ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

October 23-27, 2022, Tel Aviv

Improving Robustness by Enhancing Weak Subnets

Yong Guo, David Stutz, Bernt Schiele

max planck institut informatik

Background and Motivation

Deep networks are vulnerable to image perturbations and often yield large performance drops. We study this issue by investigating the performance of their internal sub-networks (subnets).

- It is well-known that deep networks contain some well-performing subnets, i.e., winning tickets.
- However, the role of the remaining subnets still remains unexplored.

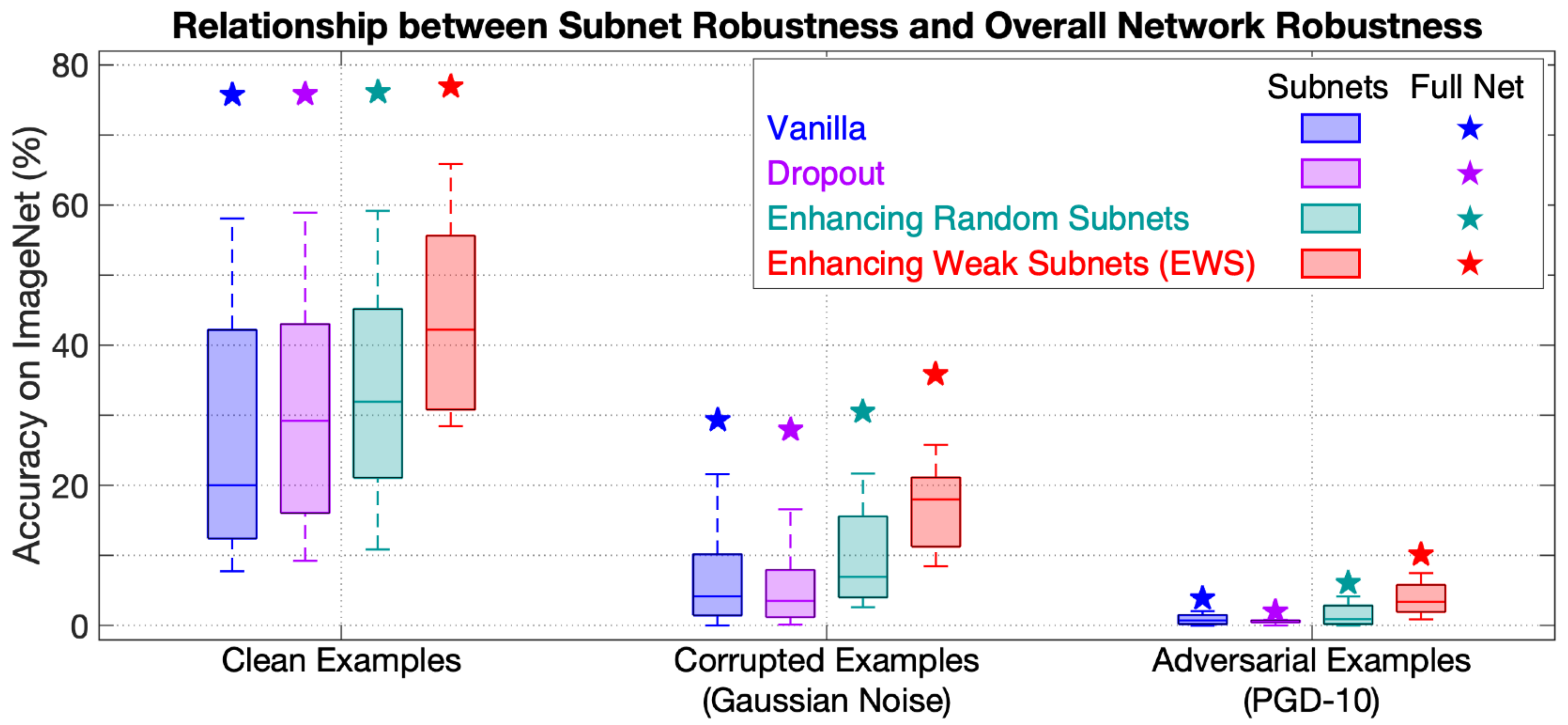


Figure 1: Performance of 1,000 randomly sampled subnets with 70% of paths/channels from a ResNet-50 on ImageNet.

Observations:

- Most subnets perform rather poorly.
- The poor subnet performance is correlated with the overall lack of robustness.

Idea: Explicitly identify and enhance these weak subnets to improve overall robustness.

Contributions

- We propose a novel robust training method which identifies and enhances weak subnets (EWS) to improve the overall robustness of the full network.
- To this end, we develop a search algorithm that obtains weak subnets by identifying particularly weak paths/channels inside the full network. Given a weak subnet, its performance is further enhanced by distilling knowledge from the full network. This approach is not only very scalable, it also adds negligible computational overhead.
- In experiments, we apply EWS on top of state-of-the-art data augmentation schemes to improve accuracy and corruption robustness on CIFAR-10/100-C and ImageNet-C [?]. Moreover, we also demonstrate the generality of our approach for improving adversarial robustness on top of recent adversarial training methods. Importantly, our approach is complementary to all these methods and improves consistently across a wide range of approaches.

EWS: Training by Enhancing Weak Subnets

Key Components of Enhancing Weak Sub-networks (EWS):

- Finding particularly weak subnets
- Enhancing weak sub-networks via knowledge distillation

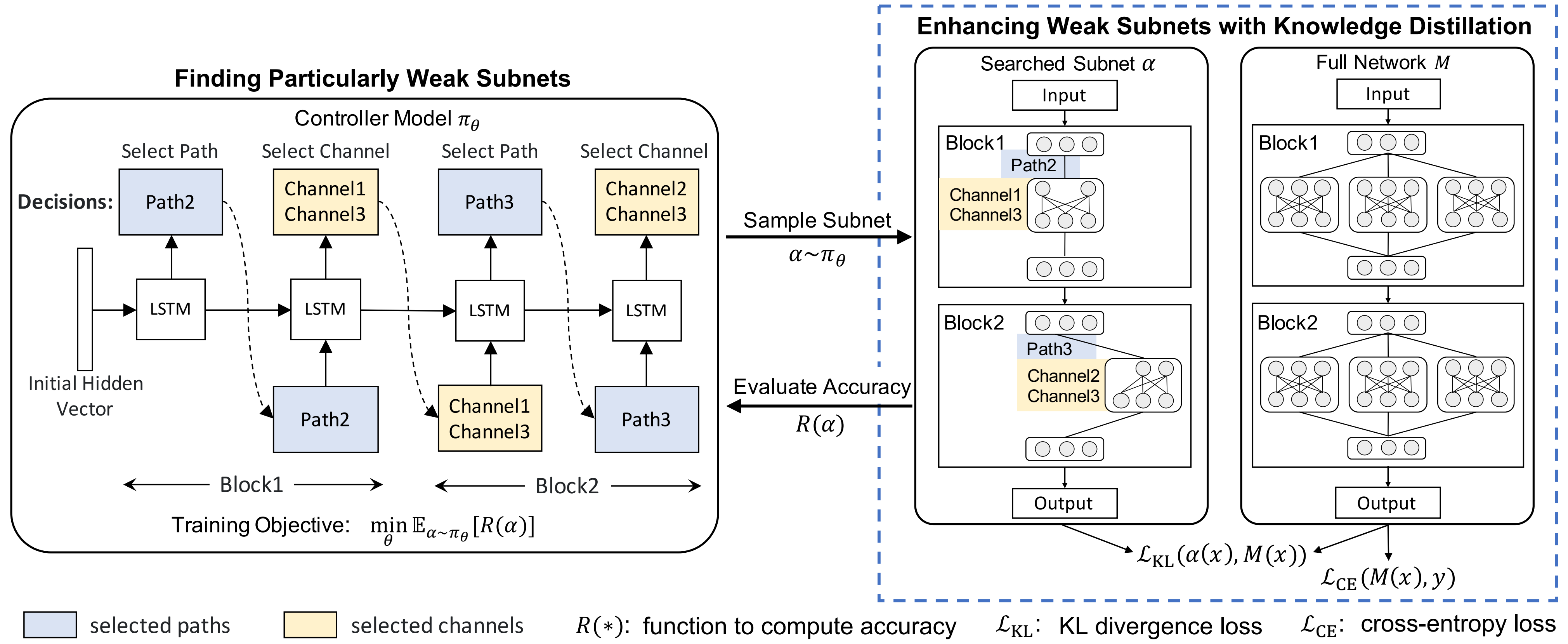

Finding Particularly Weak Subnets
Controller Model π_θ
Decisions:
Select Path
Select Channel
Select Path
Select Channel
Path2
Channel1 Channel3
Path3
Channel2 Channel3
LSTM
LSTM
LSTM
LSTM
Initial Hidden Vector
Path2
Channel1 Channel3
Path3
Block1
Block2
Training Objective: $\min_\theta \mathbb{E}_{\alpha\sim\pi_\theta}[R(\alpha)]$
Sample Subnet
$\alpha\sim\pi_\theta$
Evaluate Accuracy
$R(\alpha)$
Enhancing Weak Subnets with Knowledge Distillation
Searched Subnet α
Full Network M
Input
Block1
Path2
Channel1 Channel3
Block2
Path3
Channel2 Channel3
Output
$\mathcal{L}_{\mathrm{KL}}(\alpha(x), M(x))$
$\mathcal{L}_{\mathrm{CE}}(M(x), y)$
selected paths
selected channels
$R(*)$: function to compute accuracy
$\mathcal{L}_{\mathrm{KL}}$: KL divergence loss
$\mathcal{L}_{\mathrm{CE}}$: cross-entropy loss

Figure 2: Overview of the proposed Enhancing Weak Subnets (EWS) training method.

Training Method

Subnet Construction:

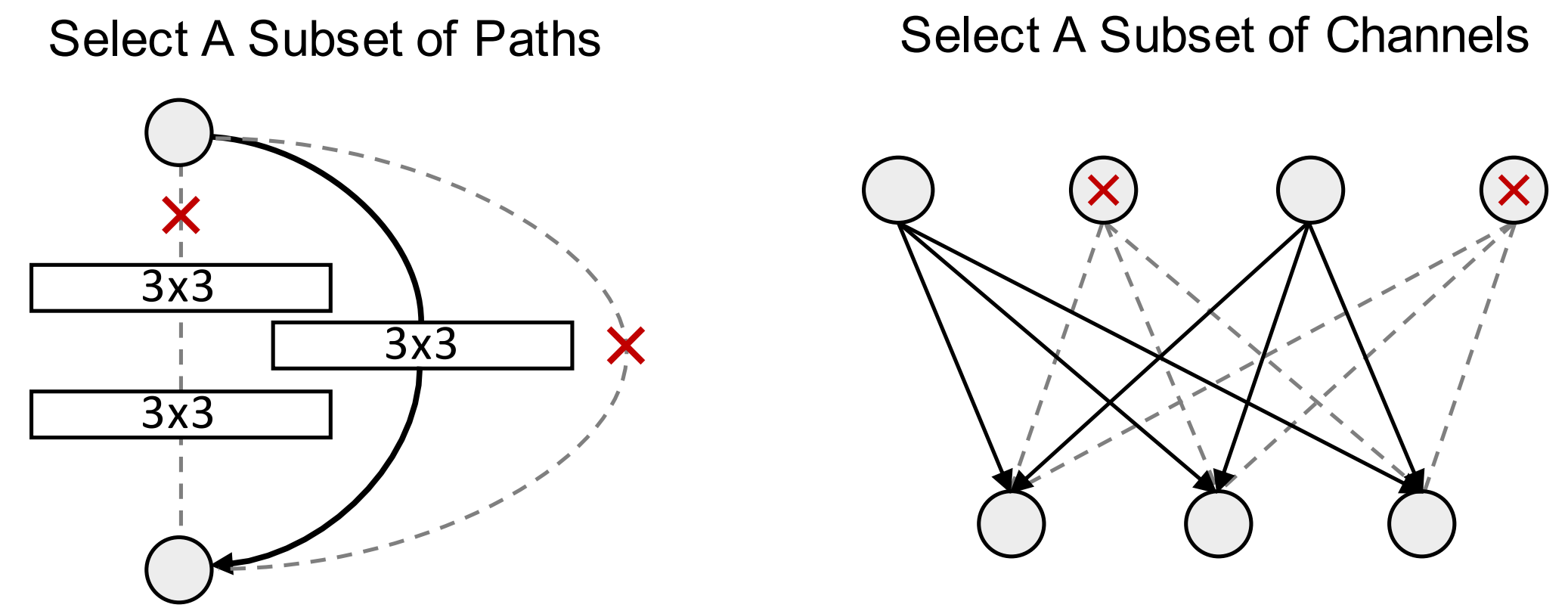


Figure 3: We select paths (left) and channels (right) in each block of the network.

Finding Particularly Weak Subnets: minimizing the expected accuracy of the selected subnets

$$\min_{\theta}\ \mathbb{E}_{\alpha\sim\pi_\theta}\left[R(\alpha)\right].$$

Enhancing Weak Sub-networks via Knowledge Distillation

$$\mathcal{L}(x,y) = \mathcal{L}_{\mathrm{CE}}\big(M(x),y\big) + \underbrace{\lambda\mathcal{L}_{\mathrm{KL}}\big(\alpha(x),M(x)\big)}_{\text{enhance weak subnet}},$$

Combining EWS with Adversarial Training

$$\min_{w}\ \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(x',y)\right] \quad \text{where } x' = \mathop{\arg\max}_{\|x-x'\|_p\le\epsilon}\ \mathcal{L}_{\mathrm{CE}}(M(x'),y).$$

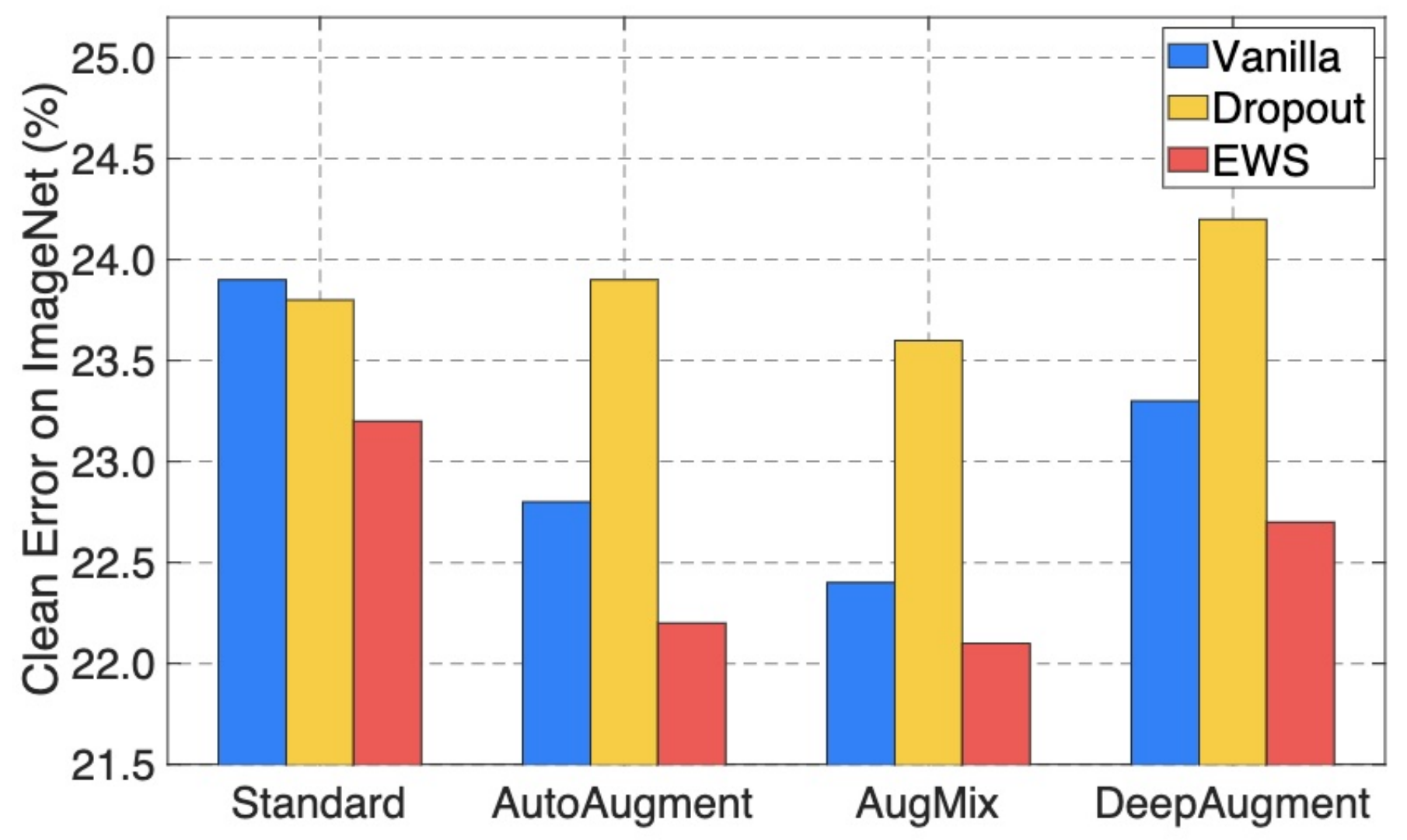
$$\mathcal{L}(x',y) = \mathcal{L}_{\mathrm{CE}}(M(x'),y) + \lambda\mathcal{L}_{\mathrm{KL}}\big(\alpha(x'),M(x')\big)$$

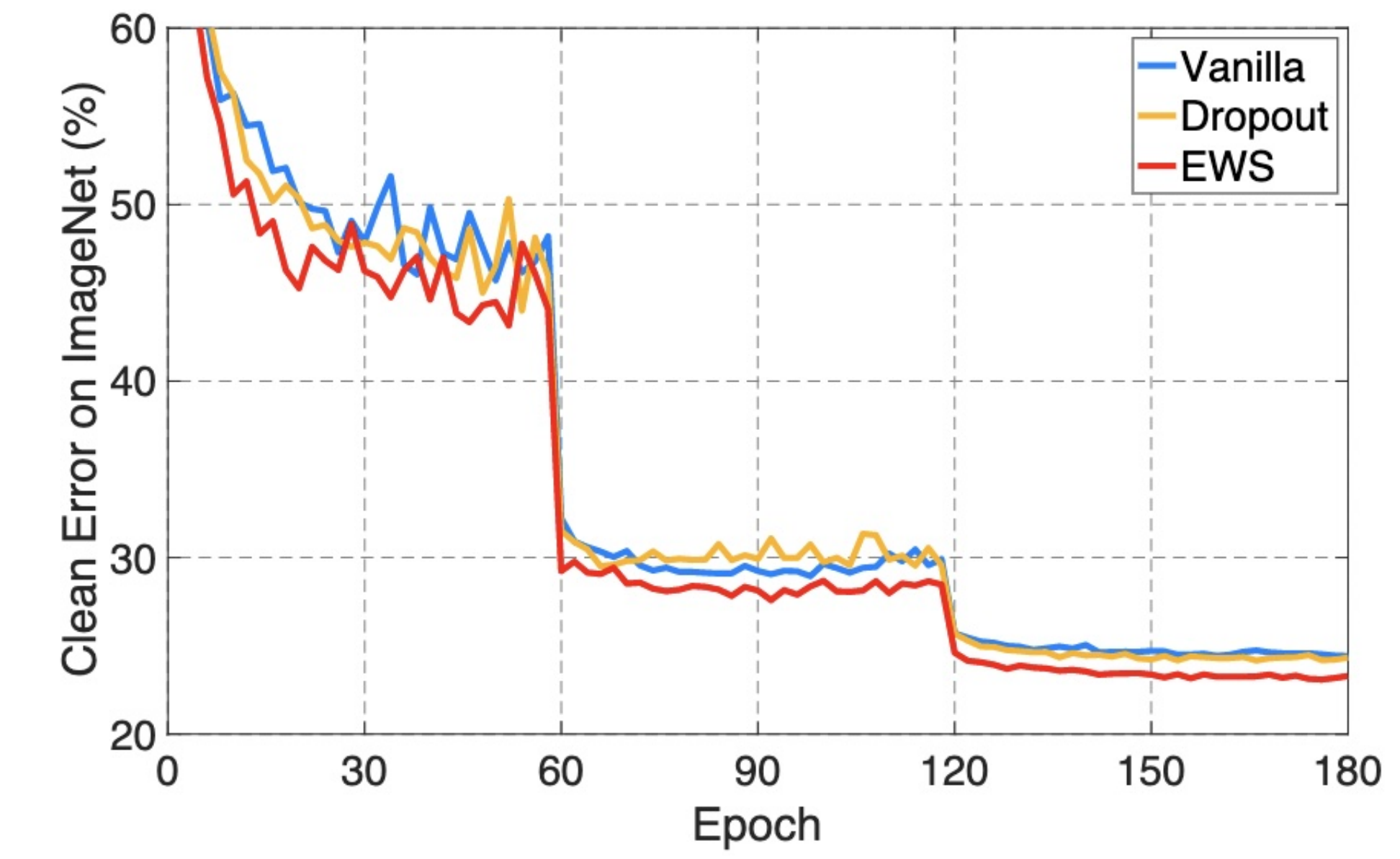
Results: Improving Corruption Robustness

- **Comparisons on CIFAR-10/100-C**

Method		CIFAR-10		CIFAR-100	
		Clean Error (%) ↓	Corruption Error (%) ↓	Clean Error (%) ↓	Corruption Error (%) ↓
Standard	Vanilla	5.32 (-0.00)	26.46 (-0.00)	23.45 (-0.00)	50.76 (-0.00)
	Dropout	5.16 (-0.16)	26.17 (-0.29)	23.19 (-0.26)	50.43 (-0.33)
	EWS	**4.44 (-0.88)**	**24.94 (-1.52)**	**22.41 (-1.04)**	**40.08 (-1.68)**
AutoAugment	Vanilla	4.05 (-0.00)	16.19 (-0.00)	23.02 (-0.00)	44.37 (-0.00)
	Dropout	3.91 (-0.14)	16.04 (-0.15)	22.84 (-0.18)	44.09 (-0.28)
	EWS	**3.23 (-0.82)**	**14.31 (-1.88)**	**22.16 (-0.86)**	**42.40 (-1.97)**
AugMix	Vanilla	4.35 (-0.00)	13.57 (-0.00)	22.45 (-0.00)	38.28 (-0.00)
	Dropout	4.19 (-0.16)	13.44 (-0.13)	22.11 (-0.34)	37.97 (-0.31)
	EWS	**3.76 (-0.59)**	**10.80 (-2.77)**	**21.81 (-0.64)**	**35.24 (-3.04)**

- **Comparisons on ImageNet**





- **Comparisons on ImageNet-C**

Method		mCE ↓	Gauss.	Shot	Imp.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Bright	Contra.	Elas.	Pixel	JPEG
Standard	Vanilla	76.5 (-0.0)	80	82	83	75	89	78	80	78	75	66	**57**	71	85	77	77
	Dropout	76.5 (-0.0)	77	79	80	78	90	79	87	**77**	77	67	58	**70**	84	75	76
	EWS	**75.1 (-1.4)**	**75**	**76**	**77**	**73**	**87**	**77**	**79**	80	**73**	**65**	58	73	**83**	**74**	**75**
AutoAugment	Vanilla	72.7 (-0.0)	69	68	72	**77**	83	80	81	79	75	64	56	70	88	57	**71**
	EWS	**71.7 (-1.0)**	**67**	**68**	**71**	78	**82**	**78**	**79**	**78**	**73**	**64**	**55**	**69**	**86**	**56**	72
AugMix	Vanilla	68.4 (-0.0)	65	66	67	70	**80**	66	66	75	72	67	58	**58**	79	69	**69**
	EWS	**67.5 (-0.9)**	**64**	**63**	**63**	**70**	81	**65**	**66**	**72**	**70**	**64**	**57**	63	**79**	**64**	70
DeepAugment	Vanilla	60.4 (-0.0)	49	50	47	59	73	65	76	64	**60**	58	51	61	76	48	67
	EWS	**58.7 (-1.7)**	**48**	**48**	**47**	**58**	**72**	**58**	**62**	**63**	62	**58**	**50**	**56**	**74**	**47**	**62**

Results: Improving Adversarial Robustness

Method		PreAct ResNet-18			WRN-28-10			WRN-34-10		
		Clean ↓	PGD-20 ↓	AA ↓	Clean ↓	PGD-20 ↓	AA ↓	Clean ↓	PGD-20 ↓	AA ↓
AT	Vanilla [56]	17.54	49.18	52.96 (-0.00)	14.89	45.17	47.81 (-0.00)	14.74	45.39	47.47 (-0.00)
	EWS	**16.85**	**47.99**	**51.84 (-1.12)**	**14.57**	**44.24**	**47.17 (-0.64)**	**14.33**	**44.04**	**46.58 (-0.89)**
	AWP [82]	19.59	46.01	51.43 (-0.00)	15.89	42.93	46.41 (-0.00)	**14.17**	41.89	45.96 (-0.00)
	AWP-EWS	**19.25**	**44.98**	**50.48 (-0.95)**	**15.81**	**41.72**	**45.58 (-0.83)**	14.21	**41.07**	**45.29 (-0.67)**
TRADES	Vanilla [89]	17.42	46.88	50.84 (-0.00)	15.50	44.11	47.40 (-0.00)	15.32	43.84	46.89 (-0.00)
	EWS	**17.10**	**45.73**	**49.67 (-1.17)**	**15.09**	**43.45**	**46.72 (-0.68)**	**14.56**	**43.13**	**46.06 (-0.83)**
	AWP [82]	18.27	45.36	49.62 (-0.00)	14.84	41.25	44.86 (-0.00)	15.55	40.85	43.90 (-0.00)
	AWP-EWS	**17.67**	**44.20**	**48.58 (-1.04)**	**14.30**	**40.40**	**44.22 (-0.64)**	**14.13**	**40.05**	**43.17 (-0.73)**
TRADES-AWP*		17.13	43.68	48.37 (-0.00)	13.37	38.51	41.97 (-0.00)	12.73	35.97	40.74 (-0.00)
TRADES-AWP-EWS*		**16.62**	**42.33**	**47.23 (-1.14)**	**12.59**	**37.60**	**41.23 (-0.74)**	**11.90**	**35.19**	**40.05 (-0.69)**