastic Data Augmentation

Randomly apply one of the predefined Image Augmentation techniques to the labeled data.

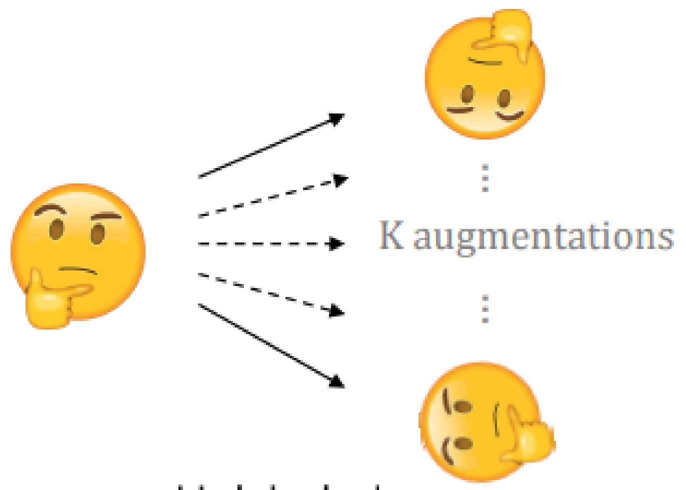
ûb,k = Augment(up): Stochastic Data Augmentation on Unlabeled Data

k∈ (1, ..., K)

Applying Stochastic Data Augmentation K times to unlabeled data



Labeled



Unlabeled

2. MixMatch

- Data Augmentation

$\hat{x}_b = \text{Augment}(x_b)$: Stochastic Data Augmentation

Randomly apply one of the predefined Image Augmentation techniques to the labeled data.

$\hat{u}_{b,k} = \text{Augment}(u_b)$: Stochastic Data Augmentation on Unlabeled Data

k∈ (1, ..., K)

Applying Stochastic Data Augmentation K times to unlabeled data

There are B labeled data and B unlabeled data in a minibatch.

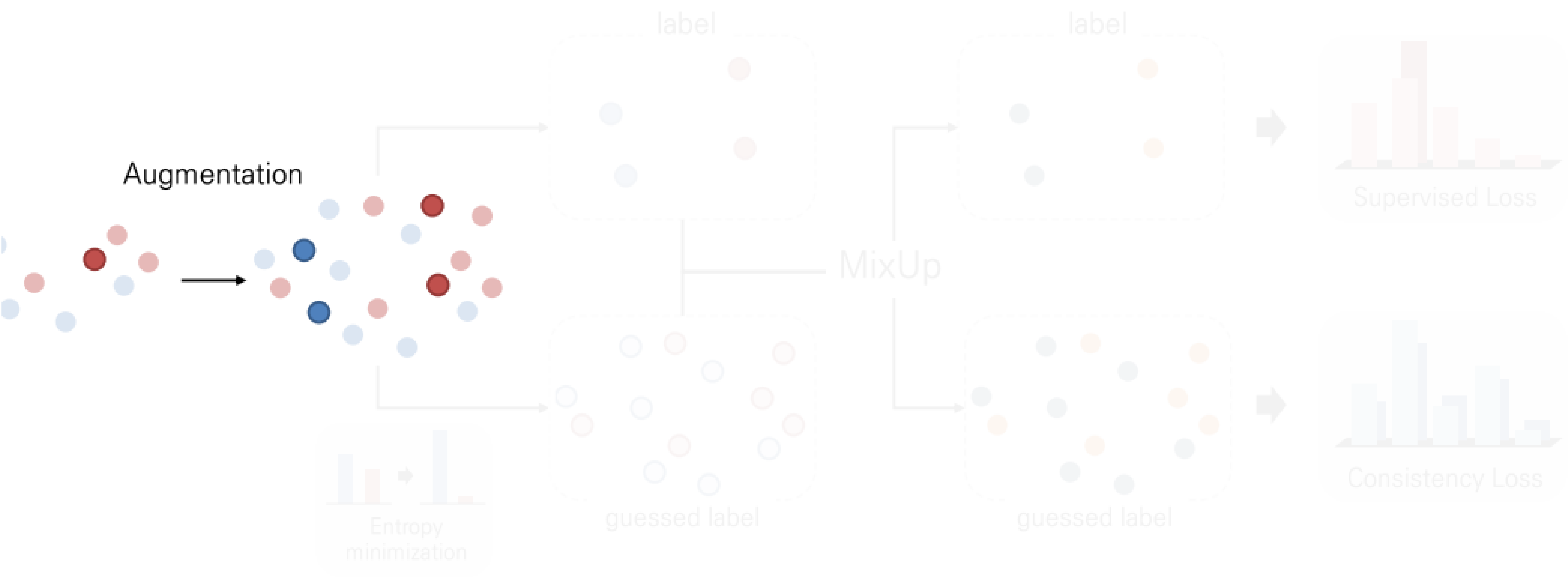
When performing Data Augmentation, B labeled data and B * k unlabeled data are generated.

X = ((b, Pb); b ∈ (1,..., B)) - Stochastic Augment applied once per data point

Û = ((ûb,k, qb); b ∈ (1, ..., B), k ∈ (1, . . ., K)) - Stochastic Augmentation applied k times per data point

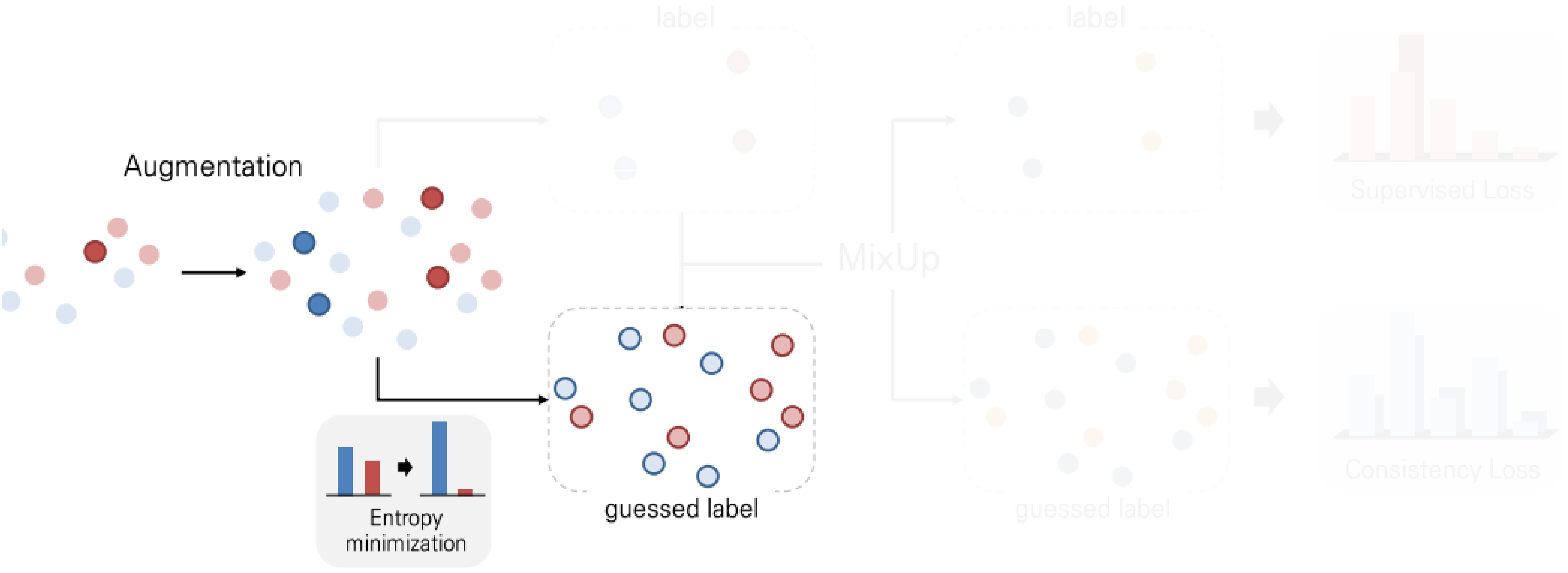
2. MixMatch

- Label Guessing & Entropy Minimization



2. MixMatch

- Label Guessing & Entropy Minimization

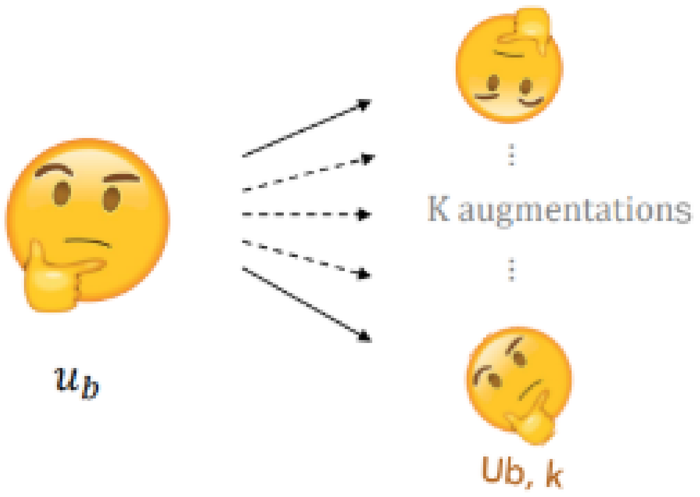


2. MixMatch

- Label Guessing & Entropy Minimization

$\hat{u}_{b,k} = \text{Augment}(u_b)$: Apply Stochastic Data Augmentation to unlabeled data

$k \in (1, \ldots, K)$



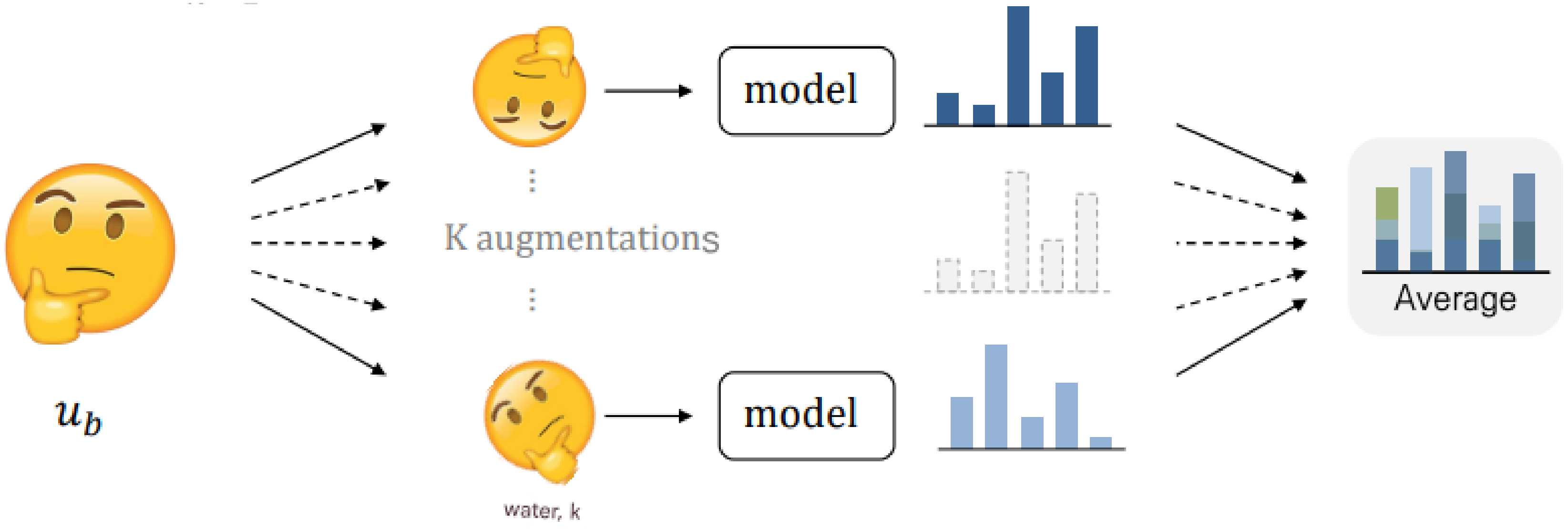
2. MixMatch

- **Label Guessing & Entropy Minimization**

$\hat{u}_{b,k} = \text{Augment}(u_b)$: Apply Stochastic Data Augmentation to unlabeled data

$k \in (1, \ldots, K)$

$$\bar{q}_b = \frac{1}{K} \sum_{k=1}^{K} \mathrm{p}_{\text{model}}(y \mid \hat{u}_{b,k}; \theta)$$: Predict label y through the model and take the average

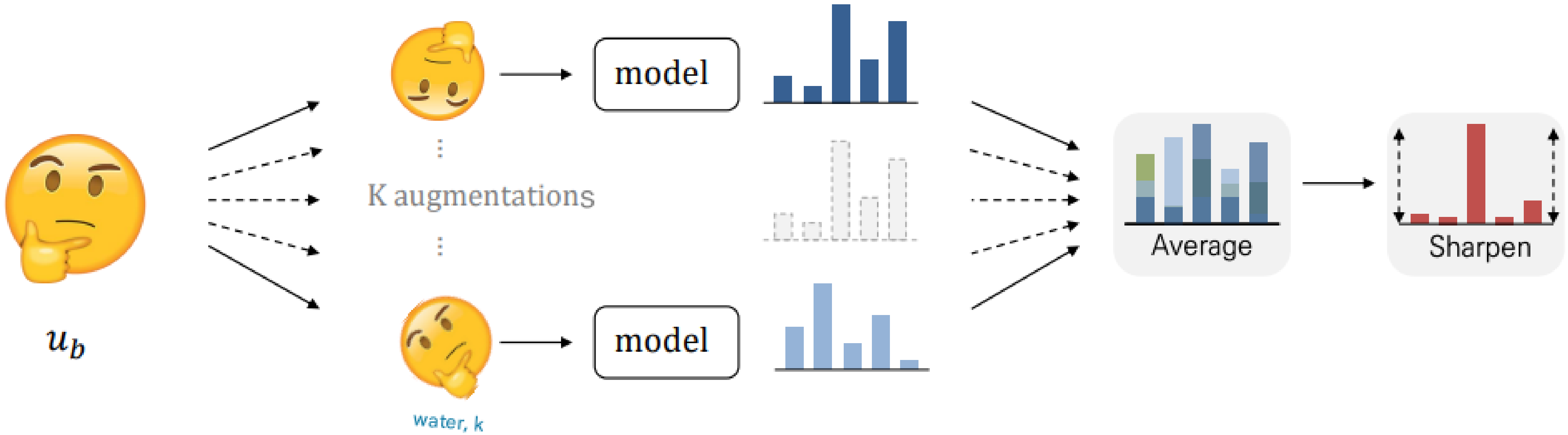


2. MixMatch

- Label Guessing & Entropy Minimization

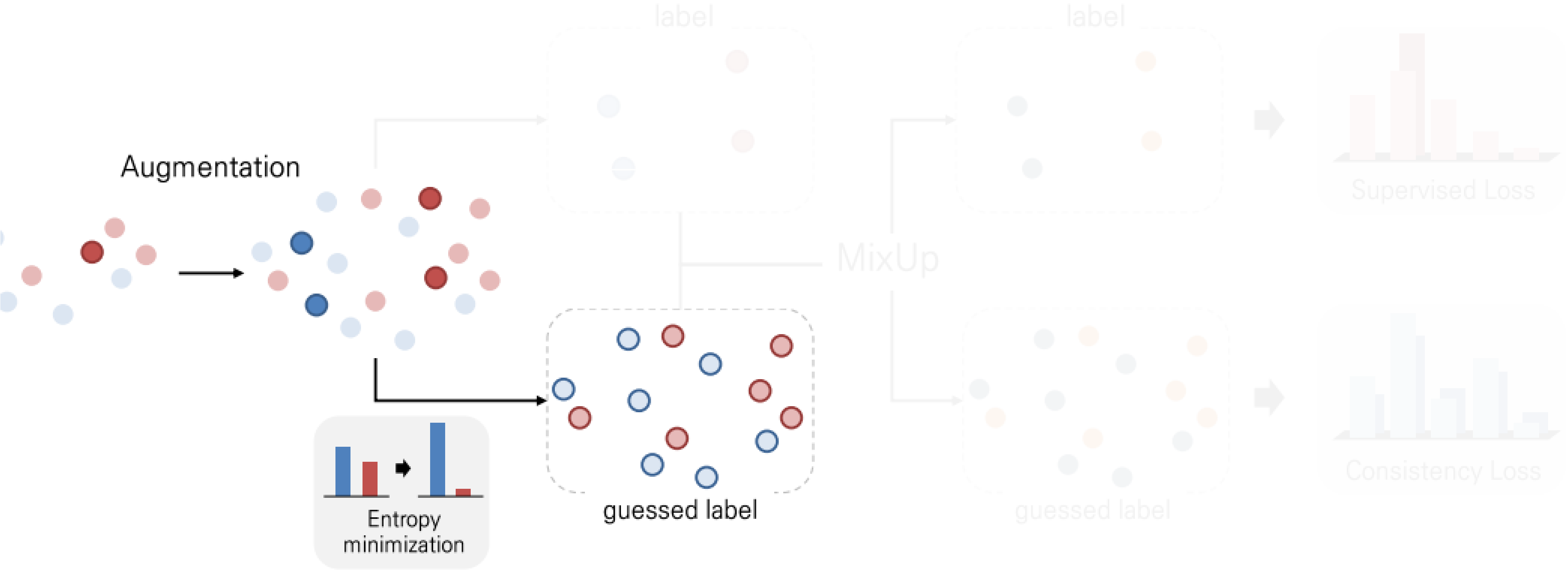
$$\operatorname{Sharpen}(p, T)_i := p_i^{\frac{1}{T}} \Bigg/ \sum_{j=1}^{L} p_j^{\frac{1}{T}}$$

: Entropy Minimization (Sharpening) using Softmax Temperature



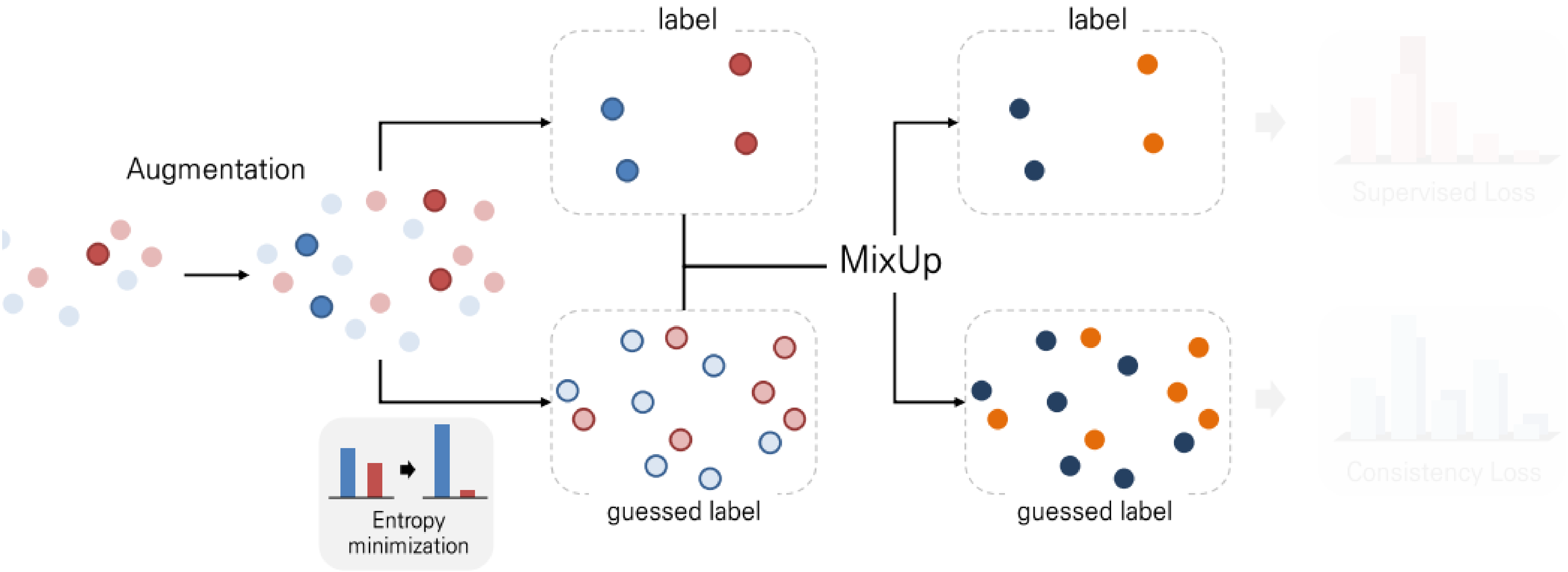
2. MixMatch

- MixUp



2. MixMatch

- MixUp



2. MixMatch

- MixUp

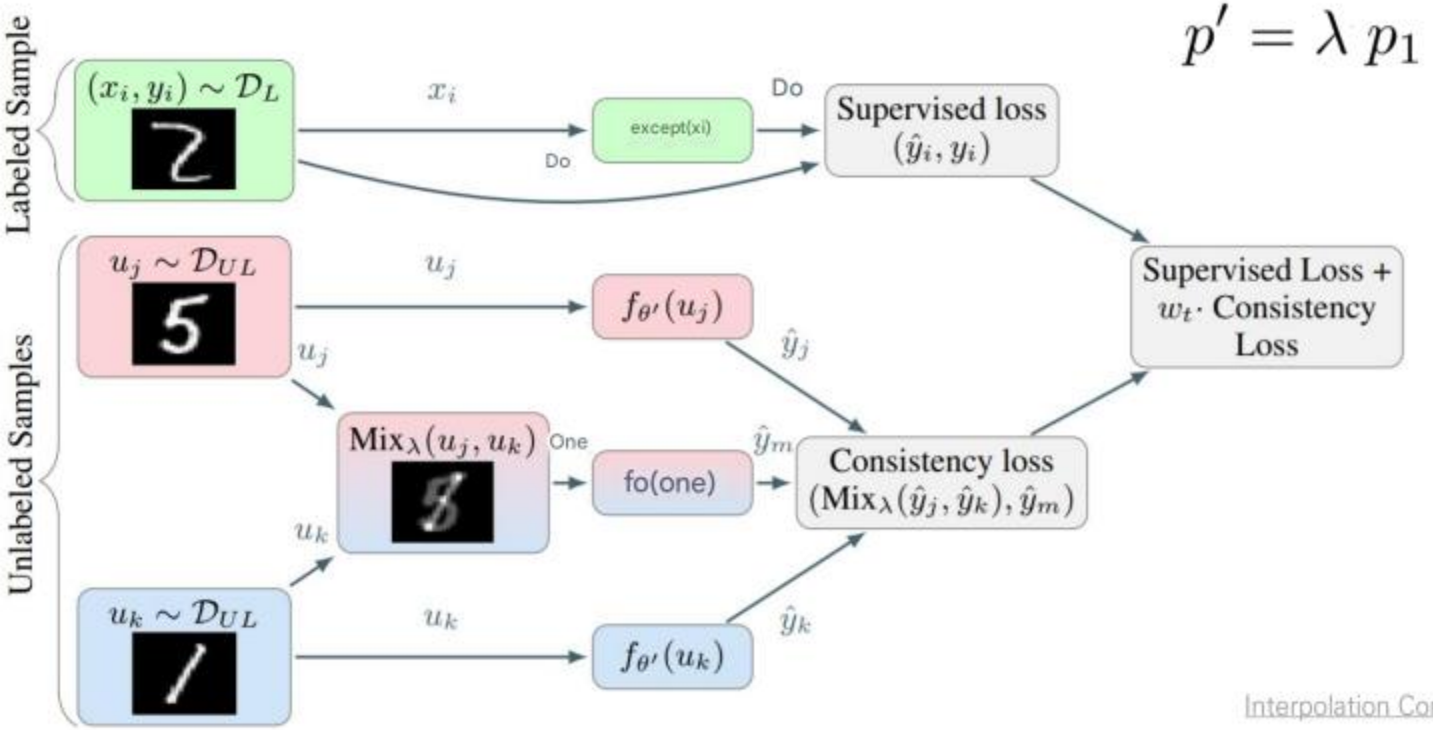
MixUp method previously used in Semi-Supervised

Perform MixUp with only unlabeled data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$

$$x' = \lambda\, x_1 + (1 - \lambda\,)x_2$$

$$p' = \lambda\, p_1 + (1 - \lambda\,)p_2$$



2. MixMatch

- MixUp

MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$

$$\lambda' = \max(\lambda, 1 - \lambda)$$

$$x' = \lambda' x_1 + (1 - \lambda') x_2$$

$$p' = \lambda' p_1 + (1 - \lambda') p_2$$

$0 \leq \lambda \leq 1$

Therefore, if we go through $\lambda' = \max(\lambda, 1 - \lambda)$

$0.5 \leq \lambda' \leq 1$

Therefore, the first term on the right side (x1, P_1) always has a coefficient greater than 0.5.

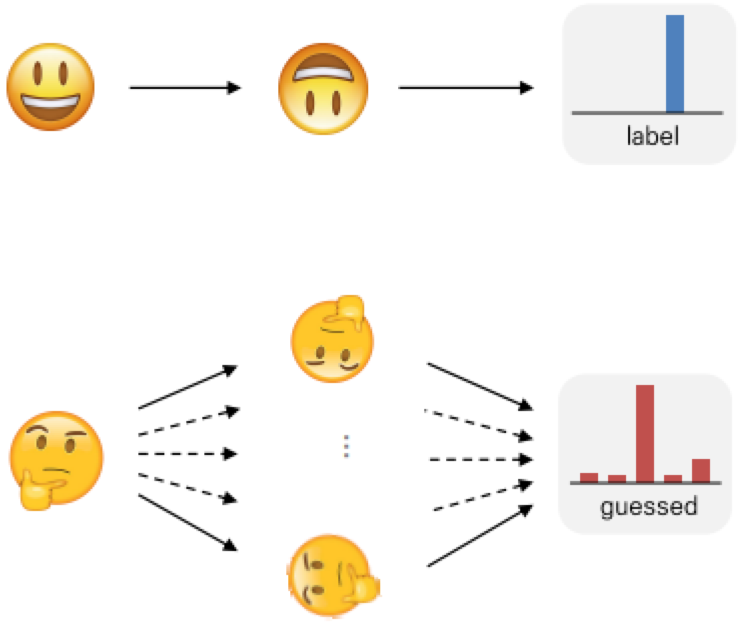
2. MixMatch

- MixUp

MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1 - \lambda)$$
$$x' = \lambda' x_1 + (1 - \lambda') x_2$$
$$p' = \lambda' p_1 + (1 - \lambda') p_2$$



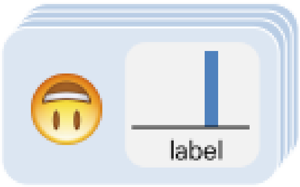
2. MixMatch

- MixUp

MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1 - \lambda)$$
$$x' = \lambda' x_1 + (1 - \lambda') x_2$$
$$p' = \lambda' p_1 + (1 - \lambda') p_2$$



$$\hat{\mathcal{X}} = \big((\hat{x}_b, p_b); b \in (1, \ldots, B)\big)$$

$$\hat{\mathcal{U}} = \big((\hat{u}_{b,k}, q_b); b \in (1, \ldots, B), k \in (1, \ldots, K)\big)$$

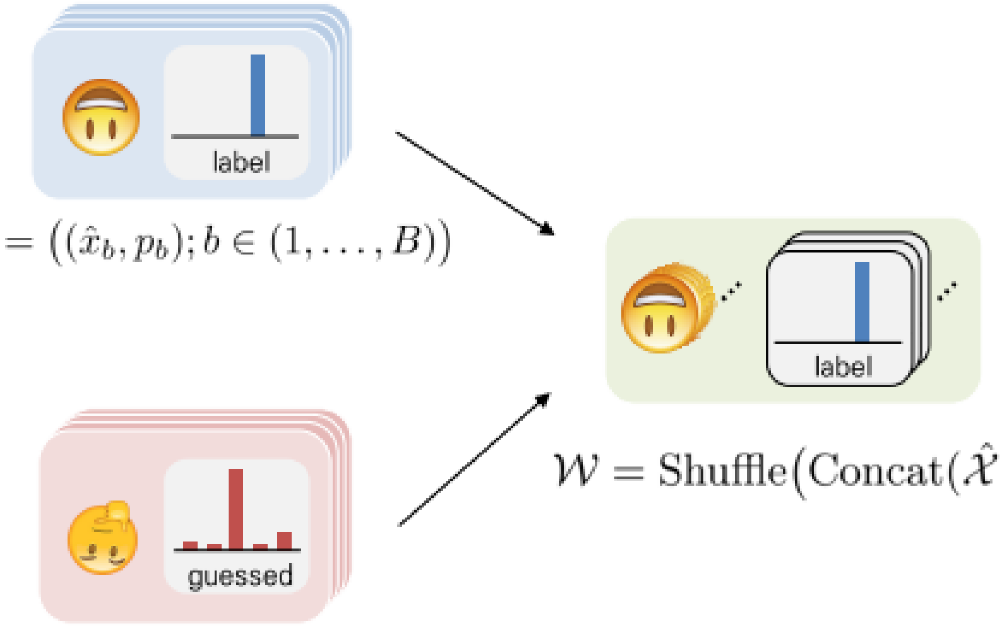
2. MixMatch

- MixUp

MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1-\lambda)$$
$$x' = \lambda' x_1 + (1-\lambda') x_2$$
$$p' = \lambda' p_1 + (1-\lambda') p_2$$



$$\hat{\mathcal{X}} = \big((\hat{x}_b, p_b); b \in (1, \ldots, B)\big)$$

$$\hat{\mathcal{U}} = \big((\hat{u}_{b,k}, q_b); b \in (1, \ldots, B), k \in (1, \ldots, K)\big)$$

$$\mathcal{W} = \mathrm{Shuffle}\big(\mathrm{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}})\big)$$

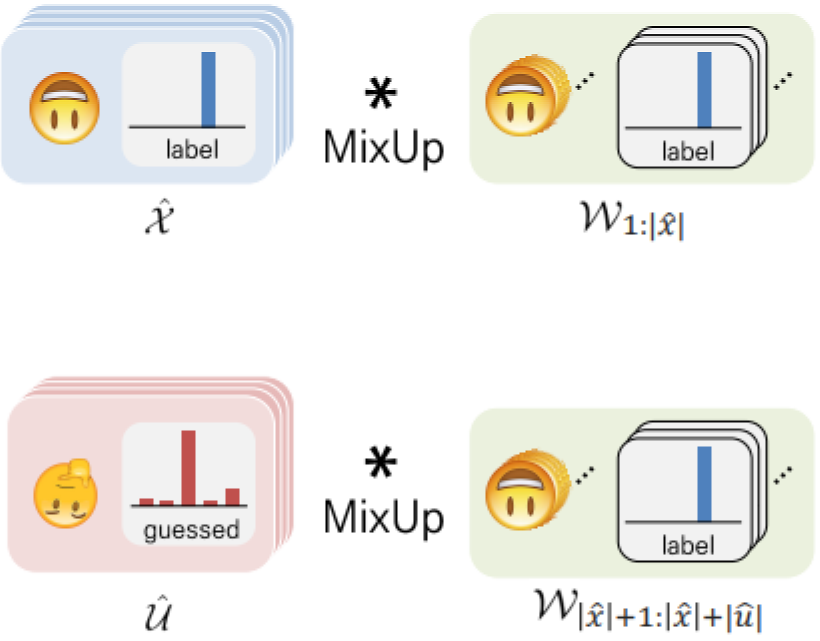
2. MixMatch

- **MixUp**

MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1 - \lambda)$$
$$x' = \lambda' x_1 + (1 - \lambda') x_2$$
$$p' = \lambda' p_1 + (1 - \lambda') p_2$$



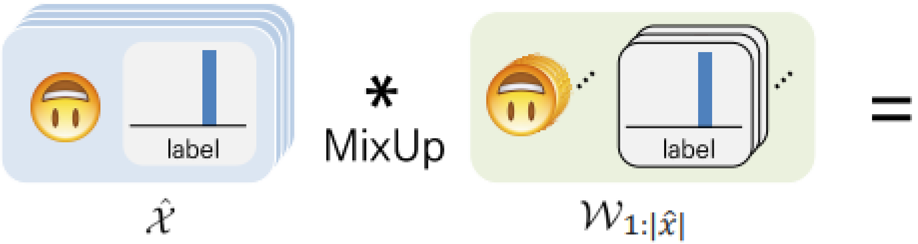
2. MixMatch

- MixUp

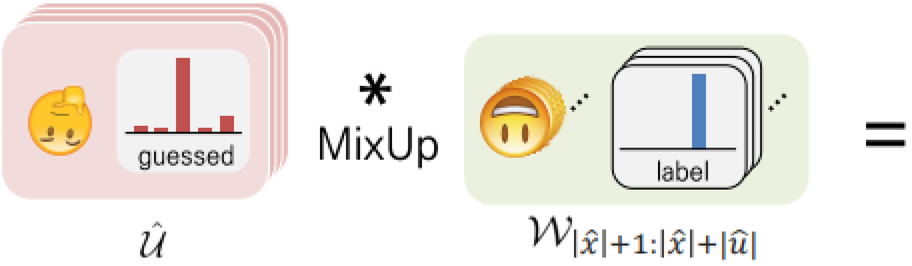
MixUp method proposed in this paper

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1-\lambda)$$
$$x' = \lambda' x_1 + (1-\lambda')x_2$$
$$p' = \lambda' p_1 + (1-\lambda')p_2$$

Perform MixUp with both Labeled and Unlabeled Data



$$x' = \lambda' x_1 + (1-\lambda')x_2$$
$$p' = \lambda' p_1 + (1-\lambda')p_2$$



$$x' = \lambda' x_1 + (1-\lambda')x_2$$
$$p' = \lambda' p_1 + (1-\lambda')p_2$$

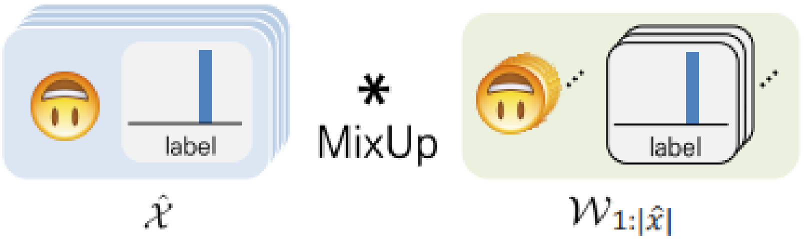
2. MixMatch

- MixUp

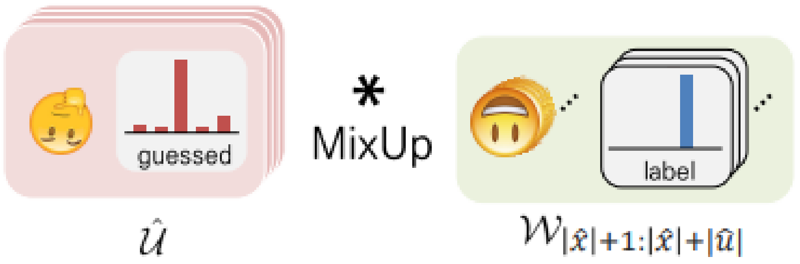
MixUp method proposed in this paper

Perform MixUp with both Labeled and Unlabeled Data

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1-\lambda)$$
$$x' = \lambda' x_1 + (1-\lambda')x_2$$
$$p' = \lambda' p_1 + (1-\lambda')p_2$$



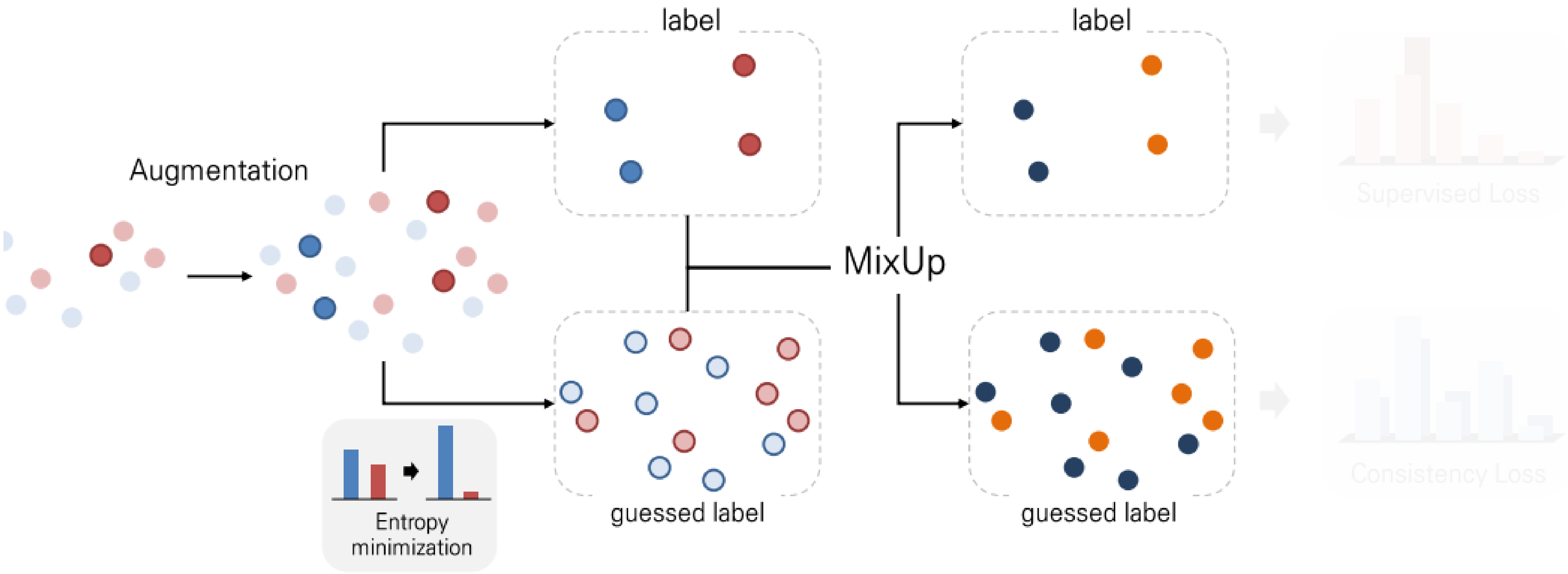
$$= \quad \mathcal{X}' = \left(\mathrm{MixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i); i \in (1, \ldots, |\hat{\mathcal{X}}|)\right)$$



$$= \quad \mathcal{U}' = \left(\mathrm{MixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}); i \in (1, \ldots, |\hat{\mathcal{U}}|)\right)$$

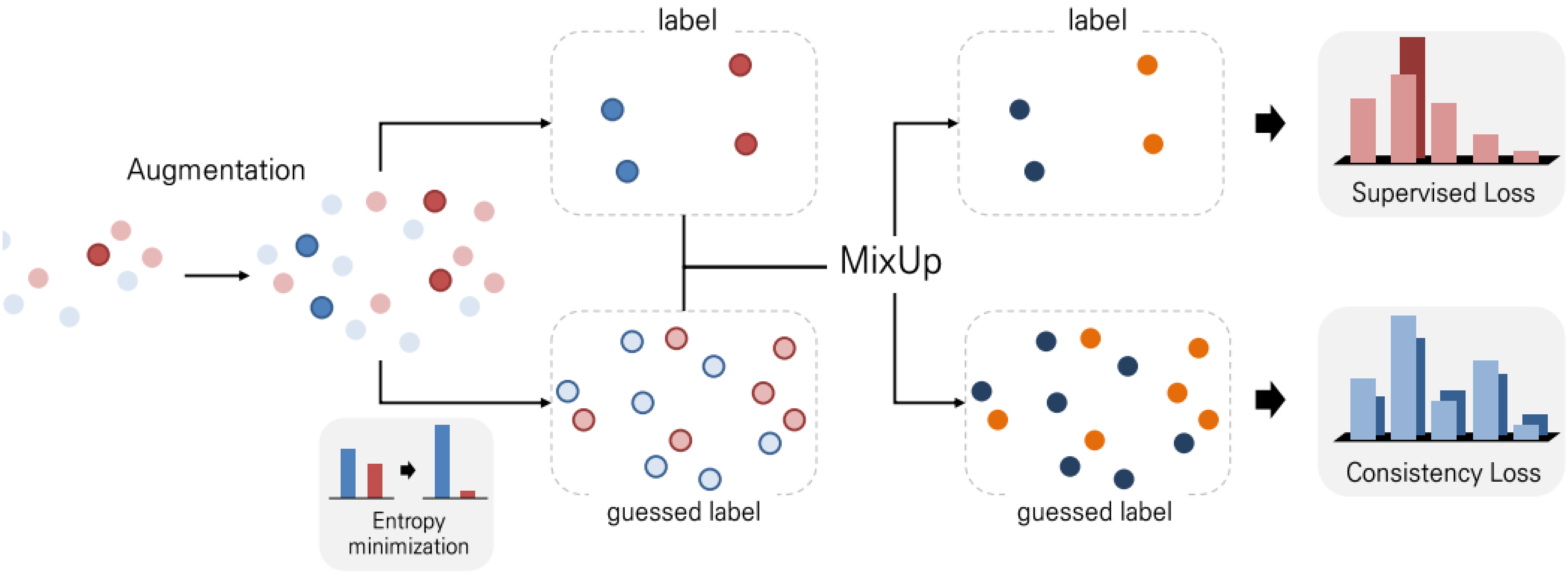
2. MixMatch

- Loss Function



2. MixMatch

- Loss Function



2. MixMatch

- ## Loss Function

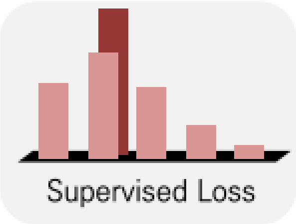
New Labeled and Unlabeled data are created through MixMatch.

$$\mathcal{X}', \mathcal{U}' = \mathrm{MixMatch}(\mathcal{X}, \mathcal{U}, T, K, \alpha)$$

Calculate Supervised Loss + Consistency Loss using generated data

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|} \sum_{x,p \in \mathcal{X}'} \mathrm{H}(p, \mathrm{p}_{\mathrm{model}}(y \mid x; \theta))$$

--- CrossEntropy ---



$$\mathcal{L}_{\mathcal{U}} = \frac{1}{L|\mathcal{U}'|} \sum_{u,q \in \mathcal{U}'} \|q - \mathrm{p}_{\mathrm{model}}(y \mid u; \theta)\|_2^2$$

------ L2 Loss -------

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}}$$

3. Experiments

- **Experiment Settings**

  Both Baseline and MixMatch use Wide ResNet-28

  We use only a portion of the entire label of the dataset and proceed with the experiment by considering the rest as unlabeled data.

  Experiment with increasing the number of labeled data

- **Baselines**

  P-Model (ICLR 2017)

  Mean Teacher (NIPS 2017)

  Virtual Adversarial Training (ICLR 2017)

  Pseudo Label

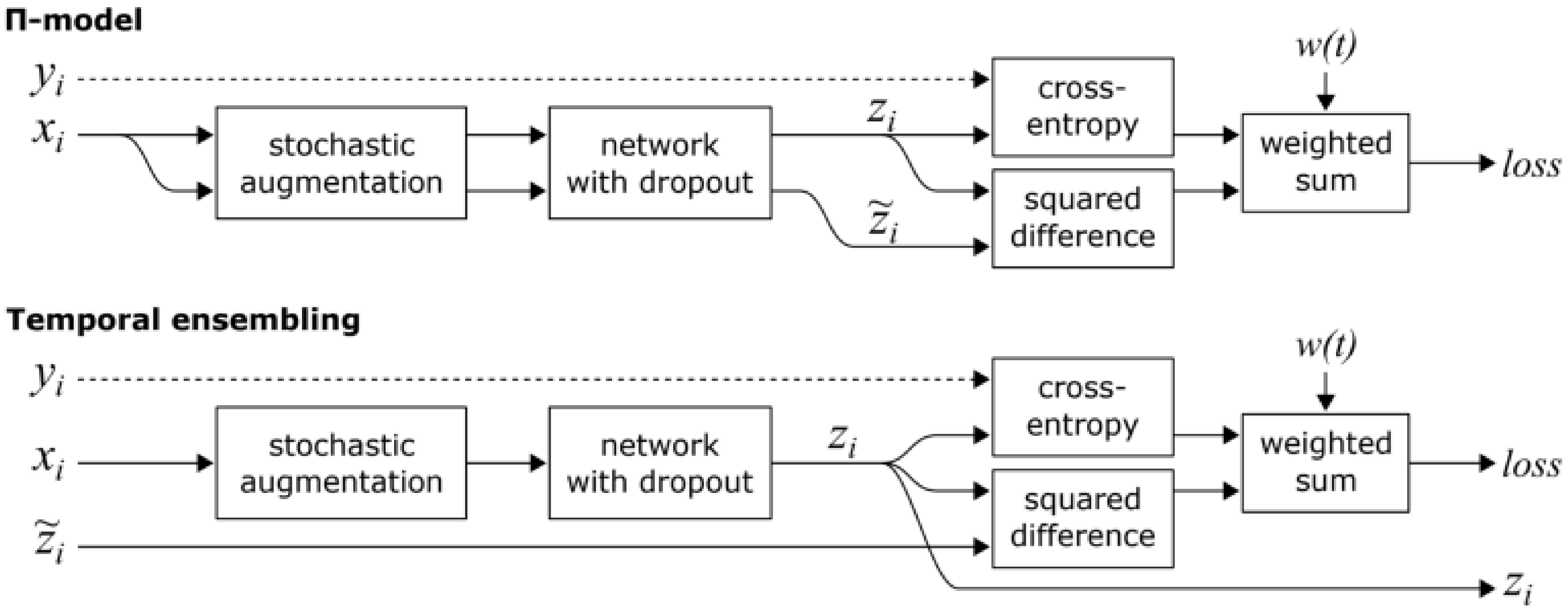
  MixUp

3. Experiments

- Baselines – Π–Model

With Labels - Cross Entropy

In the absence of labels - consistency loss between augmented data



https://arxiv.org/pdf/1610.02242.pdf

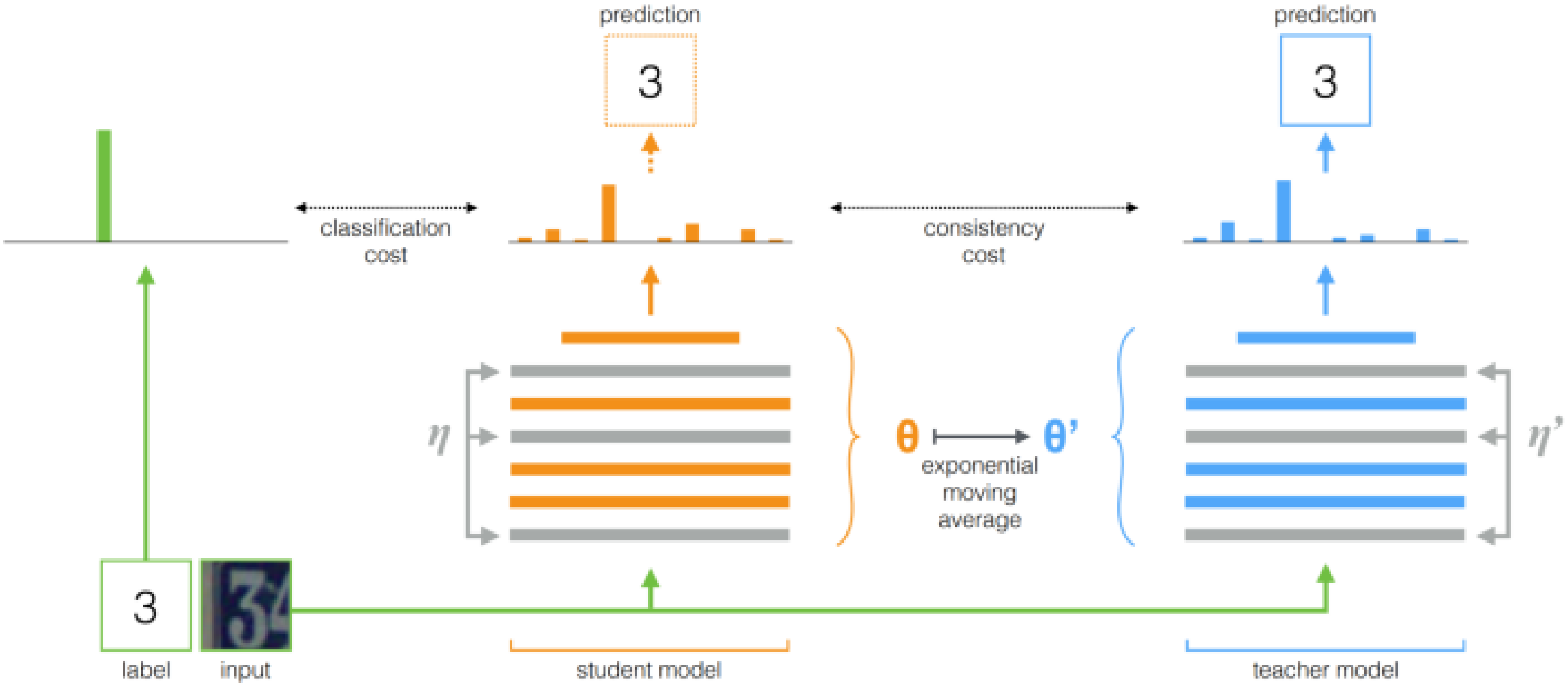
3. Experiments

- ## Baselines – Mean Teacher

Learning two models: Teacher model and Student model

The Student Model learns two loss types: supervised loss and consistency loss (with teacher).

The teacher model uses an exponential moving average of the student model's parameters.



https://arxiv.org/pdf/1703.01780.pdf

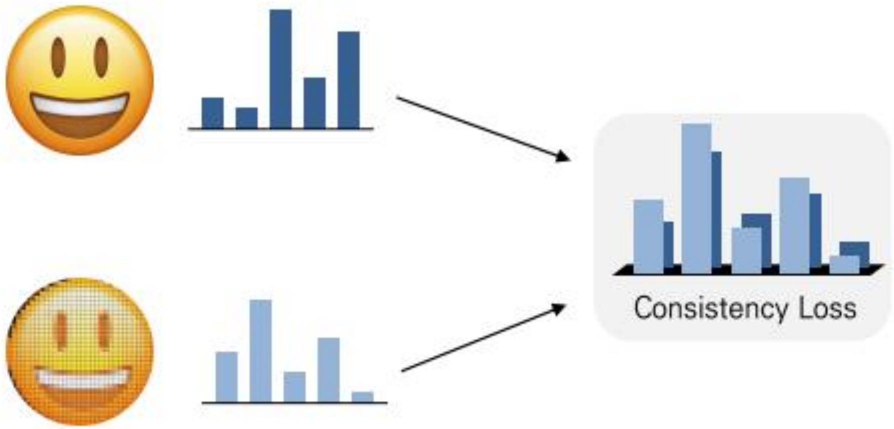
3. Experiments

▪ Baselines – VAT(Virtual Adversarial Training)

1. Create an adversarial image by adding as much noise as possible to the image.

2. Learning so that the distribution of the original image and the adversarial image are similar (Consistency Training)

Maximize KL-Divergence

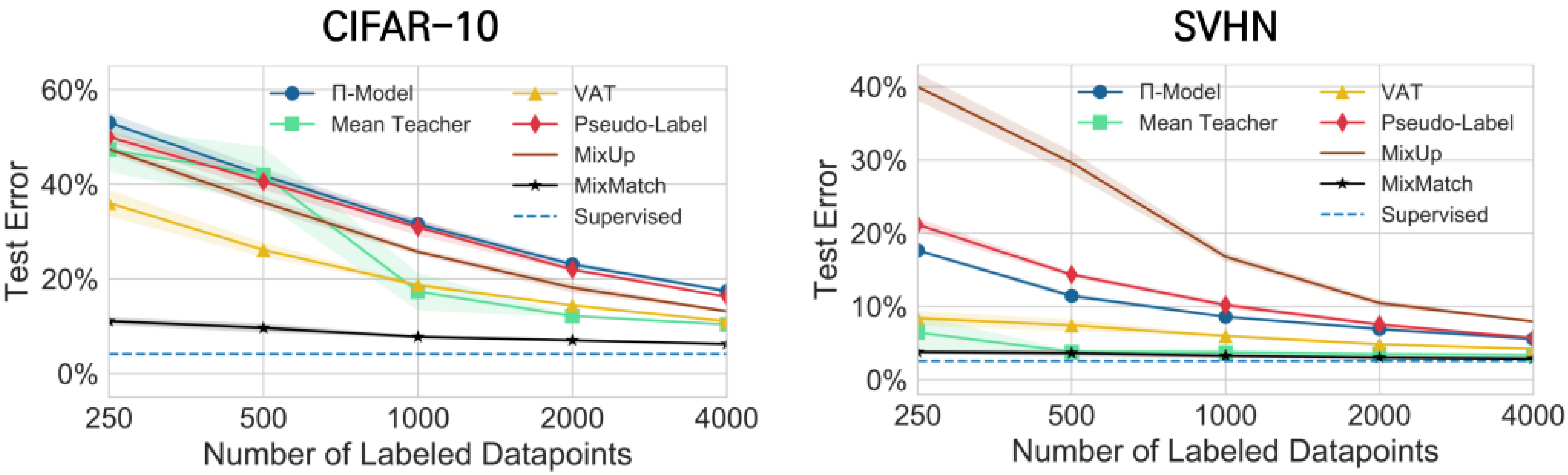


Minimize KL-Divergence

3. Experiments

- **Datasets**

  Using CIFAR-10, SVHN



3. Experiments

▪ Ablation Studies

Conduct experiments by changing model conditions

We verified that each component inside MixMatch helps to improve performance.

Ablation	250 labels	4000 labels
MixMatch	11.80	6.00
MixMatch without distribution averaging ($K = 1$)	17.09	8.06
MixMatch with $K = 3$	11.55	6.23
MixMatch with $K = 4$	12.45	5.88
MixMatch without temperature sharpening ($T = 1$)	27.83	10.59
MixMatch with parameter EMA	11.86	6.47
MixMatch without MixUp	39.11	10.97
MixMatch with MixUp on labeled only	32.16	9.22
MixMatch with MixUp on unlabeled only	12.35	6.83
MixMatch with MixUp on separate labeled and unlabeled	12.26	6.50
Interpolation Consistency Training [45]	38.60	6.81

Q & A